Polytech Nice Sophia Antipolis

Master IFI/ Ubinet track

Final report

Network sampling and discovery processes

Author:
Alina Tuholukova

Supervisors:
Konstantin Avrachenkov Giovanni Neglia

August 7, 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

Estimate an error

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 [9]

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 [9] 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) [4]

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 [6] 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 [5] 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 [9] 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

To skip or not to skip

The quality of estimation depends on the length of the chain and on the number of participants in the estimation. In order to make the chain longer we can separate the payment for coming and taking part in the testing and the payment only for providing the list of friends. This can be especially suitable for people from hard-to-reach populations, where one can obtain some amount of money without revealing needed information about him but only by pointing (or recruiting) his friends. There is a trade-off: on one hand we make the chain longer and reduce dependency between participants. On other hand we spend money on people who do not bring any information needed for research and finally there will be less participants. For now we will assume that it is not people who decide to participate or just provide the list of their friends, but researches. Thus, having fixed budget to conduct the RDS there is need to answer following questions: how much to pay for the participation in their research, how much to pay for simply providing the names of the friends and how many people invite for participation. The goal is to minimize error of the estimated parameter.

Let's say that the payment or cost of providing list of the friends is c_f and cost of participation is c_p. In this way each of n person that does one of actions gets the payment for providing friends c_f and part p of n people get additional payment c_p for also participation in test. Thus having fixed budget B:

$$B = n \cdot c_f + np \cdot c_p$$

Let's say that error of the estimated parameter is the function of n and p, $f(n, p)$ that decreases with increasing n and with decreasing p.

Insert here graph with the same p and increasing n and with the same n but increasing p.

I simulated values on the nodes according to the Gibbs distribution. Now, I want to estimated the average of these values with the help of the random walk. I use Metropolis-Hasting method to take the samples uniformly(maybe try simple random walk). The question is whether it is better to estimate the average value taking each sample or to skip some samples.

Assigning values to the nodes with the Gibbs distribution brings dependency of values between neighbors (the value of the property on one node depends on the values on its neighbors).

In this way the values of the nodes that we see on step k and on step $k + 1$ are dependent. So by skipping some nodes can decreases dependency.

The quality of estimation depends on the number of participants in the estimation and on the dependency between them. More participants there are the better is estimation. The less correlation

between participants the better. In order to make the chain longer we can separate the payment for coming and taking part in the testing and the payment only for providing the list of friends. This can be especially suitable for people from hard-to-reach populations, where one can obtain some amount of money without revealing needed information about him but only by pointing (or recruiting) his friends. There is a trade-off: on one hand we make the chain longer and reduce dependency between participants. On other hand we spend money on people who do not bring any information needed for research and finally there will be less participants. For now we will assume that it is not people who decide to participate or just provide the list of their friends, but researches. Thus, having fixed budget to conduct the RDS there is need to answer following questions: how much to pay for the participation in their research, how much to pay for simply providing the names of the friends and how many people invite for participation. The goal is to minimize error of the estimated parameter.

Let's say that the payment or cost of providing list of the friends is c_f and cost of participation is c_p. In this way each of n person that does one of actions gets the payment for providing friends c_f and part p of n people get additional payment c_p for also participation in test. Thus having fixed budget B:

$$B = n \cdot c_f + np \cdot c_p$$

Let's say that error of the estimated parameter is the function of n and p, $f(n, p)$ that decreases with increasing n and with decreasing p.

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 [3] 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%.

