Overview and comparison of random walk based techniques for estimating network averages

Konstantin Avrachenkov (Inria, France)

Ribno COSTNET Conference, 21 Sept. 2016

Motivation

Analysing (online) social networks one would like to know:

- How young is given social network?
- How many friends has an average network member?
- What proportion of population supports some political party?
- etc

Motivation

All such questions are related to the problem of estimating an average of a function $f(\cdot)$ defined on the network nodes.

Let $G = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = n$, $|\mathcal{E}| = m$, be an undirected graph representing a social network.

Then, we are interested to estimate

$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} f(v). \tag{1}$$

Constraints

Clearly, we are interested in the case when the topology of the network is not known and complete crawl of the network is not possible.

E.g., in online social networks, crawling is subject to an API limit on the number of requests per minute.

A standard Twitter account can make no more than one request per minute.

With this rate, we would crawl the entire Twitter social network in 950 years...

Snowball vs Random Walk

There are two main approaches to organize network sampling:

- Snowball sampling: sample all neighbours of a newly discovered node;
- Random Walk sampling: sample just one neighbour of a newly discovered node;

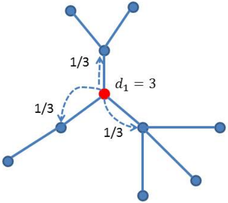
Our focus will be on the random walk based methods, since snowball sampling quickly requires excessive amount of resources and is biased by the principal eigenvector of the adjacency matrix.
(M. Newman, *Networks*, 2010; A. Maiya & T. Berger-Wolf, 2010)

Random Walk on Graph (Background)

The (discrete-time) Standard Random Walk $\{X_k,\ k = 0, 1, ...\}$ is defined by the transition probabilities

$$P(X_{k+1} = j \mid X_k = i) = p_{ij} = \begin{cases} 1/d_i, & \text{if j is a neighbour of i,} \\ 0, & \text{otherwise.} \end{cases}$$

where d_i is the degree of node i.



Random Walk on Graph (Background)

Assuming the network is connected, the stationary distribution of the standard Random Walk has a simple expression

$$\pi_i = \frac{d_i}{2m}.$$

This stationary distribution is achieved roughly after the relaxation time

$$t_{rel} = \frac{1}{1 - |\lambda_2|},$$

where λ_2 is the second largest by modulus the eigenvalue of the transition matrix P.

We also have

$$||P(X_k = i) - \pi_i|| \leq C|\lambda_2|^k, \quad k = 1, 2, \ldots$$

Metropolis-Hastings Sampling

Since the standard random walk is biased towards large degree nodes, it *might not* be a good idea to use directly the estimator

$$\hat{\mu}^{(N)} = \frac{1}{N} \sum_{k=1}^{N} f(X_k).$$

One way around this problem is to use Metropolis-Hastings chain with the following transition matrix $\mathbf{P}^{MH}$:

$$\mathbf{P}^{MH}_{ij} = \begin{cases} \frac{1}{\max(d_i,d_j)} & \text{if } j \neq i \\ 1 - \sum_{k\neq i} \frac{1}{\max(d_i,d_k)} & \text{if } j = i. \end{cases}$$

Metropolis-Hastings Sampling

By using the CLT for MCs (see e.g., Brémaud 1999), one can show the following central limit theorem for MH Chain.

Proposition

(Central Limit Theorem for MH) For MH Markov chain, it holds that

$$\sqrt{N}\left(\hat{\mu}^{(N)}_{MH}(\mathcal{G}) - \mu(\mathcal{G})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{MH}), \quad \textit{as} \quad N \to \infty,$$

where $\sigma^2_{MH} = \sigma^2_{ff} = \frac{2}{n}\mathbf{f}^T\mathbf{Z}\mathbf{f} - \frac{1}{n}\mathbf{f}^T\mathbf{f} - \left(\frac{1}{n}\mathbf{f}^T\mathbf{1}\right)^2$ *and where* $\mathbf{Z} = [\mathbf{I} - \mathbf{P} + \mathbf{1}\boldsymbol{\pi}^T]^{-1}$ *is the fundamental matrix.*

In the context of online social networks, the use of MH estimator was first proposed in (M. Gjoka et al, 2010).

