Polytech Nice Sophia Antipolis

Master IFI/ Ubinet track

Final report

Network sampling and discovery processes

Author:
Alina Tuholukova

Supervisors:
Konstantin Avrachenkov Giovanni Neglia

August 4, 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

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 [5]

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 [5] 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) [2]

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 [4] as

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

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 [3] 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 [5] 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

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

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

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

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

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

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 dependency of the value on the nodes from the values of its neighboring nodes. For instance, the study [1] 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%. Another study [6] 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!

The reason why we don't want to ignore homophility is because This is important because it affects on variance etc

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.

Moreover the level of correlation can be different depending on the network and the properties. [More examples] Because of this we would like also to be able to tune the correlation in the network

So we have to look not at the distribution of separate but at the joint distribution of the values.

4.3 Gibbs distribution

Instead of looking on distributions of the values on nodes independently, we will look at the joint distribution of values on all the nodes. In this way the value of the node depends on values of its neighbors. We have graph with n nodes. Each node i has a value $x_i \in \{1, 2, 3, ..., k\} = P$.

Algorithm

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.

Simulations

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.

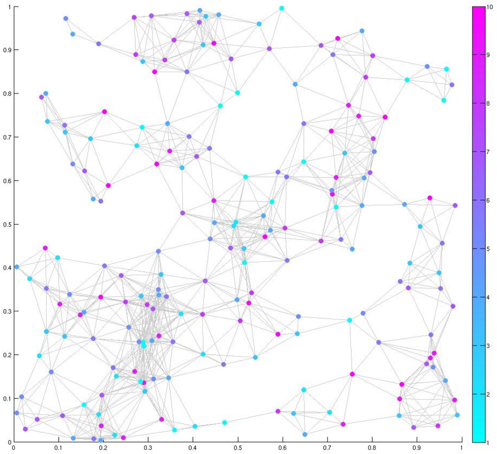

Figure 4.1: Random field

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] 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.

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

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

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

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

[6] 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.