Polytech Nice Sophia Antipolis

Master IFI/ Ubinet track

Final report

Network sampling and discovery processes

Author:
Alina Tuholukova

Supervisors:
Konstantin Avrachenkov
Giovanni Neglia

August 17, 2015

Chapter 1

Abstract

Chapter 2

Introduction to the network sampling

2.1 Motivation and challenges

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. Particularly it is interesting to analyze network structures such as online-social networks (OSNs), peer-to-peer networks (P2P) or network of individuals.

There can be variety reasons to collect information about the networks.

For example, we can be interested in estimation of total number of peers in network or number of peers that satisfy needed characteristics. This information can be used in peer-to-peer protocols. For example, peer-to-peer protocol Viceroy needs to know number of nodes in network before including the new one in it (2). Some gossip based peer-to-peer protocols require knowledge about network size in order to disseminate information (2).

OSNs possess huge amount of information about population that can be interesting for different areas of life: sociology, marketing, network engineering (3).

Another example it is human networks.

Unfortunately sampling such kind of structure is not always evident and easy. It is not always possible to identify all the nodes of the network in order to take representative subset of them for the analysis.

The simplest idea is to take node uniformly at random knowing the identity of all nodes in network (uniform sampling). This technique can provide us uniform choosing of nodes and independency of received samples. But here we can confront some problems.

Having all these advantages of P2P networks, on the other side, it is not so easy to collect needed characteristics of network what is direct in centralized systems. Moreover, the P2P networks have distributed nature, what usually implies that no node maintains the knowledge of all topology. Nevertheless, even if P2P protocol assumes existing of such a node (like BitTorrent tracker in BitTorrent protocol) it is usually regarded as its weak side.

In social networks each user has ID. So having the whole list of IDs would perfectly fit to uniform sampling technique. However, the social network owners can hide information about all IDs due to their privacy policy. Moreover, some of the IDs can be not valid.

Performing too much requests can be expensive in the meaning of resources (4). Rather than trying to find valid ID by random requests it can be more useful to choose small but representative set of nodes (3).

The other sampling techniques are based on random walking (crawling techniques). The network

is regarded as a graph. The simplest method is called the Random Tour method, where probability to go from the node to each of his neighbor is equal. It is can be shown that probability distribution is not uniform. It is biased toward the nodes that have greater number of neighbors.

The other methods remove this bias by spending less time in the nodes with greater number of neighbors. Particularly, we will regard three methods: Maximum degree method, Local degree method and Metropolis Hasting method. All crawling techniques work only on connected graphs while uniform sampling techniques can be applied even to disconnected. They also suffer from dependency of samples.

Though this network structure brings difficulties at the same time it can (naturally suggest) help to collect data from the network using chain-referral methods. The one of such examples is sampling hidden populations(e.g., drug users). Being comfortable method for finding people for studies RDS introduces some additional difficulties comparing to simple random sampling. The most important is dependency of the samples.

Then problem how to know variance (how to be able to say about confidence that result is correct).

2.2 Challenges

2.3 Goals

2.4 Contributions

The new idea to create network values in such a way that bla bla. This is general contribution that makes possible to create ... where we can control and mathematically studying properties that is very important for research purposes.

Chapter 3

Respondent-driven sampling

3.1 Motivation

In order to make correct estimates it is not enough to have just subset of individuals. We also have to know the probability of one particular individual to be selected. For example, by using telephone survey in order to collect information we automatically exclude some subsets of people (like homeless, poor) which can affect the correctness of the estimate because it is impossible to predict bias.

Examples of hidden populations: drug users, men who have sex with men, sex workers, illegal immigrants, participants in some social movements, homeless [13]

Respondent-driven sampling is a technique for estimating traits in hidden population. It is widely used for studying prevalence of HIV/AIDS among injection drug users, sex workers, men who have sex with men.

Studying prevalence of disease can help to understand and control its spreading. Unfortunately, there are difficulties with such kind of research as there is no sampling frame and members of hidden groups may not want reveal themselves.

There are several existing solutions for sampling hidden population such as snowball sampling, targeted sampling, time-space sampling, key-informant sampling. The main disadvantage of all these methods is unknown bias and variance of obtained estimation.

3.2 Technique of respondent-driven sampling

RDS begins with selecting group of initial participants that are called seeds. The procedure follows according to chain-referral model: each participant in study recruits another 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 needed size of participants is reached. During RDS participants are asked to report how many contacts they have. This process enables to collect data for making statistical analysis.

In order to study formally RDS can be regarded as Markov Chain. Assumptions:

1. Seeds are chosen proportionally to their degree in the network.

2. If individual A knows individual B than individual B knows A as well (network can be represented as undirected graph).

3. The same individual can be recruited multiple times (sampling with replacement.

4. The choice of contacts to recruit is uniformly at random.

5. Individuals know precisely their network degree.

6. Each individual is reachable from each other individual (network is connected).

For this process stationary distribution is exactly distribution proportional to network degree. So first assumption guaranties that not only first but all samples during the process are taken with probability proportional to the degree of participants in the network. In [13] this assumption 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. Without this assumption first there should be performed enough number of waves until sample can be considered drawn from stationary distribution. simulation studies about assumptions violation(sensitivity) [7]

studies of variance

In this way individuals with more friends (contacts) are more likely to be recruited. To correct this bias the responses from individuals are weighted according to their degree (number of contacts). Let $X_1, X_2, ..., X_n$ be all collected samples during RDS. Then estimate μ_f of the population mean of f is defined [10] as

$$\mu_f = \frac{1}{\sum_{i=1}^{n} 1/degree(X_i)} \sum_{i=1}^{n} \frac{f(X_i)}{degree(X_i)}$$

RDS can perform poorly if the groups of individuals form different communities. It is known fact that friends tend to have similar traits. This fact becomes a source of bias in chain-referral methods of sampling. Structure of network also affects a lot. In [9] it is shown that 'bottlenecks' between different groups in hidden population increases variance of RDS estimator. They try RDS on network structure with communities, but where individuals, that are in contact with each other, do not have similar traits and showed that such structure indeed affects on RDS estimate.

Design effect d is variance of RDS estimate over variance of estimate obtained from simple random sampling (SRS). It means that if for SDS we need n samples than to have RDS estimate with the same variance we need dn samples.

It is known fact that people tend to be friends if they share some traits: have similar age, common language, the same university.

Homophily - the tendency for individuals with similar attributes to be friends with one another. The fact that the majority of participants are recruited by other respondents and not by researchers makes RDS a successful method of data collection. However, the same feature also inherently complicates inference because it requires researchers to make assumptions about the recruitment process and the structure of the social network connecting the study population.

3.3 Estimator black and white

In [13] they introduce asymptotically unbiased estimator of the trait

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

where

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

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

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

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

3.4 Enhanced RDS

3.4.1 Comparison of RDS and independent uniform sampling

To understand what are the problems with respondent-driven sampling let's compare it with independent 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:

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

This estimator is unbiased and its variance depends on the sample size n. Let σ be the variance of the samples, then the variance of the estimator $\widehat{\mu}_{avg}$:

$$\sigma^2_{\widehat{\mu}_{avg}} = var\left(\frac{y_1 + y_2 + ... + y_n}{n}\right) = \frac{1}{n^2} var(y_1 + y_2 + ... + y_n) =$$

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

In this case the only thing that we can do in order to have better estimation is to increase sample size. 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.

Of course the way we perform sampling it is not independent uniform sampling. First, nodes are not sampled not uniformly. When each participant select another 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.

The study [8] 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 can't 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 V-H estimator does.

Second, collected values are not independent. The participants i and $i+1$ know each other, they can be friends, relatives or just acquaintances, so their traits 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.

