Social Networks 49 (2017) 67–80

Contents lists available at ScienceDirect

Social Networks

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

Stable exponential random graph models with non-parametric components for large dense networks

S. Thiemichen, G. Kauermann*

Institut für Statistik Ludwigs-Maximilians-Universität München, Germany

ARTICLE INFO

Article history:
Available online 27 December 2016

Keywords:
Exponential random graph models
Conditional independence
Subsampling
Smooth non-parametric components
Network analysis

ABSTRACT

Exponential random graph models (ERGM) behave peculiar in large networks with thousand(s) of actors (nodes). Standard models containing 2-star or triangle counts as statistics are often unstable leading to completely full or empty networks. Moreover, numerical methods break down which makes it complicated to apply ERGMs to large networks. In this paper we propose two strategies to circumvent these obstacles. First, we use a subsampling scheme to obtain (conditionally) independent observations for model fitting and secondly, we show how linear statistics (like 2-stars etc.) can be replaced by smooth functional components. These two steps in combination allow to fit stable models to large network data, which is illustrated by a data example including a residual analysis.

1. Introduction

The analysis of network data is an emerging field in statistics. It is challenging both model-wise and computationally. Recently, Goldenberg et al. (2010), Hunter et al. (2012), Fienberg (2012) published comprehensive survey articles discussing new statistical approaches and developments in network data analysis. We also refer to the monograph of Kolaczyk (2009) for a general introduction to the field, or the recent book of Lusher et al. (2013), which focuses on a specific and widely used class of network models, so-called exponential random graph models (ERGM).

In its most simple form a network consists of a set of n nodes (actors) which are potentially linked with each other through edges. These edges between the actors are thereby the focus of interest. Notationally a network can be expressed as a $n \times n$ (random) adjacency matrix $\boldsymbol{Y}$ with entries $Y_{ij} = 1$ if node i and j are connected, and $Y_{ij} = 0$ otherwise. In undirected networks one has $Y_{ij} = Y_{ji}$ while for directed links we have $Y_{ij} = 1$ if a directed edge goes from node i to node j. For the sake of readability and notional simplicity we will concentrate here on undirected networks. The term $\boldsymbol{y}$ denotes a concrete realisation of $\boldsymbol{Y}$.

A common and powerful model for network data $\boldsymbol{Y}$ was proposed by Frank and Strauss (1986) as Exponential Random Graph Model (ERGM) taking the form

$$\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}\right) = \frac{\exp\left\{\sum_{l=0}^{p} s_l(\boldsymbol{y})\theta_l\right\}}{\kappa(\boldsymbol{\theta})}, \tag{1}$$

with $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)^t$ as parameter vector and $s(\boldsymbol{y}) = (s_0(\boldsymbol{y}), \ldots, s_p(\boldsymbol{y}))^t$ as vector of statistics of the network. In Eq. (1) the term $\kappa(\boldsymbol{\theta})$ denotes the normalizing constant, that is

$$\kappa(\boldsymbol{\theta}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp\left\{\boldsymbol{\theta}^t s(\boldsymbol{y})\right\},$$

where $\mathcal{Y}$ is the set of all networks and accordingly the sum is over $2^{\binom{n}{2}}$ terms. It is therefore numerically intractable, except for very small graphs. We denote with $s_0(\boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j>i}^{n} y_{ij}$ the baseline statistic giving the number of edges in the (undirected) network, so that θ_0 serves as intercept. The interpretation of the remaining parameters θ_l, $l = 1, \ldots, p$, results through the corresponding conditional model for each single edge Y_{ij} given the remaining network $\boldsymbol{Y} \setminus Y_{ij}$, since

$$\text{logit}\left[\mathbb{P}\left(Y_{ij} = 1|\boldsymbol{Y}\setminus Y_{ij}; \boldsymbol{\theta}\right)\right] = \theta_0 + \sum_{l=1}^{p} \Delta_{ij} s_l(\boldsymbol{y})\theta_l, \tag{2}$$

where $\Delta_{ij} s_l(\boldsymbol{y}) = s_l(\boldsymbol{y} \setminus y_{ij},\ y_{ij} = 1) - s_l(\boldsymbol{y} \setminus y_{ij},\ y_{ij} = 0)$ is the so-called change statistics which is obtained by flipping the edge between nodes i and j from non-existent to existent.

