Polytech Nice Sophia Antipolis

Master IFI/ Ubinet track

Final report

Network sampling and discovery processes

Author:
Alina Tuholukova

Supervisors:
Konstantin Avrachenkov
Giovanni Neglia

September 1, 2015

Contents

1 Respondent-driven sampling **5**
- 1.1 Motivation . . . 5
- 1.2 Respondent-driven sampling . . . 6
- 1.3 Current estimators . . . 7
  - 1.3.1 Sample average . . . 7
  - 1.3.2 Volz-Heckathon estimator . . . 7
  - 1.3.3 Group estimator . . . 7
- 1.4 Problems of RDS . . . 8
  - 1.4.1 RDS is not uniform . . . 9
  - 1.4.2 RDS is not independent . . . 9
- 1.5 Enhanced RDS . . . 11

2 Network with values **15**
- 2.1 Motivation . . . 15
- 2.2 Definitions . . . 16
- 2.3 Gibbs distribution . . . 17
  - 2.3.1 Algorithm . . . 18
  - 2.3.2 Explanatory example . . . 19
  - 2.3.3 Demonstration of random graphs with values . . . 20
- 2.4 Expected energy in steady state . . . 21
  - 2.4.1 Motivation . . . 21
  - 2.4.2 Analysis . . . 23
  - 2.4.3 Results . . . 27
  - 2.4.4 Conclusions . . . 27

3 Mathematical model **29**
- 3.1 Error prediction . . . 29
  - 3.1.1 Variance prediction . . . 29
  - 3.1.2 Geometric correlation . . . 30
  - 3.1.3 Variance with skipping . . . 32
  - 3.1.4 In RDS context . . . 33
  - 3.1.5 Trying to use the result . . . 35
  - 3.1.6 General case . . . 36
  - 3.1.7 Error prediction . . . 38
  - 3.1.8 Discussion . . . 39
- 3.2 Data . . . 39

3.2.1 Data from the Project 90 39
3.2.2 Data from the Add health project 40
3.2.3 Experiments 40
3.3 Other estimators 40

Introduction

We are living in the era of information when it is crucial to collect data, to be able to analyze them and draw potentially valuable conclusions. It is particularly interesting to analyze network structures like online-social networks, peer-to-peer networks, real social networks, hidden populations.

Online-social network 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), VK (about 250 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 the sociologists and give them another instrument for the research.

Another important networks are the real ones. All people are members of numerous real social networks like network of friends, business contacts, colleagues, lovers of board games etc. The researchers are interested in studying the network structure. Add Health study [**?**] has built the networks of the students in the selected schools of the United States, which served as the foundation of numerous research [**?**].

As well as it is interesting to analyze the structure of the network graph by itself it can be useful to study the connection between the network structure and the characteristics of the users. 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 both are perhaps you became friends because you were both fond of it. One way or another this property of sharing the common characteristics between people that are in contact is usually observed in the networks and is called ***homophily***.

And actually sometimes it is the common interest that is in foundation of the network itself. The Online social network Flixster can help to meet people with similar tastes in movies, the Russian online social network Odnoklassniki helps people to find their old classmates.

This property of gathering around similar interests can help to study ***hidden populations***, the members of such population are by definition hard to reach. Examples of hidden populations are drug users, sex workers, jazz musicians. Making analysis about hidden populations can be very important. For example, studying the population of drug users or sex workers can help to understand the prevalence of some diseases. The difficulty in studying a hidden population is that it is hidden: there is no easy and quick access to the members of these populations. But once the researcher knows some of them he can take advantage of the fact that probably they are in contact with other representatives of this population. The individuals of hidden populations form the network around their interest. In this way the subset of this population can be found by "crawling" its network graph.

The sampling methods that use the contacts of known individuals of a population to find

other members are called ***chain-referral methods***. This way of sampling is different from the ideal uniform independent sampling of individuals and, because of homophily, leads to increased bias and variance of the estimators as we are going to show.

The report is organized as follows. In chapter 1 we will introduce the current method of sampling hidden populations that is called respondent-driven sampling (RDS). We will regard different estimators that are used to make inferences about the population based on collected samples. We will discuss what are the problems and challenges related to the RDS. Next we will propose the enhancement of the current RDS method that is trying to solve some of these problems.

In order to study formally introduced enhanced RDS method the chapter 2 and 3 will develop the mathematical model. Chapter 2 in particular will introduce a model to generate synthetic network with the desired level of homophily. As this problem can be applied outside the RDS context as well, it is presented in the separated chapter. In chapter 3 we will find the exact expressions for the variance and bias for the existing RDS model and the enhanced one in order to compare them later.

Finally chapter 3 presents results of the experiments that confirms the correctness of the theoretical part. In order to validate results we used the created synthetic networks as well as the real ones. We will compare the performance of the currently used RDS method to our method.

Chapter 1

Respondent-driven sampling

1.1 Motivation

In order to estimate some characteristics of a particular population researchers need to find a representative subset of this population. After analyzing the collected data they can produce general results for the whole population. However sometimes they can have difficulties in finding such subset of some populations. Examples can be the population of drug users, of people involved in homosexual relations, of sex workers, of illegal immigrants, of participants in some social movements, of homeless and so on [**?**]. The members of these populations are hard to find, they are not willing to participate in the research, that's why these populations are called *hidden populations*.

The reasons to study hidden populations can be various. Studying the prevalence of HIV can help to understand and control the spreading of some diseases. A lot of research is targeted on the studying the HIV prevalence among hidden populations like drug users, female sex workers [**?**], gay men [**?**]. Another study [**?**] studied the population of jazz musicians. Even if the jazz musicians have no reasons to hide them, it is still hard to access them with the standard sampling methods. This study considers the question important from the social side: what is more significant for the income level, the number of social connections or the level of professional activity.

In order to collect samples from the hidden population some special technique were developed. The main challenge here is that each individual has the same probability to be sampled, or if some bias exists it can be easily removed. For example, using telephone survey in order to collect information would automatically exclude some subsets of people (like homeless, poor) that don't have telephone number. This fact can affect the correctness of the estimate because it is impossible to predict the bias that we introduce by considering only individual with telephone number.

Time-space sampling tries to solve this problem by sampling from different venue-time segments. Each venue-time segment(e.g., Place Massena, Friday, 21.00-24.00) is selected with probability based on the expected number of participants at this segment. And then on the chosen venue-time segment the researchers invite the participants. However it is impossible to list all venue-time segments what results again in an unknown bias of the obtained estimation.

Another approach to sample hidden populations is using chain-referral techniques. The researchers benefit from the fact that people tend to know each other when they have common interests. A jazz musician has a lot of reasons to know another jazz musicians: they may perform

in the same clubs, they probably attend the same events, they may collaborate with each other and so on. The researcher can then find just a few representatives of the hidden population and then ask them to provide the contacts of other members of this population.

One particular chain-referral technique that is currently widely used for studying prevalence of HIV/AIDS among injection drug users, sex workers, gays is called (RDS) ***respondent-driven sampling***.

1.2 Respondent-driven sampling

RDS begins by selecting a group of initial participants that are called *seeds*. The procedure follows according to chain-referral model: each participant in the study recruits other participants. The step is called *wave*. Both participating in the research and recruiting new participants are encouraged by financial incentive. The sampling continues in this way until the needed size of participants is reached. During the interview participants are asked to report how many contacts in the hidden population they have. This process enables to collect data for making statistical analysis.

In order to study formally RDS we will model the network of people as a graph, where the individuals are represented by the nodes and the contact between two individuals is represented by the edge between the corresponding nodes. We will make following assumptions:

1. Contacts to be recruited are selected uniformly at random
2. One individual can recruit exactly one another individual
3. The same individual can be recruited multiple times
4. Seeds are chosen proportionally to their degree in the network.
5. If individual A knows individual B than individual B knows A as well (the network can be represented as undirected graph).
6. Individuals know and report precisely their network degree.
7. Each individual is reachable from each other individual (the network is connected).

When the contacts to be recruited are selected uniformly at random and one individual can recruit exactly one another individual the process of RDS recruitment can be regarded as *random walk* on the graph. In particular it will be a *standard random walk* when the same individual can be recruited multiple times and a *self-avoiding random walk* if the same individual can be recruited only once. For now we will regard only standard random walk.

For this process stationary distribution is exactly distribution proportional to the network degree. However the fourth assumption guarantees that not only the first but all samples during the process are taken with probability proportional to the degree of participants in the network. Without this assumption only the individuals selected after a large enough number of waves would be selected with a probability proportional to their degree, because the random walk could be considered to have reached its stationary distribution.

Some of these assumptions are not so restrictive. In particular it is reasonably to suppose that acquaintance is symmetric. In [**?**] the assumption that seeds are selected proportionally to their

degree is considered to be reasonable as the people that are drawn as seeds are well-known people and they have usually more contacts than on average. The other assumptions are arguable. For sure there will be some error in the reported network degree. It is also arguably that the choice of the contact to recruit is uniformly random. The sensitivity to violation of some assumptions was studied in [**?**].

1.3 Current estimators

Some unbiased estimators were developed in order to generalize the results on all population from the collected samples. We will look at some of them: Sample average (SA), Volz-Heckathon estimator (VHE) that was introduced in the paper [**?**] and estimator presented in the paper [**?**] that we will call Group estimator (GE).

1.3.1 Sample average

Let $y_1, y_2, ..., y_n$ be all collected samples during RDS. The simplest estimator of the population mean it is just a samples average:

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

But this estimator is biased. In the way that sampling is performed the individuals with more contacts are more likely to be recruited. The probability to encounter the node on the step i with the value y_i is proportional to its degree d_i. In the case when there is correlation between the quantity if interest y_i and the degree d_i this estimator will be biased towards the nodes with higher degrees.

1.3.2 Volz-Heckathon estimator

In the way that sampling is performed the individuals with more contacts are more likely to be recruited. To correct this bias the responses from individuals are weighted according to their number of contacts. Volz-Heckathon estimator corrects the bias toward the nodes with higher degrees in the following way.

Let $y_1, y_2, ..., y_n$ be all collected samples during RDS. Let denote as $d_1, d_2, ..., d_n$ the number of contacts accordingly to the observed samples. Then estimate μ of the population mean is defined as:

$$\hat{\mu} = \frac{1}{\sum\limits_{i=1}^{n} 1/d_i} \sum_{i=1}^{n} \frac{y_i}{d_i}$$

It is shown that this estimator produces asymptotically unbiased results. [put the reference]

1.3.3 Group estimator

This estimator is targeted on the estimation of the percentage of population with certain characteristics. Let us say that the people from the group A possess this characteristic and people

from the group B no. Then let $\widehat{PP_A}$ be the Group estimator, that estimates the percentage of the members of the group A.

Again each observed individual should report his number of contacts: $d_1, d_2, ..., d_n$ and whether he/she possess the asked trait. After collecting all the samples the total number of the representatives of the group A, denoted as n_A, and the total number of the representatives of the group B, denoted as n_B, are counted.

Then the estimators of the average number of contacts in the group A and in the group B are respectively:

$$\widehat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$$

$$\widehat{D}_B = \frac{n_B}{\sum_{i=1}^{n_B} \frac{1}{d_i}}$$

Observing the chain of recruitment the number of recruitment between and inside groups are counted. Let's denote r_{AB} as number of recruitments from the individual in the group A to the individual in the group B, r_{AA} as number of recruitments from the individual in the group A to the individual in the group A and in the same way r_{BB}, r_{BA}.

Then the estimates for the probability to hire person from the group B by the person from the group A and in the opposite way are respectively:

$$\widehat{C}_{A,B} = \frac{r_{AB}}{r_{AA} + r_{AB}}$$

$$\widehat{C}_{B,A} = \frac{r_{BA}}{r_{BB} + r_{BA}}$$

Finally the Group estimator $\widehat{PP_A}$ is:

$$\widehat{PP_A} = \frac{\widehat{D}_B \cdot \widehat{C}_{B,A}}{\widehat{D}_A \cdot \widehat{C}_{A,B} + \widehat{D}_B \cdot \widehat{C}_{B,A}}$$

The prove that the Group estimator produces asymptotically unbiased results can be access in [**?**].

Let's note with the VHE it is also possible to estimate the percentage of the population with the given trait. If the ith observed sample obtain this trait then the value if y_i should be set at 1, otherwise at 0.

[to put may be that this is exactly VHE estimator if each person recruits exactly one person]

1.4 Problems of RDS

To understand what are the problems with respondent-driven sampling we will compare it with the "ideal" sampling: independent uniform sampling.

First, let's look at the variance of the uniform sampling. Let $y_1, y_2, ..., y_n$ be the samples that are taken uniformly at random and all the samples are independent. Let's take the average value of the samples as an estimator of mean population value:

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

It is essential that here the samples $y_1, y_2, ..., y_n$ are **independent and uniform**. That is why this estimator is unbiased. Let σ be the variance of the samples, then the variance of the estimator $\widehat{\mu}_{SA}$:

$$\sigma^2_{\hat{\mu}_{SA}} = \text{Var}\left(\frac{y_1 + y_2 + ... + y_n}{n}\right) = \frac{1}{n^2}\text{Var}(y_1 + y_2 + ... + y_n) =$$

$$= \frac{1}{n^2}\left(\text{Var}(y_1) + \text{Var}(y_2) + ... + \text{Var}(y_n)\right) = \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + ... + \sigma^2\right) = \frac{\sigma^2}{n}$$

We can see that the variance of the estimator depends on the sample size n. In this case the only thing that we can do in order to have better estimation is to increase sample size. To have more precise result we should try to take as much samples as possible. If each sample has a cost, then sample size is restricted by the budget of the research project.

1.4.1 RDS is not uniform

Of course the way we perform sampling it is not independent uniform sampling. First, nodes are not sampled not uniformly. When each participant select other participants with the same probability among his friends there is bias towards the nodes with high degrees. So the more contacts an individual has more probable he will be invited to participate in the study.

In some cases, when we can control the probability of selecting the next participant we can achieve uniform sampling even with random walk. For example, the study [**?**] was using Metropolis-Hasting Random Walk to sample Facebook. This method requires information about user's degree and the degrees of all his neighbors. According to this information the selection probabilities are counted for all the friends and then one of them is selected by the computer. In Facebook it is feasible to do: with API requests needed information is collected and then probabilities to choose one of the user's friends are counted. Though this method gives us mechanism to sample uniformly even with random walk, it is not really possible in the situations where we cannot control selection probabilities. Like in the case with hidden populations: it is individual who decides how to hire.