Respondent Driven Sampling (RDS)

MH approach is known to be not very efficient because of frequent resampling.

D. Heckathorn and co-authors in early 2000's proposed to use the standard random walk but to unbias the estimator in the following way:

$$\hat{\mu}^{(N)}_{RDS}(\mathcal{G}) = \frac{\sum_{t=1}^{N} f(X_t)/d(X_t)}{\sum_{t=1}^{N} 1/d(X_t)} := \frac{\sum_{t=1}^{N} f'(X_t)}{\sum_{t=1}^{N} g(X_t)}, \qquad (2)$$

Respondent Driven Sampling (RDS)

Using 2D CLT for MCs from (E. Nummelin, 2002), we can show (our CSoNet'16 paper) that the RDS estimator is asymptotically consistent with a given asymptotic variance.

Proposition

The RDS estimate $\hat{\mu}_{RDS}(\mathcal{G})$ satisfies

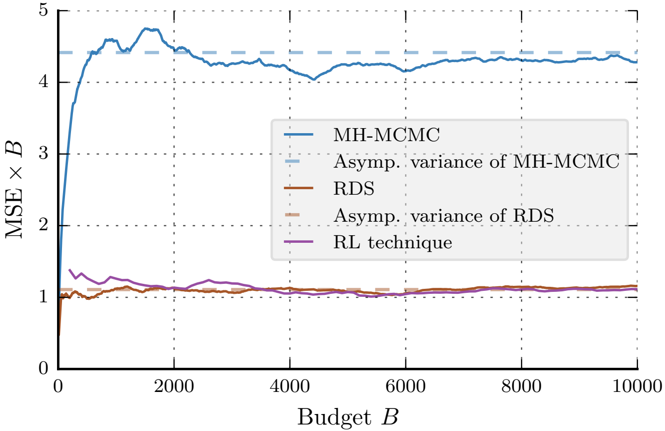
$$\sqrt{N}(\hat{\mu}^{(N)}_{RDS}(\mathcal{G}) - \mu(\mathcal{G})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{RDS}),$$

with σ^2_{RDS} given by

$$\sigma^2_{RDS} = d^2_{av}\left(\sigma^2_1 + \sigma^2_2\mu^2(\mathcal{G}) - 2\mu(\mathcal{G})\sigma^2_{12}\right),$$

where $\sigma^2_1 = \frac{1}{|E|}\mathbf{f}^T\mathbf{Z}\mathbf{f}' - \frac{1}{2|E|}\sum_x \frac{f(x)^2}{d(x)} - \left(\frac{1}{2|E|}\mathbf{f}^T\mathbf{1}\right)^2,$
$\sigma^2_2 = \sigma^2_{gg} = \frac{1}{|E|}\mathbf{1}^T\mathbf{Z}\mathbf{g} - \frac{1}{2|E|}\mathbf{g}^T\mathbf{1} - (\frac{1}{d_{av}})^2$ *and*
$\sigma^2_{12} = \frac{1}{2|E|}\mathbf{f}^T\mathbf{Z}\mathbf{g} + \frac{1}{2|E|}\mathbf{1}^T\mathbf{Z}\mathbf{f}' - \frac{1}{2|E|}\mathbf{f}^T\mathbf{g} - \frac{1}{d_{av}}\frac{1}{2|E|}\mathbf{1}^T\mathbf{f}.$

RDS Variance vs MH Variance



Tour based estimators

One very fruitful idea is to use tours for the construction of estimators. This idea goes back to the works (L. Massoulié *et al*, 2006) and (C. Cooper *et al*, 2013).

For instance, suppose we would like to estimate the number of edges in the network.

