Model-based clustering for populations of networks

Signorelli, M.; Wit, E.C.

Statistical Modelling 2020; **20(1)**: 9–29

Model-based clustering for populations of networks

Mirko Signorelli[1] and Ernst C. Wit[2,3]
[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.
[2]Institute of Computational Science, Università della Svizzera italiana, Lugano, Switzerland.
[3]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, The Netherlands.

Abstract: Until recently obtaining data on populations of networks was typically rare. However, with the advancement of automatic monitoring devices and the growing social and scientific interest in networks, such data has become more widely available. From sociological experiments involving cognitive social structures to fMRI scans revealing large-scale brain networks of groups of patients, there is a growing awareness that we urgently need tools to analyse populations of networks and particularly to model the variation between networks due to covariates. We propose a model-based clustering method based on mixtures of generalized linear (mixed) models that can be employed to describe the joint distribution of a populations of networks in a parsimonious manner and to identify subpopulations of networks that share certain topological properties of interest (degree distribution, community structure, effect of covariates on the presence of an edge, etc.). Maximum likelihood estimation for the proposed model can be efficiently carried out with an implementation of the EM algorithm. We assess the performance of this method on simulated data and conclude with an example application on advice networks in a small business.

Key words: cognitive social structure, EM algorithm, graph, mixture of generalized linear models, model-based clustering, network modelling, population of networks

1 Introduction

The last decades have witnessed growing interest in the analysis of relational data. Typically, these data come in the form of a network that displays relations between individuals or objects, and they are represented by means of a graph wherein nodes (i.e., individuals or objects) are connected by edges (i.e., relations). In some applications, especially in genetics, relations cannot be observed directly and the main task is to infer them or their strength from the data (Friedman et al., 2008; Abegaz and Wit, 2013; Vujačić et al., 2015). In this article, we are interested in cases where relations between individuals or objects are observed and the networks themselves are the data.

Address for correspondence: Mirko Signorelli, Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands.
E-mail: m.signorelli@lumc.nl

 10.1177/1471082X19871128

For a long time, network science was almost exclusively concerned with the analysis of a single network, mainly because of the difficulty in collecting relational data and of limited computing capacity. Statistical modelling of a single network (Snijders, 2011) has typically focused on certain aspects of network topology, such as degree distribution, network statistics or the presence of community structures. This has resulted into the development of a range of statistical network models that include the p_1 and p_2 models (Holland and Leinhardt, 1981; van Duijn et al., 2004), stochastic blockmodels (Holland et al., 1983; Snijders and Nowicki, 1997), ERGMs (Frank and Strauss, 1986), latent space models (Hoff et al., 2002) and the family of loglinear models proposed by Perry and Wolfe (2012).

More recently, increased computing capacities, alongside with technological advances such as the development of sensor-based measurements, the diffusion of functional magnetic resonance imaging, the invention of high-throughput technologies in biology and the advent of social media, have multiplied the availability of relational data, spurring the analysis not only of larger networks, but also of several instances of the "same" network. The latter includes multilayer networks, dynamic networks and populations of networks. The availability of such collections of several networks poses new modelling challenges. Clearly, when data on several networks are available, modelling each network separately would be inefficient: irrespective of whether we are dealing with multilayer networks, longitudinal networks, or populations of networks, we expect networks therein to be similar to a certain degree; if this is indeed the case, analysing each network separately would not only be cumbersome, but also failing to use the statistical power of the ensemble. Instead, the specification of a joint model for the collection of networks makes it possible to achieve a more parsimonious representation of the data and to borrow information across networks in the estimation process; moreover, such a model may also be employed to identify groups of similar networks. Below, we briefly review some of the solutions that to date have been proposed to tackle this problem in the presence of dynamic networks, multilayer networks or populations of networks.

Dynamic networks allow to represent the evolution of a network system over time. Snijders (2001) proposed a stochastic actor-oriented model where the decision to create or dissolve an edge is based on some covariates, as well as on the current state of the network itself. Hanneke et al. (2010) introduced a dynamic extension of ERGMs, known as temporal ERGM (TERGM). An extension of the latent space models for dynamic networks has been proposed by Sewell and Chen (2015). Matias and Miele (2017), instead, developed a dynamic stochastic blockmodel that allows group membership of units to vary over time.

Multilayer networks are collections of networks that represent different types of relationships (multiple *layers*) between a group of subjects. Two statistical models that allow to model jointly the layers of those networks are, among others, those of Stanley et al. (2016) and Paul et al. (2016). Stanley et al. (2016) proposed a multilayer stochastic blockmodel that assumes the existence of groups of networks, called strata, that share the same community structure. Paul et al. (2016), instead, introduced a multilayer stochastic blockmodel that assumes that the communities are the same in all layers but allows different block-interaction probabilities in each layer.