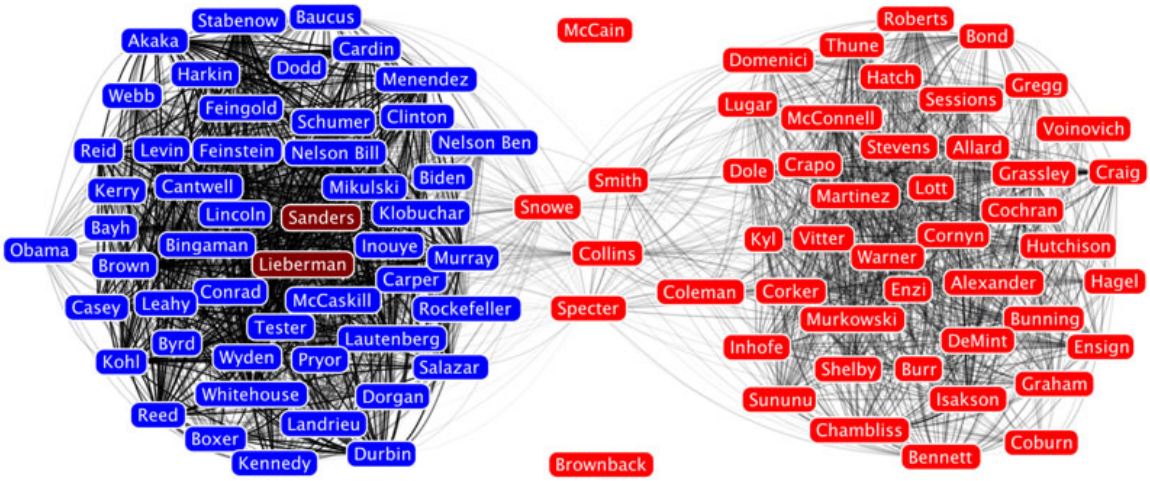

Figure 4.1: Voting patterns of U. S. Senators during 2007 [1]. The red labels represent Republicans, the blue labels represent Democrats, the brown labels represent two Independents.

Another study [10] 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 [8] 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], [2]. 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 [7] 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 [2] 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].

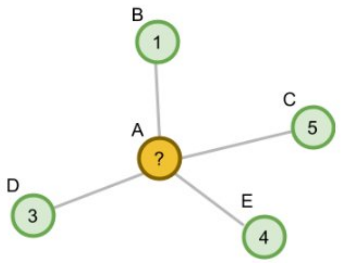

Figure 4.2: Preferential attachment graph with 200 vertices, 1 link for new arriving node and values $[1, ..., 5]$

4.2.6 Demonstration of random graphs with values

First, random geometric graph with 200 nodes and radius 0.13 was created, $RGG(200, 0.13)$. The set of values is $P = \{1, 2, ..., 10\}$. According to the first step of algorithm for each node was generated random property. The properties are depicted on the pictures as colors. Following pictures describe the properties of the graph after 2000 iterations of 2-3 steps for different temperature.

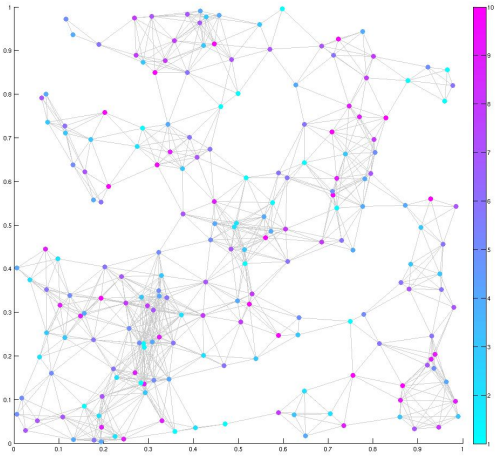

Figure 4.3: Random field

From the pictures we can observe that the level of dependency between values of the node changes with different temperature. 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 values 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:

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

On the figures ... we present correlation between values depending on this distance k for the graphs shown above.

4.3 Expected energy in steady state

I create graph and assign values to all nodes uniformly from all possible values. Then I start to change values on nodes according to Gibbs sampling having fixed temperature. I suppose that in steady state (after enough amount of steps) energy will have some expected value. Depending on temperature T we should be able to predict this energy.

If expected energy is known it can help detect convergence and indicate when it is time to stop running algorithm.