When the values $y_1, y_2, ..., y_n$ are collected with RDS, then the variance of estimator:

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

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

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

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:

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

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

3.4.2 Modified RDS

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

Observation 1 *No matter how values are correlated, the more is sample size, the better is estimation*

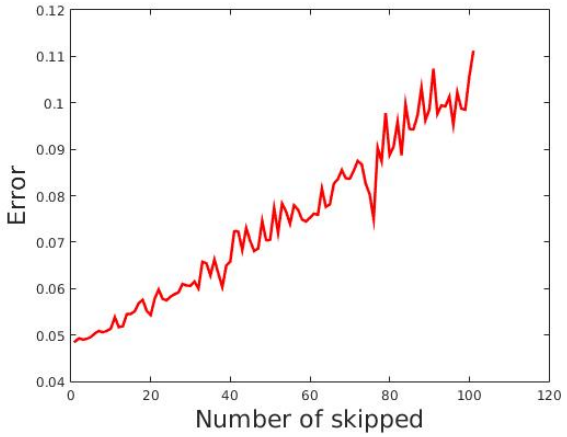
Because $f(n) \leq n$, the bigger n the less is sampling variance. The factor $f(n)$ prevents variance to decrease linearly with increasing n like in the case with independent samples. But increasing n for sure does not bring any harm. So the size of the sample should be always as big as the budget allows.

Observation 2 *Just thinning of sample doesn't help*

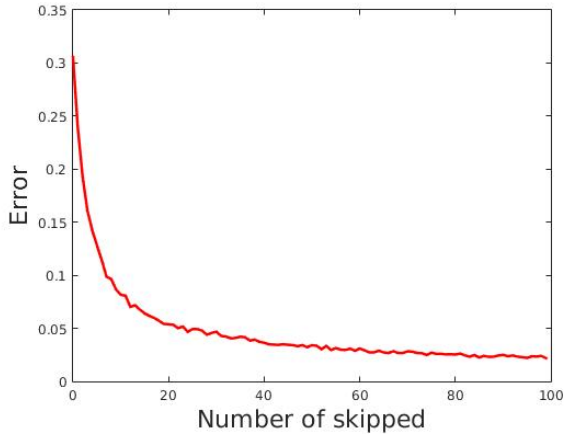
In order to reduce correlation between sampled values one could try to thin out sample. It means that instead of taking all collected values $y_1, y_2, ..., y_n$ into the estimation we can take only, let's say, each second value, $y_1, y_3, ..., y_n$. The samples $y_1, y_3, ..., y_n$ are less correlated, so we can expect some improvement. But what we observed experimentally (latter we showed 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 3.1 (a).

Observation 3 *Skipping reduces variance*

However if the size of the sample is fixed the further 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 3.1(b) we can see experimental results and further we will show it formally.



(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 3.1: Experiments on the data from Project 90

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 payed 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.

What we propose lies in the middle. We keep the idea of skipping some values, that will decrease the correlation, but we also suggest to pay less to the individuals that we don't take into the estimation. So among people that are willing to participate some of them will be asked both: to make tests (like blood test, questionnaire) and recruit another 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 payed less.

Let's say that each informator receives C_1 units of money and each participant receives 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 parameter k 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. 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 payed, 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 payed 2 euros, $C_1 = 2€$, each participant is payed 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 $k = 1$. Everybody is participant and payed 10 euros. Then with our budget we can collect 6 samples that due to homophily can be highly correlated.

On the figure 2 $k = 2$. 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? That is the question that we are going to analyze. We will quantify the error depending on the number of samples that we skip $k - 1$.

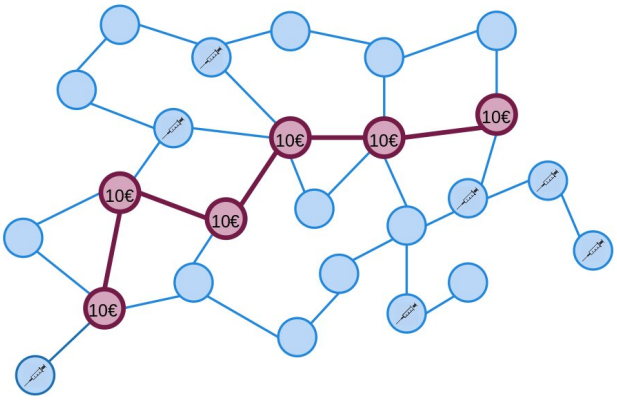


Figure 3.2: sd

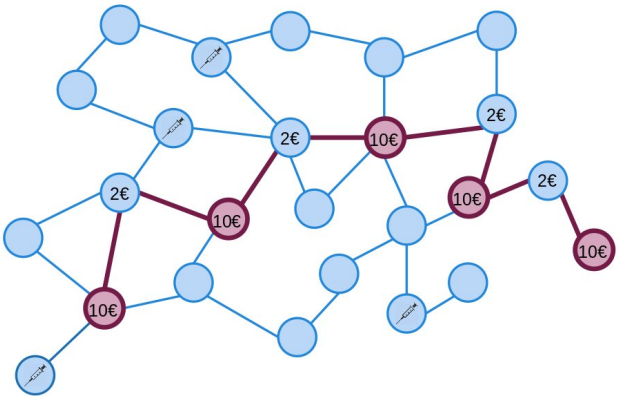


Figure 3.3: sd

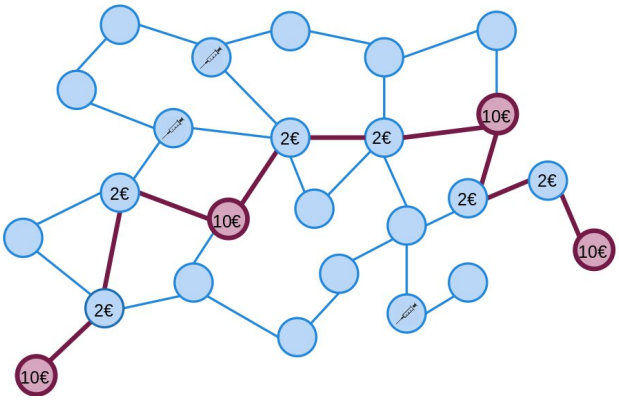


Figure 3.4: sd

Chapter 4

Mathematical model

In order to imitate the network structure

There is number of different random networks that can imitate real networks. Erdos Renui for peer-to-peer networks. Random geometric graph for sensor networks [check from PFE report]. Others for community.

4.1 Network modeling

4.1.1 Erdos-Renyi model

4.1.2 Random geometric graph

4.1.3 Preferential attachments model

4.1.4 Small world

4.2 Network with values

4.2.1 Motivation

All the random graph models give us possibility to generate only the structure of a network. The next step is to generate the values on the nodes of the obtained graph which will represent some attribute. For instance, if we have a social network the attribute can be the age, gender of a user.

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.

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 [6] 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%.

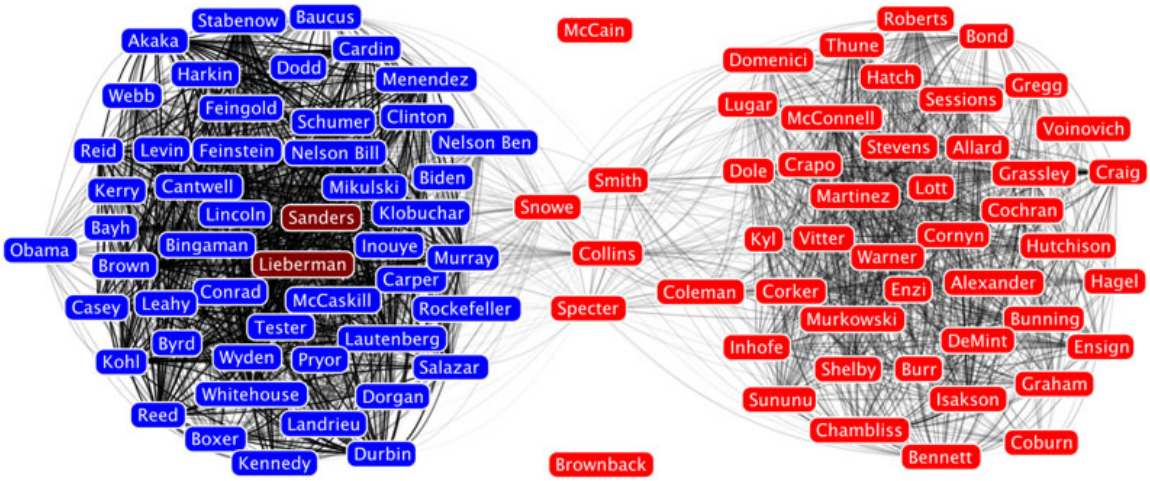


Figure 4.1: Voting patterns of U. S. Senators during 2007 [1]. The red labels represent Republicans, the blue labels represent Democrats, the brown labels represent two Independents.