Finally, a *population of networks* can be defined as a collection of independent graphs, each of which corresponds to a different statistical unit. Populations of networks arise, from example, when different individuals are asked to provide their view of relationships within a social network (Krackhardt, 1987) or when brain networks are compared across groups of patients (Taya et al., 2016). Recently, there has been a growing interest in statistical modelling of populations of networks. Sweet et al. (2014) proposed a hierarchical stochastic blockmodel that aims to infer clusters of nodes that are shared across networks. Similarly, Reyes and Rodriguez (2016) introduced a stochastic blockmodel for populations of networks that attempts to identify a unique community structure which is shared across networks. Durante et al. (2017), instead, extended the latent space model approach of Hoff et al. (2002) to populations of networks by proposing a mixture model that describes the joint density of networks in the population using few components, each of which has a different latent-space representation. Finally, Mukherjee et al. (2017) proposed to cluster graphs within a population of networks through a spectral clustering algorithm that is applied to a distance matrix that measures the distances between the graphon estimates of the graphs.

In this article, we focus on the problem of finding and characterizing clusters of graphs that are similar with respect to the effect of certain covariates of interest on the presence or absence of edges in a population of networks. Towards this aim, we propose to model the population of networks with a mixture model whose components can be any statistical network model that can be specified as a generalized linear model (GLM) or a generalized linear mixed model (GLMM); this includes, for example, the p_1 and p_2 models, degree-corrected stochastic blockmodels a priori and the loglinear network models of Perry and Wolfe (2012). The advantages of this methodological framework are that it makes it possible to describe populations of networks using some statistical models that are popular in social network analysis, it can flexibly handle the inclusion of different types (monadic, dyadic and graph-specific) of covariates and it furthermore allows to detect subpopulations of networks (if any). In Section 2, we introduce and formalize our mixture of network models and we elaborate on the specification of the components of the mixture. Model estimation is considered in Section 3, where we discuss how the proposed model can be estimated with an implementation of the expectation–maximization (EM) algorithm. In Section 4 we assess the performance of our method on simulated data, and in Section 5, we present an example application to data on advice relationships in a small manufacturing firm described by Krackhardt (1987).

2 Model specification

We consider a sample of K graphs $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K\}$, where each graph $\mathcal{G}_k = (V, E_k)$ comprises a set of edges E_k between a set of v vertices V, from a population of networks $\mathcal{S}$. We represent $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K\}$ with an array $\mathcal{Y}$ of dimension $v \times v \times K$,

where each horizontal slice $\mathbf{Y}_k$ is the adjacency matrix of graph $\mathcal{G}_k$. Therefore, an entry Y_{ij}^k in $\mathcal{Y}$ refers to the presence (and strength) or absence of edge (i, j) in the kth graph $\mathcal{G}_k$. If the graphs in $\mathcal{S}$ are undirected, each $\mathbf{Y}_k$ is symmetric and we can restrict our attention to the upper triangle of $\mathbf{Y}_k$.

In principle, one could imagine that each graph $\mathcal{G}_k$ is drawn from a different distribution $f(\mathbf{Y}|\theta_k),\ k \in \{1, \ldots, K\}$ with parameter vector θ_k:

$$\mathbf{Y}_k \sim f(\mathbf{Y}|\theta_k).$$

In the presence of many networks, however, this would result in a cumbersome modelling exercise, yielding K different models obtained from separate analyses of each graph. Since each graph is defined on the same set of vertices, it is natural to consider models with additional structure.

2.1 Specification of the mixture model

In this article, we consider the existence of clusters of graphs with similar $f(\mathbf{Y}|\theta_k)$: if any such cluster exists, we would like to borrow information among graphs within that cluster, so as to estimate a joint model for graphs belonging to that cluster rather than many separate network models. As a result, we assume that the population of networks $\mathcal{S}$ arises from $M \leq K$ subpopulations $\mathcal{S}_1, \ldots, \mathcal{S}_M$ of graph models, each with probability density function $f(\mathbf{Y}|\theta_m),\ m \in \{1, \ldots, M\}$. We denote by $Z_k \in \{1, \ldots, M\}$ the label that identifies the subpopulation of graph $\mathcal{G}_k$, such that $Z_k = m$ if $\mathbf{Y}_k \sim f(\mathbf{Y}|\theta_m)$ (i.e., $Z_k = m$ if $\mathcal{G}_k \in \mathcal{S}_m$). Since in real problems it will typically be unknown which graph belongs to which subpopulation, the vector of identifying labels $\mathbf{Z} = (Z_1, \ldots, Z_K)$ is a latent variable. Therefore, we view each graph in the sequence as a random draw from a mixture model whose components are the probability density functions $f(\mathbf{Y}|\theta_m)$:

$$\mathbf{Y}_k \sim \sum_{m=1}^{M} \pi_m f(\mathbf{Y}|\theta_m), \tag{2.1}$$

with mixing proportions $\pi_m = Pr(Z_k = m)$ denoting the prior probabilities that a graph belongs to the mth subpopulation $\mathcal{S}_m$. Clearly, we assume $\pi_m \geq 0\ \ \forall m \in \{1, \ldots, M\}$ and $\sum_{m=1}^{M} \pi_m = 1$. If we let $\Theta = (\theta_1, \ldots, \theta_M)$, the likelihood of model (2.1) is thus

$$\begin{aligned} L(\Theta|\mathcal{Y}, \mathbf{Z}) = Pr(\mathcal{Y}, \mathbf{Z}|\Theta) &= \prod_{k=1}^{K} Pr(\mathbf{Y}_k|Z_k, \Theta) Pr(Z_k|\Theta) \\ &= \prod_{k=1}^{K} \pi_{Z_k} f\left(\mathbf{Y}_k|\theta_{Z_k}\right). \end{aligned} \tag{2.2}$$

This likelihood suffers from the usual identifiability issues when considering mixture models. Each of the K components can be permuted without altering the likelihood. So, as there are $K!$ possible permutations, there exists $K!$ symmetries in the likelihood. Moreover, the possibility of empty components raises the possibility that certain parameters θ_k are not identifiable. As our aim is to find the maximum likelihood estimate (MLE), we will be satisfied with finding one of the $K!$ equivalent MLEs. The issue of empty (or near-empty) components is dealt with via information criteria to select the number of components. Although not providing any theoretical guarantees, (near) empty components will be discouraged due to the unnecessary numbers of parameters they introduce.

2.2 Specification of the components of the mixture

The way in which the probability density functions $f(\mathbf{Y}|\theta_m)$ in Equations (2.1) and (2.2) can be specified depends on the properties that are deemed relevant for the analysis of the networks at hand. If, for example, interest lies in clustering a sequence of binary graphs based on similarities in their degree distributions, $f(\mathbf{Y}|\theta_m)$ can be specified as a p_1 or p_2 model (Holland and Leinhardt, 1981; van Duijn et al., 2004). If a partition of vertices into groups or communities is available and the probabilities of interaction between vertices are believed to depend on group memberships, a stochastic blockmodel (Holland et al., 1983) can be employed to specify f. If both the degree distribution and community structure are deemed relevant, different types of degree-corrected stochastic blockmodels (Wang and Wong, 1987; Signorelli, 2017) can be considered. If one would like to cluster graphs based on the values of network statistics that reflect socially relevant patterns of interaction (e.g., transitivity), they could consider ERGMs (Frank and Strauss, 1986).

In this article, we focus our attention on network models that assume edges to be independent conditionally on the model parameters (and, potentially, on a set of unobserved random effects), so that their likelihood can be specified as that of a generalized linear (mixed) model. The motivation behind this choice is threefold. Firstly, a wide range of popular network models (among which are the p_1 and p_2 models, stochastic blockmodels a priori, degree-corrected stochastic blockmodels a priori, the family of models considered by Perry and Wolfe (2012) and the unconstrained model that we introduce in Section 2.2.3, but not ERGMs) can be specified as GLMs (McCullagh and Nelder, 1989))or as GLMMs (Breslow and Clayton, 1993). Moreover, the GLM(M) framework enables us to easily incorporate monadic, dyadic and graph-specific covariates into the network generative models. Finally, mixtures of GLM(M)s can be estimated efficiently and this ensures computational efficiency in the estimation of mixtures of network models, which we will base on an iterative algorithm that may require several iterations and, thus, could otherwise become computationally burdensome.

