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:
Available online 28 September 2016

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.

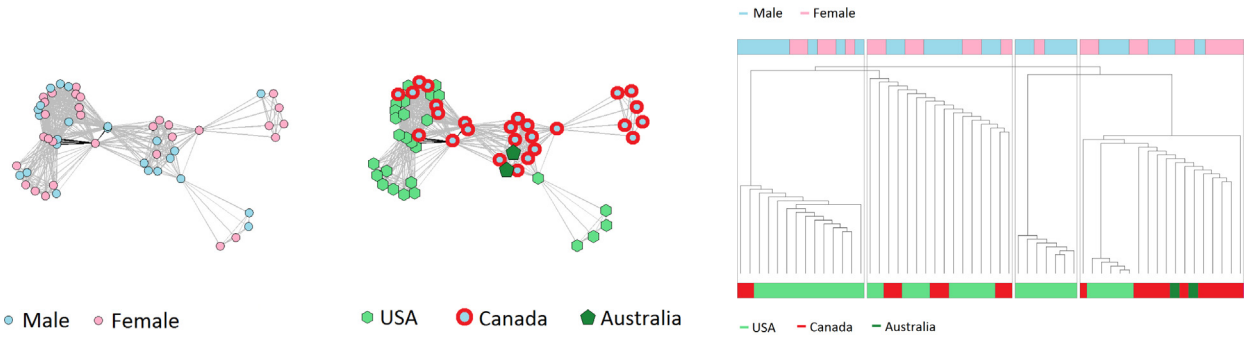
2. Literature review

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.



(a) Network and nodal gender. (b) network and nodal nationality. (c) Community detection.

Fig. 1. Community structure of the first largest component with nodal genders (blue for men, pink for women) and nationalities.

In either cases, exponential random models have been used as important tools for the statistical inference on both sides of this underlying duality, as detailed in Goldenberg et al. (2010). Indeed, they have so far provided a flexible way to deal with network features, such as propensities for homophily, mutuality, and triad closure, through choice of sufficient statistics (Robins et al., 2007; Morris et al., 2008). In recent years, they have been extended to embrace more complex settings, such as the ones associated with valued networks (Krivitsky, 2012), dynamic networks (Hanneke et al., 2010; Desmarais and Cranmer, 2012), two-mode networks (Wang et al., 2009).

A further extension of exponential random models is the one proposed in this paper, where individual properties and network structures are included in a joint sample space, as discussed in Section 4. From the best of our knowledge, only two recent contributions have undertaken a systematic analysis into this class of joint models: a working paper from Fellows and Handcock (2012) – who proposed a likelihood-based inference to approximate the joint distribution – and a recently appeared paper from Thiemichen et al. (2016) – who design a multilevel model with nodal random effects to account for heterogeneity in the network local properties.

As already mentioned in Section 1, a major drawback when dealing with this class of models is the intractability of the normalizing constant for most of the model specifications (Caimo and Friel, 2011). 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 Møller et al. (2006), 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. This algorithmic approach, known as the *auxiliary variable method* in its original formulation, has been substantially improved by Murray et al. (2006), allowing a more efficient convergence to the limit posterior distribution. This latter algorithmic procedure is adopted in described in Section 5 and combined with the simulation mechanism by Castro and Nasini (2015).

3. The co-authorship data set

The data set used in the study is composed of the scientific publications indexed in the Web of Science (WOS) database between 2009 and 2013 in the field of Neuroscience. It has been extracted from the WOS in May 2014 by conducting a search using the field of subject category (WC). As a result, a collection of 153,182 research papers was retrieved in the first step. Then, we conducted stratified random sampling to re-sample from the retrieved set of papers. The sample size was determined with a 3% sampling error and 95% of level of confidence. Table 1 shows the total number of publications and the stratified sample size per year in the studied field. Accordingly, 1060 papers in neuroscience were randomly downloaded from the WOS.

Table 1
The total number of publications and the stratified sample size, 2009–2013.

Year	# publications (%)	Stratified sample size
2009	28,819 (18.81%)	199
2010	30,154(19.69%)	208
2011	31,030 (20.26%)	214
2012	31,265 (20.41%)	218
2013	31,914 (20.83%)	221
Total	153,182	1060

Authors have been assigned to a nationality – based on the country of the institution where the author is affiliated –, and to genders – based on a filtering procedure of their first names.[3] Finally, after eliminating those papers whose gender and nationality was unclear (53 of them, 5%), our data set comprised 1007 (95%) of 1060 papers. These 1007 papers were used as our data set for further analysis, corresponding to 5385 authors.

We project the author–paper network into a one-mode *author–author network* of co-authorship. In other words, a network structure of scientific collaboration between authors was generated by connecting those authors whose names jointly appear in one or more of the 1007 articles.[4] The resulting one-mode networks comprises 207 disconnected components. The size and density of the three largest components are reported in Table 2.

Figs. 1–3 show the network plots of the three largest components of the author–author network, associated with nodal characteristics of gender and nationality. The right plots in each figure report the dendrogram of the community detection algorithm.[5] The aim was to analyze the overlap between membership in the

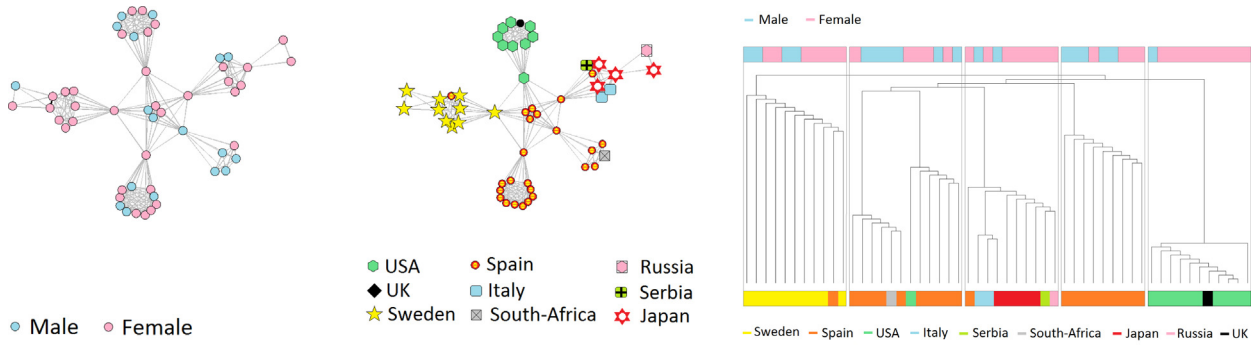
[3] In particular, 5261 (91.80%) authors have been directly classified, while 124 (2.16%) unclassified names have been assigned manually by contacting them. The gender of 346 (6.04%) authors remained unspecified. They correspond to 53 (5%) papers, which have been eliminated from the sample, resulting in a data set comprising 1007 (95%) of the 1060 papers. These 1007 papers were used as our data set for further analysis.

[4] Note that other approaches are possible. For instance, a *paper–paper network* (articles sharing authors) can be obtained where two papers are connected if and only if they share at least a common author.

[5] Co-authorship communities have been detected using a hierarchical clustering approach (Clauset et al., 2004).