Another way to remove degree bias is by reweighing samples according to their degrees, what actually the VHE estimator does, we discussed this in the previous section.

1.4.2 RDS is not independent

Second problem is that collected values are not independent. The participants i and $i+1$ know each other, they can be friends, relatives or just acquaintances, so their values y_i and y_{i+1} can be dependent. The drug users may be in contact because they go to then same drug dealer and probably buy the same kind of drugs.

The tendency of people with connections to have the similar characteristics is called *homophily*.

And indeed we encounter often a homophily in the real situations. A lot of real networks demonstrate that the value on the node depends from the values of its neighboring nodes. For instance, the study [**?**] is evaluating 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 can become obese are increased by 57%.

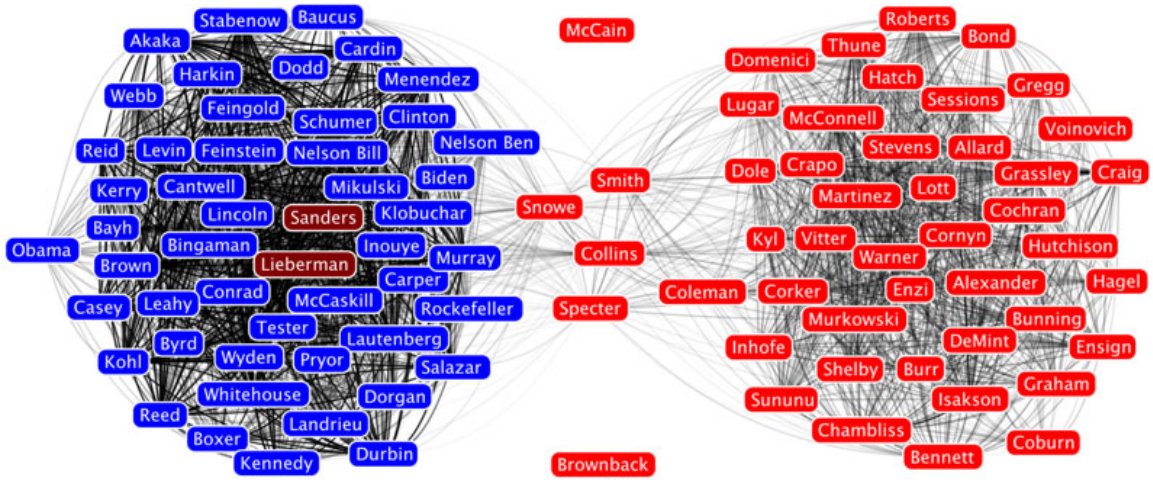


Figure 1.1: Voting patterns of U. S. Senators during 2007 [**?**]. The red labels represent Republicans, the blue labels represent Democrats, the brown labels represent two Independents.

Another study [**?**] that analyzes the data of users and their interactions in the MSN Messenger network found strong relation between users communication behavior (the number of messages exchanged, the total time of chatting, etc) and attributes such as age, gender and even the categories of the query requests!

On the figure **??** the links present the similar votes of U. S. Senators during 2007. With the red labels representing Republicans, the blue labels representing Democrats, the brown labels representing two Independents we can vividly observe the homophily in this network.

Therefore when the values $y_1, y_2, ..., y_n$ are collected with standard RDS, samples are **not independent and uniform**, and the variance of estimator is:

$$\sigma^2_{\hat{\mu}_{SA}} = \mathrm{Var}\left(\frac{y_1 + y_2 + ... + y_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(y_1 + y_2 + ... + y_n) =$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathrm{Cov}(y_i, y_j)\right) = \frac{\sigma^2}{n} + \frac{2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathrm{Cov}(y_i, y_j)}{n^2} =$$

$$= \frac{\sigma^2}{n}\left(1 + \frac{2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathrm{Cov}(y_i, y_j)}{n\sigma^2}\right) = \frac{\sigma^2}{n}\left(1 + \frac{2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathrm{corr}(y_i, y_j)}{n}\right)$$

So variance of estimator is influenced by some correlation factor $f(n)$ where$f(n) > 1$, that depends on how much the values are correlated:

$$\sigma^2_{\hat{\mu}_{SA}} = \frac{\sigma^2}{n}f(n)$$

On one hand, when we increase number of participants the factor σ^2/n decreases, but there is also correlation factor $f(n)$, the bigger nuber of participants, the bigger is this correlation factor. And then in order to improve estimation we could try reduce this correlation factor.

In some literature as in [**?**] the ratio of the variance of RDS estimate to the variance of estimate obtained from independent uniform random sampling is called design effect, d. They warn the users of RDS: to have the same variance using independent uniform random sampling we need n samples while using RDS need dn samples. We can notice that the correlation factor $f(n)$ is exactly the design effect.

1.5 Enhanced RDS

We will state two observations. Combining them we will try to improve current RDS technique.

Observation 1 *Just thinning of sample doesn't help*

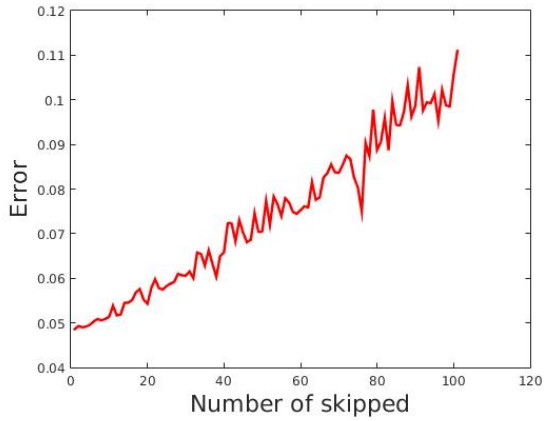
In order to reduce correlation between sampled values one could try to thin out sample. It means that instead of taking all collected values $y_1, y_2, ..., y_n$ into the estimation we can take only, let's say, each second value, $y_1, y_3, ..., y_n$. The samples $y_1, y_3, ..., y_n$ are less correlated then all samples, thus we will reduce correlation factor $f(n)$, so we can expect some improvement. In the same by taking each second values we will degrees in two times the total number of samples. It is not clear if the correlation was reduced enough to compensate this fact. What we observed experimentally (latter we will show it formally) that in general just discarding some samples with the fixed size of recruitment chain will not improve estimation. Experimental results are presented on the figure **??** (a), where we can observe the *mean squared error* of the estimator for different scenarios.

Observation 2 *Skipping reduces variance*

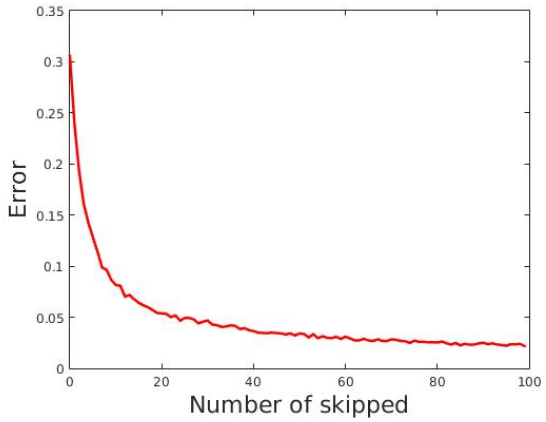
However if the size of the sample is fixed the farther are individuals in the chain of recruitment from each other the better. Let's say that we want exactly 30 participants for study and we have options. We can perform RDS until we reach 30th person and then take each one of them, $y_1, y_2, ..., y_{30}$ in the estimation. Or we can continue RDS until we reach 60th person and then take each second of them $y_2, y_4, ..., y_{60}$ in the estimation. The both variants have the same sample size but in the second scenario the values are less correlated. So we expect that the correlation term will decrease. Again on the figure **??**(b) we can see experimental results and further we will show it formally.

Keeping in mind that we have fixed budget, the second scenario implies that we don't pay to the participants $1, 3, ..., 59$ or we pay to all 60 individuals half of what we paid in the first scenario. But as the motivation to participate in the study is money for the same job people should get the same amount of money.

Now based on the stated observations we can make some conclusions. It can be useful to skip some values between sample, but without reducing sample. But if all individuals involved receive the same amount of money, skipping necessary implies reducing the sample size. At the same time some individuals do less work (individuals that are skipped do not need to come, for example, to the laboratory, and make tests). Therefore, it is reasonably to think that there may be two different incentives: for taking part in the research (making tests, filling questionnaires) and for providing the contacts of other individuals.



(a) 1000 samples are collected with RDS. The sample size is changed as we skip some values. We can observe that error grows when we try to discard each second, each third and so on nodes.



(b) In all cases the sample size is the same. The number of nodes collected with RDS is different. We observe that error decreases when the distance between samples increases

Figure 1.2: Experiments on the data from Project 90

So among people that are willing to participate some of them will be asked both: to make tests and recruit other participants, let's call them *participants* and some of them will be asked only to recruit other participants, let's call them *informators*. As the informators make less efforts they will be paid less.

Let us say that each informator receives C_1 units of money and each participant receives $C_1 + C_2$ units of money. The amount of money that we can spend on the research, the budget, is fixed and is denoted as B. Let n be the length of the chain of all the recruitments (it means that informators and participants in total are n).

Let $k-1$ be the number of skipped samples in the chain between participants (it means that we take only each kth individual as a participant participant, the rest $k-1$ individuals are just informators).

The reason to use informators is to try to reduce correlation by making bigger the distance between participants and therefore less correlation. If the informators were willing to do their job for free then, according to the third observation, we would try to have as more informators between participants as possible. But all the idea of respondent-driven sampling holds on the money incentives, without payment nobody will do anything. For this reason informators should be also paid, but less then actual participants. As we still spend part of the budget on informators the number of participants will decrease with increasing number of informators.

To understand better the idea, let's imagine that we have 60 euros budget, each informator is paid 2 euros, $C_1 = 2€$, each participant is paid 10 euros, $C_2 + C_1 = 10€$. On the figure you can see different scenarios of RDS for the same budget.

On the figure 1 the parameter k is equal to 1. It means that everybody is participant and paid 10 euros. Then with our budget we can collect 6 samples that due to homophily can be highly correlated.

On the figure 2 the parameter k is equal to 2. It means that we take one participant, one informer, one participant and so on. We can see clearly that we went deeper in the network, the

recruitment chain is longer. In this case on our budget we can collect only 5 samples, but they are less correlated than in the previous situation.

Both scenarios require the same budget, so in the terms of money they are equal. But what is better 6 more correlated samples, or 5 but less correlated samples?

If the characteristic that we try to estimate does not correlate between friends or people who have contact then it is useless to discard some values (we just pay for nothing). But if the values are highly correlating intuitively skipping can help a lot. There is a trade-off: on one hand we make the chain longer and reduce dependency between participants. On other hand we still spend money on people who do not bring any information needed for research and finally there will be less participants.

The better scenario will be the one that has the least error. If we are able to quantify the error depending on the number of samples that we skip $k - 1$, we will be able to suggest the sampling method that will bring the most precise result. That is the question that we are going to analyze formally in the next chapter.

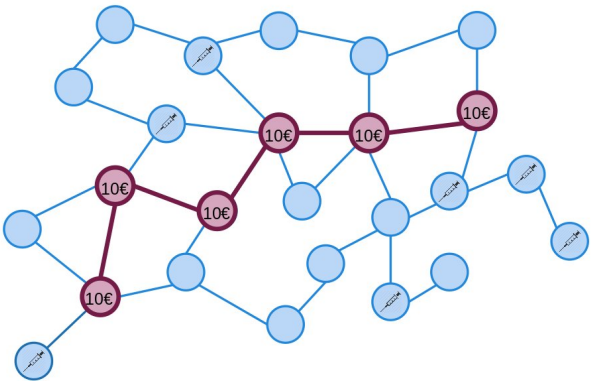


Figure 1.3: Scenario 1

Budget B	60€
C_1	2€
C_2	8€
k	1
Sample size	6
Chain size	6

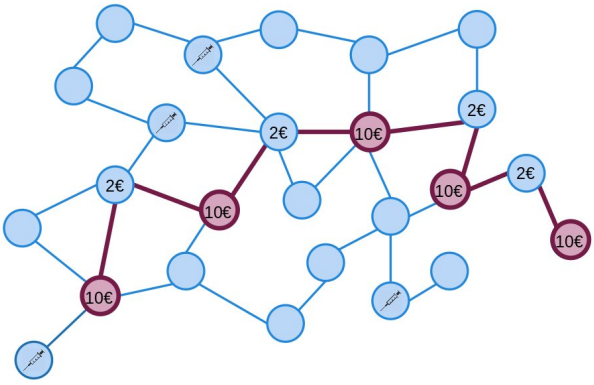


Figure 1.4: Scenario 2

Budget B	60€
C_1	2€
C_2	8€
k	2
Sample size	5
Chain size	9

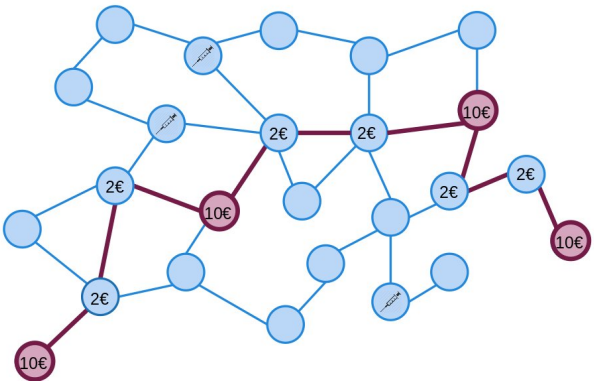


Figure 1.5: Scenario 3

Budget B	60€
C_1	2€
C_2	8€
k	3
Sample size	4
Chain size	10

Chapter 2

Network with values

2.1 Motivation

As it has been already mentioned we will model the network formed by the members of hidden population as a graph. The nodes of this graph correspond to the individuals. The edge between two nodes represents the connection between corresponding persons.

We have several possibilities to obtain graph structure. First, there are multiple available network structures that were collected from the real networks. In particular the Stanford Large Network Dataset Collection [**?**] provides the 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.

