Chapter 1

Presentation

About me

Slide 1

About internship

Slide 2

We study the techniques of network sampling. Let me motivate you with a real-world problem. So let's say that there is network of all people, connected . And some of them take drugs. We would like to study population that consume drugs (for example so understand how some diseases can spread). But there is no like a list of all people from which we can sample and they will not reveal themselves very easily So that's why they are hidden. But the researchers in this domain may know some individuals belonging to this population. Then he can ask persons that he knows to refer another individuals from this population and so on. To be more motivated participants are payed both for interviewing and for recruiting the others. So technique that is currently used for studying hidden population and it is called RDS.

Slide 3 Our goal is to improve estimates.

Problem: dependency Of course the way we perform sampling it is not just random sampling. In random sampling we know that the variance of mean is 1/n.

If estimator is unbiased the best estimator is the one which has the least variance.

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n}$$

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{n}\phi(correlation)$$

But in rds we also have some factor, that depends on how much the values are correlated (¿1).

So in order to decrease dependency we may try to skip some values. So by not taking some intermediate nodes we will decrease correlation factor but in the same time we will decrease the size of the sample.

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.

And our challenge was to study this problem formally.

Slide 4 Mathematical model For studying this problem we need some model. So for modeling the network of individuals and contact between them we have graph theory, random graphs. The RDS is random walk. And it leaves us simulating attributes on the nodes. So in order to correspond to our task the values should be assigned in such a way that they are correlated. To model this part we used Gibbs distribution. (paste formula). And here we have parameter T (temperature) and with the temperature we can kind of to tune correlation in the network.

Slide 5

First we explored a simple example when we have just a line and the correlation between nodes depends on the distance between them. so we have this formula from the number and it already shows us something. (explain the formula) The more is k

Slide 6

And them we were able to write the formula for general case. It is much more complicated but has similar structure.

(To say that we have limited budget and we pay intermediate nodes less)

About Phd (or further goals)

Slide 7 In the nearest future So the formula we have is quite complicated and requires some knowledge that we may not have (eigenvalues), Random matrix theory. Also we would like to test this method on real data. Like there is study about the network of obese people.

For PhD It matters how much is the correlation. So in reality we don't know this. So we need to be able to learn about it on the way. And with

example with hidden population it is not really possible. because the sample size is not so large so we can learn from it correlation first and then decide how much to skip. But it is more feasible with online RDS. where reward can be smaller or no reward at all.