Table 2
Number of connected components and densities of the author–author bipartite network.

	First largest	Second largest	Third largest	Total
One-mode author–author network	55 (0.34)	53 (0.17)	35 (0.36)	5385 (0.001)



(a) Network and nodal gender. (b) network and nodal nationality. (c) Community detection.

Fig. 2. Community structure of the second largest component with nodal genders (blue for men, pink for women) and nationalities.

selected communities and authors' demographic characteristics (gender, nationality).

We observe a quite accurate matching between nationality and membership of the detected communities (graphically illustrated by different colors). However, connected nodes do not seem to follow any assortative or dissortative pattern with respect to their genders. This facts will be confirmed in the probabilistic analysis of Section 6.

4. Modeling framework

In our modelling framework, a social structure is parsimoniously defined as a set $\mathcal{V}$ of N individuals, each of which characterized by a specified vector of K features, and a collection of pairwise relations among individuals $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Features and relationships are taken as categorical and static in this paper, although these assumptions can be easily relaxed, see Nasini and Martínez-de-Albéniz (2015). 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 by $\mathbf{Y}_k$ a categorical variable with m_k categories. Across the N individuals its observation is specified in terms of an $N \times m_k$ binary matrix $\mathbf{y}_k \in \mathcal{Y}_k$, where $\mathcal{Y}_k \subseteq \{0,1\}^{N \times m_k}$ is the set of all possible realizations of $\mathbf{Y}_k$ (we use the notation $y^k_{j,i} \in \{0,1\}$, i.e., whether node i has level j of attributes k). Similarly, let $\mathbf{Z}$ be the adjacency matrix of a random network with N nodes and $\mathcal{Z} \subseteq \{0,1\}^{N \times N}$ the set of its possible realizations.

Definition 1 *(Exponential family of distributions).* Let $\mathbf{X}$ be a random vector taking values in $\mathcal{X}$, and $\mathbf{x}$ a possible realization. In the exponential family of distributions the conditional probability of $\mathbf{x} \in \mathcal{X}$ takes the following form:

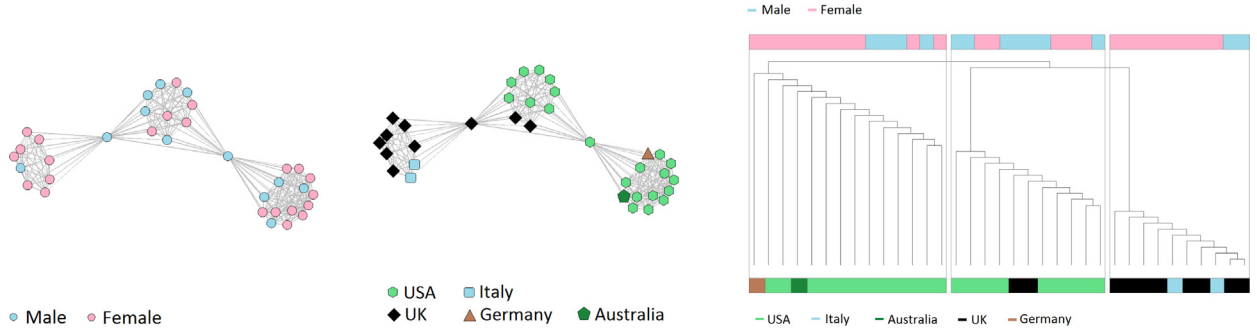
$$P(\mathbf{x}|\theta) \propto q_{\boldsymbol{\theta}}(\mathbf{x}) := h(\mathbf{x}) \exp(T(\mathbf{x})^{\mathrm{T}} \boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of *natural parameters* of the distribution, which can usually take any value in the reals, $T(\mathbf{x})$ is a vector of sufficient statistics, $h(\mathbf{x})$ is an underlying measure on the sample space $\mathcal{X}$ and the symbol $\propto$ denotes proportionality.

In the specific case where the sample space under consideration is a family of networks, the defined modeling framework is best known as exponential random graph model, as previously introduced in Section 2. As discussed in the next subsection, this class of models is able to accommodate a large variety of homophily specifications, based on the defined causal and probabilistic association between individual properties and connection structure.

4.1. The duality between exogenous properties

As already mentioned, a social structure has been described based on the distinction between two levels of characterizations: the individual features (demographic and behavioral properties), defined in the space $\mathcal{Y}$, and the connection pattern (friendship or professional partnership), defined in the space $\mathcal{Z}$.



(a) Network and nodal gender. (b) network and nodal nationality. (c) Community detection.

Fig. 3. Community structure of the third largest component with nodal genders (blue for men, pink for women) and nationalities.

In Section 1, we mentioned the epistemological dilemma about the direction of the causal association between individual properties and connection patterns. Such a duality can be translated into our statistical modeling by *the propagation of the information* from fixed (exogenous) covariates to uncertain (endogenous) variables. The causal interpretation would result in assuming either individual properties to *cause* the appearance of a connection or to expect the latter to *cause* the appearance of similarity between connected nodes.

In the first case, nodal characteristics are statistically treated as exogenous properties (covariates and explanatory variables), while inference is made on the probability distribution of pairwise connections – this is coherent with the classical regression analysis (Hair et al., 2006). Based on the previously defined exponential family of distributions, the inclusion of exogenous individual properties results in an ERGM with assortativity measures:

$$P(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y}^{(0)}) \propto \begin{cases} \exp\left[\boldsymbol{\beta}^{\mathrm{T}} B(\mathbf{z}) + \boldsymbol{\gamma}^{\mathrm{T}} G(\mathbf{y}^{(0)}, \mathbf{z})\right] & \text{if } \mathbf{z} \in \mathcal{Z} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{y}^{(0)}$ is the observed vector of individual properties, $B(\mathbf{z})$ is a vector of sufficient statistics which accounts for combinatorial properties of the network structure $\mathbf{z}$ (such as the clustering coefficient, the assortativity coefficient, the average path length, etc.), but independent of nodal exogenous properties, $G(\mathbf{y}^{(0)}, \mathbf{z})$ is a vector of sufficient statistics which internalizes the interaction between the explanatory nodal characteristics $\mathbf{y}^{(0)}$ and connection variables $\mathbf{z}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are corresponding vector of parameters.

By contrast, in the case the connection pattern is regarded to induce individual properties, interpersonal ties are statistically treated as exogenous information, while inference is made on the probability distribution of individual properties. Based on the previously defined exponential family of distribution, the inclusion of exogenous connections results in a multivariate regression model with network dependency between individual observations:

$$P(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{z}^{(0)}) \propto \begin{cases} h(\mathbf{y}) \exp\left[\boldsymbol{\alpha}^{\mathrm{T}} A(\mathbf{y}) + \boldsymbol{\gamma}^{\mathrm{T}} G(\mathbf{y}, \mathbf{z}^{(0)})\right] & \text{if } \mathbf{y} \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{z}^{(0)}$ is the observed network structure, $A(\mathbf{y})$ is a vector of sufficient statistics which accounts for properties of the categorical variables $\mathbf{y}$ (such as the number of nodes per each level of each categorical variable, number of associated categories, etc.) only, $G(\mathbf{y}, \mathbf{z}^{(0)})$ is a vector of sufficient statistics which internalizes the interaction between nodal characteristics $\mathbf{y}$ and the exogenous connection variables $\mathbf{z}^{(0)}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are corresponding vector of parameters. Note that, in this view, individual features are not independent of each other and, in particular the model's cross-sectional structure internalizes social influence across individuals in a static setting.