Another study [14] 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 query requests!

On the figure 4.1 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.

The reason why we don't want to ignore homophily is because the way of performing sampling and the property of homophily together influence on the sampling variance. Further we will count formally the sampling variance for 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 [12] 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 in the same way as we can tune the ??? density in Erdos-Renui graph.

4.2.2 Definitions

We have graph with n nodes. To each node i will be assigned random value X_i from the set $V, V = \{1, 2, 3, ..., k\}$.

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 distribution should take into account the values of the node's neighbors as well.

Let's denote $(X_1, X_2, ..., X_n)$ as $\bar{X}$. We will 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 will 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 distribution we will take Gibbs distribution, that originally comes from physics [put a reference].

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 each possible configuration $\bar{x}$ we will associate the number that is called global energy of the graph and 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's turn our attention to the one node i. We will define local energy on the node i as:

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

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

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

4.2.3 Gibbs distribution

Now let's consider the following probability distribution:

$$p(\bar{x}) = \frac{e^{-\frac{\varepsilon(\bar{x})}{T}}}{\sum\limits_{\bar{x}' \in |V|^n} e^{-\frac{\varepsilon(\bar{x}')}{T}}} \tag{4.1}$$

where T is temperature, $T > 0$.

The reason why it is interesting to look at this distribution follows from the theorem [Theorem 2.1, p. 260], [5]. When random field has distribution [reference on the formula] then the probability that the node has particular value depends on the values of its neighboring nodes.

This means that for particular node i whatever the values are assign to the vertices probability that it will have values from the V:

$$p(X_i = x_i | X_{N_i} = x_{N_i}) = p(X_i = x_i | X_{V \setminus i} = x_{V \setminus i})$$

This property is called Markov property.

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

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

The temperature parameter T is very important, it plays the role of the tuner of the correlation level in the network. Later we will show some examples for better understanding.

not mine: it favors states of small energy, especially when the temperature is small.

[maybe write about standard application for image processing]

Gibbs distribution found many interesting applications in real-world problems. Particularly it lies in the basement of the proposed in [11] distributed algorithm for channel selection of the Access Points. The channels should be selected in such way that interference in the network is minimized.

4.2.4 Algorithm

Practically speaking direct sampling from the distribution 4.1 is not so easy. Let's 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 take .... to sample.

In order to produce samples from this distribution we are using Gibbs sampler. The idea of Gibbs sampler is to change the configuration $\bar{x}$ with time k: $\bar{x}^k$. From the book [5] the $\bar{x}^k_{k>0}$ is regarded as Markov Chain where probability.

The stationary distribution of this Markov Chain is exactly 4.1. Following algorithm after enough amount of step will produce a sample from the distribution 4.1.

1. Create random configuration of properties on all nodes.
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 $x \in P$ count the local energy on chosen node i as
$$E_i(x) = \sum_{j|i\sim j} (x - x_j)^2$$
4. Choose a new value x_i according to probability
$$\frac{e^{-\frac{E_i(x)}{T}}}{\sum\limits_{x' \in P} e^{-\frac{E_i(x')}{T}}}$$
   where T is temperature.
5. Continue 2-3 needed number of iterations.

4.2.5 Explanatory example

Let's say that we have the graph to which nodes we want to assign values $1, 2, 3, 4, 5$. Now let's look only at the vertex A its neighbors B, C, D, E which have assigned values $1, 5, 3, 4$. And now it is turn of A to chose a value.

With different values the node A will have different local energy. Let's summarize it in the table [reference].

[to transpose]

Value	Energy
1	29
2	15
3	9
4	11
5	21

As we said the values that bring node to small local energy are favorable. In this example, the highest probability will be for the values 3.

But also we have temperature parameter T. To feel the impact of temperature we will present the distribution of values for different values of 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 is 1. And when the temperature is really high the choose of value will not almost depend on the values of its neighbors. That confirms that the distribution favors the values that bring local energy of the node to minimal.

To understand this better let's look at the node i [put the figure and reference to it].

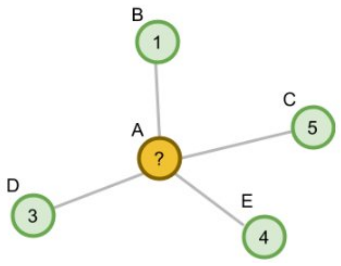
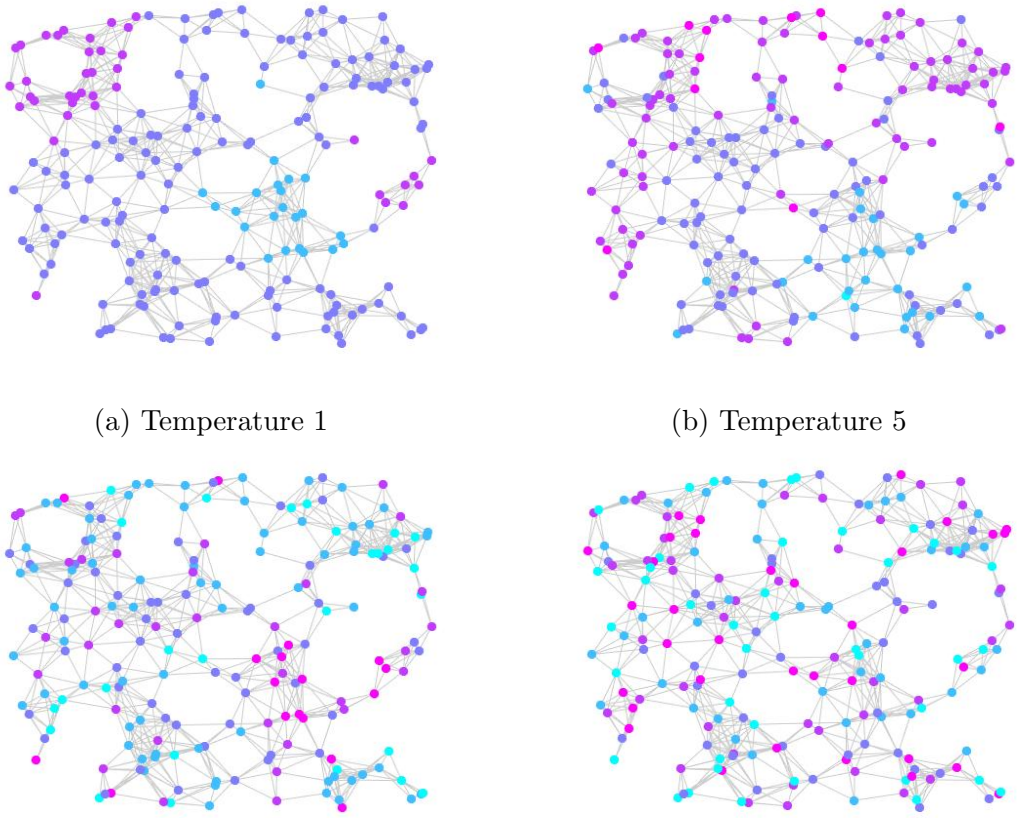


Figure 4.2: Preferential attachment graph with 200 vertices, 1 link for new arriving node and values $[1, ..., 5]$



(a) Temperature 1

(b) Temperature 5

(c) Temperature 20

(d) Temperature 1000

Figure 4.3: RGG(200, 0.13) with generated values for different temperature

4.2.6 Demonstration of random graphs with values

In order to show that chosen model has sense 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.