Consider the first return time to node i

$$T_i^+ = \min\{t > 0 : \; S_t = i \; \& \; S_0 = i\}.$$

The expected value of the first return time is given by

$$E[T_i^+] = \frac{1}{\pi_i} = \frac{2m}{d_i}.$$

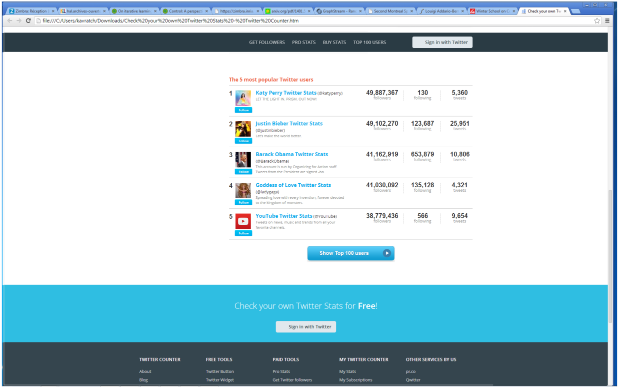
Tour based estimators

Let $R_k = \sum_{j=1}^{k} T_k$ be the time of the k-th return to node i. Then, we can use the following estimator for the number of edges

$$\hat{m} = \frac{d_i R_k}{2k}.$$

This idea can be easily extended to estimate a large variety of network characteristics.

Twitter as example



The 5 most popular Twitter users

		followers	following	tweets
1	Katy Perry Twitter Stats (@katyperry) LET THE LIGHT IN. PRISM. OUT NOW!	49,887,367	130	5,360
2	Justin Bieber Twitter Stats (@justinbieber) Let's make the world better.	49,102,270	123,687	25,951
3	Barack Obama Twitter Stats (@BarackObama) This account is run by Organizing for Action staff. Tweets from the President are signed -bo.	41,162,919	653,879	10,806
4	Goddess of Love Twitter Stats (@ladygaga) Spreading love with every invention, forever devoted to the kingdom of monsters.	41,030,092	135,128	4,321
5	YouTube Twitter Stats (@YouTube) Tweets on news, music and trends from all your favorite channels.	38,779,436	566	9,654

Check your own Twitter Stats for **Free**!

TWITTER COUNTER	FREE TOOLS	PAID TOOLS	MY TWITTER COUNTER	OTHER SERVICES BY US
About	Twitter Button	Pro Stats	My Stats	pr.co
Blog	Twitter Widget	Get Twitter followers	My Subscriptions	Qwitter

Twitter as example

Assuming that a rough estimation of the number of users is $500 \cdot 10^6$ and the average number of followers per user is 10, the expected return time from the nodes like "Katy Perry" or "Justin Bieber" is about $2 \cdot 10 \cdot 500 \cdot 10^6 / 50 \cdot 10^6 = 200$.

To obtain a decent error ($\leq 5\%$), we need about 1000 samples, and hence in total about 200000 operations. This is orders of magnitude less than the size of the Twitter follower graph!

Tour based estimators

In the tour-based estimators we return just to one node. Of course, hitting a set of several nodes should be much easier.

The problem is that the process becomes not Markovian...

Fortunately, there are at least two solutions to this problem.

Super-node idea

The first solution is... to change the problem...

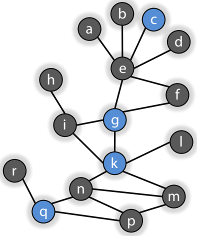


Figure: Original network

$\Rightarrow$

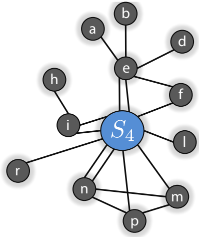


Figure: Modified network with a super-node

Super-node idea

Using the results from (F. Chung, 1997), we can show that

Proposition

The random walk on the modified graph with super-node has a smaller mixing time with respect to the walk on the original graph.

Super-node idea

The super-node can be grown in time (with some care!) to include for instance large degree nodes found during the tours.

More details in (K.A., B. Ribeiro and J. Sreedharan, ACM Sigmetrics 2016).

Super-node idea

The second idea is based on reinforcement learning (our CSoNet 2016 paper).

Define $Y_n := X_{\tau_n}$ for $\tau_n :=$ successive times to visit some set of nodes $\mathcal{V}_0$.

Then $\{(Y_n, \tau_n)\}$ is a semi-Markov process on $\mathcal{V}_0$.