Therefore, we shall specify the mixture model in (2.1) as a mixture of GLMs (Grün and Leisch, 2008) by assuming that the value of each edge y_{ij}^k is drawn from an exponential family distribution and that a transformation of the conditional

expectation of Y_{ij}^k is linear in the parameters:

$$g\left[E\left(Y_{ij}^k|\mathbf{x}_{ijk},\, \theta,\, Z_k = m\right)\right] = \mathbf{x}_{ijk}^T\theta_m,$$

where g is a link function and $\mathbf{x}_{ijk}$ is a vector associated to θ_m that can contain *monadic* (i.e., node-specific) covariates for nodes i and j, *dyadic* (i.e., edge-specific) covariates for edge (i, j) and graph-specific covariates for graph $\mathcal{G}_k$. Extensions to mixtures of GLMMs are straightforward and will be used in the application. The density of graph $\mathcal{G}_k$ can then be obtained as $f\left(\mathbf{Y}_k|\theta_{z_k}\right) = \prod_{i<j} f\left(y_{ij}^k|\theta_{z_k}\right)$ if $\mathcal{G}_k$ is undirected, or as $f\left(\mathbf{Y}_k|\theta_{z_k}\right) = \prod_{i\neq j} f\left(y_{ij}^k|\theta_{z_k}\right)$ if it is directed. Hereinafter, we shortly introduce three network models that we will use in Section 4 to illustrate our method.

2.2.1 p_1 model

In social network analysis, the popularity of individuals is often regarded as one of the possible determinants of the formation of relations in a network. This reflects the idea that in certain social settings, individuals may be more likely to relate to popular individuals than to isolated ones: for example, if you live in a small village in the heart of the Alps, you are more likely to interact with popular figures such as the mayor and the priest, rather than with a woodsman who lives in a remote cottage in the middle of the woods. This idea is at the basis of the p_1 model (Holland and Leinhardt, 1981), a simple network model that assumes that the probability of an edge between any two nodes i and j depends (only) on the expected degrees of the two nodes. If, for example, a population of binary undirected networks is considered, we can specify a mixture of p_1 models by letting $y_{ij}^k|z_k \sim \text{Bern}\left(\pi_{ij}^{z_k}\right)$, where $\text{logit}\left(\pi_{ij}^{z_k}\right) = \theta^{z_k} + \alpha_i^{z_k} + \alpha_j^{z_k}$ and $\sum_{i=1}^{v} \alpha_i^{z_k} = 0$.

2.2.2 Stochastic blockmodel

Besides popularity, group membership of nodes is another factor that can shape the way in which relations are formed. Real networks often feature the presence of *communities* of nodes whose members are highly connected with each other and tend to form sporadic connections with members from other communities. For example, it has been shown that parliamentarians tend to collaborate more frequently with members from their own parliamentary group rather than with those from other political groups (Signorelli and Wit, 2018). In general, group membership typically induces a so-called *community structure* in networks, wherein nodes from the same community are closely tied to each other and sporadically linked to nodes from other communities. The effect of community membership on the formation of relations is usually modelled with stochastic blockmodels (Holland et al., 1983; Snijders and Nowicki, 1997). Let $\mathcal{P}$ denote a partition of V into $p < v$ groups and denote by $C : V \to \mathcal{P}$ a community-assignment function, so that $C(i)$ is the community that node i belongs to. In stochastic blockmodels, the probability of an edge between nodes i and j depends on the communities that the two nodes belong to: $y_{ij}^k|z_k \sim \text{Bern}\left(\pi_{ij}^{z_k}\right)$,

where

$$\text{logit}\left(\pi_{ij}^{z_k}\right) = \theta_{C(i)C(j)}^{z_k}. \tag{2.3}$$

Depending on whether the community-assignment function is known or not, it is possible to distinguish stochastic blockmodels *a priori* (Holland et al., 1983), wherein community labels are known and interest lies in the reconstruction of relationships between communities, from stochastic blockmodels *a posteriori* (Snijders and Nowicki, 1997). In this work, we focus on the simpler a priori stochastic blockmodel, which is computationally cheap and, thus, can be easily incorporated into the iterative estimation procedure proposed in Section 3.1. Mixtures of stochastic blockmodels a posteriori, instead, are considered in the works of Stanley et al. (2016) and Reyes and Rodriguez (2016).

2.2.3 Unconstrained network model

The p_1 model and stochastic blockmodel described above are two examples of simple and thrifty statistical network models that can be employed to model commonly observed features of real networks such as heterogeneity in node degrees and community structure. These models comprise a number of parameters that is considerably lower than the number of nodes pairs and, thus, they allow a very parsimonious description of networks; however, in reality these models are likely to be often too simplistic. It may thus be desirable to consider more complex statistical models, which can improve model fit and enable a more realistic description of the complex structure of a network. For example, it is possible to combine the aforementioned models into a degree-corrected stochastic blockmodel (Wang and Wong, 1987) that can account for degree heterogeneity and community structure at the same time, or to incorporate covariates into stochastic blockmodels (Signorelli and Wit, 2018). A further example of how to combine different statistical network models into a more realistic one can be found in the example application that we provide in Section 5, where we will specify a network model that combines features of the p_2 model and of the stochastic blockmodel, and that furthermore accounts for the effect of some monadic covariates on the formation of advice relationships.