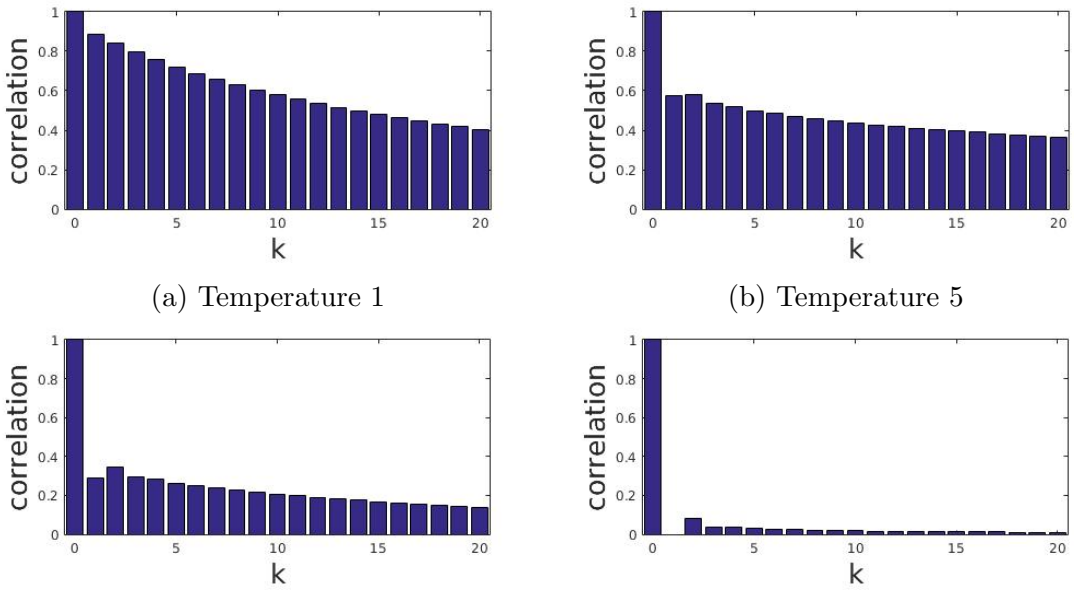
On the figure 4.3 presented the same random geometric graph with 200 nodes and radius 0.13, $RGG(200, 0.13)$ where the values $V = \{1, 2, ..., 5\}$ are chosen according to the Gibbs distribution. The values are depicted on the pictures as colors.

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.

In order to give more formal illustration we can look at the correlation between values of the nodes that we see during the random walk on the graph.

Let's denote as $Y_0, Y_1, ..., Y_i, ...$ the random nodes that we observe during the random walk. The value on the node Y_i we will denote as $f(Y_i)$. Let's say that we start random walk with stationary distribution. Then covariance between two values $f(Y_i)$ and $f(Y_{i+k})$ depends only on the distance k in the sequence $Y_0, Y_1, ..., Y_i, ..., Y_{i+k}, ...$ between them:

$$cov(f(Y_i), f(Y_{i+k})) = cov(f(Y_0), f(Y_k))$$



(a) Temperature 1

(b) Temperature 5

(c) Temperature 20

(d) Temperature 1000

Figure 4.4: Correlation of the values of the nodes observed during the random walk depending on the difference in their order for different temperature

On the figure 4.4 we present correlation between values depending on this distance k for the graphs shown above. In this way we can assure us that the higher is temperature the less correlated values of neighbors are.

4.3 Expected energy in steady state

4.3.1 Motivation

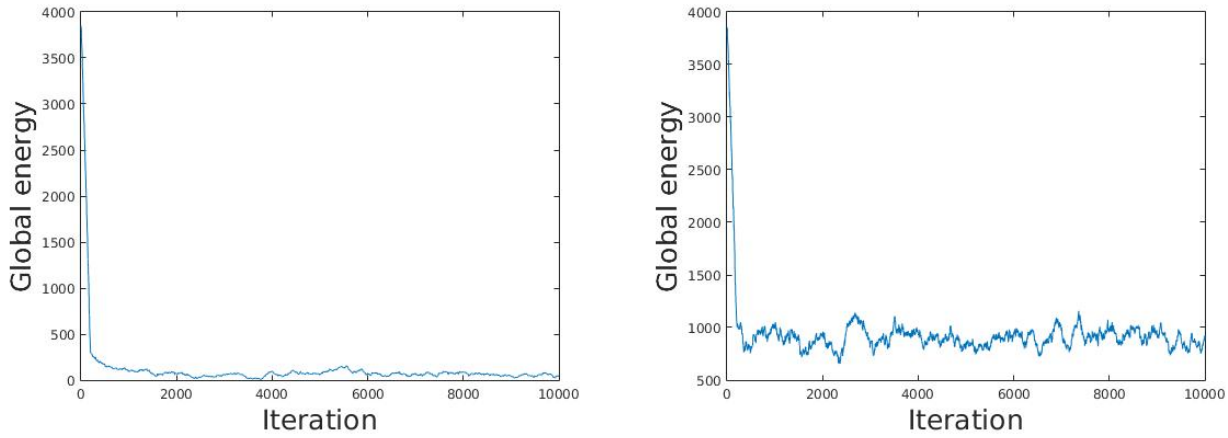
The question that is still not clear in the algorithm is when to stop it. How many steps are enough to perform in order to claim that achieved configuration is indeed sampled from the Gibbs distribution?

Let's look as global energy of the graph is changing during the steps of algorithm. For each configuration $\bar{x}$ we can count the global energy of the graph as:

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

First values are assigned to all nodes uniformly from all possible values. We can observe on the figure 4.5 that the energy is high. 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. Knowledge about this expected value and its variation can indicate us when it is time to finish the algorithm.

There is another reason why we would like to predict energy. We saw previously that varying temperature T we can change how strongly values of the neighbors are related: low temperature brings high correlation of values and high temperature brings almost random assignment of values.



(a) Temperature 2

(b) Temperature 10

Figure 4.5: Energy changing with iterations of the algorithm

However it is impossible to use only temperature T as a metric of values correlation. Temperature by itself does contain a lot of information. For the graph RGG(200, 0.13) the temperature 200 we can consider as high (because values are not really correlated) but for the graph RGG(2000, 0.06) it is not the case. We can't judge the level of correlation only by temperature, we should take into account number of nodes, number of edges, structure of graph, possible values that can be assigned to nodes as well.

On the figure 4.5 we can see that starting from the same energy there is different expected global energy and its variation for different temperature. So we see that we start at the maximum energy and then come to its expected value for the fixed temperature.

Then the more appropriate metric can be following number: in how much times energy decreases from its initial value (that corresponds to the random configuration) to its stable value (the one that we want to be able to predict).

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 this purpose we were interested in finding dependency of the global energy from the temperature.

4.3.2 Analysis

We would like to know the expected value of the global energy. If expected energy is known it can help detect convergence and indicate when it is time to stop running algorithm.