4.2. A joint model for endogenous individual properties and network structures

We consider here sample spaces which involve both families of networks and nodal properties verifying linear constraints. We start by defining the sample space under consideration as $\mathcal{X} = \mathcal{Z} \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_K \subseteq \{0, 1\}^{N \times N + \sum_{k=1}^{K} m_k}$ – the set of network structures among N individuals, taking K categorical properties, as illustrated in Fig. 4.

In the exponential family of distributions, the joint probability of individual properties and network patterns can be defined as the natural extension of (2) and (3) when everything is regarded to be endogenous:

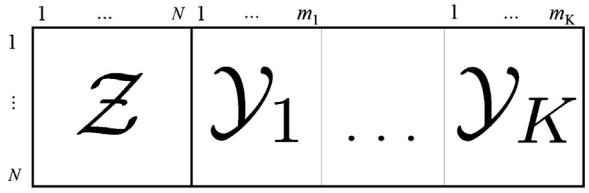


Fig. 4. Sample space.

$$P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \begin{cases} \exp\left[\boldsymbol{\alpha}^{\mathrm{T}} A(\mathbf{y}) + \boldsymbol{\beta}^{\mathrm{T}} B(\mathbf{z}) + \boldsymbol{\gamma}^{\mathrm{T}} G(\mathbf{y}, \mathbf{z})\right] & \text{if } \mathbf{x} \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $A(\mathbf{y})$, $B(\mathbf{x})$ and $G(\mathbf{y}, \mathbf{z})$ are defined as in (2) and (3).

The specification of the sample space $\mathcal{X}$ can incorporate both network and nodal properties, in accordance with our modeling assumptions and our need to control specified combinatorial properties (Castro and Nasini, 2015). In other words, $P(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ if $\mathbf{x}$ does not satisfy a set of feasibility constraints. As illustration, three possible sample spaces can be constructed by exogenously fixing the degree sequence, the number of edges and the size of each categorical level. They are specified in term of the solution sets of systems of linear constraints. Note that intersections of these sets give rise to hybrid sample spaces with complex combinatorial structures. Formally, the three examples can be written as follows:

Fixed degree sequence

$$\begin{cases} \sum_{k=1}^{m_k} y_{h,r}^k = 1 \quad k = 1 \ldots K, \quad r = 1 \ldots n \\ \sum_{r=1}^{n} z_{rs} = d_s \quad r \in \mathcal{V} \end{cases}$$

Fixed number of edges

$$\begin{cases} \sum_{h=1}^{m_k} y_{h,r}^k = 1 \qquad k = 1 \ldots K, \quad r = 1 \ldots n \\ \sum_{(r,s) \in \mathcal{V} \times \mathcal{V}} z_{rs} = d \end{cases}$$

Fixed categorical levels

$$\begin{cases} \sum_{h=1}^{m_k} y_{h,r}^k = 1 \quad k = 1 \ldots K, \quad r = 1 \ldots n \\ \sum_{r=1}^{m} y_{h,r}^k = f_k \quad k = 1 \ldots K, \end{cases}$$

An exact method to sample from $\mathcal{X}$ will be discussed in the next section based on the simulation mechanism by Castro and Nasini (2015).

5. Estimation method

As noted by Murray et al. (2006) and by Caimo and Friel (2011), the intractability of the normalizing constants of most random network models entails a "double intractability" of the posterior distribution when the model is embedded in a Bayesian framework. This is also true for model (4). MCMC algorithms are often used to draw samples from distributions with intractable normalization constants. However, they do not apply to a doubly-intractable constant.

Consider the kernel of the probability function (4) and let $\mathbf{x}^{(0)} \in \mathcal{X}$ be the observed data set – the co-authorship network structure

$\mathbf{z}^{(0)}$, the nodal genders $\mathbf{y}^{(0),1}$, the nodal nationalities $\mathbf{y}^{(0),2}$. Given a prior distribution $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, apply the Bayes rule:

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{x}^{(0)}) = \frac{P(\mathbf{x}^{(0)}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}{\int_{\alpha,\beta,\gamma} P(\mathbf{x}^{(0)}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma}}$$

Since both $P(\mathbf{x}^{(0)}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{x}^{(0)})$ can only be specified under proportionality conditions, Murray et al. (2006) proposed a MCMC approach which overcomes the drawback to a large extent, based on the simulation of the joint distribution of the parameter and the sample spaces, conditioned to the observed data set $\mathbf{x}^{(0)}$, that is to say, $P(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{x}^{(0)})$. We follow the same approach. Our application of the Metropolis-Hastings method (Bolstad, 2009) to simulate from such distribution is summarized in Algorithm 1.

Algorithm 1. Exchange algorithm of Murray et al. (2006).

1: Initialize $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
2: **repeat**
3: Draw $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ from $h(.|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
4: Draw $\mathbf{x}'$ from $P_x(.|\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$
5: Accept $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ with probability $\min\left\{1, \frac{P_x(\mathbf{x})}{P_x(\mathbf{x}')} \times \frac{P(\mathbf{x}'|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})P(\mathbf{x}^{(0)}|\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')\pi(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')}{P(\mathbf{x}^{(0)}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})P(\mathbf{x}'|\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}\right\}$
6: Update $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
7: **until** Convergence

The distribution $h()$ is used to simulate candidate points from the posterior and it is here assumed to be symmetric; $P_x(\mathbf{x})$ is the probability of generating the point $\mathbf{x}$, whose expression is (7).

Note that in step 3 of Algorithm 1 a new value of the parameters $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ is randomly proposed and in step 4 a sample from $\mathcal{X}$ is simulated with probability given in (4).

