Social Networks 46 (2016) 76–86

Contents lists available at ScienceDirect

Social Networks

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

Focused model selection for social networks

Eugen Pircalabelu*, Gerda Claeskens

ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Available online 11 April 2016

Keywords:
Variable selection
Social network
Focused information criterion
Exponential random graphs
Network based models

ABSTRACT

We present a focused selection method for social networks. The procedure is driven by a *focus*, the main quantity we want to estimate well. It represents the statistical translation of a research hypothesis into parameters of interest. Given a collection of models, the procedure estimates for each model the mean squared error of the estimator of the focus. The model with the smallest such value is selected. We present focused model selection for (i) exponential random graph models, (ii) network autocorrelation models and (iii) network regression models to investigate existing relations in social networks. Worked-out examples illustrate the methodology.

1. Motivation

Social network analysis aims at understanding and explaining regularities and structures that describe relations linking individuals or any other social units such as organizations, political parties, etc. We present a methodology for model selection in the context of network based parameter estimation that is based on the *focused information criterion* (FIC) introduced and studied under various statistical contexts in the works of Claeskens and Hjort (2003, 2008a), Hjort and Claeskens (2006), Zhang and Liang (2011), Rohan and Ramanathan (2011), Claeskens (2012), Behl et al. (2014) among others. More recently in the works of Pircalabelu et al. (2015a,b) the FIC has been applied to estimate probabilistic graphical models. The goal of the present manuscript is to extend the application of FIC to social network models.

In the focused selection procedure, the focus, which is a function of the model parameters, plays a central role. Throughout the paper we denote the focus by μ. The focused information criterion is constructed to select from a set of models that model where the focus is optimally estimated in terms of *mean squared error* (MSE), which is the sum of the estimator's variance and its squared bias. The FIC selection procedure makes it possible to select explanatory models for focuses especially suited for social network analysis. Often, researchers are not interested in the whole network, but rather in quantities that summarize or describe phenomena such as actor centrality, edge prediction or strength of interpersonal influence between actors. For these quantities of interest (i.e., focuses) that can be formulated as functions of parameters of the underlying model, the FIC may be used to select models that estimate those focuses with small MSE. Since the true MSE is in general unknown, it needs to be estimated. The FIC value is such an estimated MSE, sometimes modulo some constants that do not depend on the models. See Section 4 for more details on the relation between FIC and MSE. First, we specify the focus and a list of plausible explanatory models. Next, we estimate the focus parameter and its associated MSE (or FIC) value. Finally, we select as chosen model for this focus that model of the list which has the smallest FIC value.

Unlike the classical information criteria, such as Akaike's information criterion (AIC, Akaike, 1973) and the Schwarz Bayesian information criterion (BIC, Schwarz, 1978) the focused information criterion allows to select a model that is *directed* towards the particular focus. That is, the FIC will select a model that performs well in MSE sense to the estimation of the focus, the quantity of interest. Different focuses, thus different interests, might lead to different models being selected, which is more informative since the selected model is determined based on specific research interests.

To motivate the focused selection approach for model selection for social networks we use three model classes, namely the exponential random graph models (ERGM), network autocorrelation models (NAM) and network regression models (NRM). See Section 2 for details regarding model specifications.

We start from the 'Florentine families' dataset (Breiger and Pattison, 1986; Padgett, 1994) which consists of a social network that records the marriage ties between 16 influential Florentine families (which family is linked with which other family), a social network that records the business ties between the 16 families, the wealth of each family, the number of seats on the civic council for each family, and the total number of ties linking the family to any of the other 116 families from Florence. The networks are represented

* Corresponding author.
E-mail addresses: eugen.pircalabelu@kuleuven.be (E. Pircalabelu), gerda.claeskens@kuleuven.be (G. Claeskens).

http://dx.doi.org/10.1016/j.socnet.2016.03.002
0378-8733/

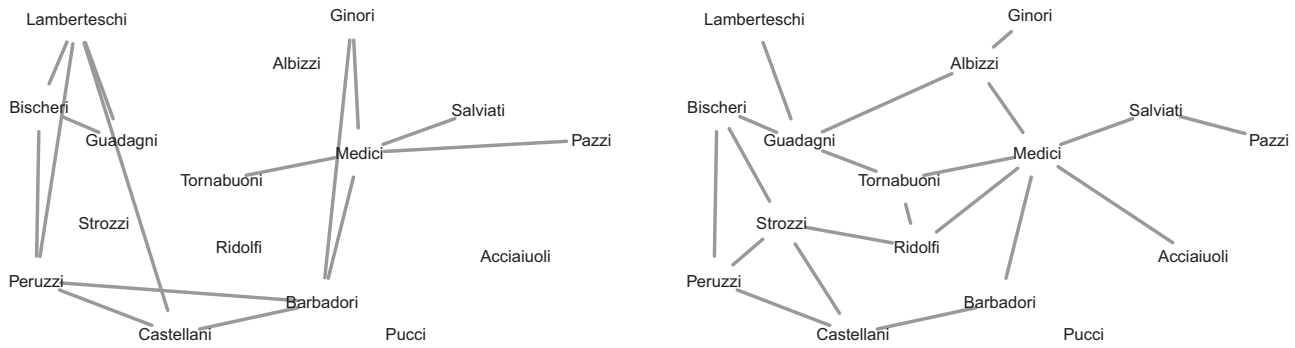

Fig. 1. Florentine families data. The left panel displays the business network while the right panel displays the marriage ties between the families.

by two adjacency matrices containing the value 1 on positions (i, j) and (j, i) if members from family i have married (or have business ties with) members from family j. A value of 1 is translated into a tie (edge) between families i and j to denote that the two families have marriage or business ties linking them together. Fig. 1 shows both the marriage and business ties that link together the Florentine families.