The other possibility is to use random graph theory to generate the random graph that will imitate the network structure. There is number of different random graphs. In particular we used Erdős-Rényi graph, random geometric graph, preferential attachments graph.

Our goal is to study mathematically the error of the estimator of the presented enhanced RDS method. Its purpose is to estimate the average value of some attribute on the nodes. For instance, if we have a social network the attribute can be the age, gender of a user. But for now there is nothing to estimate. The Stanford Large Network Dataset Collection contains only real network structures (without attributes on the nodes). All the random graph models give us possibility to generate only the structure of a network. However for the purposes of our problem each node should maintain the value of the attribute, that is going to be estimated.

Of course there are the real network structures where the nodes have some data, but they are scarce. The Project 90 data[**?**], Add health project data[**?**] are two of them. But this is not enough for study purposes. We can not justify our results based only on few examples. What would satisfy us is following. In the same way we can generate numerous random graphs with desired properties, we wanted to have mechanism to generate the values on the nodes of the obtained graph which will represent needed attribute. Having already some graph structure we want to generate the values of the nodes, that will imitate the node's attribute. Let us analyze what do we want from these values and what are the challenges to be encountered in this task.

The simplest idea is to assign values randomly to the nodes independently of the graph structure. For example, we could assign the age of the user according to the uniform distribution or normal distribution or any distribution we want our attribute to be distributed. This approach has an explicit weakness: it does not take into account the homophily, the tendency of people

with connections to have the similar characteristics.

As we already explained the property of homophily is frequently encountered in the real networks. The reason why we do not want to ignore homophily is because the way of performing sampling and the property of homophily together influence on the sampling variance and bias. Further we will count formally the sampling variance for the given network and attribute of the nodes.

So for study purposes we would like to assign values to the nodes of the network in such a way that the value of the node depends on values of its neighbors.

The other point is that the level of correlation can be different within the same network but for different attributes.

The study [**?**] was investigating how the binge drinking is influenced by the position of student in the network of students. The students were labeled according to belonging to one the group: member of a bingering group, liasons, isoletes, etc. The researchers looked for the relation of the episodes of binge drinking per fixed period of time and the student's label. They found strong dependency while the students were young, but not when they became adults. So for the same network the different attributes: binge drinking in school and binge drinking after school have different level of dependency of the friend's behavior.

Regarding what was said above we would like also to be able to tune the correlation in the network.

To sum up we have two goals to be achieved:

First The values on the nodes of the network should have the property of homophily

Second We should have the mechanism to control the level of homophily in the network

In the proceeding we will develop the technique that deals with these two problems.

2.2 Definitions

First we will give some definitions and then we will introduce the algorithm of creating synthetic network with values, where the level of homophily can be tuned.

Let us imagine that we already have the graph with n nodes. To each node i we want to assign a random value X_i from the set of values $V, V = \{1, 2, 3, ..., k\}$.

Instead of looking on distributions of the values on nodes independently, we will look at the joint distribution of values on all the nodes. The desired distribution should take into account the values of the node's neighbors as well.

Let us denote $(X_1, X_2, ..., X_n)$ as $\bar{X}$. We call $\bar{X}$ as a ***random field***. When random variables $X_1, X_2, ..., X_n$ take values $x_1, x_2, ..., x_n$ respectively we call $(x_1, x_2, ..., x_n)$ a ***configuration*** of a random field and we will denote it as $\bar{x}$. As the basement of our distribution we will take Gibbs distribution [**?**].

For simplicity, instead of writing $p(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ where $x_1, x_2, ..., x_n \in V$ we will write $p(\bar{X} = \bar{x})$ or just $p(\bar{x})$.

For the random field $\bar{X}$ we will associate the random number that is called ***global energy*** of the random field that is counted in the following way:

$$\varepsilon(\bar{X}) = \sum_{i \sim j, i \leq j} (X_i - X_j)^2$$

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

Let us turn our attention to the one node i. Then the ***local energy*** on the node i will be defined as:

$$\varepsilon_i(X_i) = \sum_{j|i\sim j} (X_i - X_j)^2$$

Then we can rewrite the expression of the global energy knowing the local energies on all the nodes:

$$\varepsilon(\bar{X}) = \frac{1}{2} \sum_i \varepsilon_i(X_i)$$

When the random field $\bar{X} = (X_1, X_2, ..., X_n)$ takes the configuration $\bar{x} = (x_1, x_2, ..., x_n)$ then the global energy of the configuration $\bar{x}$ is:

$$\varepsilon_i(x_i) = \sum_{j|i\sim j} (x_i - x_j)^2$$

and local energy on the node i is:

$$\varepsilon_i(x_i) = \sum_{j|i\sim j} (x_i - x_j)^2$$

2.3 Gibbs distribution

Now let's consider the following probability of random field $\bar{X}$ to take the configuration $\bar{x}$:

$$p(\bar{x}) = \frac{e^{-\frac{\varepsilon(\bar{x})}{T}}}{\sum\limits_{\bar{x}' \in |V|^n} e^{-\frac{\varepsilon(\bar{x}')}{T}}} \tag{2.1}$$

where T is temperature, $T > 0$.

The reason why it is interesting to look at this distribution follows from the theorem 2.1, p. 260, book [**?**]: *when random field has distribution* **??** *then the probability 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 the node i. If the L is the subset of nodes then X_L will denote the set of random variables of the corresponding nodes. The the last property can be formulated in the following way:

$$p(X_i = x_i | X_{N_i} = x_{N_i}) = p(X_i = x_i | X_{\{1,2,...,n\}\setminus i} = x_{\{1,2,...,n\}\setminus i})$$

This property is called *Markov property* and it is exactly what we want from the values on the nodes: to be dependent of the values on the neighboring nodes.

Moreover for each node i, knowing values of its neighbors, we can write the distribution of values as following:

$$p(X_i = x_i) = \frac{e^{-\frac{\varepsilon_i(x_i)}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\varepsilon_i(x')}{T}}}$$

It is the temperature parameter T that plays very important role of the tuner of homophily level (or the correlation level) in the network. Later we will show some examples for better understanding.

Gibbs distribution found many interesting applications in real-world problems. In particular it lies in the basement of the proposed in [**?**] distributed algorithm for channel selection of the Access Points. The channels should be selected in such way that interference in the network is minimized and the developed algorithm to achieve this uses Gibbs distribution.

2.3.1 Algorithm

Practically speaking direct sampling from the distribution **??** is not so easy. We can just notice that the number of possible configuration $\bar{x}$ is $|V|^n$ (where $|V|$ is size of the values set and n is number of the nodes in graph) as to each from n nodes we need to assign the value from the set V. In this way for the graph with just 100 nodes and 10 possible values it would make up 10^{100} possible configurations. Another difficulty is that in order to sample from derived distribution probability for each of 10^{100} possible configurations should be counted. This would require huge precision from the computer.

That is why we need to develop another way to produce samples from this distribution. For this purpose we are using Gibbs sampler [**?**].

We will regard each configuration $\bar{x}$ as a state of the Markov chain. We will denote as $\bar{x}^k$ the state at the step k. Let us regard the process with following transition probabilities. There is positive probability to transit from one state to another if the corresponding configurations differ only in the value of one node and the values of all other nodes are the same. Let us take two configurations $(x_1, x_2, ...x_i, ..., x_n)$ and $(x_1, x_2, ...x'_i, ..., x_n)$ which differ only in the value of the node i. Whatever the value was on the node i in the first configuration the probability to transfer form the first state to the second is:

$$\frac{e^{-\frac{\varepsilon_i(x'_i)}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\varepsilon_i(x')}{T}}} \tag{2.2}$$

And what is the most interesting, the stationary distribution of this Markov Chain is exactly **??**. So now we need only to run described process for enough amount of time. When the Markov Chain achieves its stationary distribution the process can be stopped. The configuration that corresponds to the final state will be taken from the distribution **??**.

Therefore following algorithm after converging will produce a sample from the distribution **??**.

1. Create random configuration of properties on all nodes.

   On this step the random values are assigned to the nodes according to uniform distribution independently from each other.

2. Choose the node i

   - according to some distribution $q = q_1, ..., q_n$ or
   - visiting each node consequently (periodic Gibbs sampler)

3. For each value $x \in V$ count the local energy on chosen node i as

$$\varepsilon_i(x) = \sum_{j|i\sim j} (x - x_j)^2$$

4. Choose a new value x for the node i according to probability

$$\frac{e^{-\frac{\varepsilon_i(x)}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\varepsilon_i(x')}{T}}} \tag{2.3}$$

   where T is temperature.

5. Continue 2-4 needed number of iterations.

Let us note that in the expression **??** the summation contains only $|V|$ terms. Moreover the energy used in the formula is local, and not global. It means that in order to count this probability the node needs information only about values of its neighbors.

2.3.2 Explanatory example

Presented algorithm lets us assign the values to the nodes according to the distribution **??**. The value on each node will depend from the values of its neighbors. In this way we will imitate the homophily in the network. That was one of our goals. The other goal was to control the level of homoplily. Sometimes we want the value of attribute to be very dependent from the neighbors, sometimes no. The role of a homophily level "tuner" plays the parameter temperature T.

To understand the influence of the temperature on the value distribution **??** we will look at the following example.

Let us say that we have the graph to which nodes we want to assign values $1, 2, 3, 4, 5$. Now let us look only at the vertex A its neighbors B, C, D, E which have assigned values $1, 5, 3, 4$ respectively (figure **??**). And now it is turn of A to choose a value.

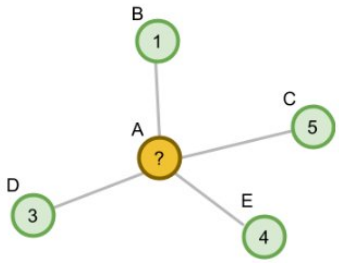


Figure 2.1: Node A and its neighbors

With different values the node A will have different local energy. Let us summarize it in the table.

Value on the node A	1	2	3	4	5
Corresponding energy	29	15	9	11	21

We can see that with different values the local energy on the node A will be different. And according to the distribution **??** the values that bring low energy is more preferable: the less is the energy that is caused by particular value the more probable it will be chosen. In this example the lowest energy corresponds to the value 3, so it will have the highest chance to be selected.

To feel the impact of temperature we will present the distribution of values for different temperature T in the next table.

Temperature	$p(A=1)$	$p(A=2)$	$p(A=3)$	$p(A=4)$	$p(A=5)$
0.1	0.0000	0.0000	1.0000	0.0000	0.0000
1	0.0000	0.0022	0.8789	0.1189	0.0000
10	0.0483	0.1957	0.3566	0.2920	0.1074
100	0.1769	0.2035	0.2161	0.2118	0.1917
10000	0.1998	0.2000	0.2002	0.2001	0.1999

We can see that when the temperature is 0.1 the probability to choose the value 3, that has the lowest energy, is 1. So the values of the neighbors B, C, D, E indeed impact a lot on the value of the node A. In this case, the value of A is dictated by the values of its neighbors. This is exactly what we wanted to achieve: that characteristic of the person is influenced by its contacts. In this case we can say that the level of homophily is high. However we may want to have such dependency but not so strong.

We can try to play with the temperature parameter. As the temperature increases we can observe more "randomness". Thus when the temperature is 1 the value 3 is still very probable, but there is also some positive probability that values $2, 4$ will be picked. And the more we increase temperature the higher becomes probability to choose the value different from 3. We can interpret this as the person still depends on the connections but has also some "free choice".

When the temperature is really high the choice of value will not almost depend on the values of its neighbors.

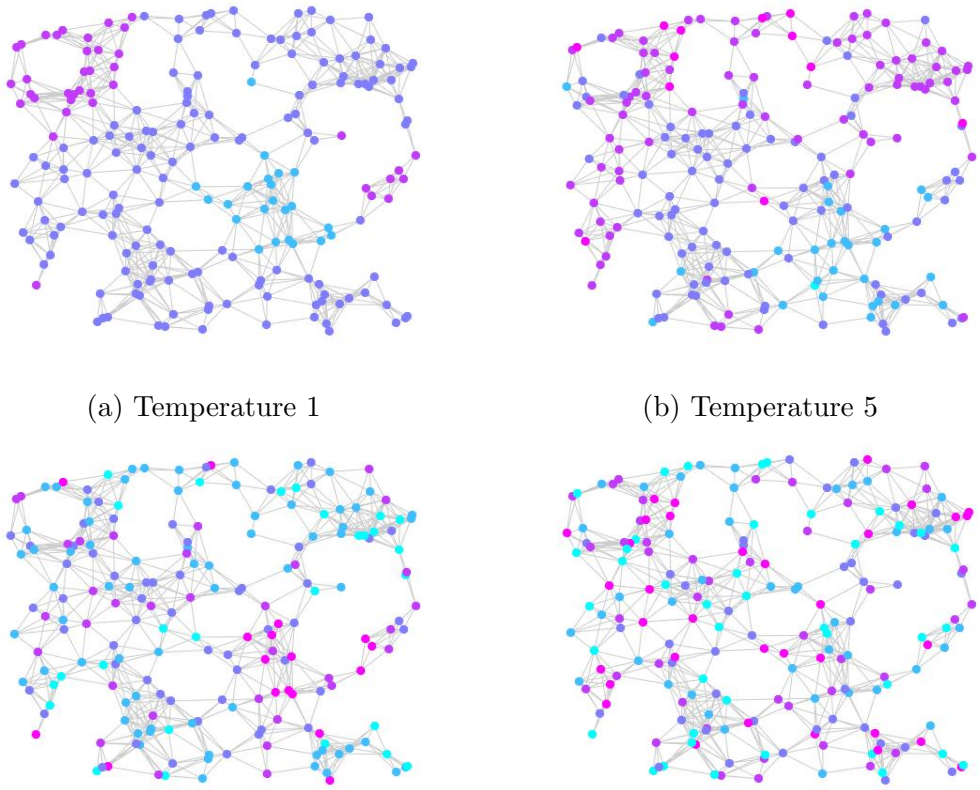
With this example we observed that the distribution favors the values that bring local energy of the node to minimal and the lower the temperature is the more favorable they are. For us it means that the low temperature corresponds to the high level of homophily. If we want to decrease the level of homophily we should increase the temperature.

2.3.3 Demonstration of random graphs with values

In order to demonstrate that proposed model works first we will show the generated graphs with values to see the result visually and then we will look at the ways to measure level of values dependency in the graph.