* Corresponding author at: Institut für Statistik, Ludwigs-Maximilians-Universität München, Ludwigstr. 33, 80539 München, Germany.
E-mail address: goeran.kauermann@stat.uni-muenchen.de (G. Kauermann).

http://dx.doi.org/10.1016/j.socnet.2016.12.002
0378-8733/

Exponential random graph models are numerically unstable, in particular if the number of actors n gets large, and (simple) linear network statistics like the number of 2-stars or triangles are included. Hence, for large networks one is faced with two relevant problems. First, the model itself is in its (simple) linear formulation notoriously unstable leading to either full or empty networks. This issue is usually called degeneracy problem, see, for example, Schweinberger (2011), Chatterjee and Diaconis (2013). Secondly, the estimation is per se numerically demanding or even unfeasible since numerical simulation routines are too time consuming. We aim to tackle both problems in this paper. First, we propose the use of stable statistics which are derived as smooth, non-parametric curves. Secondly, instead of fitting the model to the entire network we propose to draw samples from the network adjacency matrix $\boldsymbol{y}$ such that estimation in each sample is numerically (very) easy.

We emphasize that it is the combination of the two ideas that allows to fit Exponential Random Graph Models to large and sufficiently dense networks. That is to say that only in large networks the sampling approach is useful and feasible to provide sufficient information for estimation. Moreover, especially in large networks we are faced with instability problems where the proposed smooth, non-parametric statistics naturally stabilise the model. Hence, even though the two ideas proposed in this paper are separate, only their combination makes them really beneficial for network data analysis.

Schweinberger (2011) denotes network statistics (and the corresponding ERGM) as unstable if the statistics is not at least of order $O(n^2)$. In fact he shows that any k-star or triangle statistics is unstable leading to an odd behaviour of model (1). Effectively, unstable networks are either complete (i.e. have all possible edges) or empty (i.e. all nodes are unconnected) unless for a diminishing subspace of the parameter space for n increasing. If n gets large it is therefore advisable to replace the statistics in model (1) by stable statistics of order $O(n)$. A first proposal in this direction are alternating star and alternating triangle statistics as proposed in Snijders et al. (2006), or geometrically weighted statistics as proposed in the context of curved exponential random graph models, see Hunter and Handcock (2006). Hunter (2007) shows that from a modelling point of view the alternating statistics are equivalent to geometrically weighted degree or geometrically weighted edgewise shared partners, respectively. Both approaches stabilize the models but for the price of less intuitive interpretations of the parameter estimates. We propose an alternative by making use of non-parametric models based and the technique of smoothing (see, e.g., Ruppert et al., 2003). The non-parametric model thereby maintains the interpretability of the ERGM based on the conditional model (2). To motivate our idea we start with the conditional model (2) and replace the linear terms through non-linear smooth components. This leads to the conditional non-parametric model

$$\text{logit}\left[\mathbb{P}\left(Y_{ij}=1|\boldsymbol{Y}\backslash Y_{ij}\right)\right]=\theta_0+\sum_{l=1}^{p} m_l(\Delta_{ij}s_l(\boldsymbol{y})), \tag{3}$$

where $m_l(\cdot)$ are smooth functions which need to be estimated from the data. Models of type (3) have been proposed in a simple regression framework as generalized additive models, see, e.g., Hastie and Tibshirani (1990), or Wood (2006), but apparently the structure here is more complex as we are tackling network data. We additionally need to postulate that functions $m_l(\cdot)$ are monotone and bounded which in turn leads to stable network statistics in the definition of Schweinberger (2011). We make use of penalized spline smoothing which also allows to accommodate constraints on the functional shape leading to stable network models. In fact, assuming $m_l(\cdot)$ to be monotone and bounded, we may derive a non-parametric exponential random graph model from (3) which takes the form

