Social Networks 48 (2017) 202–212

Contents lists available at ScienceDirect

Social Networks

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

Conditionally exponential random models for individual properties and network structures: Method and application

Stefano Nasini[a,*], Víctor Martínez-de-Albéniz[b], Tahereh Dehdarirad[c]

[a] IESEG School of Management (LEM CNRS 9221), Lille/Paris, France
[b] IESE Business School, University of Navarra, Barcelona, Spain. Supported by the European Research Council –Ref. ERC-2011-StG 283300-REACTOPS and by the Spanish Ministry of Economics and Competitiveness (Ministerio de Economía y Competitividad) – Ref. ECO2014-59998-P
[c] University of Barcelona, Barcelona, Spain

ARTICLE INFO

Article history:

Keywords:
Exponential random models
Social networks
Homophily
Bibliometrics
Bayesian inference
MCMC

ABSTRACT

Exponential random models have been widely adopted as a general probabilistic framework for complex networks and recently extended to embrace broader statistical settings such as dynamic networks, valued networks or two-mode networks. Our aim is to provide a further step into the generalization of this class of models by considering sample spaces which involve both families of networks and nodal properties verifying combinatorial constraints. We propose a class of probabilistic models for the joint distribution of nodal properties (demographic and behavioral characteristics) and network structures (friendship and professional partnership). It results in a general and flexible modeling framework to account for homophily in social structures. We present a Bayesian estimation method based on the full characterization of their sample spaces by systems of linear constraints. This provides an exact simulation scheme to sample from the likelihood, based on linear programming techniques. After a detailed analysis of the proposed statistical methodology, we illustrate our approach with an empirical analysis of co-authorship of journal articles in the field of neuroscience between 2009 and 2013.

1. Introduction

Homophily is a widely studied characteristics of social networks, which is often associated with different forms of human self-segmentation, in terms of demographic and behavioral characteristics (Bell, 2014). It is typically defined as the tendency of individuals to associate with similar others and has been analyzed in a vast range of network studies (McPherson et al., 2001). For instance, in the field of marketing, researchers are interested in analyzing how demographic clusters purchase goods and services which are similar by themselves. In the field of bibliometrics, demographic and behavioral characteristics of co-authors are studied with the aim of analyzing the pattern of collaborations in a given scientific community (Teixeira da Silva, 2011; Haeussler and Sauermann, 2013).

Dealing with homopily in terms of the *association* between individual characteristics and connection patterns entails an epistemological concern, resulting from the direction of *causality* in the observed nodal similarities. In fact, in the presence of homophily, we could either assume individual properties to *cause* and *affect* the appearance of a connection or to expect the latter to *drive* and *boost* the appearance of similarity between connected nodes. In practice, when combined with data, our concept of *causality* must be cast in the language of probability and in particular in the specification of random vectors (endogenous to the model) and fixed parameters or covariates (exogenous to the model). From this practical outlook, modeling the joint distribution of nodal properties (demographic and behavioral characteristics) and network structures (friendship and professional partnership) allows endogenizing the underlying duality of network self-similarity to a large extent. From a purely phenomenological viewpoint, a probabilistic framework where both connections and nodal properties are regarded as random vectors allows any *causality statements* to be translated into *information statements*, with no need for predefined assumptions regarding the direction of causality (Chater et al., 2006; Dawid et al., 2004). In fact, in the language of probability, the very naive hypothesis that a link is affected by the nodal properties of its endpoints implies that $P(link|nodal\ properties) \neq P(link)$, which leads to $P(link\ \&\ nodal\ properties) \neq P(link) \cdot P(nodal\ properties)$ and by Bayes' rule yields that $P(nodal\ properties|link) \neq P(nodal\ properties)$. As a result, the existence of a connection between two

* Corresponding author.
E-mail addresses: S.Nasini@ieseg.fr (S. Nasini), valbeniz@iese.edu (V. Martínez-de-Albéniz), tdehdari@gmail.com (T. Dehdarirad).

http://dx.doi.org/10.1016/j.socnet.2016.09.001
0378-8733/

nodes changes the probability of observing given properties in its corresponding endpoints. Thus, despite being a suitable representation of causes and effects, conditional probability entails symmetric arguments which invert the direction of causality. By contrast, the joint probability explicitly assumes statistical uncertainty on both sides, which is the natural condition in many social settings.

Our aim is to design a joint distribution for the association between nodal properties and connection patterns by a fully endogenous definition of nodal similarities. With this approach, we are able to capture social influence – cross-sectional dependencies between individual features are driven by their connection patterns – and social selection – connections are driven by individual features. Specifically, an exponential random model is proposed to characterize this joint distribution (Lusher et al., 2012; Caimo and Friel, 2011; Robins et al., 2007), allowing for a direct inclusion of both (i) nodal similarities as a collection of sufficient statistics, and (ii) constraints in the sample space of the so-defined multidimensional random variable. The latter represents an important capability when the researcher is interested in controlling for the presence of exogenous influences (number of connections, number of nodes with specified properties, etc.), whose effect she/he wishes to isolate from the rest of the model dependencies.