On the figure **??** presented 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. The values are depicted on the pictures as colors.

From the pictures we can observe that the level of dependency between values of the node changes with different temperature. When temperature is 1 we can distinctly distinguish clusters. With increasing temperature, 5 and 20, the values of neighbors are still similar but with more and more variability. When temperature is very high then the values seem to be assigned randomly.



(a) Temperature 1

(b) Temperature 5

(c) Temperature 20

(d) Temperature 1000

Figure 2.2: RGG(200, 0.13) with generated values for different temperature

In order to show formally that the level of homophily decreases when temperature increases we will look at the correlation between values of the nodes that we see during the random walk on the graph.

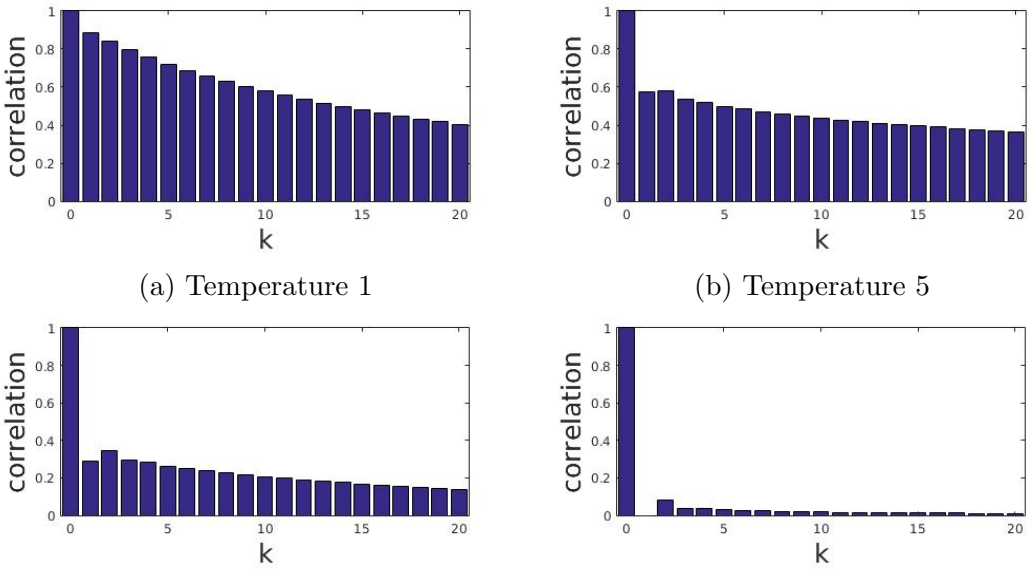
Let $Y_0, Y_1, ..., Y_i, ...$ be the sequence of values on the nodes that we observe during the random walk. Let us say that we start random walk with stationary distribution, so the values $Y_0, Y_1, ..., Y_i, ...$ have the same stationary distribution. Then correlation between two values Y_i and Y_{i+k} depends only on the distance k in the sequence $Y_0, Y_1, ..., Y_i, ..., Y_{i+k}, ...$ between them [**?**]:

$$\text{corr}(Y_i, Y_{i+k}) = \text{corr}(Y_0, Y_k)$$

On the figure **??** we present correlation between values depending on this distance k for the graphs shown above. For each k the height of the bar will correspond to the $\text{corr}(Y_0, Y_k)$. In this way we can assure us that the higher is temperature the less correlated values of neighbors are.

2.4 Expected energy in steady state

The question that is still not clear about the algorithm is when to stop it. How many steps are enough to perform in order to claim that achieved configuration is indeed sampled from the Gibbs distribution? In order to answer this question we should try to understand if we can detect



(a) Temperature 1

(b) Temperature 5

(c) Temperature 20

(d) Temperature 1000

Figure 2.3: Correlation of the values of the nodes observed during the random walk depending on the difference in their order for different temperature

somehow that the process described in the subsection **??** has reached its stationary state.

For this purpose we can try to use the notion of the global energy that can be counted for each configuration. Maybe it can signalize us when it is safe to stop the algorithm.

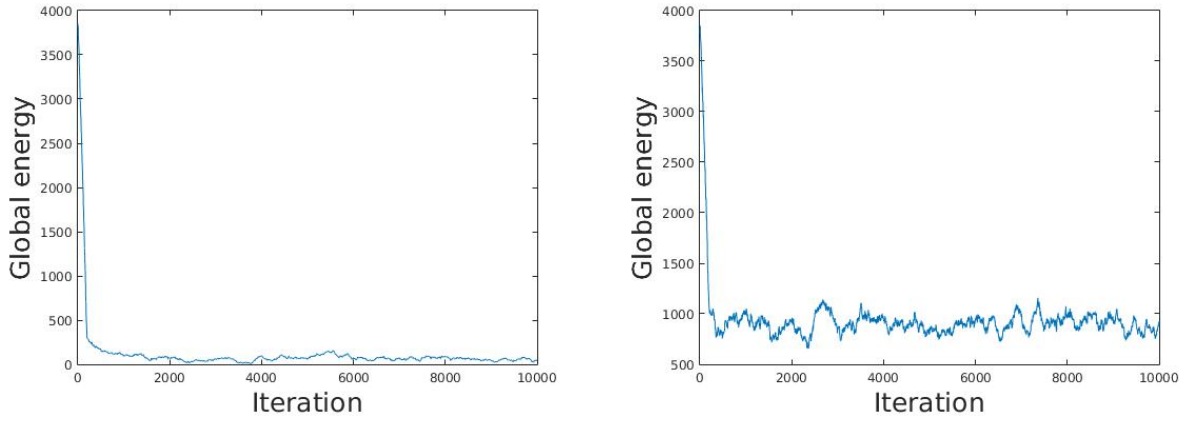
Let us look how the global energy of the configuration is changing during the steps of algorithm. For each configuration $\bar{x}$ t the global energy is:

$$\varepsilon(\bar{x}) = \sum_{i \sim j, i \leq j} (x_i - x_j)^2$$

For each iteration of the algorithm when the configuration is changed the global energy is counted. The results are presented on the figure **??** .

At the beginning of the algorithm values are assigned to all nodes uniformly from all possible values. We can observe on the figure **??** that the energy for this first random configuration is the highest. As values on the nodes are changing according to Gibbs sampling the energy decreases (with some variation). After about 200 steps (it means that up to this time each node updated its state once) we can see that the changes of energy do not exceed some thresholds. So after some time energy comes to some value, stabilizes and does not change a lot. We will call this value ***stationary global energy***. Knowledge about its expected value and variance can indicate us when it is time to finish the algorithm. For simplicity we will refer to the expected value of stationary global energy as to ***expected global energy***. If we can predict what is the expected global energy, we have the idea when the algorithm can be stopped.

There is another reason why we would like to predict energy. We saw previously that by varying the temperature T we can change how strongly values of the neighbors are correlated: low temperature brings high correlation of values and high temperature brings almost random assignment of values. However it is impossible to use only temperature T as a metric of values correlation. Temperature by itself does not contain a lot of information. For the graph RGG(200,



(a) Temperature 2 (b) Temperature 10

Figure 2.4: Energy changing with iterations of the algorithm

0.13) the temperature 200 can be considered as high (because values are not really correlated) but for the graph RGG(2000, 0.06) it is not the case. We can not judge the level of correlation only by temperature, we should take into account also the other factors as number of nodes, number of edges, structure of graph, possible values that can be assigned to nodes.

That is why we need another metric. And again the knowledge about expected global energy can help us. On the figure **??** we can see two illustrations of energy changing with time for different temperature. On the both pictures we see the result of the same algorithm, for the same graph, starting from the same initial configuration, but using different temperatures, and we observe that it comes to different expected global energy.

Then the more appropriate metric for the level of homoplily can be following number: in how much times energy decreases from its initial value (that corresponds to the random configuration) to its stable value (expected global energy).

The problem here is that the algorithm needs the temperature parameter. So after deciding that we want to decrease the initial energy in 5 times we still need to understand which temperature will bring the system to this target energy.

For these purposes we were interested in finding dependency of the global energy from the temperature.

2.4.1 Analysis

For the reasons discussed above we would like to know the expected value of the global energy.

Let the random variable $\varepsilon(\bar{X})$ be total energy of the graph with random field $\bar{X} = (X_1, X_2, ..., X_n)$. We can write expected energy by definition as:

$$E[\varepsilon(\bar{X})] = \sum_{\bar{x}} p(\bar{x}) \cdot \varepsilon(\bar{x}) \tag{2.4}$$

where probability of one particular configuration $\bar{x}$ is

$$p(\bar{x}) = \frac{e^{-\frac{\varepsilon(\bar{x})}{T}}}{\sum\limits_{\bar{x}' \in |V|^n} e^{-\frac{\varepsilon(\bar{x}')}{T}}} \tag{2.5}$$

Both in **??** and in **??** formulas summation is over all possible configurations and the number of all possible configuration is huge, $|V|^n$. One of the ways to calculate such kind of expressions would be following: using Gibbs sampling, run the algorithm until convergence large enough amount of times, and then average the result. In this way we can build empirical dependency of expected energy from the temperature. In order to have these empirical results we need to perform the algorithm large amount of times and for different temperature. Such simulations can take a lot of time, especially because we do not know when it is safe to stop algorithm. So actually we would like to have some theoretical results (at least approximated, just to have a notion about energy dependency from the temperature).

First, let us rewrite global energy as:

$$\varepsilon(\bar{X}) = \sum_{i \sim j, i \leq j} (X_i - X_j)^2$$

Then the expected global energy of a random field is:

$$E[\varepsilon(\bar{X})] = \sum_{i \sim j, i \leq j} E[X_i - X_j]^2$$

In order to count this expression we need to know the distribution of values for each node, $X_1, X_2, ..., X_n$. Moreover, we need to know the correlation between random variables X_i, X_j for all pairs i, j.

In reality both these requirements are difficult to satisfy.

In fact, it seems that we know the distribution of values on the nodes. So the first demand should be easy. We have already written that the values on the node i are distributed in the following way:

$$p(X_i = x) = \frac{e^{-\frac{\varepsilon_i(x)}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\varepsilon_i(x')}{T}}} \tag{2.6}$$

But if we look closer at the local energy on the node i, $\varepsilon_i(x)$, we will notice that it implies that the values on the neighboring nodes are known:

$$\varepsilon_i(x) = \sum_{j | i \sim j} (x - x_j)^2$$

It means that distribution **??** is actually conditional on the values of the neighbors. We can rewrite it as:

$$p(X_i = x) = \frac{e^{-\frac{\varepsilon_i(x)}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\varepsilon_i(x')}{T}}} = \sum_{a_1, ... a_j \subset V} \prod_{j | j \sim i} p(X_j = a_j) \frac{e^{-\frac{\sum\limits_{j | j \sim i} (x - a_j)^2}{T}}}{\sum\limits_{x' \in V} e^{-\frac{\sum\limits_{j | j \sim i} (x' - a_j)^2}{T}}}$$

We can see that this expression includes also probabilities for other nodes to have some particular values. So the probability to have a value on the node i depends on the values of its neighbors that are also dependent from their neighbors and so on. It means that it can be very difficult or even impossible to write the expression for the distribution of the values on the node in explicit way.

The second demand of knowing correlation for all pairs X_i, X_j require to know the joint distribution of these random variables. That brings us to the same problem: no way to write explicitly the expression of this distribution.

That's why in order to calculate expected energy at least approximately we will make some assumptions.

The same expected value

The experiments showed us that the expected value of the each variable $X_1, X_2, ..., X_n$ is the same and equals to the average value of the values from the set V.

$$E[X_i] = av_V$$

We didn't show it formally, that's why we will write that this is an assumption. But we have reasons to believe that it is true and can be proved.

Assumptions about no correlation between neighbors

We will make an assumption that values are assigned independently of each other. Of course it is not true. The distribution of the values on a node takes into the account the values of its neighbors. By making this assumption we will make some error, but we can get the approximated expression. If the approximation is close to the reality it can give us at least the idea about energy-temperature dependency.

Let N_i be the number of the neighbors of the node i. So if there is no correlation between the values assigned to the nodes, then we can write global energy as:

$$E[\varepsilon(\bar{X})] = \sum_{i\sim j, i\leq j} E[X_i - X_j]^2 = \frac{1}{2}\sum_{i\sim j} E[X_i - X_j]^2 =$$

$$= \frac{1}{2}\sum_{i\sim j}\left(\mathrm{Var}(X_i - X_j) + E[(X_i - X_j)]^2\right) = \frac{1}{2}\sum_{i\sim j}\left(\mathrm{Var}(X_i - X_j) + (E[X_i] - E[X_j])^2\right) =$$

$$= \frac{1}{2}\sum_{i\sim j}\left(\mathrm{Var}(X_i) + \mathrm{Var}(X_j) - 2\mathrm{Cov}(X_i, X_j)\right) = \frac{1}{2}\sum_{i\sim j}\left(\mathrm{Var}(X_i) + \mathrm{Var}(X_j)\right) =$$

$$= \frac{1}{2}\sum_{i\sim j} 2N_i\mathrm{Var}(X_i) = \sum_{i\sim j} N_i\mathrm{Var}(X_i) \tag{2.7}$$

Special case

Let us look at the case when the values are assigned to the nodes according to the same distribution in independent way (again there is no correlation between assigned values). Let m be the number of edges in the graph. Then the expression for energy becomes:

$$E[\varepsilon(\bar{X})] = \sum_{i \sim j} N_i \mathrm{Var}(X_i) = 2m\mathrm{Var}(X_i)$$

where X_i and X_j are random variables with the same distribution.

In particular we can compute expected energy in this way for initial random configuration where the values are assigned to the nodes independently and uniformly from all possible values in V. When X_i is distributed uniformly in V the variance of X_i is:

$$\mathrm{Var}(X_i) = \frac{|V|^2 - 1}{12}$$

Then the expected energy of the graph on random field $\bar{X}$ is

$$E[\varepsilon(\bar{X})] = m\frac{|V|^2 - 1}{6}$$

This value is the expected global energy of the initial configuration. It means that there is no need to conduct numerous experiments in order to count the average global energy of initial configuration. This formula gives us exact answer.