We can write expected energy by definition as

$$E[\varepsilon(\bar{x})] = \sum_{\bar{x}} p(\bar{x}) \cdot \varepsilon(\bar{x}) \tag{4.2}$$

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

$$p(\bar{x}) = \frac{e^{-\frac{\varepsilon(\bar{x})}{T}}}{\sum\limits_{\bar{x}' \in |V|^n} e^{-\frac{\varepsilon(\bar{x}')}{T}}} \tag{4.3}$$

Both in 4.3.2 and in 4.3.2 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: use Gibbs sampling, run the algorithm until convergence large enough amount of times, average the result. But actually it is the opposite of what we are trying to do. In this way we can build empirical expected energy dependency of the temperature. Simulations can take a lot of times. But actually I wanted to have an explicit formula.

Let $\varepsilon(\bar{x})$ be total energy of the graph on configuration $\bar{x}$. We can write 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 field is

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

[write more that it is difficulty to count]

Difficult to generalize for any graph, because there is correlation moreover difficult to write the distribution of values for one particular node. later we will show why. In order to calculate expected energy at least approximately we will make some assumptions.

The same expected value

$$E[X_i] = av_V$$

Assumptions about no correlation between neighbors

Because difficult [describe difficulties] we will make an assumption that values are assigned independently of each other. Let N_i be the number of the neighbors of the node i. So if there is no correlation between value assigned to the nodes, then we can write global energy as:

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

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

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

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

Special case

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

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

where x_i and x_j are random variance with the same distribution. Interesting fact that it reminds us the famous formula for energy!

Particularly, we can compute expected energy in this way for 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:

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

Then the expected energy of the graph on random configuration $\bar{x}$ is

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

Assumptions about values distribution

To answer the question: what is the expected energy of the field after reaching steady state, we need to know the distribution of values on the nodes.

After reaching stationary state probability that the node i will have value $x \in P$ is

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

where $\varepsilon_i(x)$ is local energy on the node i that is counted as $\varepsilon_i(x) = \sum\limits_{j|i \sim j} (x_i - x_j)^2$.

But this probability actually is conditional on the values of the neighbors. Then the expression becomes:

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

So the probability to have some particular value on the node i depends on the values of its neighbors that are also dependent from their neighbors and so on. In order to simplify expression for values distribution on the node i we will make some assumptions. First, let's say that all neighbors of the node i have the value $av_V, av_V = average(V)$. Let us denote the set of neighbors of the node i as N_i. If for all $j \in N_i : x_j = av_P$ then $p(x_j = av_P) = 1$. With this assumption probability that the node i will have value $x \in P$ is

$$p(X_i = x) = \frac{e^{-\frac{N_i(x-av_P)^2}{T}}}{\sum\limits_{x' \in P} e^{-\frac{N_i(x'-av_P)^2}{T}}}$$

Then we can count expected energy as:

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

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

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

where d is average degree of the graph.
Then expected energy is counted in the following way:

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

4.3.3 Results

On the following examples: red line represents predicted energy for given temperature, blue line represents calculated energy after running algorithm.

Even if the result is not precise it can give us intuition.

4.4 Error prediction

4.4.1 Variance prediction

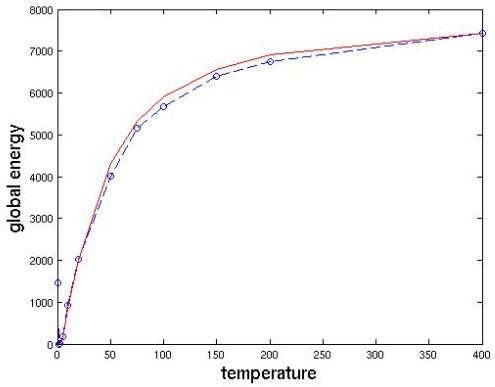
Let $Y_1, Y_2, Y_3, ..., Y_n$ be the samples that are taken during the random walk. Let's take the average value of the samples as an estimator of mean population value:

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

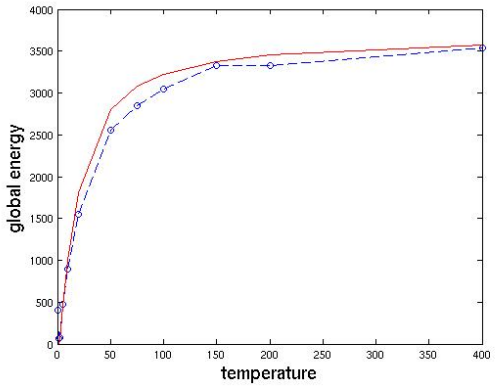
We noted that the variance of the estimator $\widehat{\mu}_{avg}$ is influence by the correlation factor as the random variables $Y_1, Y_2, Y_3, ..., Y_n$ are correlated:

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

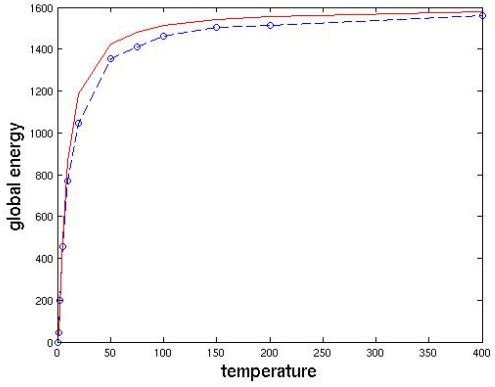
Let's look what will happen with variance in the suggested modified RDS.



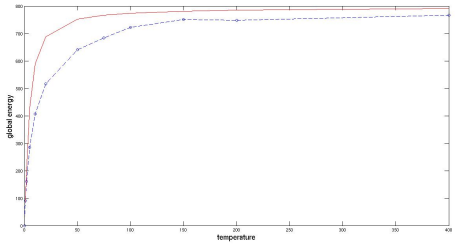
(a) Random ER graph with 200 vertices and values $[1, ..., 5]$ p = 0.1



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



(c) Grid on torus graph with 200 = 20x10 vertices and values $[1, ..., 5]$



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

Figure 4.6: Energy prediction for different gr

We have budget B. n individuals that we see during the RDS receive C_1 units of money for recruiting another individuals, each kth person from them take also part in the study and receives C_2 additional units of money.

In this way $\frac{n}{k+1}$ individuals from n are participants. Then we can write:

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

From this equality we can see that the length of the chain n is:

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

And then the number of participants is:

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

So 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 also depends on k. Saying all this we can write that variance of estimator as:

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

Now we have to find the expression for the correlation factor.

4.4.2 Simple example

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:

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

So correlation between the nodes that are at the distance 1 in the chain have correlation ρ, at distance 2 have correlation ρ^2 an so on. Then we can write the variance of the mean estimator as:

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

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

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

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

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

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

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

In the end we have the following expression:

$$var\left[\bar{Y}\right] = \frac{\sigma^2}{n}\frac{1 - \rho^2 - 2\rho/n + 2\rho^{n+1}/n}{(1-\rho)^2} \tag{4.4}$$

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.

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

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

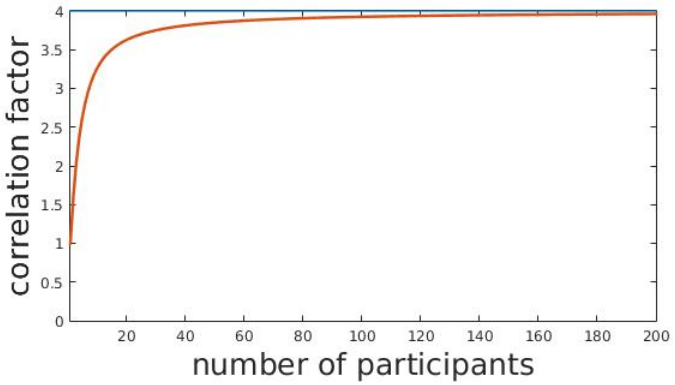


Figure 4.7: Correlation factor depending on the number of the participants when ρ is 0.6

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

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

Approximation is close when is n big enough. To compare approximated expression with original one, look at the figure 4.8 where parameter ρ is 0.6. As it is reasonable to suppose that the sample size is bigger then 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.