Exponential families possess good properties that typically simplify the statistical inference of parameters. But as we explain in Section 4, the inclusion of nodal similarities as sufficient statistics for this joint distribution entails the impossibility of a complete characterization of the probability density (mass) function, due to the intractability of the normalizing constant. This represents one of the strongest barriers to the numerical optimization of the likelihood function and legitimates the use of approximation approaches – such as the Monte Carlo maximum likelihood of Geyer and Thompson (1992) and pseudo-likelihood estimation of Strauss and Ikeda (1990).

As suggested by Caimo and Friel (2011), this drawback can be overcome by embedding the defined model into a Bayesian estimation framework, which reformulate the estimation problem based on the ability of simulating from the posterior distribution. We build on Murray et al. (2006), which proposed a MCMC method to simulate from this class of distributions, allowing a flexible estimation of the effect of nodal similarity – which is the main scope of this paper. As it will be accurately discussed in Section 4.2, this estimation approach can be further exploited to accommodate sample spaces characterized by systems of linear constraints, based on the simulation mechanism by Castro and Nasini (2015).

We illustrate our method through the analysis of co-authorship of over a thousand journal articles between 2009 and 2013 in the neuroscience research community. The two reasons behind the choice of this empirical application are supported by (i) the relevance of homophily in scientific collaborations (Teixeira da Silva, 2011; Haeussler and Sauermann, 2013) and (ii) the growing interest in this new line of applications of exponential random models (Goldenberg and Moore, 2005; Wimmer and Lewis, 2010). The practical goal is to jointly study demographic and behavioral characteristics of co-authors, along with their pattern of collaborations in a given scientific community.[2]

Previous studies on co-authorship networks adopted a variety of statistical approaches (Newman, 2003, 2004a), with the purpose of identifying the structure of scientific partnerships and the role played by individual characteristics. The majority of these methodological contributions focus on modelling the structure of scientific co-authorship, based on the projection of a two-mode network (author–paper network) into a one-mode structure of co-authorship (author–author network), where links represent co-authors, i.e., authors sharing common papers, as described by Leydesdorff and Wagner (2008). We use a similar approach in this paper, by considering a set $\mathcal{V}$ of N authors with connection structure $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. We denote by $\mathcal{K}$ a set of K categorical properties (in our application, $\mathcal{K} = \{genders, nationalities\}$) defined for each author in $\mathcal{V}$ and assume the nodal similarities to reflect the overlap of authors' categorical statuses, with respect to the properties in $\mathcal{K}$.

As a result, the statistical application of the proposed probabilistic framework provides substantial insights into the level of *homophily* in co-authorship networks, in terms of specific socio-demographic characteristics, while accounting for relevant network features based on observed nodal properties. Specifically, our approach is able to simultaneously generate the following statistical insights:

- *estimate authors' collaboration pattern based on their demographic and behavioral properties*;
- *estimate authors' demographic and behavioral properties based on their pattern of connections*.

In other words, the proposed modeling approach connects nodal (individual) properties with network structure in a fully probabilistic way, so that information flows in both directions and one can be used to predict the other.

The rest of the paper is organized as follows. We review the literature in Section 2. The data set is then introduced and described in Section 3, along with the relevant descriptive statistics for both individual properties and connection patterns. The model is described in Section 4 and the estimation procedure discussed in Section 5. The numerical results are presented in Section 6 and suggest that the initial modeling decision concerning the direction of the causal association plays an important role in the resulting estimation. Section 7 concludes.

2. Literature review

Homophily, as the tendency of individuals to associate with similar others, has been observed in a vast range of network studies (McPherson et al., 2001). In their seminal paper, Lazarsfeld et al. (1954) discriminated between *status homophily* and *value homophily*. The first one consists to the tendency of individuals with similar social status characteristics to connect with each other. By contrast, value homophily, refers to a more general similarity between demographic and behavioral properties of connected nodes.

Statistical approaches to account for the observed homophily in social networks have been generally based on the ability to reproduce the observed correlations between nodal properties. In this respect, two well-established streams of contributions should be mentioned within the network analytics literature: (i) a vast class of models for assortative mixing (Newman, 2003), with particular attention to the analysis of degree assortativity (Newman, 2004a; Buccafurri et al., 2015), and assortative patterns based on exogenous properties (Pelechrinis and Wei, 2016); (ii) spatially-based models to relate attributes of connected individuals (Winsborough et al., 1963; Carley, 1986; Robins et al., 2001).

In the first stream of literature, the design of the network formation is based on nodal similarities with respect to either exogenous nodal quantities, or to endogenous network properties at the nodal level (such as nodal centrality indexes). In the second stream of literature individual attributes are modeled as a result of a network influence process, where the network structure is taken as exogenous.

[2] Co-authorship networks are designed to represent collaborations between scholars, which are established based on observed joint publications.