Assumptions about values distribution

We saw that it is impossible to write explicitly the distribution of the values on one node. In order to simplify the expression for values distribution on the node i we will also make some assumptions. First, let's say that all neighbors of the node i have the value $av_V, av_V = average(V)$. If for all $j \in N_i : X_j = av_V$ then $p(X_j = av_V) = 1$. With this assumption probability that the node i will have value $x \in V$ is

$$p(X_i = x) = \frac{e^{-\frac{N_i(x - av_V)^2}{T}}}{\sum\limits_{x' \in V} e^{-\frac{N_i(x' - av_V)^2}{T}}} \tag{2.8}$$

Then combining expressions **??** and **??** we can write expected energy as:

$$E[\varepsilon(\bar{X})] = \sum_{i \sim j} N_i Var(X_i) = \sum_{i \sim j} N_i \left(E[x_i]^2 - (E[x_i])^2\right) = \sum_{i \sim j} N_i \left(\frac{\sum\limits_{i \in V} i^2 e^{-\frac{N_i(i - av_V)^2}{T}}}{\sum\limits_{x' \in P} e^{-\frac{N_i(x' - av_V)^2}{T}}} - av_V^2 \right)$$

To simplify even more we can assume that each node has the same following distribution of values:

$$p(X_j = x) = \frac{e^{-\frac{d(x - av_P)^2}{T}}}{\sum\limits_{x' \in P} e^{-\frac{d(x' - av_P)^2}{T}}} \tag{2.9}$$

where d is average degree of the graph.

Then expected energy is counted in the following way:

$$E[\varepsilon(\bar{x})] = 2mVar(x_i) = 2m \left(\frac{\sum\limits_{i \in P} i^2 e^{-\frac{d(i-av_P)^2}{T}}}{\sum\limits_{x' \in P} e^{-\frac{d(x'-av_P)^2}{T}}} - av_P^2 \right)$$

2.4.2 Results

On the figure **??** we can observe predicted energy with assumptions including that the values on the nodes are distributed as in **??** when we consider for each vertex its number of neighbors (it is called prediction ALL-N), predicted energy with assumptions including that the values on the nodes are distributed as in **??** when for each vertex we take the average degree (it is called prediction AV-D) and the expected energy counted with simulations.

We can see that both predictions coincides well with the experiments. The prediction **??** is more accurate where each vertex has distribution of values according to its degree. As the number of neighbors influences on the distribution and prediction **??** ignores it, it performs worse than prediction **??**.

Even if the result is approximated it can give us great intuition about energy-temperature dependency and save time.

To understand the practical significance of this result, let us look at the part of Facebook graph **??** where the values are generated with the Gibbs sampler. It took almost 4 hours to build the energy-temperature dependency and less then 1 second to build both predictions. The prediction **??** is very accurate. Having this prediction there is no difficulty to understand what temperature we need to take to reduce the initial energy let's say in five times.

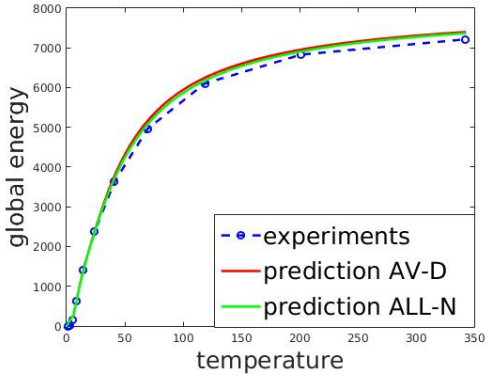
2.4.3 Conclusions of the section

In this chapter we challenged us with problem to generate the values on the given graph structure that satisfy our two requirements. First, values should imitate homoplily in the network: the value of a node should be correlated with values of its neighbors. For this purpose we introduced the algorithm that allows to assign values to the nodes according to Gibbs distribution. Gibbs distribution has exactly the property that we need: the value of a node is dependent of its neighbors.

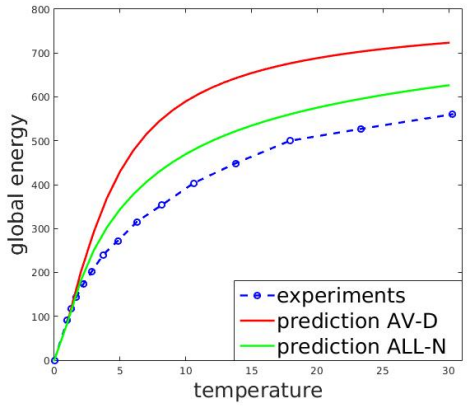
Our second requirement, was to be able to control the homophily level in the network. We showed how the temperature parameter deals with this problem. Changing this parameter we can tune the level of correlation of the values.

Finally, we discussed the problem of energy prediction that has two goals. First, it gives us the notion about when to stop the algorithm. Second, it enables us to introduce another metric to have an idea about correlation level in the network. We presented the theoretical results for energy prediction that are approximated ones. However this approximation is close enough to solve the two desired goals.

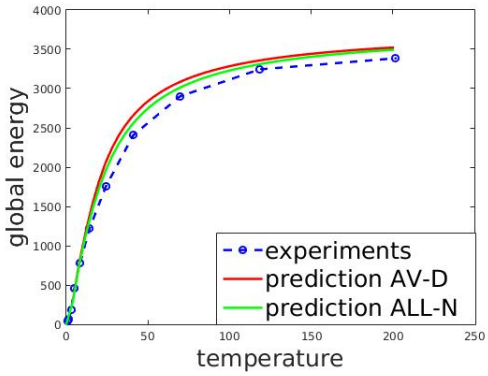
The problem about generating synthetic network with values was important for the goals of my internship subject. However it is general result that can be applied wherever the researchers need to imitate not only the network but the values on the nodes as well.



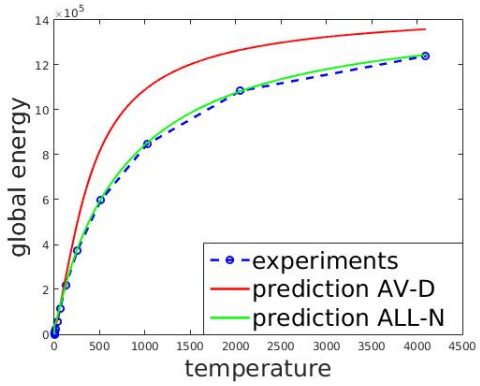
(a) Random ER graph with 200 vertices and values $[1, ..., 5]$ p = 0.1



(b) Preferential attachment graph with 200 vertices, 1 link for new arriving node and values $[1, ..., 5]$



(c) Random geometric graph with 200 vertices, radius 0.13 and values $[1, ..., 5]$



(d) Part of Facebook graph [**?**] with generated values from the Gibbs sampler

Figure 2.5: Energy prediction for different graphs

Chapter 3

Mathematical model

3.1 Error prediction

Now, when we have the network, where each node maintain the value we can begin to study formally described in the section **??** method of enhanced RDS.

Due to homoplily in the network and the way the respondent-driven sampling is performed the variance of the estimator will be different from the case if it was just independent uniform sampling. Studying variance is important for multiple reasons. It essential for building confidence intervals. It is the factor that influences on the error of estimator. When we have multiple estimators and we know the error of each of them we have the instrument to compare them.

In this section we will take the sample average (SA) as the estimator for the population mean. In order to decrease the variance of the estimator we will increase the distance between samples as it was described in the section **??**. We will look at the different possibilities of sampling hidden populations in the conditions of limited budget. For each scenario we will have the different error of the estimator. We are going to choose the scenario, that has the minimum error of the estimator.

3.1.1 Variance prediction

Let $Y_1, Y_2, Y_3, ..., Y_n$ be the samples that are taken during the random walk. We will take the average value of the samples as an estimator of mean population value:

$$\widehat{\mu}_{SA} = \frac{Y_1 + Y_2 + ... + Y_n}{n}$$

We noted that the variance of the estimator $\widehat{\mu}_{SA}$ is influence by the correlation factor as the random variables $Y_1, Y_2, Y_3, ..., Y_n$ are correlated:

$$\sigma^2_{\widehat{\mu}_{SA}} = \frac{\sigma^2}{n} f(n)$$

Let's look what will happen with variance in the suggested enhanced RDS.

To remind the notation, we denote budget as B, n individuals that we see during the RDS receive C_1 units of money for recruiting other individuals, each kth person from them take also part in the study and receives C_2 additional units of money.

In this way $\frac{n}{k}$ individuals from n are participants. Then the following equality should be true:

$$B = nC_1 + \frac{n}{k}C_2$$

From this equality we can see that with the fixed budget B and skipping $k-1$ individuals between participants, the length of the chain n can be:

$$n = \frac{kB}{kC_1 + C_2}$$

When each kth person is a participant, it means that for real we have only values $Y_k, Y_{2k}, ..., Y_{\frac{kB}{kC_1+C_2}}$ that we can take for the estimation. In this way the number of participants m is:

$$m = \frac{B}{kC_1 + C_2}$$

Then for each scenario with the fixed budget B depending on the k the estimator is:

$$\widehat{\mu}_{SA}^k = \frac{Y_k + Y_{2k} + ... + Y_{\frac{kB}{kC_1+C_2}}}{\frac{B}{kC_1+C_2}}$$

We can see that the number of participants depends on k and the bigger is k the less participants we have. Also we observed experimentally that the bigger is k (with the fixed sample size) the less is the correlation. So actually correlation factor depends on k. We can write the variance of estimator in the next form:

$$\sigma^2_{\hat{\mu}_{SA}^k} = \frac{\sigma^2}{\frac{B}{kC_1+C_2}} f(k)$$

Now we have to find the expression for the correlation factor $f(k)$.

3.1.2 Geometric correlation

First we will find correlation for simple example and then we will generalize it for any kind of graph.

For now we will not care about graph structure. Let's assume that our collected samples $Y_1, Y_2, ..., Y_n$ are correlated in the known way:

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

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. We will refer to this model as to *geometric model*. Then we can write the variance of the mean estimator as:

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

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

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

$$= \frac{\sigma^2}{n}\left(n + 2n\frac{\rho-\rho^n}{1-\rho} - 2\rho\sum_{i=0}^{n-2}(\rho^{i+1})'\right) =$$

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

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

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

In the end we have the following expression:

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{\sigma^2}{n}\frac{1-\rho^2-2\rho/n+2\rho^{n+1}/n}{(1-\rho)^2} \tag{3.1}$$

From here we can get that correlation factor is:

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

It can be shown that this factor $f(n)$ is increasing function of $n, (n > 1)$ and it has its minimum 1 when $n = 1$. It is clear, when there is only one individual there is no correlation, because there is only random variable Y_1. When new participants are invited, the correlation increases due to homophily as we explained earlier.

Let's look what happens to the correlation factor when n goes to infinity:

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

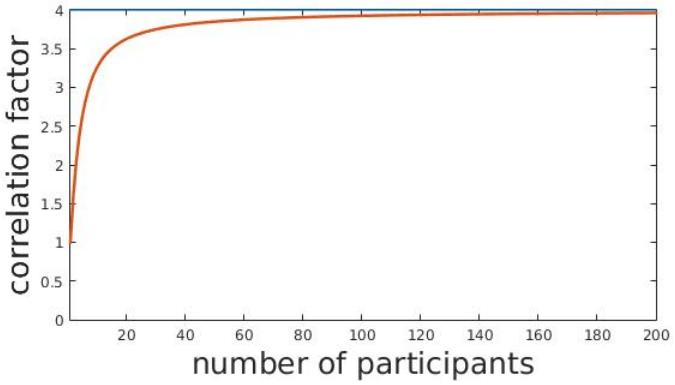


Figure 3.1: Correlation factor depending on the number of the participants when ρ is 0.6

Using the following approximation the expression for the sample variance becomes much more easier:

$$\mathrm{Var}_{approx}\left[\bar{Y}\right] = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$$

Approximation is close when n is big enough. To compare approximated expression with original one, look at the figure **??** where 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 is much simpler and some facts that are very difficult to prove for real Var $\left[\bar{Y}\right]$ become easier for the approximation.

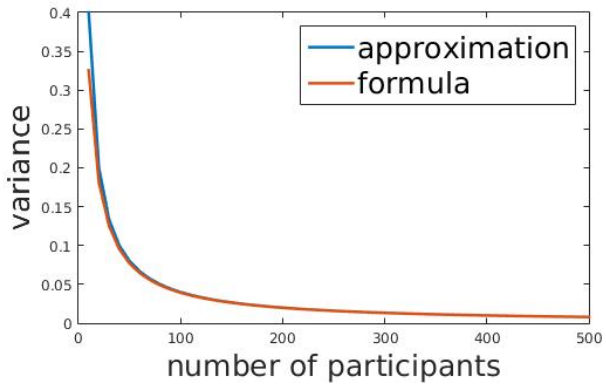


Figure 3.2: $\rho = 0.6$

On the figure **??** we can compare the variance for different level of correlation.

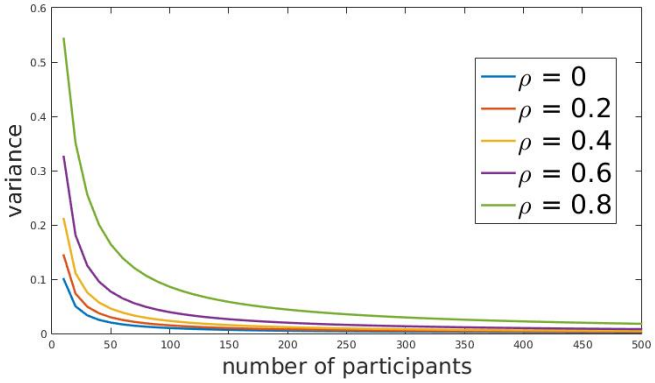


Figure 3.3: $\rho = 0.6$

3.1.3 Variance with skipping

Let's say that now we collected nk samples $Y_1, Y_2, Y_3, ..., Y_{kn}$ that are correlated in the same way as in previous subsection. We will take each k sample and look at the variance of the next random variable:

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

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

$$\text{corr}(Y_{ik}, Y_{(i+h)k}) = \rho^{kh}$$

If we introduce new random variables $Z_1, Z_2, ..., Z_n$ such that $Z_1 = Y_k, Z_2 = Y_{2k}, Z_3 = Y_{3k}, ..., Z_n = Y_{nk}$, $\bar{Z} = \bar{Y}^k$ and denote ρ^k as r, $r = \rho^k$. Then:

$$\text{corr}(Z_i, Z_{i+h}) = \text{corr}(Y_{ik}, Y_{(i+h)k}) = \rho^{kh} = r^h$$

To sum up we have random variables $Z_1, Z_2, ..., Z_n$ where the correlation between any two of them depends on the distance in the chain in the following way: $\text{corr}(Z_i, Z_{i+h}) = \rho^{kh} = r^h$. For this problem we already know the variance of $\bar{Z}$:

$$\text{Var}\left[\bar{Z}\right] = \frac{\sigma^2}{n}\frac{1 - r^2 - 2r/n + 2r^{n+1}/n}{(1-r)^2}$$

Or approximation:

$$\text{Var}\left[\bar{Z}\right] \simeq \frac{\sigma^2}{n}\frac{1+r}{1-r}$$

Let's return to the previous notation and then we get:

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

Or approximated expression:

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

3.1.4 In RDS context

Returning to the RDS context we can finally write the variance for different scenarios with the same budget. On the same budget B, taking each k node as a participant we will collect samples $Y_k, Y_{2k}, ..., Y_{\frac{kB}{kC_1+C_2}}$, where number of participants is $m = \frac{B}{kC_1+C_2}$.