Clearly, this is a computationally intensive procedure, whose main source of numerical effort is embedded in line 4. Since $P(.|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \exp\left[\boldsymbol{\alpha}^{\mathrm{T}}A(\mathbf{y}) + \boldsymbol{\beta}^{\mathrm{T}}B(\mathbf{z}) + \boldsymbol{\gamma}^{\mathrm{T}}G(\mathbf{y}, \mathbf{z})\right]$, a point from the sample space can be drawn by a Metropolis-Hasting approach.The corresponding acceptance probability of passing from $\mathbf{x} = [\mathbf{y}, \mathbf{z}]$ to $\mathbf{x}' = [\mathbf{y}', \mathbf{z}']$ is

$$P_A(\mathbf{x}, \mathbf{x}') = \min\left\{1, \frac{P_x(\mathbf{x})}{P_x(\mathbf{x}')} \times \frac{\exp\left[\boldsymbol{\alpha}^{\mathrm{T}}A(\mathbf{y}') + \boldsymbol{\beta}^{\mathrm{T}}B(\mathbf{z}') + \boldsymbol{\gamma}^{\mathrm{T}}G(\mathbf{y}', \mathbf{z}')\right]}{\exp\left[\boldsymbol{\alpha}^{\mathrm{T}}A(\mathbf{y}) + \boldsymbol{\beta}^{\mathrm{T}}B(\mathbf{z}) + \boldsymbol{\gamma}^{\mathrm{T}}G(\mathbf{y}, \mathbf{z})\right]}\right\} \tag{5}$$

Thus, in the presence of combinatorial constraints in the sample space $\mathcal{X}$, line 4 of Algorithm 1 entails the solution of a sequence of linear programs. Castro and Nasini (2015) provide detailed analysis of how to make this generating process more efficient by exploiting the matrix structure of the associated linear program. More specifically, simulating from those sample spaces $\mathcal{X}$ coincides with generating extreme points of algebraically defined polytopes. Castro and Nasini (2015) show that many systems of linear constraints characterizing families of complex networks are defined by totally unimodular coefficient matrices, allowing the correct generation of conditional random networks by specialized interior-point methods. Namely, given a sample space $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^q : D\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$, where q is the dimension of $\mathcal{X}$, D is a coefficient matrix and $\mathbf{b}$ a right-hand term, we can formulate the linear program:

$$\begin{aligned} \min \quad & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{s.t.} \quad & D\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq 0. \end{aligned} \tag{6}$$

Letting $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ the Lagrangian multipliers of $D\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq 0$, Castro and Nasini (2015) prove that knowing the probability of the objective gradient $\mathbf{c}$, say $P_c(\mathbf{c})$, the probability of the optimal point in the sample space is

$$P_x(\mathbf{x}) = \int\int P_c(D^T\boldsymbol{\mu} - \boldsymbol{\lambda}) \left\| \begin{matrix} D & & \\ & D^T & I \\ \Lambda & & X \end{matrix} \right\| d\boldsymbol{\lambda} d\boldsymbol{\mu} \tag{7}$$

where Λ and X are diagonal matrices made up with the components of $\boldsymbol{\lambda}$ and $\mathbf{x}$, respectively.

This result is particularly important when estimating the natural parameter of the defined exponential random model, as the ability to correctly simulate from $\mathcal{X}$ is required by some of the most applied algorithms which deal with intractable normalizing constants $Z(\boldsymbol{\theta})$.

6. Numerical results and analysis

In this section different specifications of the described class of models are numerically analyzed, with the aim of assessing their probabilistic properties under different statistical settings. Let $\mathbf{b} = [b_1 \dots b_K]$ be the vector containing the number of observed levels of individual properties in the data set and $\mathcal{Y} = \{\mathbf{y} \in \{0, 1\}^{N \times \sum_{k=1}^{K} m_k} : \sum_{h=1}^{m_k}\sum_{r \in \mathcal{V}} y_{h,r}^k = \mathbf{b}\}$ the corresponding sample space of all individual profiles of N nodes verifying constant demographic aggregate quantities. Likewise, let $d \in \mathbb{R}$ be the number of observed connections in our data set and $\mathcal{Z} = \{\mathbf{z} \in \{0, 1\}^{N \times N} : \sum_{(r,s) \in \mathcal{E}} z_{rs} = d\}$.

The following conditional specification of models (2) and (3) are taken into account:

$$P(\mathbf{y}|\mathbf{z}^{(0)}, \mathbf{b}, \boldsymbol{\gamma}) \propto \begin{cases} \exp\left[\sum_{k \in \mathcal{K}} \gamma_k \sum_{(r,s) \in \mathcal{V} \times \mathcal{V}} z_{rs}^{(0)} \left(\sum_{h=1}^{m_k} y_{h,r}^k y_{h,s}^k\right)\right] & \text{if } \mathbf{y} \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$P(\mathbf{z}|\mathbf{y}^{(0)}, d, \boldsymbol{\gamma}) \propto \begin{cases} \exp\left[\sum_{k \in \mathcal{K}} \gamma_k \sum_{(r,s) \in \mathcal{V} \times \mathcal{V}} z_{rs} \left(\sum_{h=1}^{m_k} y_{h,r}^{(0),k} y_{h,s}^{(0),k}\right)\right] & \text{if } \mathbf{z} \in \mathcal{Z} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Beside, the specification of the joint model (4) is the following:

$$P(\mathbf{x}|d, \mathbf{b}, \boldsymbol{\gamma}) \propto \begin{cases} \exp\left[\sum_{k \in \mathcal{K}} \gamma_k \sum_{(r,s) \in \mathcal{V} \times \mathcal{V}} z_{rs} \left(\sum_{h=1}^{m_k} y_{h,r}^k y_{h,s}^k\right)\right] & \text{if } \mathbf{x} \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The notations $y_{j,i}^{(0),k}$ and $z_{is}^{(0)}$ refer to the realization of $y_{j,i}^k \in \{0, 1\}$ and $z_{is} \in \{0, 1\}$ in the observed data.

The vector of sufficient statistics contains the association between edges and the two nodal properties. The corresponding natural parameters γ_k account for how strongly the observed homophily deviates from full randomness (uniformity over $\mathcal{X}$), for $k \in \mathcal{K}$. Thus, in the case of uniform distribution within the sample space $\mathcal{X}$, we should have $\gamma_k = 0$; hence any non-zero value of the natural parameters entails a deviation from such independence. Specifically, if $\gamma_k = 0$, (10) is reduced to the product between the probability mass functions of K multinomial random

variables Multinom$(1/N \ldots 1/N, b_k)$ and the Erdos-Renyi random graph model with fixed number of edges d.

When a flat distribution is adopted as an improper prior probability for $\gamma_1 \ldots \gamma_{\mathcal{K}}$, the following family of posterior distributions is obtained:

$$P(\boldsymbol{\gamma}|\mathbf{x}^{(0)}) \propto \frac{1}{Z(\boldsymbol{\gamma})} \exp\left[\sum_{k\in\mathcal{K}} \gamma_k \sum_{(r,s)\in\mathcal{V}\times\mathcal{V}} z_{rs}^{(0)} \left(\sum_{h=1}^{m_k} y_{h,r}^{(0),k} y_{h,s}^{(0),k}\right)\right], \tag{11}$$

where $Z(\gamma_k)$ is the partition function of the selected likelihood (8), (9), and (10), which can be explicitly defined as

$$Z(\boldsymbol{\gamma}) = \begin{cases} \sum_{\mathbf{y}\in\mathcal{Y}} \exp\left[\sum_{k\in\mathcal{K}} \gamma_k \sum_{(r,s)\in\mathcal{V}\times\mathcal{V}} z_{rs}^{(0)} \left(\sum_{h=1}^{m_k} y_{h,r}^{k} y_{h,s}^{k}\right)\right] & \text{for model (8)} \\ \sum_{\mathbf{z}\in\mathcal{Z}} \exp\left[\sum_{k\in\mathcal{K}} \gamma_k \sum_{(r,s)\in\mathcal{V}\times\mathcal{V}} z_{rs} \left(\sum_{h=1}^{m_k} y_{h,r}^{(0),k} y_{h,s}^{(0),k}\right)\right] & \text{for model (9)} \\ \sum_{\mathbf{x}\in\mathcal{X}} \exp\left[\sum_{k\in\mathcal{K}} \gamma_k \sum_{(r,s)\in\mathcal{V}\times\mathcal{V}} z_{rs} \left(\sum_{h=1}^{m_k} y_{h,r}^{k} y_{h,s}^{k}\right)\right] & \text{for model (10)} \end{cases} \tag{12}$$

Thus, the immediate effect of the specification of the conditional information on the posterior distribution is entirely based on the partition function that each specification provides. In particular, they only differ for the integration regions $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ of the sample spaces defined by the assumed conditional information. When the dimensionality of the sample space is large (as in the case of $\mathcal{Y} \times \mathcal{Z}$), this implies posterior distributions whose density is mainly concentrated around small values (as the the partition function grows quickly with the homophily parameter γ_k).