Super-node idea

Let $\xi := \min\{n > 0 : X_n \in \mathcal{V}_0\}$ and define

$$\begin{aligned} T_i &:= E_i[\xi], \\ h(i) &:= E_i\left[\sum_{m=1}^{\xi} f(X_m)\right], \;\; i \in \mathcal{V}_0. \end{aligned}$$

Then (see e.g., Ross, 2013), the Poisson equation for the semi-Markov process (Y_n, τ_n) is

$$V(i) = h(i) - \beta T_i + \sum_{j \in \mathcal{V}_0} p_Y(j|i) V(j), \;\; i \in \mathcal{V}_0. \tag{3}$$

Here β is the desired stationary average of f.

Super-node idea

The reinforcement learning algorithm for the solution of the Poisson equation works as follows:

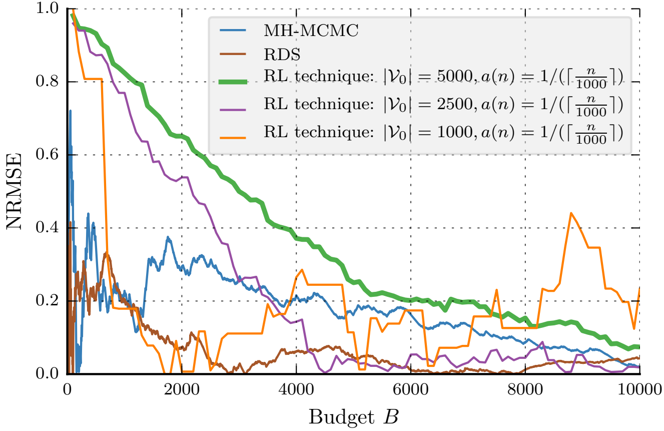
Let $\{z\}$ be IID uniform on $\mathcal{V}_0$. For each $n \geq 1$, generate an independent copy $\{X_m^n\}$ of $\{X_m\}$ with $X_0^n = z$ for $0 \leq m \leq \xi(n) :=$ the first return time to $\mathcal{V}_0$.

A reinforcement learning step is then

$$V_{n+1}(i) = V_n(i) \ + a(n)\mathbb{I}\{z = i\}\times \left[\left(\sum_{m=1}^{\xi(n)} f(X_m^n)\right) - V_n(i_0)\xi(n) + V_n(X_{\xi(n)}^n) - V_n(i)\right], (4)$$

where $a(n) > 0$ are stepsizes satisfying $\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty$.

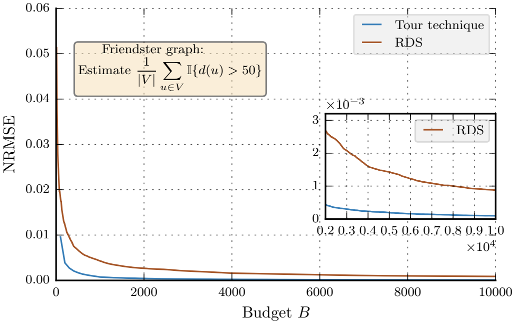
Super-node idea



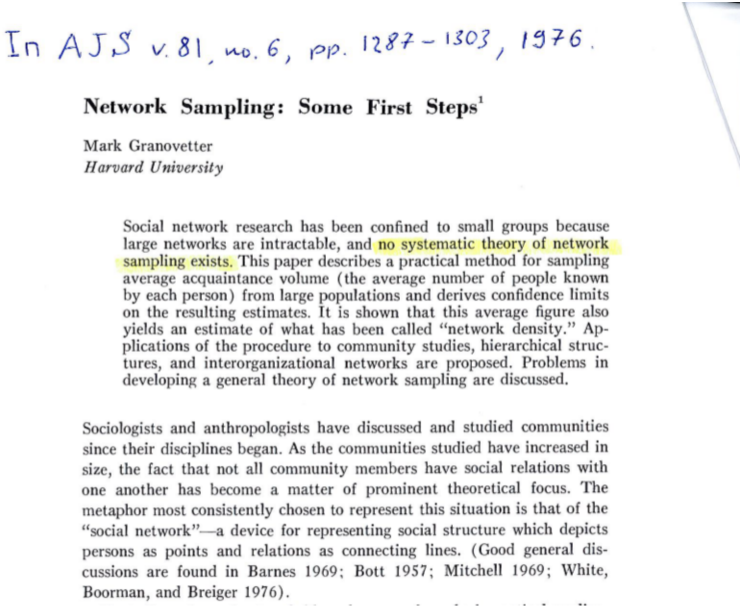
More details in (K.A., V. Borkar, A. Kadavankandy and J. Sreedharan, CSoNet 2016).