And if the samples are correlated as in the precvious section then variance of the estimator, depending on k, will be:

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

Or approximated:

$$\sigma^2_{\hat{\mu}_{SA}}(k) = \frac{\sigma^2}{\frac{B}{kC_1+C_2}}\frac{1+\rho^k}{1-\rho^k} \tag{3.2}$$

We can regard this expression as the function of k. Now in order to decide, how much samples to skip between the participant, we need to find the minimum of it.

On the figure **??** we can observe how the variance **??** changes for different level of dependency among values.

First let's observe what happens to the factors of the variance when we increase the k. The number of participants decreases, so the left factor in the expression **??** $\frac{\sigma^2}{\frac{B}{kC_1+C_2}}$ increases. In the same time the participants are less correlated, the right factor $\frac{1+\rho^k}{1-\rho^k}$ decreases. And the behavior of the variance depends on which factor is "stronger".

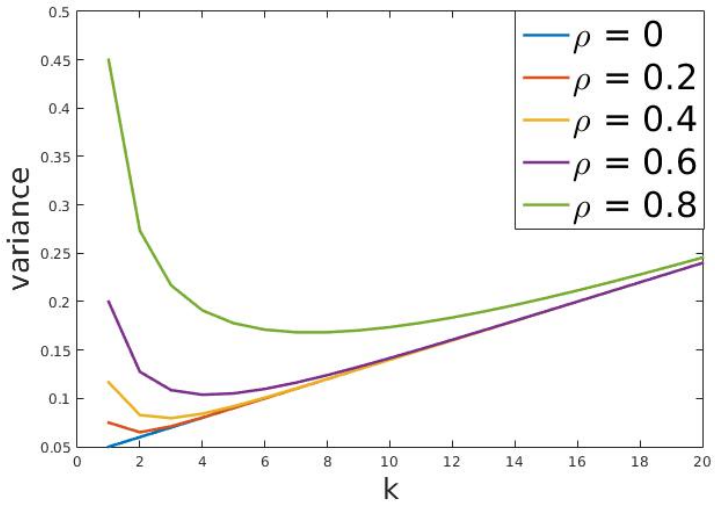


Figure 3.4: Variance with the formula **??** when $B = 100, C_1 = 1, C_2 = 4$

Let's for now concentrate on the case when $\rho = 0.8$. Starting when the $k = 1$ we observe that variance decreases when k increases. It means that correlation is very high and by skipping some values we reduce it significantly for the variance. When $k = 7$ we observe that function reaches its minimum. This result says that in these settings by taking only each seventh individual as a participant we will obtain the minimum error. As k increases further, the variance also start to increase slowly. It signals us that by skipping more that 7 persons, we will not decrease significantly the correlation, Trying to skip more then seven individuals we will just waste budget on skipping without any purpose.

We have a trade-off here. If we skip too many values we can drastically decrease the number of participants. And the opposite, if we don't skip at all the correlation between values can be too high.

We desire to find the value k when the variance is minimal. This values k depends on the level of correlation, that is expressed here by parameter ρ. Thus we observe that when we take lower values of ρ the desired value of k decreases. This coincides with our intuition: the lower is dependency, the less values we need to skip. Finally we see, that in case of no correlation ($\rho = 0$) it is useless to skip.

By studying the variance function, we can notice some interesting properties. In particular, the observations that we stated in the scetion **??** were based on the experiments and human logic. Now, we can show it formally.

Observation 1 *Just thinning of sample doesn't help*

Just thinning means that when all the samples are collected $y_1, y_2, ..., y_n$ we can try to take only k part of them. However to collect each sample we spend equal amount of money, let's denote it as $(C_1 + C_2)$. It means that we should look on function:

$$f(k) = \frac{\sigma^2}{B/(k(C_1 + C_2))} \frac{1 + \rho^k}{1 - \rho^k}$$

This function is strictly increasing when $k \geq 1$. It means that just thinning the sample can only increase the variance.

Observation 2 *Skipping reduces variance*

Here we have the fixed sample size n. However we can increase the distance k between samples as much as we want for free, $C_1 = 0$. Then we should look at the function:

$$f(k) = \frac{\sigma^2}{n} \frac{1+\rho^k}{1-\rho^k}$$

which is decreasing function.

3.1.5 Trying to use the result

Next we tried to use the derived result. In order to use the function **??** to find the desired k which minimizes the variance we need to define parameter ρ. What we tried is to take as ρ is the average correlation between the immediate neighbors when the graph and the values on its nodes are given.

However, the suggested k did not coincide with the k found with the experiments (look the figure **??**). The value k that is suggested by formula was always underestimated comparing with the real one. The explanation could be following.

According to geometrical model the values $Y_1, Y_2, ..., Y_n$ have correlation $\text{corr}(Y_i, Y_{i+h}) = \rho^h$. Correlation between values Y_1, Y_{101} is of the power 100. However when we walk on the graph the values Y_1 and Y_{101} can appear to be the values of the immediate neighbors. And when in reality the values of the corresponding nodes can be pretty much correlated, according to the geometrical model their correlation is so small that can be neglected.

The correlation between values is underestimated, except for the correlation of the two consecutive samples Y_i, Y_{i+1}. Therefore the correlation factor and its influence are underestimated. And when in reality skipping one more value could significantly reduce correlation, the geometrical model can show that it is useless.

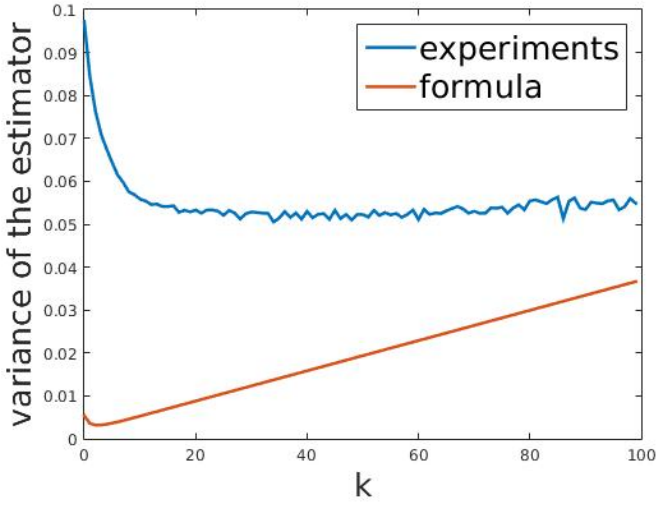


Figure 3.5: Variance of the estimator according to the formula **??** and counted with the experiments scenarios for different k. The graph is the part of Facebook network [**?**] with generated values from the Gibbs sampler

Therefore, the model is too simplified to find the exact result. However, its application can be very important. Knowing that the result from the model is underestimated, it can give us the *minimal* number of node to skip. And the result may not be the best possible, but comparing to the standard scenario without skipping has lower error. So this model will give us the lower bound on the value k.

In the next section we will generalize it on any graph with any values.

3.1.6 General case

Due to the fact that the correlation between samples was oversimplified the results of geometric model were not exactly the same that in reality. This result could give us the lower bound on the value k, but we challenged us to find it precisely.

In the previous example we were able to write the variance of the estimator because the correlation between all the random samples was known.

So our first goal is to generalize formula for the variance of the estimator when the values $Y_1, Y_2, ..., Y_n$ are collected with the random walk on the graph.

Let $f = (f_1, f_2, ..., f_n)$ be the values of the attribute on the nodes $1, 2, ..., n$. First, let P be the transition matrix of the random walk. We consider that probability for the individual to choose any of his friend is the same. Therefore, if the random walk visits the node i then one of the neighbors of this node will be chosen with probability $\frac{1}{d_i}$. Then we can write the transition matrix:

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

The stationary distribution of the random walk is:

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

Let Π be the matrix that consist of n rows, where each row is the vector π. Let $f = (f_1, f_2, ..., f_n)$ be the values of the attribute on the nodes $1, 2, ..., n$.

We consider also that chain starts from initial distribution π. Then covariance between the random values Y_i and Y_j depends only on $j - i$:

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(Y_0, Y_{j-i})$$

And then the formula for the variance:

$$\mathrm{Var}(Y_0) = < f, f >_\pi - < f, \Pi f >_\pi$$

and for the covariance:

$$\mathrm{Cov}(Y_0, Y_h) = < f, (P^h - \Pi) f >_\pi$$

where $< a, b >_c$ is weighted scalar product and if $a = (a_1, ..., a_n), b = (b_1, ..., b_n), c = (c_1, ..., c_n)$ then:

$$< a, b >_c = \sum_{i=1}^{n} a_i b_i c_i$$

To see, how these formulas were derived, consult 6 chapter of the book [**?**].

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

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

$$= \frac{1}{n^2}\left(n(< f, f >_\pi - < f, \Pi f >_\pi) + 2\sum_{i=1}^{n}\sum_{j|i<j}^{n} f, (P^{j-i} - \Pi)f >\right) \tag{3.3}$$

In this way we know the variance of the SA estimator, the expression is quite cumbersome. Another problem is that practically speaking computing the large powers of the matrix P can take a lot of time. Therefore, we will try to simplify this expression.

Let's look at auxiliary matrix P^*. Let D be the matrix $n \times n$ where $d_{ii} = \pi_i$ and $d_{ij} = 0$ if i is different form j. Auxiliary matrix P^* is build in the following way:

$$P^* = D^{\frac{1}{2}} P D^{-\frac{1}{2}}$$

Then if $\lambda_i (i = 1..r)$ are eigenvalues, v_i are corresponding right eigenvectors and u_i are corresponding left eigenvectors of the matrix P^*:

$$P^h - \Pi = \sum_{i=2}^{r} \lambda_i^h v_i u_i^T \tag{3.4}$$

For more explanation refer to the chapter 6 of the book [**?**].

Using formulas **??** and **??** and reasoning similar to the case with the geometric model we derived following formula for the variance of the estimator SA:

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{1}{n}\sum_{i=2}^{r} \frac{1 - \lambda_i^2 - 2\frac{\lambda_i}{n} + 2\frac{\lambda_i^{n+1}}{n}}{(1-\lambda_i)^2} < f, v_i >_\pi^2 \tag{3.5}$$

It means that for general model the correlation factor $f(n)$ is:

$$f(n) = \frac{1}{\mathrm{Var}(Y_0)}\sum_{i=2}^{r} \frac{1 - \lambda_i^2 - 2\frac{\lambda_i}{n} + 2\frac{\lambda_i^{n+1}}{n}}{(1-\lambda_i)^2} < f, v_i >_\pi^2$$

Again as in the geometric model the expression **??** for the variance can be simplified if we take instead the correlation factor it limiting version when n approaches infinity:

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{1}{n}\sum_{i=2}^{r} \frac{1 + \lambda_i}{1 - \lambda_i} < f, v_i >_\pi^2$$

Finally when we know the expression for estimator we can go back to the problem of sampling hidden population. Then for the different scenarios with different number of samples that are skipped between the participants $k - 1$ we get the variance:

$$\text{Var}\left[\bar{Y}^k\right] = \frac{1}{\frac{B}{kC_1+C_2}} \sum_{i=2}^{r} \frac{1-\lambda_i^{2k} - 2\frac{\lambda_i^k}{\frac{B}{kC_1+C_2}} + 2\frac{\lambda_i^{k\left(\frac{B}{kC_1+C_2}+1\right)}}{\frac{B}{kC_1+C_2}}}{(1-\lambda_i)^{2k}} < f, v_i >_\pi^2 \tag{3.6}$$

or simplified version:

$$\text{Var}\left[\bar{Y}^k\right] = \frac{1}{\frac{B}{kC_1+C_2}} \sum_{i=2}^{r} \frac{1+\lambda_i^k}{1-\lambda_i^k} < g, v_i >_\pi^2$$

Interestingly, the expression for the variance in general case has the same structure as for the geometric model. Therefore, all the proved observations for the geometric model are true for the general model. Moreover the interpretation of the derived formula is the same. The are tow factors, left and right, that "compete" with each other. If we try to decrease the lest factor, we will increase the right 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 just to find it with search.

On the figure **??** we can see that the variance computed with experiments coincides with the variance **??**. The graph is the same and values of its nodes are the same as in figure **??**, but now prediction is correct.