$$\mathbb{P}\left(\boldsymbol{Y}=\boldsymbol{y}|\theta_0, m_l(\cdot), l=1,\ldots,p\right) = \frac{\exp\left\{s_0(\boldsymbol{y})\theta_0+\sum_{l=1}^{p}\sum_i\sum_{j>i} y_{ij}m_l(\Delta_{ij}s_l(\boldsymbol{y}))\right\}}{\kappa(\theta_0, m_l(\cdot), l=1,\ldots,p)} \tag{4}$$

Apparently, model (4) appears rather complex due to its semi-parametric structure and estimation looks like a challenging task. We will argue, however, that smoothing techniques can easily be applied and estimation becomes feasible by making use of sampling strategies in the network adjacency matrix $\boldsymbol{y}$ leading to numerically simple likelihoods and in fact consistent (though not efficient) estimates. Note that model (4) comprises 2-star or triangle statistics in that

$$m_l(\Delta_{ij}s_l(\boldsymbol{y})) = \sum_{\substack{k=1\\k\neq j}}^{n} y_{ik} + \sum_{\substack{k=1\\k\neq i}}^{n} y_{jk} \quad \text{or}$$

$$m_l(\Delta_{ij}s_l(\boldsymbol{y})) = \sum_{k=1}^{n} y_{ik}y_{kj}$$

Concerning geometrically weighted statistics as proposed by Snijders et al. (2006) and Hunter (2007), geometrically weighted degree (GWD) falls in formulation (4), while geometrically weighted edge-wise shared partners (GWESP) do not. This is due to the violation of Markov independence of the later as the change statistic $\Delta_{ij}s_l(\boldsymbol{y})$ does not only depend on the direct neighbourhood of y_{ij} but the rest of the network as well.

Estimation in exponential random graph models is cumbersome and numerically demanding as it requires simulation based routines. Snijders (2002) suggests the calculation of $\partial\kappa(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ in the score equation resulting from (1) using stochastic approximation. Hunter and Handcock (2006) propose to use MCMC methods in order to obtain the maximum likelihood estimate. The approach is extended and improved in Hummel et al. (2012). In a recent paper Caimo and Friel (2011) develop a fully Bayesian estimation routine by incorporating the so-called exchange algorithm from Murray et al. (2006) which circumvents the calculation or approximation of the normalisation constant for the price of extended MCMC sampling. A general survey of available routines for fitting Exponential Random Graph Models is given in Hunter et al. (2012). In fact, if the network is large, MCMC based routines readily become numerically infeasible. As aforementioned, we will therefore make use of subsampling the network adjacency matrix and fit the model to subsamples that allow for simple likelihoods. We follow ideas of Koskinen and Daraganova (2013). In fact, for models with k-stars or triangles only, the edges follow a Markovian independence structure by conditioning on parts of the network (see Frank and Strauss, 1986, or Whittaker, 2009). This is exemplified in a simple network with four nodes in Fig. 1. Conditioning on edges Y_{12}, Y_{14}, Y_{23}, and Y_{34} we find that Y_{13} and Y_{24} are conditionally independent, which

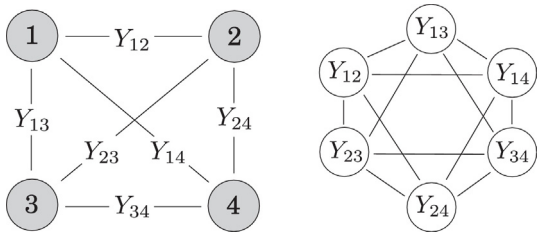

Fig. 1. Visualisation of the induced Markov independence graph (right) for an exponential random graph model for a simple 4-node network (left).