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

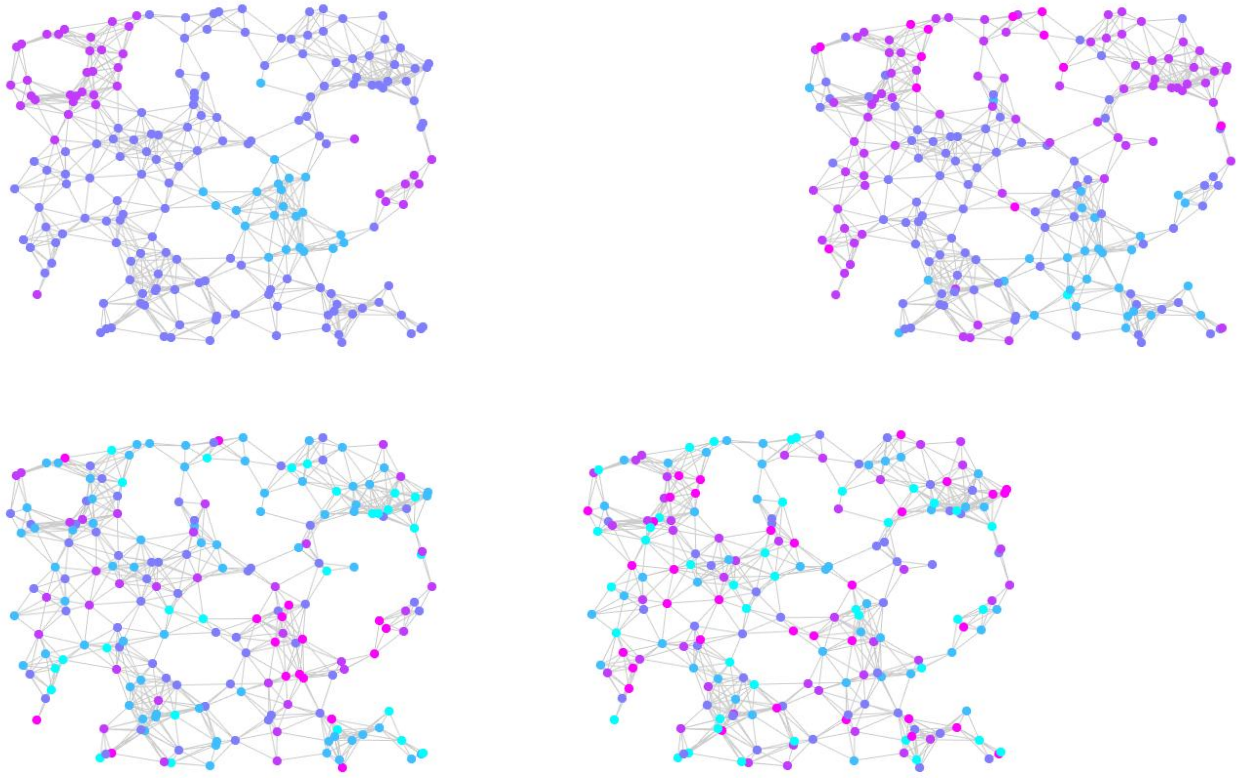

Figure 4.4: DFT, ruptureasdaosia poaisodiaposd

of values. However it is impossible to use only temperature T as a metric of values correlation. Temperature by itself does bring a lot of information. We should take into account structure of graph, possible values that can be assigned to nodes as well.

More appropriate metric can be energy of the graph or better, in how much times energy decreases from random configuration to configuration chosen from Gibbs distribution.

We can write expected energy by definition as

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

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

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

Both in 1 and 2 formulas summation is over all possible configurations and the number of all possible configuration is huge, $|P|^n$. One of the ways to calculate the expected energy is with Gibbs sampling, running algorithm until convergence and calculating then energy but it is the opposite of what we are trying to do.

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

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

In order to calculate expected energy at least approximately we will make some assumptions. But first let's notice that if values are assigned to nodes independently and from the same distribution, then expression for energy becomes

$$E[\varepsilon(\bar{x})] = mE[(x_i - x_j)^2] = m\left(Var(x_i - x_j) + E[(x_i - x_j)]^2\right)$$

$$= m\left(Var(x_i - x_j) + (E[x_i] - E[x_j])^2\right)$$

$$= m\left(Var(x_i) + Var(x_j) - 2Cov(x_i, x_j)\right) = 2mVar(x_i)$$

where x_i and x_j are random variance with the same distribution, m is the number of edges in the graph.

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 P.

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

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

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 $\varepsilon_i(x) = \sum\limits_{j|i \sim j} (x_i - x_j)^2$.