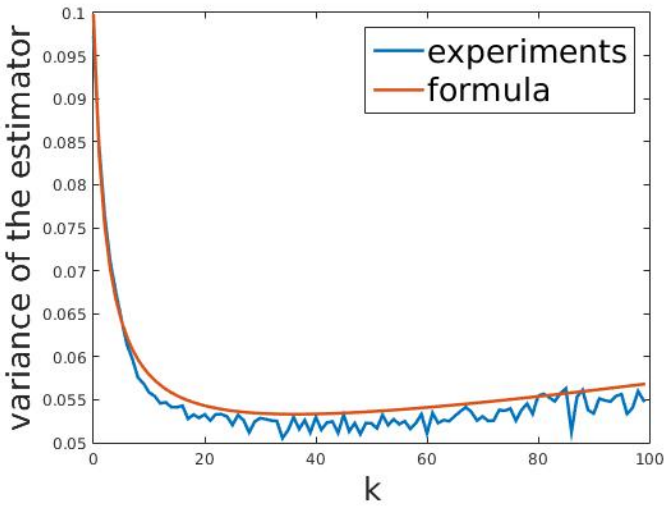


Figure 3.6: Variance of the estimator according to the formula **??** and counted with the experiments scenarios for different k. The graph is the part of Facebook network [**?**] with generated values from the Gibbs sampler

3.1.7 Error prediction

However variance it is not the only source of the error. We should also consider bias: the difference between expected value of the estimator and the real value. We can then compute bias of the estimator $\hat{\mu}_{SA}$ as follows:

$$\text{Bias}(\hat{\mu}_{SA}) = E[\hat{\mu}_{SA}] - \mu_{SA} =< f, \pi > -\mu_{SA}$$

Then the mean squared error of the estimator, $MSE(\hat{\mu}_{SA})$, can be written with variance and bias as:

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

We should note, that for all the scenarios with different k the bias is the same. This bias appears due to the way the sampling is performed: random walk visits the nodes with more connections more frequently. And this is the way of sampling for all the scenarios.

3.1.8 Discussion

The formula **??** implies that the graph and the values on the nodes are known. In this case we can correctly predict the error of the estimator for different scenarios. However, the researchers usually do not possess this information.

The expression **??** may be simplified for the graphs for which the distribution of eigenvalues is known. Like this it will be applicable to some specific graphs, but it will lose generality.

There is no way to predict this value correctly without knowing the graph and the values of the nodes. The variance depends both on the graph structure and homoplily level in the network. When researchers start respondent-driven sampling they may have no clue about graph structure and what is the level of homophily in it. Therefore they can have only guesses about the error of the estimator that they will get. Other option is try to learn more about these two factors.

3.2 Data

To validate our theoretical results we performed numerous simulations. For the graph structure with values on the nodes we used different sources of the data.

First, we used random graphs with values generated with the Gibbs sampling algorithm. Apart from this we used the real network structures like part of Facebook [**?**], but with values again assigned with the Gibbs sampling algorithm.

The real network structures where the nodes have some data are scarce. There are just several sources like this. We used data from the Project 90[**?**], data from the project Add health[**?**].

3.2.1 Data from the Project 90

Project 90 [**?**] studied how the network structure influences on the HIV prevalence. Besides the data about social connections the study collected some data about drug user, such as race, gender, whether he/she is sex worker, pimp, sex work client, drug dealer, drug cook, thief, retired, housewife, disabled, unemployed, homeless.

For our experiments we took the largest connect component from the available data, which consists from 4430 nodes and 18407 edges.

On the figure **??** you can see the graph structure built with the Gephi tool [**?**], where the attribute gender is depicted with color for every node.

3.2.2 Data from the Add health project

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a huge study that began with surveying students from the 7-12 grades in the United States during 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 includes information about social, economic, psychological and physical status of the students and other.

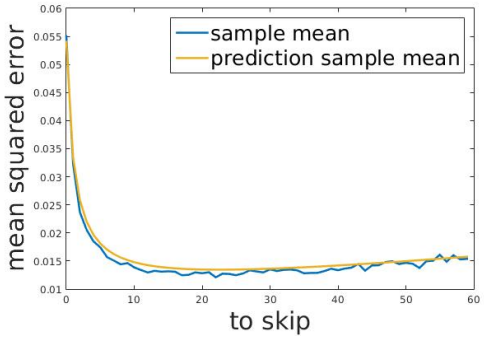
The social network of students' connections was built based on the reported friends by each participant. Each of the students was asked to provide the names of 0-5 male and 0-5 female friends. Then the network structure was built that now can help to analyze if some characteristics of the students indeed are influenced by their friends.

Though this data are very valuable they are not in the free access. Part of them are actually available but the ids of the student and of his/her friends are masked that makes impossible to recreate the network. However the part of the data can be accessed through the link [**?**] but only with few attributes for students, such as: sex, race, grade in school and, if whether they belong to the middle or high school.

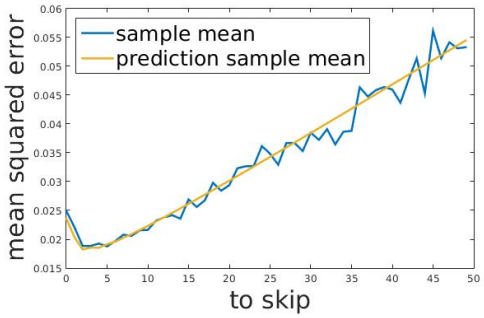
There are several networks available for different communities. We took the graph with 1996 nodes and 8522 edges. It is presented on the figure **??**, built again with the Gephi tool [**?**], where the attribute race is depicted with color for every node.

3.2.3 Experiments

On the following figures we can observe that the experiments results are very close to the theoretical. Looking on these figures we can notice that in all cases is would be useful to skip some nodes. However this number differs for different attributes. The reason is the level of homophily changes depending on the attribute, even if the graph structure is the same.



(a) Project 90 race

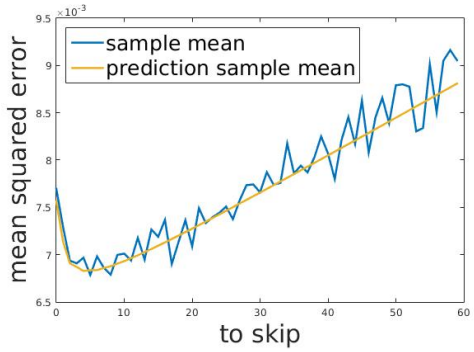


(b) Add health race

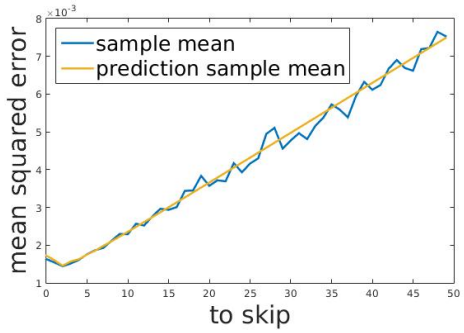
3.3 Other estimators

In the section **??** we got acquainted with following estimators: Sample average, Volz-Heckathon estimator, Group estimator.

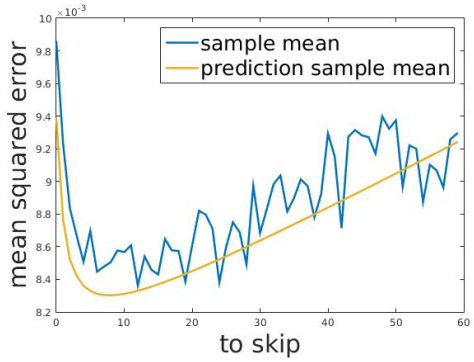
For the enhanced RDS method we used only the Sample average estimator. We applied Sample average estimator to the different scenarios when some of the nodes are skipped. We



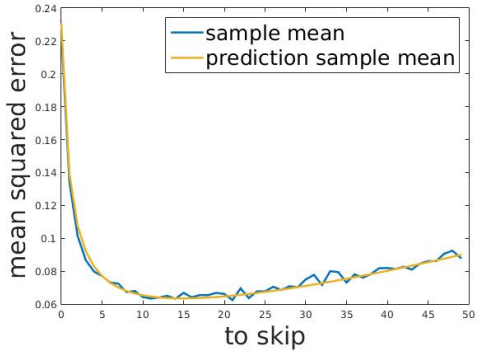
(a) Project 90 sex



(b) Add health sex



(a) Project 90 drug dealer



(b) Add health grade

Figure 3.9: Predicted and empirical mean-squared error of the estimator for the different scenarios

saw that the error of estimator differs for each scenario, as for each scenario the sample size and correlation factor is different. The purpose of this study was to suggest the best number of nodes to skip between taking two samples. And we actually observed on the real data that indeed is can be useful to skip.

However, as we mentioned, the Sample average estimator (and therefore estimators for all scenarios) suffers from the bias toward the nodes with higher degrees.

Volz-Heckathon estimator is unbiased. However the problem of dependency between samples remains. Therefore the skipping may also improve Volz-Heckathon estimator. That's why in our experiments for each k we will apply the both estimators.

The error of group estimator will be counted only for the case when $k = 1$ as this estimator uses the information about who hired whom, what would be not correct to use with skipping.

On the following figures you can see the performance of the estimators for different k. The group estimator was applied to the cases when the attribute to estimate has only two values (like gender) and only when $k = 1$.

First, we can notice that the Group estimator corresponds to the Volz-Heckathon estimator when $k = 1$. Indeed in the case when each person hires exactly one person these estimators coincide. As part of this work we considered only this way of recruitment. Therefore in all our experiments the Volz-Heckathon estimator when $k = 1$ is exactly the same as Group Estimator.

Second, we can observe that also Volz-Heckathon estimator benefits from skipping some nodes. This is true for all the examples. Defining the exact error of Volz-Heckthon estimator when some

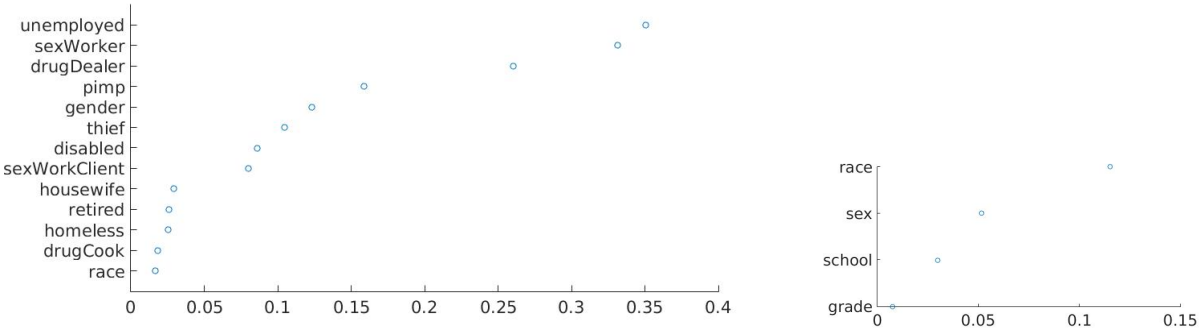
values are discarded can be one of the future works.

We also see that there is no estimator that perform better in all cases. For example, Sample average estimator has less error for the all traits of Add health graph, when Volz-Heckathon estimator has better results for the traits such as disabled, drug dealer, sex worker, thief for the Project 90 graph.

As stated in [**?**] the advantage to use VHE appears only when the estimated attribute depends on the degree of the node. Indeed, our experiments show the same result.

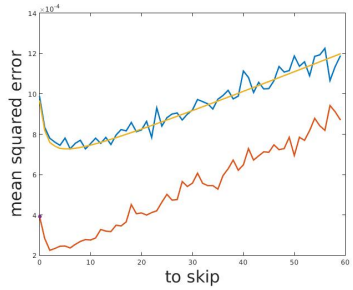
On the figure **??** the correlation between the nodes degree and value for all attributes of two datasets is presented.

Let's observe the correlation between node degree and attributes for the Project 90. For the attribute 'disabled' and all other attributes that have bigger correlation with degree the VHE performs better. For the rest of the attributes, where the correlation with not degree is not so big, the SA performs better than VHE. None of the estimators have the best performance everywhere.

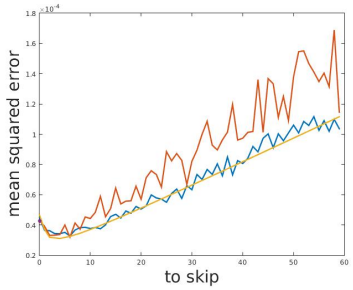


(a) Project 90 (b) Add health

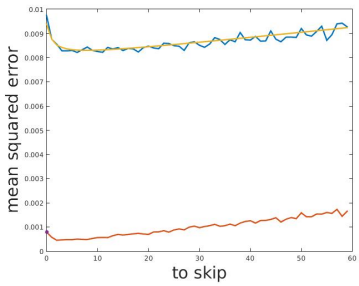
Figure 3.10: Correlation between node degree and its value