[??? f(n) ¡ n]

Observation 1 *No matter how values are correlated, the more is sample size, the better is estimation*

From 4.4.2 we can see it.

If random variables $Y_1, Y_2, ..., Y_n$ were independent then the variance of $\bar{Y}$ would be $var_{ind}[\bar{Y}] = \frac{\sigma^2}{n}$.

But we consider random variables $Y_1, Y_2, ..., Y_n$ that are dependent with known correlation and the variance in this case is bigger.

$$var\left[\bar{Y}\right] = var_{ind}[\bar{Y}]\frac{1+\rho}{1-\rho} = var_{ind}[\bar{Y}]\left(1 + \frac{2\rho}{1-\rho}\right) > var_{ind}[\bar{Y}]$$

On the figure 4.9 we can compare the variance for different level of correlation.

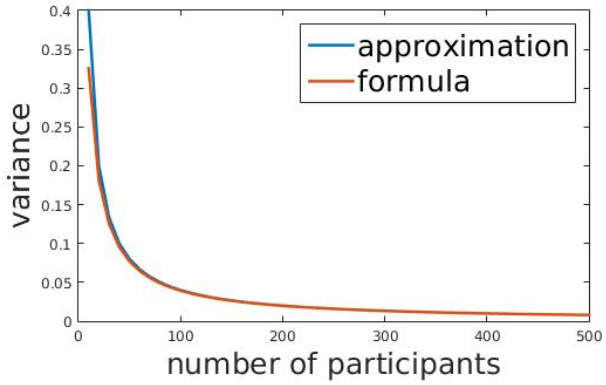


Figure 4.8: $\rho = 0.6$

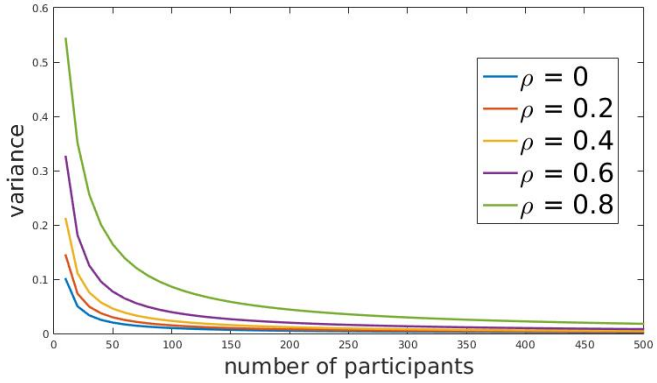


Figure 4.9: $\rho = 0.6$

4.4.3 Variance with skipping

Let's look at the variance of the next random variable:

$$\bar{Y}^k = \frac{Y_1 + Y_{1+k} + Y_{1+2k} + ... + Y_{1+(n-1)k}}{n}$$

where $corr(Y_{1+ik}, Y_{1+(i+h)k}) = \rho^{kh}$.

If we introduce new random variables $Z_1, Z_2, ..., Z_n$ such that $Z_1 = Y_1, Z_2 = Y_{1+k}, Z_3 = Y_{1+2k}, ..., Z_n = Y_{1+(n-1)k}$ and $r = \rho^k$, $\bar{Z} = \bar{Y}^k$. Then:

$$corr(Z_i, Z_{i+h}) = corr(Y_{1+(i-1)k}, Y_{1+(i+h-1)k}) = \rho^{kh} = r^h$$

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

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

Or approximation:

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

Let's return to the previous notation and get [check with script]:

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

or:

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

4.4.4 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 have $m = \frac{B}{kC_1+C_2}$ samples. Then variance, depending on k will be:

$$\sigma^2_{\hat{\mu}_{avg}}(k) = \frac{\sigma^2}{m}\frac{1-\rho^{2k}-2\rho^k/m+2\rho^{k(n+1)}/m}{(1-\rho^k)^2}$$

Or:

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

[should put analysis of function: convex, one minimum]

Let's understand what happens when we increase the k. If we denote m as the number of participants then on the budget B with

So we have some kind of trade-of here. And for sure it is not evident. 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. If we take 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.

, m individuals from the n is the number of participants. So both to informators and participants that are n in total we pay C_1 units of money for recruiting new people, but also to the m participants we pay C_2 units of money for participating in the research. So the budget is distributed in the following way:

$$B = nC_1 + mC_2$$

Here $m = \frac{n}{k+1}$ as we take each $k+1$ node as a participant. So having budget B and skipping each k node allows as to perform $n = \frac{(k+1)B}{(k+1)C_1+C_2}$ steps with $m = \frac{B}{(k+1)C_1+C_2}$ number of participants.

And then variation is actually the function of k.

With decreasing k what happens and say that it is not evident what is better. And our challenge was to study this problem formally. Can we quantify this number that we should skip. We want to solve it analytically. To obtain this result we used several simplifications. On the figures 4.6 presented experimentation results.

Studying variance of estimator is important for the construction of confidence interval and testing of hypothesis. The goal is to minimize variance. Let's look on the next function of k:

$$f(k) = \frac{\sigma^2}{\frac{B}{kC_1+C_2}} \frac{1+\rho^k}{1-\rho^k}$$

.

Observation 2 *Just thinning of sample doesn't help*

It means that we should look on function

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

it is increasing

Observation 3 *Skipping reduces variance*

If we fix m and will increase k:

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

.

is decreasing function.

4.4.5 General case

In the previous example we were able to write the variance of the estimator because the correlation between all the collected samples was known. We would like to generalize formula: where RDS is random walk on the arbitrary 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 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 X_i and X_j depends only on $j - i$:

$$cov(X_i, X_j) = cov(X_0, X_{j-i})$$

And then the formula for the variance

$$var(X_0) =< f, f >_\pi - < f, \Pi f >_\pi$$

and for the covariance:

$$cov(X_0, X_h) =< f, (P^h - \Pi)f >_\pi$$

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 more, how these formulas were derived, consult 6 chapter of the book [5].

For now we can write the formula for the variance of the estimator as:

$$var\left[\bar{X}\right] = \frac{1}{n^2}\left(nvar(X_i) + 2\sum_{i=1}^{n}\sum_{j|i<j}^{n} cov(X_i, X_j)\right) =$$

$$= \frac{1}{n^2}\left(n(< f, f >_\pi - < f, \Pi f >_\pi) + 2\sum_{i=1}^{n}\sum_{j|i<j}^{n} f, (P^{j-i} - \Pi)f >\right) \tag{4.5}$$

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 λ_i are eigenvalues, v_i are right eigenvectors and u_i are left eigenvectors of the matrix P^*:

$$P^h - \Pi = \sum_{i=2}^{n} \lambda_i^n v_i u_i^T \tag{4.6}$$

For more explanation refer to the chapter 6 of the book [5].

Using formulas 4.4.5 and 4.4.5 and reasoning similar to the one in the previous section, when we had know variance we will derive following formula for the variance of the estimator:

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

Again variance can be simplified:

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

Using the same reasoning as for simple case

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

Function: variance of mean with skipping having fixed budget and payments (simplified)

$$f(k) = \frac{1}{\frac{B}{kC_1+C_2}} \sum_{i=2}^{r} \frac{1+\lambda_i^k}{1-\lambda_i^k} < g, v_i >_\pi^2$$

[put interpretation of the formula, correlation factor and so on]. [put the graph]

4.4.6 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 $\hat{\mu}_{avg}$ as following:

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

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

$$MSE(\hat{\mu}_{avg}) = bias(\hat{\mu}_{avg})^2 + var(\hat{\mu}_{avg})$$

4.5 Data

To validate our theoretical results we used different sources of the data. First, we used random graphs with values generated with the algorithm. Then we used real network structures like part of Facebook [**?**] and assigned values with the algorithm.