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

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 $av_P, av_P = average(P)$. 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

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

Now let's 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:

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

Show that energy is less then in random configuration
E

Actually I predict energy as

$$E[\varepsilon(\bar{x})] = 2mVar(x_i)$$

when I consider that all nodes are identically distributed. And as

$$E[\varepsilon(\bar{x})] = \sum_{i=1..n} N_i Var(x_i)$$

when distribution of values depends on the nodes degree.

$$E[(x_i - x_j)^2] = 2Var(x_i)$$

We could approximate this distribution as normal with mean $\frac{R}{2}$ and variance $\frac{T}{2d}$.

If $x_i \sim N(\frac{R}{2}, \frac{T}{2d})$ and $x_j \sim N(\frac{R}{2}, \frac{T}{2d})$ then the differences $x_i - x_j$ is distributed normally with mean 0 and variance $\frac{T}{d}$.

$$p(x_i - x_j = x) = \frac{e^{-\frac{dx^2}{2T}}}{\sum\limits_{x' \in P} e^{-\frac{dx'^2}{2T}}}$$

With increasing T approximation gets worse. In reality (in more real approxim)

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

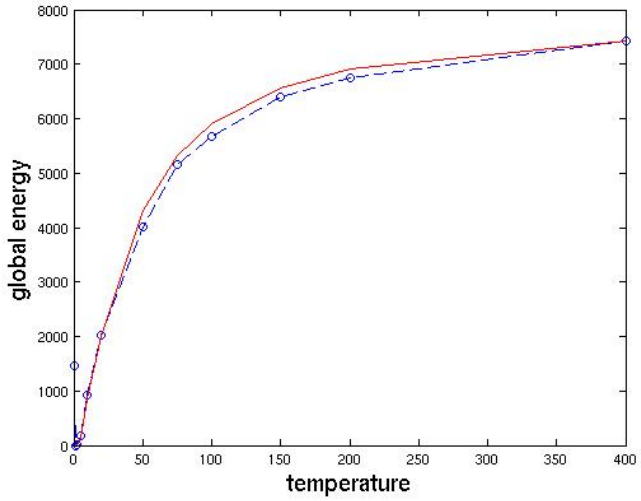

Figure 4.5: Random ER graph with 200 vertices and values $[1, ..., 5]$ p = 0.1

$$P(\bar{a}) = \frac{e^{-\frac{\varepsilon(\bar{a})}{T}}}{\sum\limits_{\bar{a}'} e^{-\frac{\varepsilon(\bar{a}')}{T}}}$$

$$\varepsilon(\bar{a}) = \sum_{i \sim j, i \leq j} (a_i - a_j)^2$$

$$E[\varepsilon(\bar{a})] = 2m\sigma^2$$

$$m = \sum_{i \in V} d_i$$

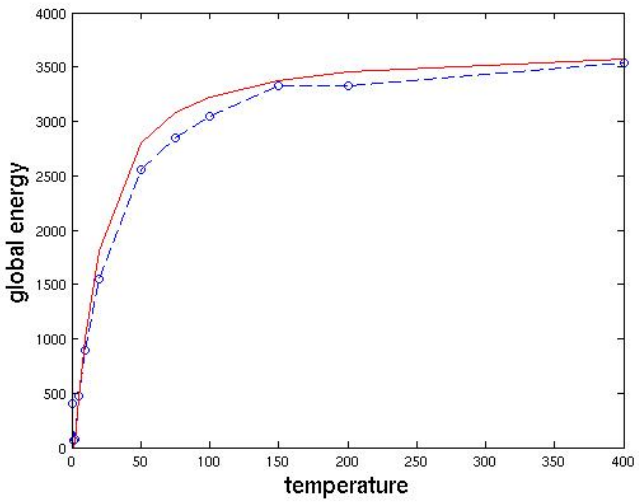

Figure 4.6: Random geometric graph with 200 vertices, radius 0.13 and values $[1, ..., 5]$

4.4 Error prediction

Studying variance of estimator is important for the construction of confidence interval and testing of hypothesis.

Let's look on the network graph where nodes have correlated values. Now let's assume that correlation between the nodes depends on the distance between them in the following way: nodes at the distance 1 have correlation ρ, at the distance 2 correlation ρ^2 and so on. If the distance between nodes i and j is k then $corr(X_i, X_j) = \rho^k$.