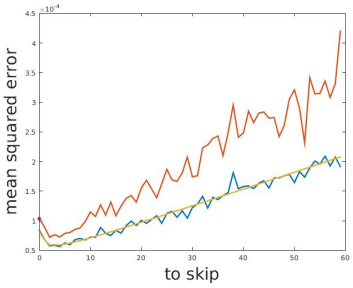
(a) Project 90 disabled



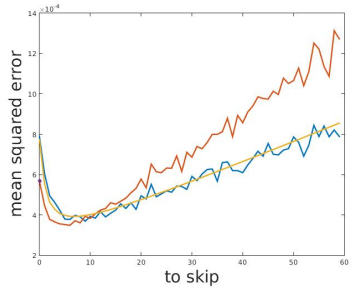
(b) Project 90 drug cook



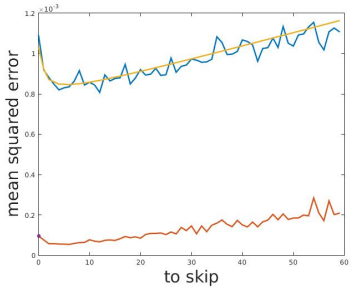
(c) Project 90 drug dealer



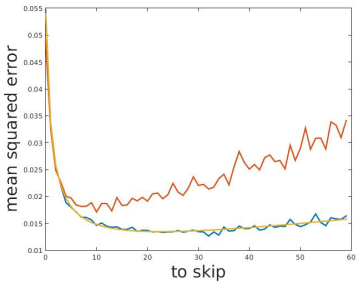
(d) Project 90 homeless



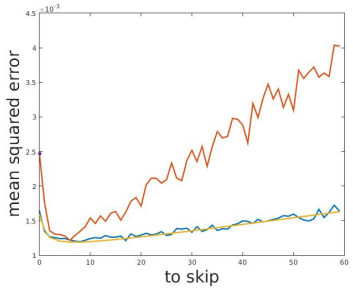
(e) Project 90 housewife



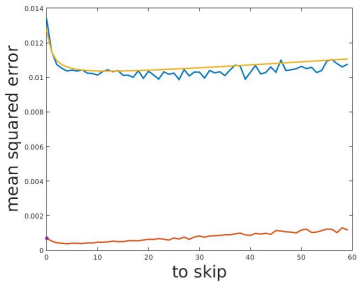
(f) Project 90 pimp



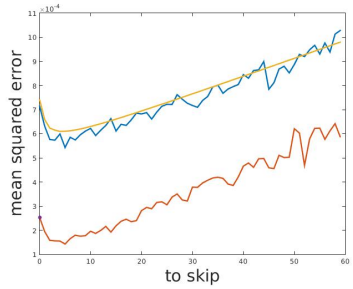
(g) Project 90 race



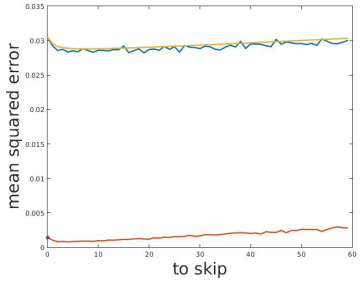
(h) Estimators sex work client



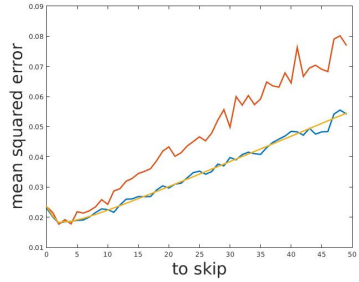
(i) Project 90 sex worker



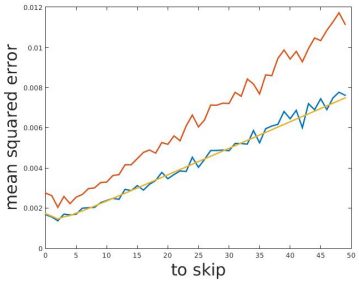
(j) Project 90 thief



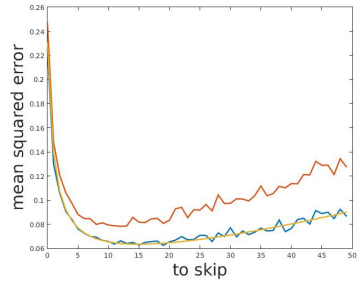
(k) Project 90 unemployed



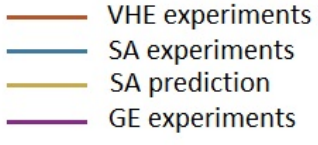
(l) Add health race



(m) Add health gender



(n) Add health grade



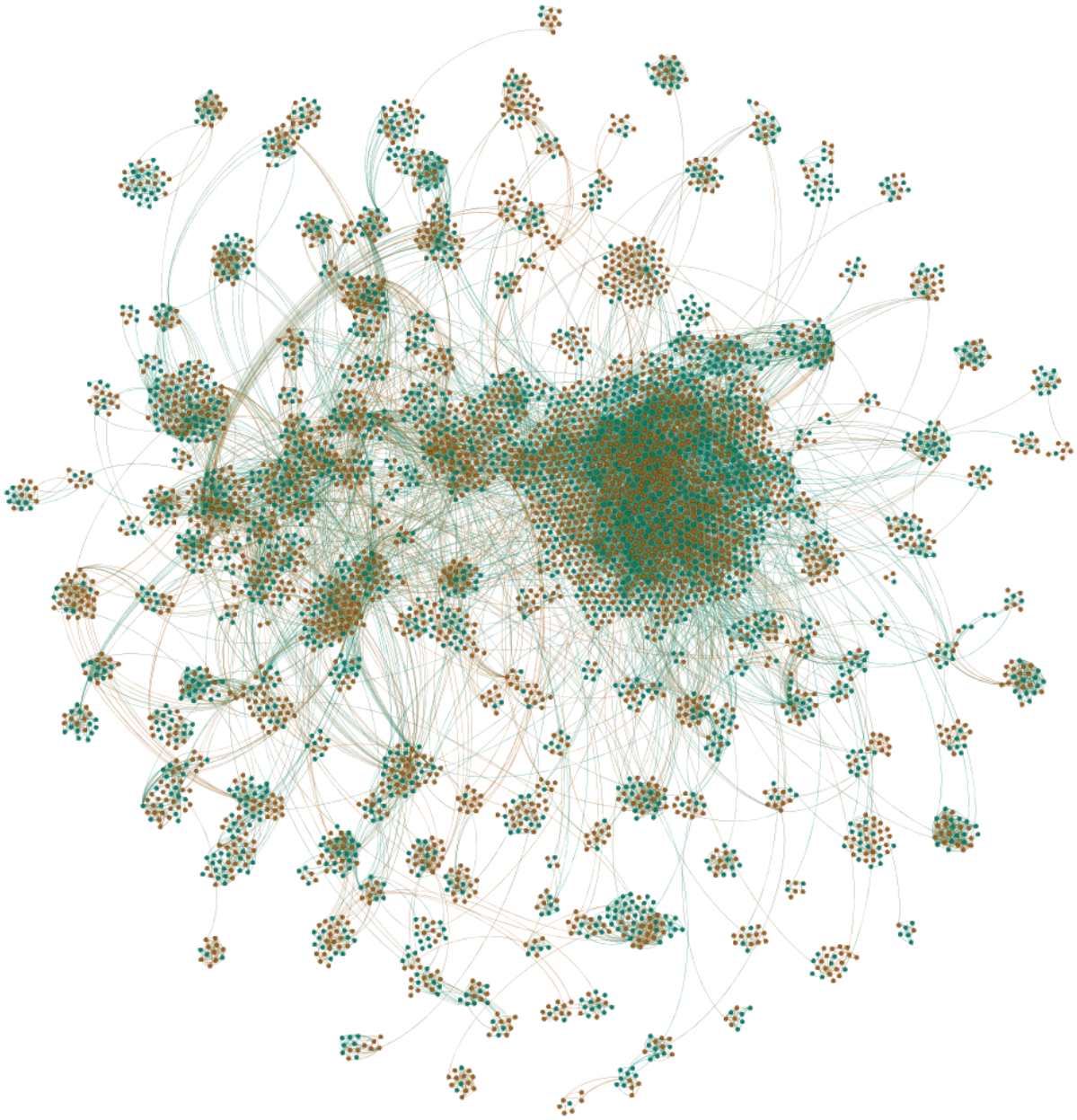


Figure 3.12: Graph from the Project 90 data. The colors of the node represent the gender: brown nodes correspond to the male participant, green nodes correspond to the female participants

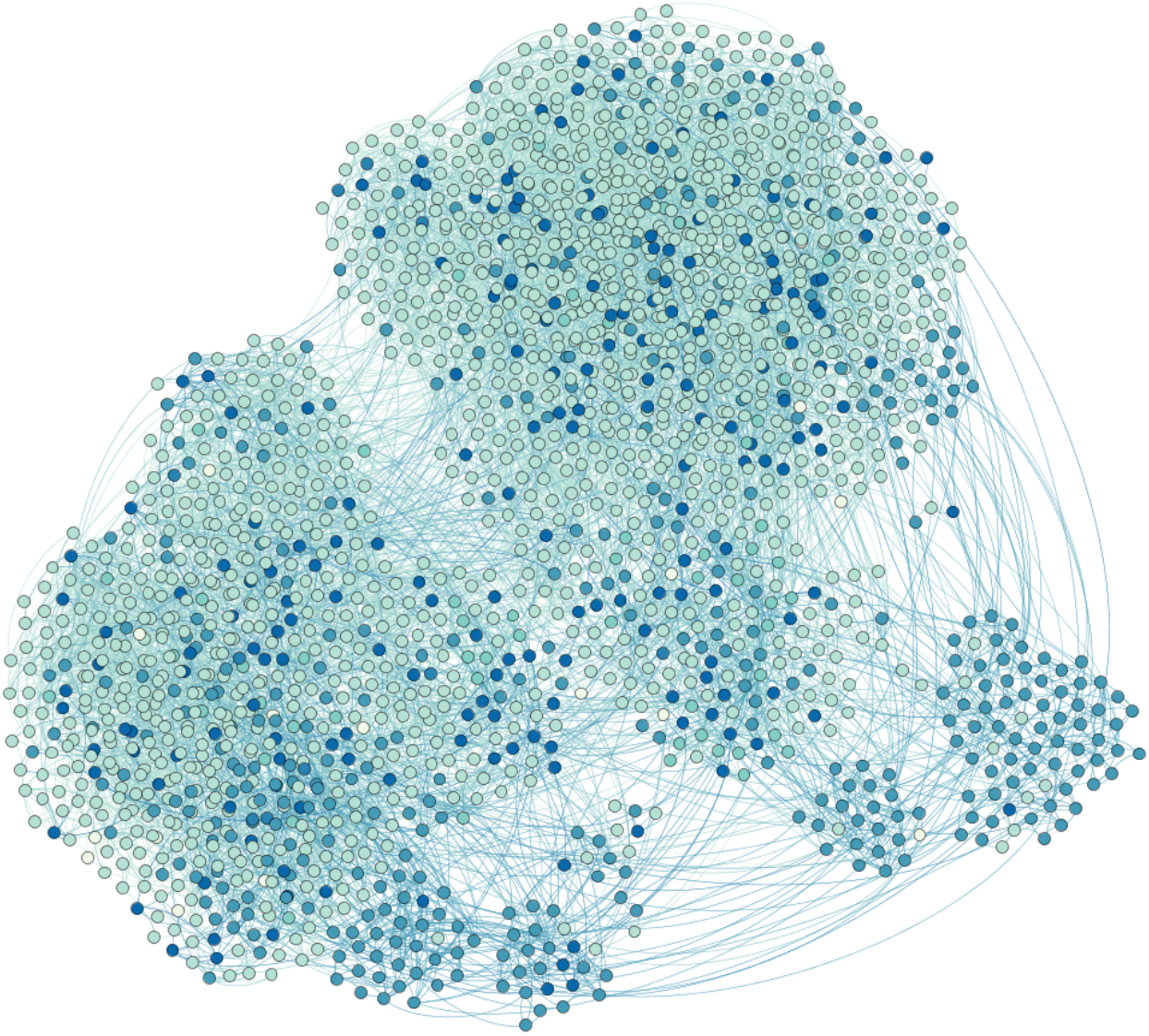


Figure 3.13: Graph from the Add Health data. The colors of the node represent the race

Conclusion

In this work we regarded the sampling technique that is called respondent-driven sampling. This technique can be very useful in the cases when the members of the population are hard to find. The fact that members of the population form a network allows researcher to use their connections to reach other members.

We observed that this way of sampling is not uniform independent sampling. The way of sampling and the presence of homophily in the network influence on the error of the estimator. The researchers should keep attention that particularly the variance of the estimator is greater than in the uniform independent sampling.

The level of the homoplily in the network influences on the correlation between samples. One way of decreasing the correlation is thinning out the sample. We showed formally that just thinning the sample will increase the error of the estimator. Instead we proposed to enhanced respondent-driven sampling that allows to decrease the correlation between samples without reducing drastically sample size.

In order to study the performance of the enhanced RDS formally we needed to find the expression for the error of the estimator. The enhanced RDS method looks on the different scenarios when the different number of nodes is skip. Scenario with the lowest error of the estimator is the best one and should be applied.

The challenge that we encountered during the study is 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 needed the mechanism of assigning values to the nodes that imitates the property of homophily in the network (the value of the node should depend from the values of its neighbors). Created algorithm allows not only to imitate homophily but also to control its level in the network. This result is general and can be applied surely not only in the context of sampling hidden population. This mechanism allows to create the network with controllable level of dependency between neighbors, that can find its application in different areas.

Using created mathematical model we targeted to find the error of the estimator for different scenarios of enhanced RDS. The scenario with the minimal error would be the best and recommended for applying. First we regarded the geometric model where correlation between samples is established.

Though this model turned out to be too simplified to produce the exact result it can give us the lower bound on the desired result.

After regarding the simple geometrical model we were able to generalize the result for any kind of graph with any kind of correlation. Theoretical results were then validated with the numerous experiments. For the experiments we used the synthetic networks with values generated according to created model and real data from the school survey Add Health and Project 90.

Finally, we compared two estimators: Sample average and Volz-Hechathon estimator. The

exact error for different scenarios of enhanced RDS method was found only for Sample average estimator. However, experiments showed that also VHE benefits from this enhancement of RDS. After comparison VHE and SA extimator we came to the following conclusion: for the attributes that are correlated with the degree of the node the Volz-Hechathon estimator will produce more correct results. In the cases when there is no such correlation Sample average performs better. We saw that both estimators benefit from the introduced enhanced RDS method.

In any case, using any estimator, the enhanced RDS technique shows better results than the standard RDS, especially when the level of homophily is high. Regarding that the real networks show the presence of this property the enhanced RDS technique can reduce the error of the estimates.

Bibliography

[1] Computer science university of maryland. http://www.cs.umd.edu/hcil/science20. Accessed: 2015-08-07.

[2] Gephi - the open graph viz platform. http://gephi.github.io/.

[3] 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.

[4] The national longitudinal study of adolescent to adult health. http://www.cpc.unc.edu/projects/addhealth. Accessed: 2015-07-01.

[5] The office of population research at princeton university. https://opr.princeton.edu/archive/p90/. Accessed: 2015-07-01.

[6] Stanford large network dataset collection. https://snap.stanford.edu/data/. Accessed: 2015-07-01.

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

[8] 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.

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

[10] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *arXiv preprint arXiv:0906.0060*, 2009.

[11] Sharad Goel and Matthew J Salganik. Respondent-driven sampling as markov chain monte carlo. *Statistics in medicine*, 28(17):2202–2229, 2009.

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

[13] 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.

[14] 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.

[15] Bruno Kauffmann, François Baccelli, Augustin Chaintreau, Vivek Mhatre, Konstantina Papagiannaki, and Christophe Diot. Measurement-based self organization of interfering 802.11 wireless access networks. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pages 1451–1459. IEEE, 2007.

[16] 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.

[17] Michael Pollard, Harold D Green, David P Kennedy, Myong-Hyun Go, and Joan S Tucker. Adolescent friendship networks and trajectories of binge drinking. 2013.

[18] 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.

[19] Matthew J Salganik and Douglas D Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.

[20] Parag Singla and Matthew Richardson. Yes, there is a correlation:-from social networks to personal behavior on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 655–664. ACM, 2008.

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