The network structures where the nodes have some data are scares. But there are several sources like this: data from the Project 90[4], data from the project Add health[3].

4.5.1 Data from the Project 90

Project 90 [4] studied how the network structure influences on the HIV prevalence. Besides the data about social connection 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 4.10 you can see the graph structure built with the Gephi tool [2], where the attribute gender is depicted with color for every node.

4.5.2 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 going survey student as they grow. 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. 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 communities that have two schools, the school code (explain better about sister school).

On the figure **??** you can see the graph structure built with the Gephi tool [2], where the attribute race is depicted with color for every node.

4.5.3 Simulated form Gibbs distribution

4.5.4 Experiments

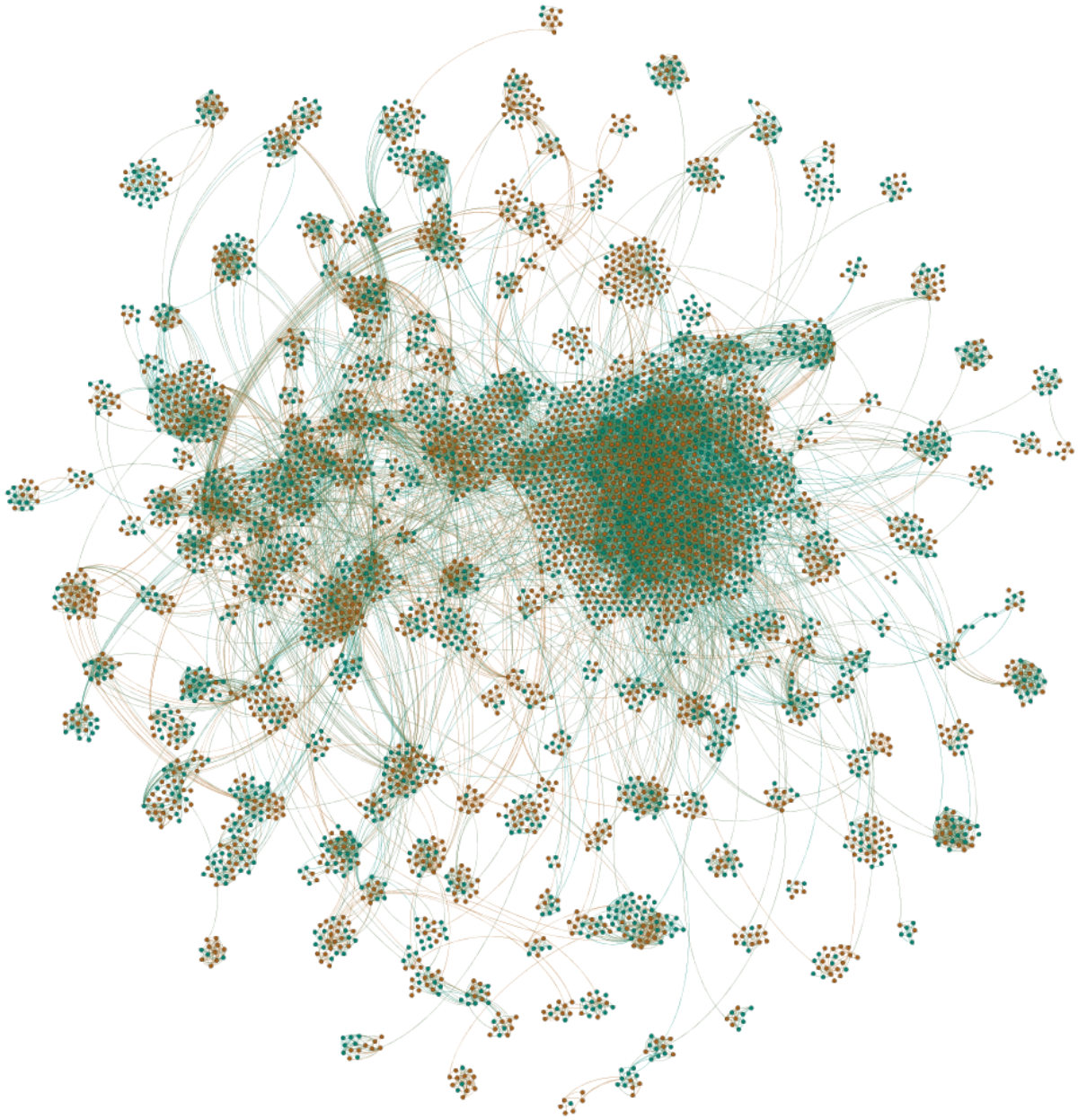


Figure 4.10: Graph from the Project 90 data. The colors of the node represent the gender: brown nodes correspond to the male participant, green nodes correspond to the female participants

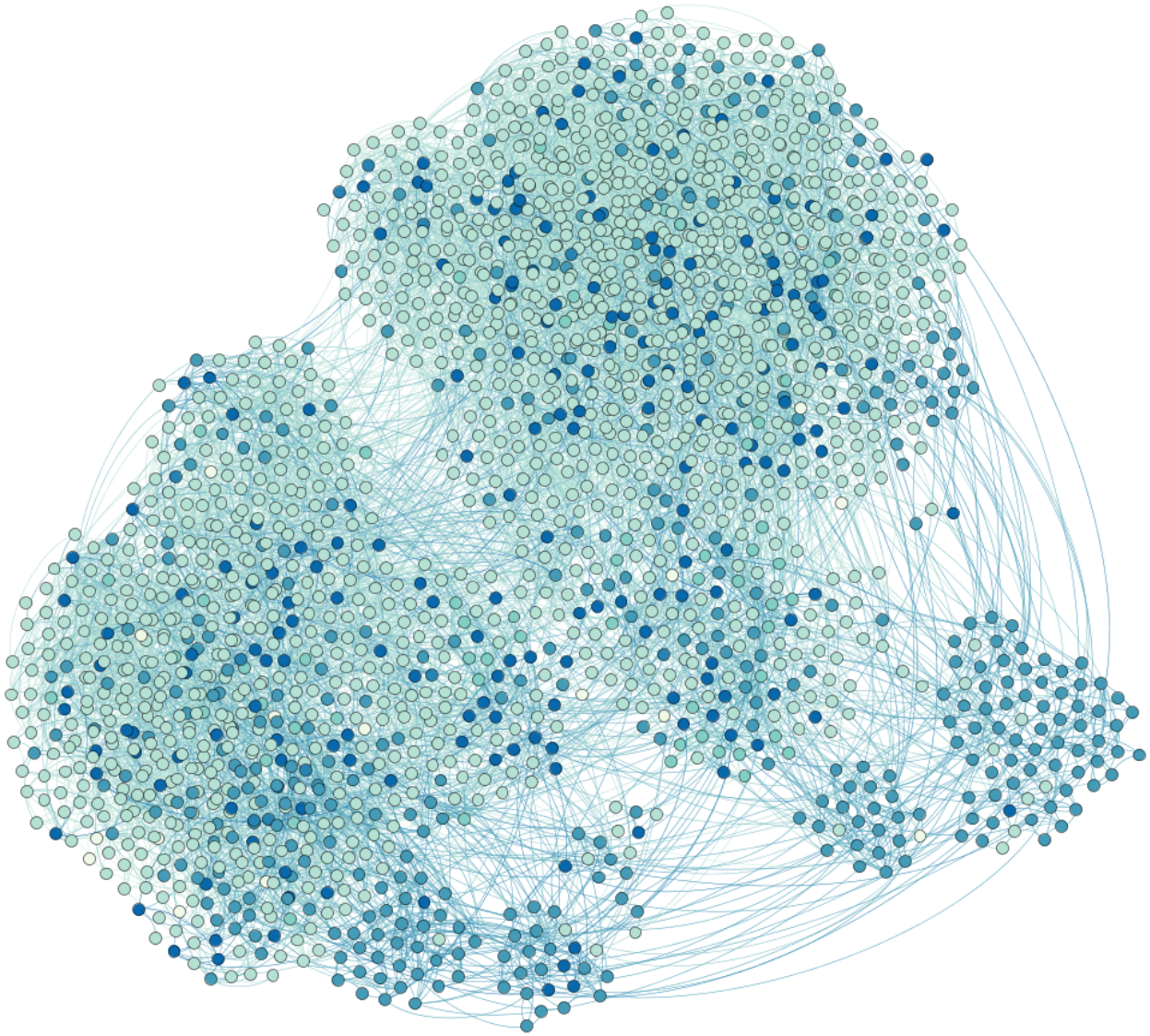


Figure 4.11: Graph from the Add Health data. The colors of the node represent the race: brown nodes correspond to the male participant, green nodes correspond to the female participants

Chapter 5

Comparing to other methods

We use mean as a estimator. Two commonly used estimators of sampling variance in RDS are the Salganik bootstrap estimator (SBE) and the Volz-Heckathon estimator (VHE). These estimators try to take into account the correlation between neighbors in the referral chain.