Tour estimator with super-node

A very promising idea is to combine tour-type RDS estimators with super-node:



Historical interlude



In AJS v. 81, no. 6, pp. 1287–1303, 1976.

Network Sampling: Some First Steps[1]

Mark Granovetter
Harvard University

Social network research has been confined to small groups because large networks are intractable, and no systematic theory of network sampling exists. This paper describes a practical method for sampling average acquaintance volume (the average number of people known by each person) from large populations and derives confidence limits on the resulting estimates. It is shown that this average figure also yields an estimate of what has been called "network density." Applications of the procedure to community studies, hierarchical structures, and interorganizational networks are proposed. Problems in developing a general theory of network sampling are discussed.

Sociologists and anthropologists have discussed and studied communities since their disciplines began. As the communities studied have increased in size, the fact that not all community members have social relations with one another has become a matter of prominent theoretical focus. The metaphor most consistently chosen to represent this situation is that of the "social network"—a device for representing social structure which depicts persons as points and relations as connecting lines. (Good general discussions are found in Barnes 1969; Bott 1957; Mitchell 1969; White, Boorman, and Breiger 1976).

Conclusion

Unfortunately, there is yet no unifying theory for network sampling.

Fortunately, there is a number of interesting open questions:

- Can we do a theoretical analysis for NMSE which includes both bias and asymptotic variance?
- Can we analyse main sampling schemes on some typical random graph models? Is there the best sampling algorithm for a given random graph model?
- Is there a substantial effect of network function on the performance of an estimator?
- Is there a scheduling approach better than snowball and random walk? Some compromise or generalization?
- Is there a better use of information in case of subsampling?

References

- Avrachenkov, K., Ribeiro, B. and Sreedharan, J.K. Inference in OSNs via Lightweight Partial Crawls. In Proceedings of the 2016 ACM Sigmetrics 2016.
- Avrachenkov, K., Borkar, V.S., Kadavankandy, A. and Sreedharan, J.K. Comparison of Random Walk Based Techniques for Estimating Network Averages. In Proceedings of CSoNet 2016.
- Avrachenkov, K., Neglia, G. and Tuholukova, A. Subsampling for Chain-Referral Methods. In Proceedings of ASMTA 2016.
- Brémaud, P. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

References

- Chung, F.R., 1997. *Spectral Graph Theory*. American Math. Soc., 1997.
- Cooper, C., Radzik, T. and Siantos Y. Fast Low-Cost Estimation of Network Properties Using Random Walks. In Proceedings of WAW 2013.
- Gjoka, M., Kurant, M., Butts, C.T. and Markopoulou, A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In Proceedings of IEEE Infocom 2010.
- Granovetter, M. Network Sampling: Some First Steps. *American Journal of Sociology*, 81(6), pp.1287-1303, 1976.

References

- Massoulié, L., Le Merrer, E., Kermarrec, A.M. and Ganesh, A. Peer Counting and Sampling in Overlay Networks: Random Walk Methods. In Proceedings of PODC 2006.
- Maiya, A.S. and Berger-Wolf, T.Y. Sampling Community Structure. In Proceedings of WWW 2010.
- Newman, M. *Networks: An Introduction*. Oxford university press, 2010.
- Nummelin, E. MC's for MCMC'ists. *International Statistical Review*, 70(2), pp.215-240, 2002.
- Salganik, M.J. and Heckathorn, D.D. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34(1), pp.193-240, 2004.

Thank you!

Any questions and suggestions are welcome.