First, let's look at the line where nodes are correlated as described above. Then $corr(X_i, X_{i+h}) = \rho^h$. Now let's start to collect the values along the line starting from the first node, $X_1, X_2, ..., X_n$. Then we can count variation of the mean of $X_1, X_2, ..., X_n$.

$$var\left[\frac{X_1, X_2, ..., X_n}{n}\right] = var\left[\bar{X}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov(X_i, X_j) =$$

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

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

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

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

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

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

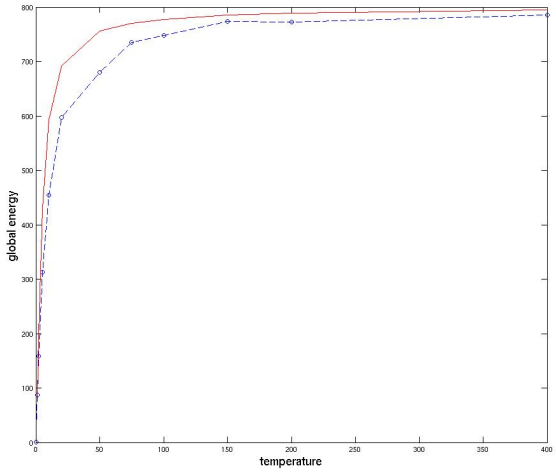

Figure 4.7: Chain graph with 100 vertices and values $[1, ..., 5]$

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

Let's simplify a bit expression for variance by approximated one.

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

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

$$var\left[\bar{X}\right] = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$$

Approximation is especially good with big n and ρ.

If random variables $X_1, X_2, ..., X_n$ were independent then the variance of $\bar{X}$ would be $var_{ind}[\bar{X}] = \frac{\sigma^2}{n}$.

But we consider random variables $X_1, X_2, ..., X_n$ that are dependent with known correlation and the variance in this case is bigger.

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

The less is correlation between nodes the closer are variances $var\left[\bar{X}\right]$ and $var_{ind}[\bar{X}]$.

Variance with skipping

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

$$\bar{X}^k = \frac{X_1 + X_{1+k} + X_{1+2k} + ... + X_{1+(n-1)k}}{n}$$

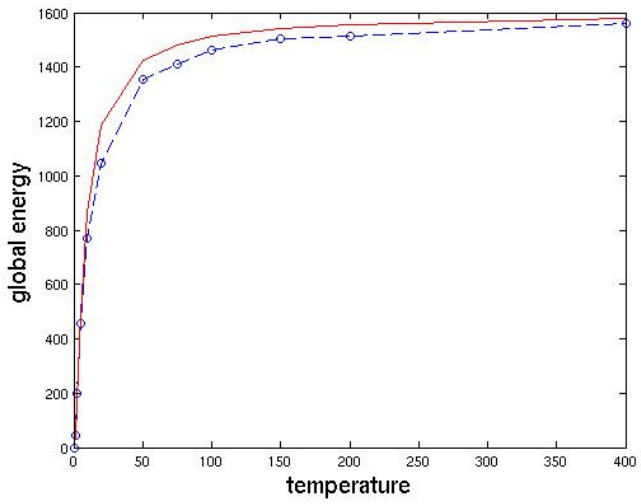

Figure 4.8: Grid on torus graph with 200 = 20x10 vertices and values $[1, ..., 5]$

So $corr(X_{1+ik}, X_{1+(i+h)k}) = \rho^{kh}$. Now let's introduce new random variable $Y_1, Y_2, ..., Y_n$ such that $Y_1 = X_1, Y_2 = X_{1+k}, ..., Y_n = X_{1+(n-1)k}$ and $r = \rho^k$, $\bar{Y} = \bar{X}^k$. Then $corr(Y_i, Y_{i+h}) = corr(X_{1+(i-1)k}, X_{1+(i+h-1)k}) = \rho^{kh} = r^h$.

To sum up we have random variables $Y_1, Y_2, ..., Y_n$ where $corr(Y_i, Y_{i+h} = \rho^{kh} = r^h$. But we already know that

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

Then

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

In RDS context
B - budget
C_1 - cost of one step of walk (individuals just provide the correct number of their contacts)
C_2 - cost of participation (cost of interview with individuals)
n - number of steps
m - number of participants from n
The next equality should be true:

$$B = n \cdot C_1 + m \cdot C_2$$

If we want to skip k steps between taking the node as a participant then

$$B = nC_1 + \frac{n}{k+1}C_2$$

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.