The remaining parts of this section provide a detailed numerical analysis of the three model specifications (8), (9), and (10). A simulation exercise is carried out in Section 6.1 to study the statistical properties of the obtained posterior distribution, when the number of nodes and nodal attributes vary. Subsection 6.2 applies the described model specification (8), (9) and (10) to the second largest component of the co-authorship network, introduced in Section 3.

6.1. Simulation exercise

A simulation exercise is carried out in this subsection to analyze the behavior of the Bayesian posterior distribution when the number of individuals and nodal attributes vary. The variability of the Bayesian posterior distribution is an important indicator of our knowledge of the homophily parameter γ_k.

We consider $\gamma_k = 1/K$, for each $k \in \mathcal{K}$, as predefined natural parameters of (8), (9) and (10).[6] Each of the K attributes is supposed to have three levels, so that the corresponding sample spaces are subsets of $\{0,1\}^{3K}$, $\{0,1\}^{N(N-1)/2}$ and $\{0,1\}^{3K+N(N-1)/2}$ for (8), (9) and (10) respectively. Three replicates of simulation are carried out for all combinations of $N \in \{50, 100, 150\}$ and $K \in \{5, 10, 15\}$, resulting in 27 different runs.

Tables 3–5 report the average standard deviation and coefficient of variation of the model parameters' Bayesian posterior distribution, for each of the nine combinations of N and K (averaged over all replicates), for (8), (9) and (10) respectively.

[6] We choose $1/K$ so that (i) similarity between nodes increases the likelihood of a link, and (ii) similarity measures are scaled between 0 and 1 regardless of the value of K.

Table 3
Average standard deviation (on the left part of each cell) and coefficient of variation (on the right side of each cell) of the posterior distributions for different combination of N and K, based on the model specification (8).

# Nodes	# Attributes		
	$K=5$	$K=10$	$K=15$
$N=50$	0.134–0.947	0.144–1.343	0.149–1.358
$N=100$	0.068–0.814	0.072–1.419	0.074–1.370
$N=150$	0.022–0.127	0.043–0.804	0.044–1.008

Table 4
Average standard deviation (on the left part of each cell) and coefficient of variation (on the right side of each cell) of the posterior distributions for different combination of N and K, based on the model specification (9).

# Nodes	# Attributes		
	$K=5$	$K=10$	$K=15$
$N=50$	0.538–2.236	0.588–2.612	0.621–2.816
$N=100$	0.613–2.379	0.676–4.878	0.683–4.209
$N=150$	0.619–3.025	0.644–4.386	0.705–5.591

It can be seen that when the network structure is fixed, in model specification (8), the increase in the number of nodes has a positive effect in reducing the posterior variability of γ_k. We can interpret it, based on the increase in the number of terms added in the computation of the partition function $Z(\boldsymbol{\gamma})$, so that when the number of nodes grows large the posterior distribution concentrates around small values. By contrast, when the network structure is random and the nodal attributes are exogenous, in model specification (9), larger networks are associated with less accurate estimation of the homophily parameter γ_k. The same is true when the number of nodal attributes is taken into account. In this case, a heuristic interpretation is that the presence of a large amount of individual properties might result in a mixed combination of their effects, and a higher uncertainty in the single effect of each of them.

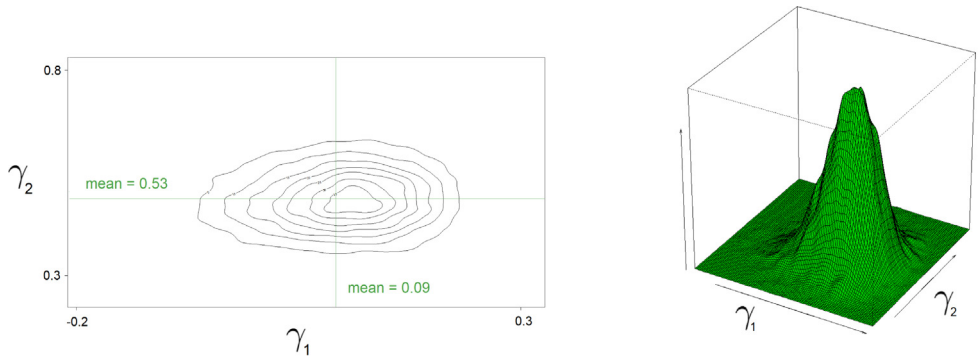
When the numerical results in Table 5 are taken into account, two different behaviors of the model specification (10) can be observed: (i) similar to (8), the increase in the number of nodes has a positive effect in reducing the posterior variability; (ii) similar to the model specification (9), the increase in the number of attributes is associated with higher variability.

6.2. Empirical application to the co-authorship data set

To assess the ability of the defined probabilistic approaches to analyze the underlying duality of homophily and network self-similarity in the context of bibliometric analysis, we proposed an empirical application of models (8), (9) and (10) to the second largest component of the co-authorship network in the neuroscience community, as shown in Figs. 1 and 2. The dimensions are $N = 54$ (nodes), $K = 2$ (categorical properties), $m_1 = 2$ (genders), $m_2 = 9$ (nationalities). The sample space is defined as the Cartesian product between the set of N node undirected networks with fixed number of edges and the set of all possible realization of

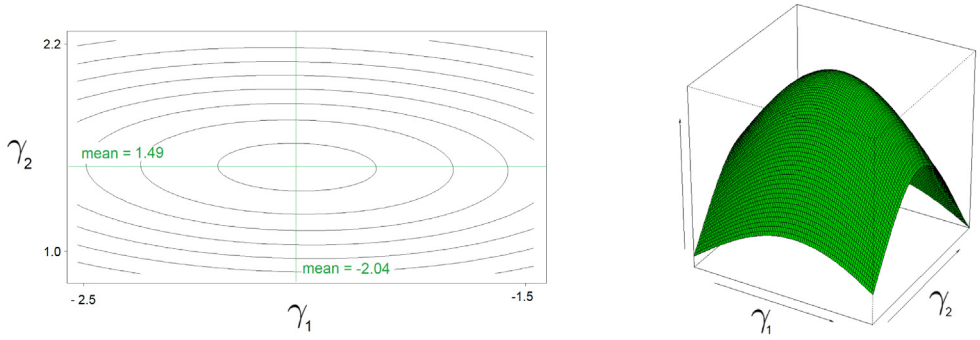
Table 5
Average standard deviation (on the left part of each cell) and coefficient of variation (on the right side of each cell) of the posterior distributions for different combination of N and K, based on the model specification (10).

