Social Networks xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Snowball sampling for estimating exponential random graph models for large networks

Alex D. Stivala [a,*], Johan H. Koskinen [b], David A. Rolls [a], Peng Wang [a], Garry L. Robins [a]

[a] Melbourne School of Psychological Sciences, The University of Melbourne, Australia
[b] The Mitchell Centre for SNA, and Social Statistics Discipline Area, University of Manchester, United Kingdom

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Exponential random graph model (ERGM)
Snowball sampling
Parallel computing

ABSTRACT

The exponential random graph model (ERGM) is a well-established statistical approach to modelling social network data. However, Monte Carlo estimation of ERGM parameters is a computationally intensive procedure that imposes severe limits on the size of full networks that can be fitted. We demonstrate the use of snowball sampling and conditional estimation to estimate ERGM parameters for large networks, with the specific goal of studying the validity of inference about the presence of such effects as network closure and attribute homophily. We estimate parameters for snowball samples from the network in parallel, and combine the estimates with a meta-analysis procedure. We assess the accuracy of this method by applying it to simulated networks with known parameters, and also demonstrate its application to networks that are too large (over 40 000 nodes) to estimate social circuit and other more advanced ERGM specifications directly. We conclude that this approach offers reliable inference for closure and homophily.

1. Introduction

Exponential random graph models (ERGMs), first introduced by Frank and Strauss (1986), are a class of statistical model that are useful for modelling social networks (Lusher et al., 2013). Since their introduction, a body of work has been developed around ERGM theory and practice, including the introduction of new specifications for modelling social networks (e.g., Snijders et al., 2006; Robins et al., 2007; Goodreau, 2007), and more sophisticated methods for estimating ERGM parameters (e.g., Snijders, 2002; Handcock et al., 2008; Wang et al., 2009; Caimo and Friel, 2011; Hummel et al., 2012). Originally, the most common method for estimating ERGM parameters was maximum pseudo-likelihood (Strauss and Ikeda, 1990). More recently, Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) (Corander et al., 1998, 2002; Snijders, 2002; Hunter and Handcock, 2006) has become the preferred method (Robins et al., 2007). These techniques have several advantages over maximum pseudo-likelihood: if the estimation does not converge, a degenerate model is likely (a situation that maximum pseudo-likelihood does not indicate); converged estimates can be used to produce distributions of graphs in which the observed graph is typical for all effects in the model; reliable standard errors for the estimates are obtained (Robins et al., 2007); and point estimates are more accurate (Snijders, 2002).

Such MCMCMLE techniques require the generation of a distribution of random graphs by a stochastic simulation process. This process, which requires a number of iterations to "burn in" the Markov chain, as well as 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 a practical time. Furthermore, this process is inherently sequential (although several chains can be run in parallel, they each must be burned in), which limits the ability to take advantage of the parallel computing power available in modern high performance computing resources.

In this paper, we show how to fit ERGMs to certain large networks where model fitting using standard MCMC procedures would be impractical or impossible. The key idea takes advantage of recent developments in conditional estimation for ERGMs (Pattison et al., 2013) to take multiple snowball samples, estimate ERGM parameters for each sample in parallel, and combine the results with meta-analysis.

To the best of our knowledge, the work of Xu et al. (2013) is the first to take a similar approach. Xu et al. use a special data-intensive supercomputer to estimate an ERGM for a Twitter

* Corresponding author at: Melbourne School of Psychological Sciences, The University of Melbourne, Victoria 3010, Australia. Tel.: +61 3 8344 7035; fax: +61 3 9347 6618.
E-mail address: stivalaa@unimelb.edu.au (A.D. Stivala).

http://dx.doi.org/10.1016/j.socnet.2015.11.003
0378-8733/

"unfollow" network with over 200 000 nodes, estimating each of nearly 400 samples in parallel (by running statnet (Handcock et al., 2008) independently on each sample), and combining the results with meta-analysis (Snijders and Baerveldt, 2003). However, as Pattison et al. (2013) show, simply estimating the parameters of snowball samples without taking account of the snowball sampling structure, and assuming they are estimates of the full network, can lead to quite incorrect estimates. The issue is that, for a large class of models, standard parameter estimates for a graph are dependent on the number of nodes N and do not scale up in a consistent manner as N increases (Rolls et al., 2013; Shalizi and Rinaldo, 2013). Further, the Xu et al. (2013) method is applied only to the single Twitter unfollow network, for which the true values are not known, so there can be no comparison of true and estimated parameters, and therefore the reliability of the parameters obtained from their meta-analysis could not be assessed.

The motivations for fitting ERGMs to data are several. Usually, the aim is to infer whether certain well-established network processes that lead to tie creation are consistent with the data, and to parse apart different processes that might be operating simultaneously. This can be done by parameterizing competing explanatory processes and then inferring which of these are significant (Lusher et al., 2013). But, further, with precise parameter estimates, simulation of the model results in a distribution of graphs that can be interpreted as consistent with the data (at least in regards to the fitted effects). This distribution can be treated as the range of plausible graphs in a population of networks, from which a number of conclusions may be drawn. For instance, the population might relate to school classrooms or to communities of drug users (Rolls et al., 2013).

