Multivariate Effective Sample Size for Network Sampling Approaches

Haema Nilakanta, Zack W. Almquist, and Galin L. Jones, University of Minnesota

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Understanding Network Samples

Much of the network literature has focused on measurements which can only occur on complete network data; however there exist many networks which can only be accessed via sampling methods either because of the scale of the network (e.g., Facebook),privacy, or the population of interest cannot be enumerated (e.g., the homeless).

We focus on studying and comparing the sampling approaches:

- Simple Random Walk
- Metropolis-Hastings Random Walk

We create a framework for better assessing network sample quality by utilizing multivariate MCMC output analysis methods to get a more well-rounded understanding of the sample.

Goal:

1. Obtain a sample of m nodes via a random walk process
2. Estimate multiple network properties with the sample
3. Assess the multivariate sample quality

What Is a Network?

A network is a random graph, $G = (V, E)$ where,

- V = set of vertices or nodes
- E = set of edges or ties
- n = number of nodes or network size
- $|\mathrm{E}|$ = number of edges
- d_i = degree or number of edges for node i

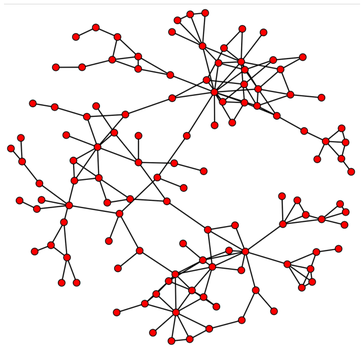

Figure 1: Example of a well connected undirected high school social network. Well connected component of the faux.mesa.high network from the ergm R package [1], $n = 120$ and $|E| = 185$.

Example: NYU Facebook Network

We applied the SRW and MH to obtain mean estimates of the 2005 New York University Facebook network [2]. We removed 56 nodes in order to make this network well connected. The resulting network had 21,623 nodes (users) and 715,673 edges (undirected online friendships).

Degree (# of friends)	Clustering Coefficient (% 3-friendship groups)	Prop Female	Prop Major=209
66.20	0.19	0.55	0.06

Table 1: True mean values of interest of NYU FB network.

We estimate the mean degree, mean clustering coefficient (ranges between 0 and 1; 0=no clustering and 1=network is fully connected), proportion of female users, and proportion of users with a specific major (labeled as 209).

Measures for assessing sample quality:
We implement the multivariate MCMC output analysis tools in the context of network sampling. This allows us to compute quantities such as:

- Means and multivariate standard errors using the multivariate batch means estimator [3]
- Multivariate effective sample size (ESS): equivalent number of samples an *iid* procedure would give with the same standard error
- $100(1-\alpha)\%$ confidence region and coverage probability
- Sampling stopping rule for some threshold $\varepsilon > 0$
- Number of unique nodes

Results of 1000 replications of a single chain, at random starting nodes, α=ε=0.05

	Stop Step	ESS	Coverage Probability	Unique Nodes	Efficiency	Precision
SRW	13237.00 (31.96)	10549.77 (25.79)	0.92 (0.00)	8179.76 (12.48)	✓	✓
MH	90099.44 (359.93)	6553.81 (8.78)	0.84 (0.00)	16967.55 (16.53)		

Table 2: Stop step (terminating sample size) , ESS, coverage probabilities, and number of unique nodes sampled by stopping time for $\varepsilon = 0.05$. Replications = 1000 and standard errors in parentheses. The SRW is more efficient and generates more precise estimates.

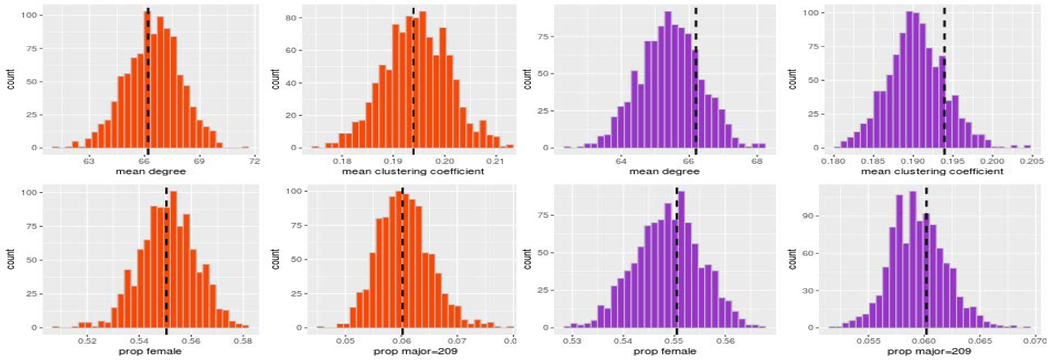

Figure 2: Mean estimates from SRW (orange) and MH (purple) at termination. Replications=1000. Black dashed line indicates true mean or proportion.

Methods

Simple Random Walk (SRW)
Probability of moving between node i and j if they are connected is $1/d_i$ and 0 otherwise. The stationary distribution of the SRW is $\pi_i = \frac{d_i}{2|E|}$. This walk is biased towards high degree nodes. To counter the bias we add weights, such that for any general function $h\colon V \to \mathbb{R}$, we write a mean estimate from the SRW as,

$$\hat{\mu}^{SRW} = \frac{1}{m}\frac{\sum_{t=0}^{m-1}\frac{h(V_t)}{d_{V_t}}}{\sum_{t=0}^{m-1}\frac{1}{d_{V_t}}}$$

Metropolis-Hastings Random Walk (MH)
To deal with oversampling high degree nodes we use the MH. Where we allow the walk to reject a move during the walk. The stationary distribution of the MH is $\pi_i = \frac{1}{n}$. Meaning each node is equally likely to be sampled, which implies,

$$\hat{\mu}^{MH} = \frac{1}{m}\sum_{t=0}^{m-1} h(V_t)$$

Discussion

By computing the multivariate standard errors from the random walk samples we obtain a better understanding of the sample quality at a more heuristic level. This gives researchers and practitioners knowledge about the:

- Relationship between estimates
- Effective number of samples they obtained from the correlated sampling procedure
- Confidence in the sample estimates
- How long to run the sampling process

We are continuing to study the relationship between network structure and chain convergence rates, to find if there are ways to identify the optimal sampling strategy for the networks of interest.

Acknowledgements & References

This research is supported by the University of Minnesota's Graduate School Interdisciplinary Doctoral Fellowship and the Minnesota Population Center

Contact information: nilak008@umn.edu

[1] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, Morris, and Martina. ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3):1–29, 2008.
[2] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social Structure of Facebook Networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.
[3] JM Flegal, J Hughes, and D Vats. mcmcse: Monte Carlo Standard Errors for MCMC. Riverside, CA and Minneapolis, MN. R package version, 2015.