Chapter 1

Presentation

During my PFE project I was studying network sampling techniques based on the random walks. The reason to use random walk for sampling is that sometimes there is no sampling frame (or it is very difficult to obtain it), it means there is no the list of population, but we can reach any individual through the chain of contacts.

One of such examples is studying traits in hidden population. The popular solution is applying technique that is called respondent-driven sampling. It is widely used for studying prevalence of HIV/AIDS among injection drug users, sex workers, men who have sex with men.

So what is this technique about. RDS begins with selecting group of initial participants that are called seeds. Each seed participates in the study and then recruits another participants. Participants are paid both for participation (interviewing) and recruitment of others. 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.

Problems. RDS can perform poorly if the groups of individuals form different communities. (Structure of network also affects a lot). It is known fact that friends tend to have similar traits: similar age, common language, the same university. This fact becomes a source of bias in chain-referral methods of sampling. Success of RDS due to financial reward. But money is restricted source.

To overcome this problems one could: collect all the data (as much as possible). But each interview has a cost. to skip but trade-off

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.

So the question is how to select participants to have the best estimates.

Mathematical model For modeling individuals and contact we use graphs. In order to study formally RDS can be regarded as Markov Chain (with some assumptions).

But during research we collect information from the individuals. So the nodes have to have attributes. We can assign values randomly, but it does not really shows reality. So apart of bias that gives graph structure there is also another source. Friends tend to have similar traits. How to assign attributes in the similar way as in real connections?

Difference between random sampling and RDS?

Assigning values according to the Gibbs distribution. We have graph with n nodes. Each node i has a value $X_i \in \{1, 2, 3, ..., k\} = P$. Then let

$$\bar{x}$$

be a configuration where

.

And probability to have configuration $\bar{x}$ is:

$$P(\bar{x}) = \frac{e^{-\frac{E(x)}{T}}}{\sum\limits_{x' \in P} e^{-\frac{E(x')}{T}}}$$

$$\frac{e^{-\frac{E_i(x)}{T}}}{\sum\limits_{x' \in P} e^{-\frac{E_i(x')}{T}}}$$

Create random configuration of properties on all nodes. Choose the node i

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

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

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

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. Continue 2-3 needed number of iterations.

Bibliography

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

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

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

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