With very large data, however, the second motivation is often of less concern, because the idea of a "population" of large data is not always coherent. (There is for instance only one world wide web, not a population.) In this case, the interest is more typically on understanding the network processes within the data, such as closure and homophily. In this article, then, we are most interested in the validity of statistical inference for our procedure and hence we focus on type I and type II errors in our results.

2. Exponential random graph models

Under a homogeneity assumption whereby parameters are equated for all structurally identical subgraphs, an ERGM is a probability distribution with the general form

$$Pr(X = x) = \frac{1}{\kappa} \exp\left(\sum_A \theta_A z_A(x)\right) \tag{1}$$

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,
- κ is a normalizing constant to ensure a proper distribution.

In the present work we will be using only undirected graphs, so the matrix X is symmetric. Assumptions about which ties are independent, and therefore the configurations A allowed in the model, determine the class of model.

In the simplest case, where all tie variables are assumed to be independent, the ERGM reduces to a Bernoulli random graph distribution, otherwise known as a simple random graph or Erdős-Renyi random graph (Gilbert, 1959). In such a model only one configuration is used, an edge between two nodes, with the network statistic $z_L(x)$, the number of edges in the network, and the corresponding parameter θ_L.

The Markov dependence assumption, that two tie variables are conditionally independent unless they have a node in common, leads to the class of *Markov random graphs* (Frank and Strauss, 1986). In such models, the subgraph configurations include stars (in which a node has ties to two or more other nodes) and triangles (three mutually connected nodes). Stars can be further categorized as 2-stars, a subset of three nodes in which one node is connected to each of the other two, 3-stars, a subset of four nodes in which one node is connected to each of the other three, and so on, in general giving k-stars. Note that configurations are nested inside each other, for example a triangle contains three 2-stars. Associated with these is the *alternating k-star* statistic (Snijders et al., 2006), which is a weighted sum of the number of k-stars from $k = 2$ to $k = N - 1$ (where N is the number of nodes), with the sign alternating:

$$z_{AS} = \sum_{k=2}^{N-1} (-1)^k \frac{S_k}{\lambda^{k-2}} \tag{2}$$

where S_k is the number of k-stars and $\lambda \geq 1$ is a damping parameter which reduces the impact of higher order stars as it is increased. The alternating star parameter provides modelling flexibility in fitting node degree distributions, and alleviates model degeneracy. Throughout we use $\lambda = 2$, as suggested by Snijders et al. (2006) and modelling experience.

A more general class of model is based on *social circuit dependence* (Snijders et al., 2006; Robins et al., 2007) and often parameterized with *higher order parameters* such as the *alternating k-triangle* and *alternating k-two-path* (or alternating two-path) statistics. A k-triangle is a combination of k individual triangles which all share one edge, useful for modelling transitivity in the network. The alternating k-triangle statistic was defined in Snijders et al. (2006), and can be expressed as:

$$z_{AT} = 3T_1 + \sum_{k=1}^{N-3} (-1)^k \frac{T_{k+1}}{\lambda^k} \tag{3}$$

where T_k is the number of k-triangles. Again, we set the damping parameter to be $\lambda = 2$ throughout.

The k-two-path configuration is the number of distinct paths of length two between a pair of nodes, equivalent to the k-triangle configuration without the common (or "base") edge. Analogous to the alternating k-star and alternating k-triangle statistics, the alternating k-two-path statistic was defined in Snijders et al. (2006), and can be expressed as:

$$z_{A2P} = P_1 - \frac{2P_2}{\lambda} + \sum_{k=3}^{N-2} \left(\frac{-1}{\lambda}\right)^{k-1} P_k \tag{4}$$

where P_k is the number of k-two-paths. We use $\lambda = 2$ throughout.

These configurations are illustrated in Fig. 1. Software to fit and simulate ERGMs using these configurations includes PNet (Wang et al., 2009) and statnet (Handcock et al., 2008).

The alternating k-triangle and alternating k-two-path statistics can also be expressed in terms of edgewise and dyadic shared partners as the "geometrically weighted edgewise shared partner" (GWESP) and "geometrically weighted dyadic shared partner" (GWDSP) statistics, respectively (Hunter, 2007). The statnet software package (Handcock et al., 2008) uses GWESP and GWDSP rather than alternating k-triangle and alternating k-two-path statistics by default (Hunter, 2007).

All the configurations discussed so far have been structural, without the consideration of nodal attributes. In addition we wish to consider how an attribute (covariate) on a node can affect the

Please cite this article in press as: Stivala, A.D., et al., Snowball sampling for estimating exponential random graph models for large networks. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2015.11.003