Modeling Large Social Networks via Snowball Samples

Alex D. Stivala[1], Johan H. Koskinen[2], David A. Rolls[2], Peng Wang[2], Garry L. Robins[2], and Alessandro Lomi[3]

[1] Melbourne School of Psychological Sciences, The University of Melbourne, Australia.
stivalaa@unimelb.edu.au
[2] The Mitchell Centre for SNA, and Social Statistics Discipline Area, University of Manchester, U.K.
[3] Social Network Analysis Research Center, University of Lugano, Switzerland

Exponential random graph models (ERGMs) are a class of statistical models for complex network structures [3]. ERGMs have been used to study social networks, communication networks and organizational structures, and have been applied widely across the social sciences, from studies of animal social behavior to criminal networks to archaeology, as well as epidemiology and public health. Despite their broad appeal, computational problems have limited the applications of ERGMs to the analysis of relatively small social networks.

Under a homogeneity assumption whereby all structurally identical subgraphs are equally probable, an ERGM is a probability distribution with the general form $\Pr(X = x) = \frac{1}{\kappa} \exp\left(\sum_A \theta_A z_A(x)\right)$, where $X = [X_{ij}]$ is a 0-1 matrix of random tie variables, x is a realization of X, A is a configuration (a small set of nodes and a subset of ties between them), $z_A(x)$ is the network statistic for configuration A, θ_A is a model parameter corresponding to configuration A, and κ is a normalizing constant to ensure a proper distribution.

Markov chain Monte Carlo techniques for estimating ERGM parameters [5] are based on the generation of a distribution of random graphs by a stochastic simulation process. This process, which requires both a number of iterations to "burn in" the Markov chain and a large number of iterations to generate samples that are not too auto-correlated, is computationally intensive, and scales (at least) quadratically in the number of nodes in the network. This limits the size of networks to which an ERGM can be fitted in practical time. Furthermore, this process is inherently sequential, which limits the ability to take advantage of parallel computing.

In this work, we fit ERGMs to networks far larger than previously possible, by taking multiple snowball samples. Snowball sampling [1] is a technique to generate a sample of nodes in a network using the network structure itself. We then estimate ERGM parameters for each in parallel using conditional estimation, and combine the results with meta-analysis. The first work to take a similar approach was [9], in which 400 snowball samples are estimated in parallel and the results combined with meta-analysis [7]. However, as shown by [6], an ERGM specification for a subgraph sample cannot be projected to give predictions on the graph from which it was drawn, so this combined estimate might be quite incorrect. One way to handle this problem is described in [2], which requires knowing the size of the full network, and that estimation over the entire set of random tie variables is feasible. Another way is to use conditional estimation based on the snowball sampling structure [4], which is the technique we use.

We show that boot-strapped approaches to estimating confidence intervals in the meta-analysis improve estimation, and investigate the validity of statistical inference by repeating the estimation many times on networks simulated from known models (with 5000 and 10 000 nodes), checking the estimates against the known parameters to check bias. These networks have a parsimonious model with only four parameters, reflecting processes related to density, degree, network closure and connectivity. We also investigate Type I and Type II statistical errors, finding that estimates can be obtained without major bias and with reasonable Type I and Type II errors.

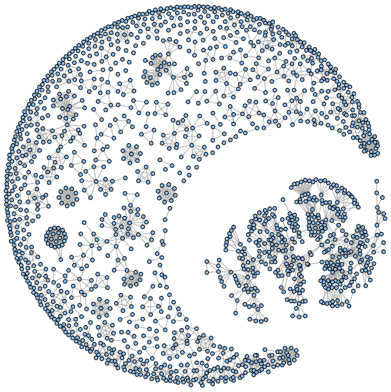

Figure 1: The network science co-authorship network.

As well as the simulated networks, we apply our method to some empirical collaboration networks, of sizes ranging from 1589 to over 40 000 nodes. The network science collaboration network that we analyse to illustrate our approach is depicted in Fig 1. For this network, the estimation took approximately 9.5 days to converge using currently available software [8], and only 1.5 hours using 20 parallel tasks (total CPU time 3.6 hours) with snowball sampling and conditional estimation. We illustrate the general value of our approach also in the analysis of larger interpersonal and interorganizational networks.

References

[1] Leo A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, pages 148–170, 1961.

[2] Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, 2010.

[3] Dean Lusher, Johan Koskinen, and Garry Robins, editors. *Exponential Random Graph Models for Social Networks*. Structural Analysis in the Social Sciences. Cambridge University Press, New York, 2013.

[4] Philippa E. Pattison, Garry L. Robins, Tom A. B. Snijders, and Peng Wang. Conditional estimation of exponential random graph models from snowball sampling designs. *Journal of Mathematical Psychology*, 57(6):284–296, 2013.

[5] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

[6] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41(2):508–535, 2013.

[7] T. A. B. Snijders and Chris Baerveldt. A multilevel network study of the effects of delinquent behavior on friendship evolution. *Journal of Mathematical Sociology*, 27(2–3):123–151, 2003.

[8] P. Wang, G. Robins, and P. Pattison. *PNet: program for the simulation and estimation of exponential random graph (p^*) models*. Department of Psychology, The University of Melbourne, 2009.

[9] B. Xu, Y. Huang, and N. Contractor. Exploring twitter networks in parallel computing environments. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment (XSEDE13): Gateway To Discovery*, page 20. ACM, July 2013.