# Nodes	# Attributes		
	$K=5$	$K=10$	$K=15$
$N=50$	0.139–0.737	0.139–1.431	0.141–1.848
$N=100$	0.091–0.255	0.086–1.503	0.098–1.690
$N=150$	0.016–0.094	0.041–0.797	0.042–0.860



(a) Contour plot of the empirical posterior (b) Bidimensional empirical posterior

Fig. 5. Marginal posterior of (γ_1, γ_2), corresponding to the second largest component of the co-authorship data set, for model (8).



(a) Contour plot of the empirical posterior (b) Bidimensional empirical posterior

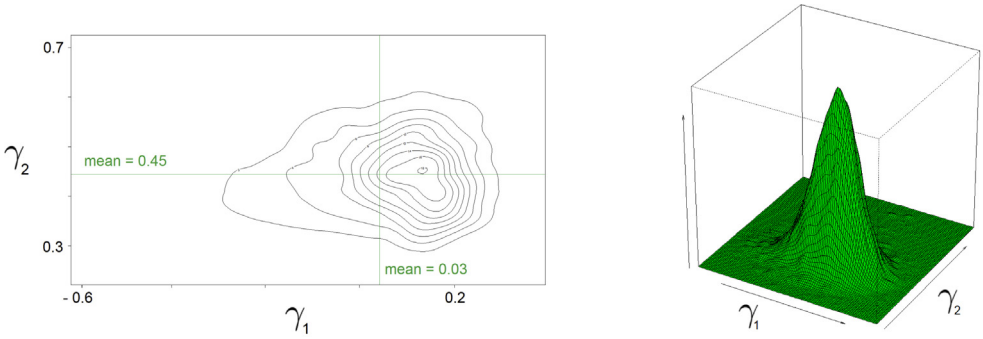
Fig. 6. Marginal posterior of (γ_1, γ_2), corresponding to the second largest component of the co-authorship data set, for model (9).

2 categorical variables with 2 and 9 levels and fixed number of individuals per level.

The three model specifications (8), (9) and (10) are compared: the contour plot in Figs. 5–7 show the estimated marginal posterior of (γ_1, γ_2), corresponding to the *gender* and the *nationality* effect, for the second largest component of the co-authorship data set in Section 3. These results have been obtained by a chain with 100,000 MCMC iterations.

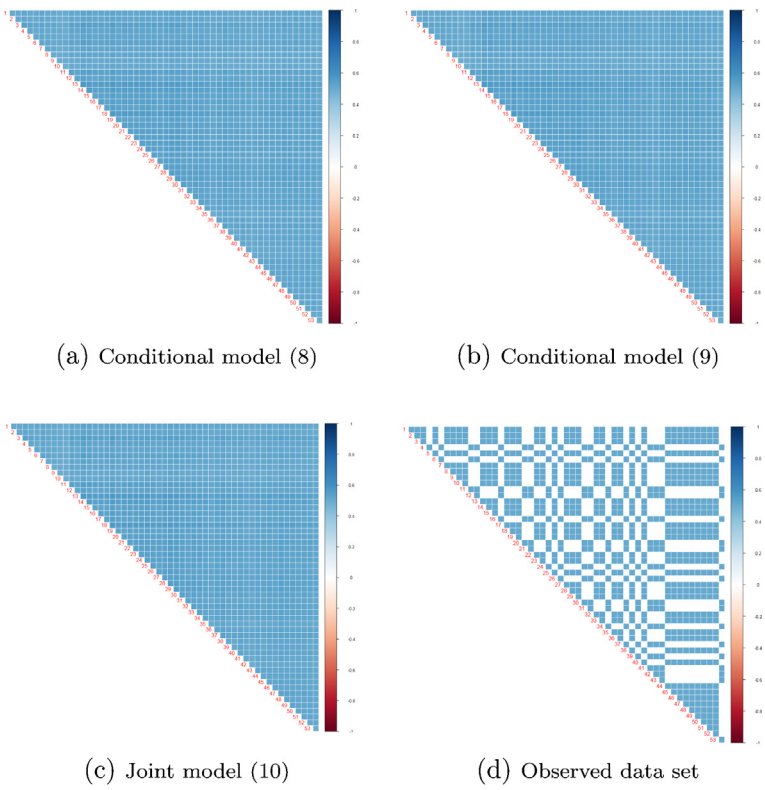
Both conditional models and the joint model agree in a positive expected effect of the nationality – the posterior expectation of γ_2 is 0.53 for model (8), 1.49 for model (9) and 0.45 for model (10) – as graphically reported in Figs. 5–7.

To interpret this result for each of the three models consider a feasible configuration of $\mathbf{z}$ and $\mathbf{y}$. For model (8), consider two pairs of connected nodes, (r, s) and (r', s'), i.e., $z_{rs} = z_{r's'} = 1$. Model (8) makes a statement about the likelihood of having r and s being similar. Specifically, a change of nodal attributes from $y^2_{h,r} = y^2_{h,r'} = 0$, $y^2_{h,s} = y^2_{h,s'} = 1$ (r is different from s and r' from s') to $y^2_{h,r} = y^2_{h,s} = 1$, $y^2_{h,r'} = y^2_{h,s'} = 0$ (r and s are similar but r', s' are still different) results in an increase of the likelihood by a factor $e^{0.53} = 1.70$, i.e., the probability of observing this new configuration is 70% higher. In other words, it is 70% more likely to observe similarity than difference between connected nodes, provided that the overall balance across nationality occurrences within the network remains unchanged. A similar interpretation is valid for model (9), where the nodal attributes are fixed. Consider now two pairs of nodes, one similar and one different, one connected and another disconnected. It is 343% more likely ($e^{1.49} - 1 = 3.43$) to have the two similar nodes connected, than the two different nodes connected. The same



(a) Contour plot of the empirical posterior (b) Bidimensional empirical posterior

Fig. 7. Marginal posterior of (γ_1, γ_2), corresponding to the second largest component of the co-authorship data set, for model (10).



(a) Conditional model (8) (b) Conditional model (9)

(c) Joint model (10) (d) Observed data set

Fig. 8. The values of gender similarities $\sum_{h=1}^{m_k} z_{rs} y_{h,r}^k y_{h,s}^k$, $k = 1$, for the $54 \times 53/2$ pairs of nodes.

reasoning is valid for model (10). The parameter γ_2 is thus a strong driver of the likelihood between two given configurations of $(\mathbf{y}, \mathbf{z})$.

As far as the gender similarity is concerned, it can be noted a much larger variability on the posterior of γ_1, suggesting a lack of information about the parameter, which translates into a higher uncertainty of the predictive posterior.