Then variance:

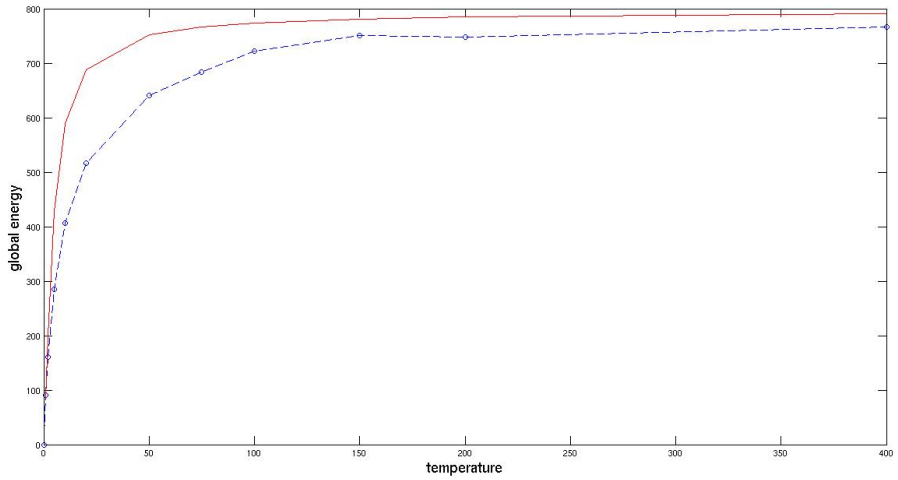

Figure 4.9: Preferential attachment graph with 200 vertices, 1 link for new arriving node and values $[1, ..., 5]$

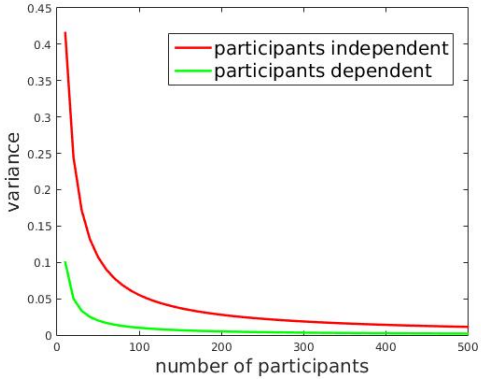

Figure 4.10: $\rho = 0.7$

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

The goal is to minimize variance. Let's look on the next function of k:

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

.

It has minimum when k is a solution for the following equation.

$$2C_1 \log(\rho)\rho^k k - C_1\rho^{2k} + 2C_2 \log(\rho)\rho^k + C_1 = 0$$

I don't know if there is explicit expression for the solution.
Check if the second derivative is always positive.

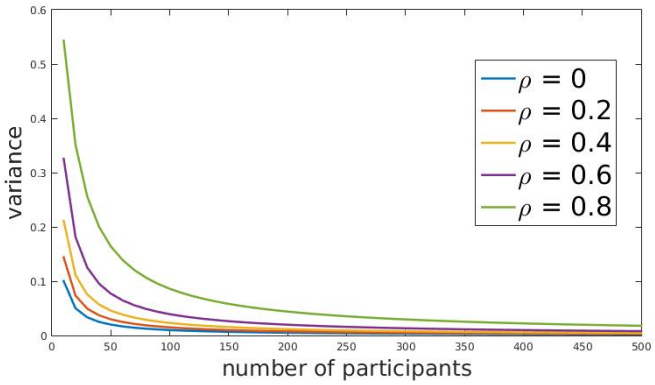

Figure 4.11: $\rho = 0.7$

Now, let's imagine that I have graph and I know ρ. I try to find k experimentally and using k from equation.

I will take as ρ covariance between neighbors on the graph with field. =(It did not work (only in ER graph, but I am not sure)

General case

Variance of mean in general case

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

Variance of mean simplified

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

Variance of mean with skipping

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

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

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

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

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

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

.

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

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

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

Chapter 5

Results

Chapter 6

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.

Chapter 7

Conclusion

Bibliography

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

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

[3] 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.

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

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

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

[7] 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.

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

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

[10] 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.