The data represent the ties formed around 1430, a period when the Strozzi and the Medici were considered to be adversaries. It was a period where the Medici family was powerful (Padgett and Ansell, 1993) and an alliance through marriage between the two families would have been improbable. In 1434 the Strozzi family have been driven into exile by Cosimo de' Medici, but due to the radical change in the political context in Florence, in 1508 the two families eventually formed an alliance through marriage (Bullard, 1979). The observed marriage network which is analyzed here does not however, have such a tie present, since it pertains to the period around 1430.

In light of this information, it is of interest to select explanatory models that best express the log odds ratio of such a tie being formed in 1430 between the two families. The log odds ratio is our first *focus* parameter. To estimate this focus, we use an exponential random graph model (ERGM) described in Section 2. A list of $32 = 2^5$ (all possible combinations of predictors) such models was considered, where the narrow, simplest model contained only one parameter, the edges parameter which acted similar to an intercept, and the most complex (full) model contained four extra parameters. All other models, were in between the simplest and the full model. We postpone the proper definition of the predictors to Section 5 as it suffices here to show what the method offers in practice. The aim is to select the suitable collection of parameters that provides the lowest FIC value for the estimated focus, in this case, the log odds ratio. Table 1 shows the two best ranking models using FIC, as well as the full and the narrow model (the latter which coincides with the best scoring AIC and BIC models). The best scoring FIC model to estimate the log odds ratio contains as predictor the wealth of the Florentine families, whereas the second best model suggests adding also the change statistic with respect to the number of triangle configurations and the Gwesp summary statistic (see Section 5). The AIC and BIC selected model (which is insensitive to the focus specification) suggests the usage of the simplest, narrow model without any additional predictors.

With the FIC we can easily change the focus of the analysis. Using the same ERGM class of models and the same list of potential models, but focusing now on the parameter associated with the triangle predictor (this is the second focus) instead of the log odds ratio, we obtain a different ranking of the models as shown in Table 2. The transitive triangle is an important summary measure for social networks, because it expresses the inclination for actors to form homogenous groups. If actor a has a tie with actor b and b has a tie with c, then under a transitive triangle assumption also a and c will be connected, implying that 'friends of friends are also friends', and as such this summary measure is an important one when describing social networks. When selecting a model that minimizes the MSE expression for the triangle parameter, the best FIC model contains as predictors the change in the transitive triangles statistic, the change in the Gwesp statistic and the wealth of the families showing that different focuses (which embody different research questions) might need different explanatory models.

We now mention some related research. Model (or parameter) selection for social networks is often performed by formal hypothesis testing as in Anderson et al. (1999), Leenders (2002), Robins et al. (2007), by assessing goodness of fit measures as in Goodreau (2007), Hunter et al. (2008a), Wang et al. (2013a), Wang et al. (2013b) or Shore and Lubin (2015), and by using information criteria as in Leenders (2002), Goodreau (2007), Hunter et al. (2008a), Stadtfeld et al. (2011) and Austin et al. (2013) among many other references. Saito et al. (2010) proposed model selection based on average Kullback–Leibler divergence. Bayesian model selection can be found in Koskinen (2004a,b), Zijlstra et al. (2005), Rodríguez (2012) and Caimo and Friel (2013).

2. Social network models

Of all social sciences, sociology and anthropology have been at the forefront of social network analysis, due to the ease with which studying small communities and the interaction between its members, can be reflected to a certain degree by graphical objects. The consequence of such graph oriented representation is that by using basic properties and notions developed for graphs, one can now describe, summarize and also quantify social relations.

A social network consists of a set of units (represented graphically by a set of nodes, one node per unit) and the social connections that exist between the units. Most often the complex relation between the units is reduced to a 'presence or absence' decision, although sometimes one may reduce it to a number that reflects in a way the intensity of the relationship rather than a crude presence/absence representation.

The types of social network models, as summarized nicely in O'Malley and Marsden (2008) vary in complexity and flexibility and reflect different research interests. We use the following three types of models.

(i) Exponential random graph models (ERGM), see Holland and Leinhardt (1981), and Wasserman and Pattison (1996). Local structures in the form of meaningful subgraphs model the global structure of the network. For example, one may use the propensity of forming a triadic configuration (unit i connects with units j and k, and as a transitive result also j and k connect) as a predictor for modeling marriage ties among families. To