After estimating the model parameter $\boldsymbol{\gamma}$, a sample of 10,000 elements from the posterior predictive distribution of $\mathcal{X}$ has been simulated and the nodal similarities have been computed for each model specification (8), (9) and (10). Figs. 8 and 9 show the observed and expected gender and nationality similarities for each of the $54 \times 53/2$ pairs of authors, respectively. On the one hand, Fig. 8 suggests that neither the conditional models (8) and (9) nor the joint model (10) are able to predict gender similarities. On the other hand, Fig. 9 clearly illustrates the resemblance between the observed matching in nodal nationalities and the ones expected under the three estimated models (based on the predictive posterior). Note however that the models cannot be compared directly, because each makes different assumptions on which elements are uncertain. For example, model (8) uses actual realizations of network links, and lets nodal attributes be uncertain. In contrast, model (9) uses actual nodal attributes and uncertain network structure. Model (10) has uncertainty in both. As a result, model (10) will naturally perform worst because it makes mistakes on nodes and links, while the others can only make mistakes on one of them. However, model (10) requires no exogenous information so it can be used for out-of-sample forecasting purposes.

Using the estimated model (10), Tables 6 and 7 report the expected proportions of edges associated with different combinations of genders and nationalities respectively. Only few proportions result to be consistently different – e.g., 12% of the observed collaborations in the second largest component are between scholars from the USA and Spain, while the estimated percentage is 3.1%. The estimated model seems to be able to effectively capture the assortative mixing of the data, i.e., the association between nodal properties and connections, along with the total amount of each individual categories and network collaboration density.

Table 6
Expected proportions of edged for each combination of genders. The observed proportions are reported within parenthesis.

	Male	Female	Total
Male	0.40 (0.48)	0.46 (0.40)	0.86 (0.88)
Female	– –	0.14 (0.12)	
Total		0.60 (0.52)	

In Fig. 10 a graphical comparison between the distribution of estimated network properties and observed ones (red line) is provided. We focus on assortativity coefficient and clustering coefficient.[7] We see that, despite not including any transitivity factor in our model specification, the predicted posterior generates networks with a clustering coefficient similar to the real one and in particular much higher than the clustering coefficient under a Bernoulli model (which is equal to the density, 0.17 in this case, vs. 0.93 observed). From the viewpoint of the bibliometric interpretation, the network assortativity with respect to nationalities is able to account for most of the observed level of clustering, so that only a residual amount of local triangles are due to a purely structural tendency for transitivity.

[7] The assortativity coefficient is defined as the Pearson correlation between degrees, for connected nodes. The local clustering coefficient is defined as the average (over all nodes) of the number of edges of a given node's neighborhood relative to the total possible number of edges, e.g., if the neighborhood of i has d_i nodes, then it is equal to $\frac{\sum_{j_1,j_2=1}^{n} z_{j_1 j_2}}{d_i(d_i-1)/2}$.

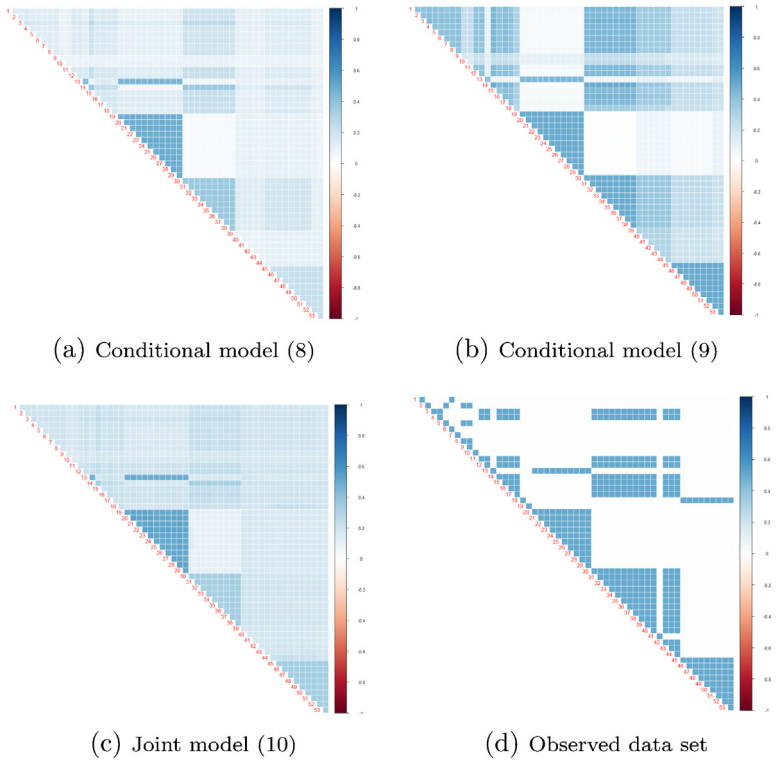


Fig. 9. The values of nationality similarities $\sum_{h=1}^{m_k} z_{rs} y_{h,r}^k y_{h,s}^k$, $k=2$, for the $54 \times 53/2$ pairs of nodes.

Table 7
Expected proportions of edges for each combination of nationalities. The observed proportions are reported within parenthesis.

	Italy	USA	Spain	Sweden	S. Africa	Japan	Serbia	Russia	UK
Italy	0.004 (0.004)	0.004 (0.000)	0.014 (0.017)	0.001 (0.025)	0.000 (0.008)	0.007 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
USA		0.260 (0.235)	0.120 (0.031)	0.012 (0.000)	0.006 (0.000)	0.060 (0.004)	0.005 (0.000)	0.000 (0.000)	0.005 (0.023)
Spain			0.267 (0.329)	0.020 (0.024)	0.010 (0.009)	0.111 (0.064)	0.010 (0.201)	0.000 (0.000)	0.010 (0.000)
Sweden				0.001 (0.017)	0.001 (0.006)	0.010 (0.000)	0.000 (0.001)	0.000 (0.008)	0.000 (0.000)
S. Africa					0.001 (0.006)	0.002 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Japan						0.085 (0.140)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Serbia							0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Russia								0.000 (0.000)	0.000 (0.000)
UK									0.000 (0.000)

To summarize, for the second largest component of the co-authorship data set, the estimated results confirm the null effect of the gender similarity on the author's connections, along with a positive effect of their nationalities. The estimated model seems to properly fit the empirical observation, as suggested by Tables 6 and 7. The distribution of the assortativity coefficient and the local clustering coefficient reveal a close matching to the observed ones, suggesting the ability of the model to capture both individual and structural properties of the co-authorship data.

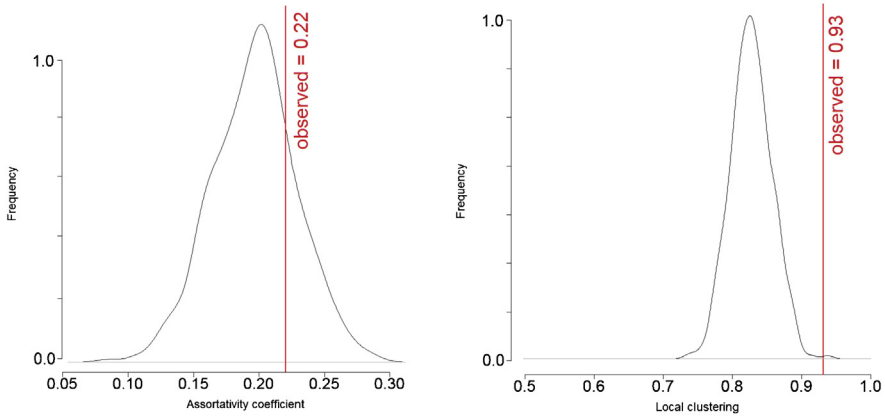


Fig. 10. Empirical distribution of the network assortativity (left plot) and the local clustering (right plot) under the estimated joint model.