Let $X_1, X_2, ..., X_n$ be all collected samples during RDS. Then RDS estimate μ_f of the population mean of f is defined [10] as

$$\mu_f = \frac{1}{\sum_{i=1}^{n} 1/degree(X_i)} \sum_{i=1}^{n} \frac{f(X_i)}{degree(X_i)}$$

Our estimator (sample mean):

$$\mu_{f2} = \sum_{i=1}^{n} \frac{f(X_i)}{n}$$

Estimator μ_f indeed performs better than estimator μ_{f2} when values of the nodes depend on the degree of the node.

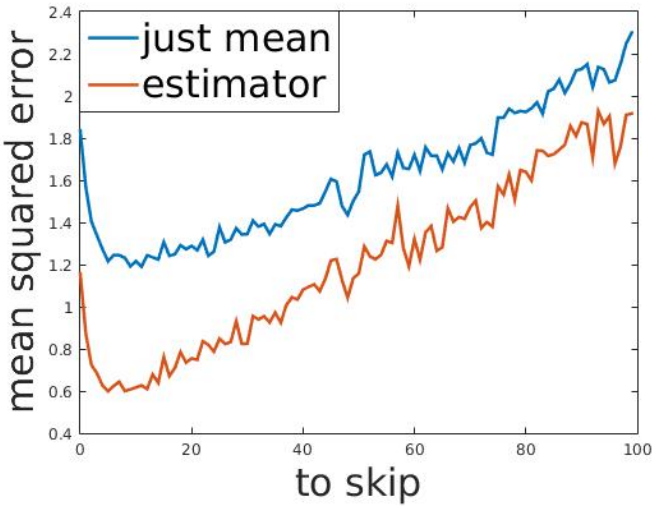


Figure 5.1: RGG(200, 0.13), measuring degree

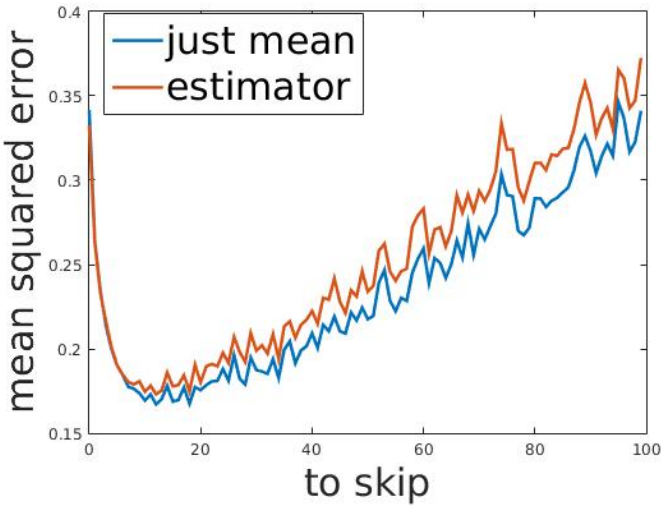


Figure 5.2: RGG(200, 0.13), measuring values

As stated also [10] the advantage to use estimtor1 appears only when the needed for estimation trait depends on the degree of the node. To support this statement they compare particularly the standard error of the sample mean and RDS estimate on the data sets from Project 90 and Add Health. The results are presented on the figure [put pictures].

But why it so much better? Why it is so suspicious? Possible explanation: degree weights are used to correct the fact that we see high degree nodes more times. Then for sure, if the values are related to the degree of the node it is useful. But if not it may do worth? but... it should not

Another explanation: so we give less weight to the nodes with high degree. But If they are more representative (like in graph with Gibbs field)....?

try: separate bias and variance

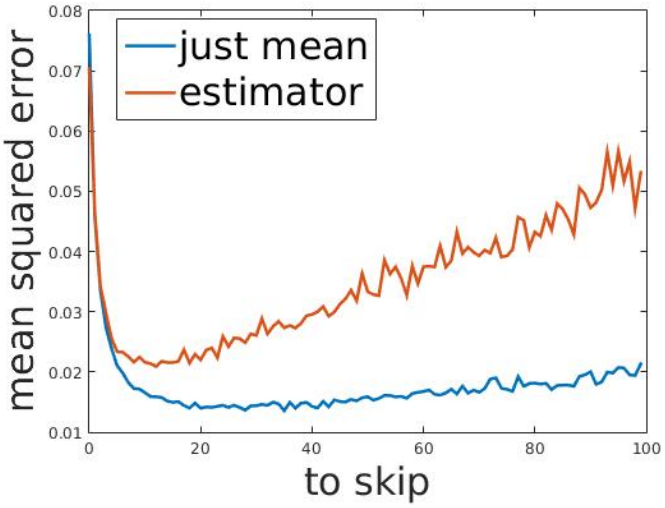


Figure 5.3: Project 90, measuring ??? race

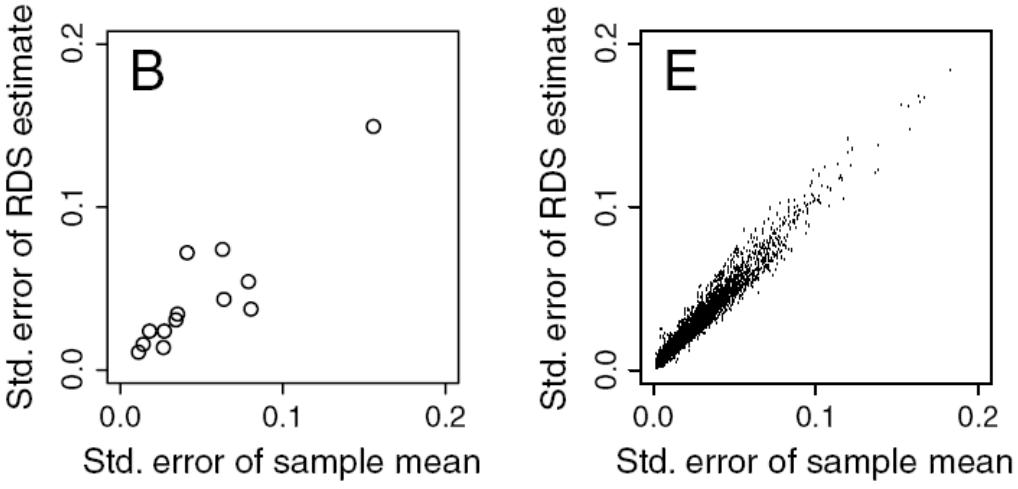


Figure 5.4: Comparison of standard error of RDS estimator and sample mean estimator on Project 90 data (left) and Add Health data (right) [10]

Chapter 6

Conclusion

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

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

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

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

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

[8] 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.

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

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

[11] 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.

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

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

[14] 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.