7. Conclusion

This paper presents an exponential random model for author's characteristics and collaboration pattern in bibliometric networks, which allows to combine the analysis of multivariate data with the study of assortative patterns of nodal similarities in networks. Our model is able to handle simultaneously uncertainty both in nodal properties and connections, and internalizes network homophily without the need of knowing whether structure or connection properties are an input or an output (i.e., directional causality). We propose a Bayesian estimation framework and a specialized MCMC algorithm to simulate from a "doubly intractable" posterior distribution. We show that the model accounts for relevant network features based only on the observed nodal properties: this provides a deeper understanding of the linkage between individual and social properties and a substantial insight into the level of *homophily* in co-authorship networks.

Our results suggest several lines of work for future research. We could compare the model fit for different specifications of the sample space $\mathcal{X}$. We could also study the inclusion of further nodal properties, such as age, principal keywords or number of received citations. Finally, our approach was derived on a projection of a dynamic two-mode network into a static one-mode network. Extending our methods to dynamic two-mode networks might reveal further insights on the interaction between nodal features and bipartite connections, e.g., author–paper links.

References

Bell, D.R., 2014. Location is (still) Everything: The Surprising Influence of the Real World on How We Search, Shop, and Sell in the Virtual One. Houghton Mifflin Harcourt.

Bolstad, W.M., 2009. Markov Chain Monte Carlo Sampling from Posterior. John Wiley and Sons, Inc, pp. 127–157.

Buccafurri, F., Lax, G., Nocera, A., 2015. A new form of assortativity in online social networks. Int. J. Hum.–Comput. Stud. 80, 56–65.

Caimo, A., Friel, N., 2011. Bayesian inference for exponential random graph models. Soc. Netw. 33 (1), 41–55.

Carley, K., 1986. An approach for relating social structure to cognitive structure. J. Math. Sociol. 12 (2), 137–189.

Castro, J., Nasini, S., 2015. Mathematical programming approaches for classes of random network problems. Eur. J. Oper. Res. 245 (2), 402–414.

Chater, N., Tenenbaum, J.B., Yuille, A., 2006. Probabilistic models of cognition: conceptual foundations. Trends Cogn. Sci. 10 (7), 287–291.

Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70 (6), 066111.

Dawid, A.P., et al., 2004. Probability, causality and the empirical world: a Bayes-de Finetti-Popper-Borel synthesis. Stat. Sci. 19 (1), 44–57.

Desmarais, B.A., Cranmer, S.J., 2012. Statistical mechanics of networks: Estimation and uncertainty. Phys. A: Stat. Mech. Appl. 391 (4), 1865–1876.

Fellows, I., Handcock, M.S., 2012. Exponential-Family Random Network Models, arXiv preprint arXiv:1208.0121.

Geyer, C.J., Thompson, E.A., 1992. Constrained Monte Carlo maximum likelihood for dependent data. J. R. Stat. Soc. Ser. B (Methodological) 54 (3), 657–699.

Goldenberg, A., Moore, A.W.,2005. Bayes net graphs to understand co-authorship networks? In: Proceedings of the 3rd International Workshop on Link Discovery. ACM, pp. 1–8.

Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M., 2010. A survey of statistical network models. Found. Trends® Mach. Learn. 2 (2), 129–233.

Haeussler, C., Sauermann, H., 2013. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. Res. Policy 42 (3), 688–703.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L., 2006. Multivariate Data Analysis, Vol. 6. Pearson Prentice Hall, Upper Saddle River, NJ.

Hanneke, S., Fu, W., Xing, E.P., et al., 2010. Discrete temporal models of social networks. Electron. J. Stat. 4, 585–605.

Krivitsky, P.N., 2012. Exponential-family random graph models for valued networks. Electron. J. Stat. 6, 1100–1128.

Lazarsfeld, P.F., Merton, R.K., et al., 1954. Friendship as a social process: a substantive and methodological analysis. Freedom Control Modern Soc. 18 (1), 18–66.

Leydesdorff, L., Wagner, C.S., 2008. International collaboration in science and the formation of a core group. J. Inf. 2 (4), 317–325.

Lusher, D., Koskinen, J., Robins, G., 2012. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications. Cambridge University Press.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: Homophily in social networks. Ann. Rev. Sociol., 415–444.

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K., 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93 (2), 451–458.

Morris, M., Handcock, M.S., Hunter, D.R., 2008. Specification of exponential-family random graph models: terms and computational aspects. J. Stat. Softw. 24 (4), 1548.

Murray, I., Ghahramani, Z., MacKay, D.J.C.,2006. MCMC for Doubly-Intractable Distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). AUAI Press, pp. 359–366.

Nasini, S., Martínez-de-Albéniz, V., 2015. Pairwise influences in dynamic choice: method and application. Working paper. IESE Business School.

Newman, M.E., 2004a. Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. U. S. A. 101 (Suppl 1), 5200–5205.

Newman, M.E.J., 2003. Mixing patterns in networks. Phys. Rev. E 67 (2), 026126.

Pelechrinis, K., Wei, D., 2016. VA-index: quantifying assortativity patterns in networks with multidimensional nodal attributes. PLOS ONE 11, 1–13, 01.

Robins, G., Pattison, P., Elliott, P., 2001. Network models for social influence processes. Psychometrika 66 (2), 161–189.

Robins, G., Pattison, P., Kalish, Y., Lusher, D., 2007. An introduction to exponential random graph (p^*) models for social networks. Soc. Netw. 29 (2), 173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P., 2007. Recent developments in exponential random graph (p^*) models for social networks. Soc. Netw. 29 (2), 192–215, Special Section: Advances in Exponential Random Graph (p^*) Models.

Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. J. Am. Stat. Assoc. 85 (409), 204–212.

Teixeira da Silva, J.A., 2011. The ethics of collaborative authorship. EMBO Rep. 12 (9), 889–893.

Thiemichen, S., Friel, N., Caimo, A., Kauermann, G., 2016. Bayesian exponential random graph models with nodal random effects. Soc. Netw. 46 (1), 11–28.

Wang, P., Sharpe, K., Robins, G.L., Pattison, P.E., 2009. Exponential random graph (p^*) models for affiliation networks. Soc. Netw. 31 (1), 12–25.

Wimmer, A., Lewis, K., 2010. Beyond and below racial homophily: Erg models of a friendship network documented on facebook1. Am. J. Sociol. 116 (2), 583–642.

Winsborough, H.H., Quarantelli, E., Yutzy, D., 1963. The similarity of connected observations. Am. Soc. Rev. 28 (6), 977–983.