Clearly, more realistic network models may require a larger set of parameters and this could increase the complexity of maximum likelihood estimation for model (2.1) and computing time. To illustrate this, we consider the extreme scenario of a mixture of saturated network models, where the number of parameters is equal to the number of edge pairs multiplied by the number of subpopulations of graphs, namely $Mv(v-1)/2$ in undirected graphs and $Mv(v-1)$ in directed graphs. This model simply assumes that $y_{ij}^k | z_k \sim \text{Bern}\left(\pi_{ij}^{z_k}\right)$, leaving the probabilities $\pi_{ij}^{z_k}$ unconstrained. It represents the most complex model that can be specified to model relations within a population of networks with $M \leq K$ subpopulations, and it does not make any restrictive assumption about which factors affect the creation of edges. As such, in practice this model may represent a useful starting point in the analysis of the population of networks: in particular, its generality can be exploited at an initial

stage of the analysis to choose the number of subpopulations M in the mixture and to identify some important patterns in the data; information gathered from this complex model could then be exploited to further refine the analysis by specifying a simpler network model that accounts for the most important effects that are believed to affect the presence of edges. We provide an example of this modelling approach in Section 5.

3 Model estimation

We propose to estimate the unknown parameter vector Θ of the mixtures of network models described in Section 2 with maximum likelihood. Since the likelihood function $L(\Theta|\mathcal{Y}, \mathbf{Z})$ in Equation (2.2) depends both on the observed graphs $\mathcal{Y}$ and on the unobserved vector $\mathbf{Z}$, such likelihood can be maximized by implementing the EM algorithm as illustrated below.

3.1 EM algorithm

The EM algorithm (Dempster et al., 1977) represents a popular choice for the estimation of mixture models. The algorithm allows the maximization of a likelihood $L(\theta|\mathbf{y}, \mathbf{z})$ that depends both on observed data $\mathbf{y}$ and on latent data $\mathbf{z}$, and it consists of successive iterations of two steps, respectively called expectation (E) and maximization (M). The expectation step requires the computation of the conditional expectation of the likelihood $L(\theta|\mathbf{y}, \mathbf{z})$ given the current estimate of θ and the observed data $\mathbf{y}$, whereas the maximization step updates the parameter estimates by maximizing the expected likelihood determined in the E step. We propose the following implementation of the EM algorithm for the maximization of (2.2):

1. Choose a starting point for the algorithm made by the initial probabilities $p^0_{km} = Pr(Z_k = m) \in [0, 1]$ for $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, M\}$, with $\sum_{m=1}^{M} p^0_{km} = 1 \; \forall k$. Denote by $\mathbf{P}^0$ the $K \times M$ matrix which collects these probabilities;
2. Given $\mathbf{P}^0$, estimate the parameters of the mixture of GLMs with weights given by $\left(p^0_{1m}, \ldots, p^0_{Km}\right)$ for the mth component, and obtain $\hat{\Theta}^0 = \left(\hat{\theta}^0_1, \ldots, \hat{\theta}^0_M\right)$;
3. For $t = 1, 2, 3, \ldots$ until convergence is reached:
   a. **E step**. Given $\hat{\Theta}^{t-1}$, derive $\mathbf{P}^t$ as

$$p^t_{km} = \frac{f(\mathbf{Y}_k|\hat{\theta}^{t-1}_m)}{\sum_{j=1}^{M} f(\mathbf{Y}_k|\hat{\theta}^{t-1}_j)}.$$

   b. **M step**. Given $\mathbf{P}^t$, estimate a mixture of GLMs with weights given by $\left(p^t_{1m}, \ldots, p^t_{Km}\right)$ for the mth component, and obtain $\hat{\Theta}^t$.

In principle, it is possible to initialize the EM algorithm introduced above with any matrix of initial probabilities $\mathbf{P}^0$. However, it is possible to reduce the number of iterations and facilitate convergence to the true MLE by considering multiple sensible initial guesses of the cluster memberships. Therefore, we consider three different cluster initializations by means of three network similarity measures combined with the partition around medoids (PAM) clustering method (Reynolds et al., 2006). The first similarity measure is the Jaccard index (Jaccard, 1912). The second is given by the L^1 distance between the adjacency matrices (note that for binary graphs, this is equivalent to the L^2 distance). The third similarity measure is obtained by first computing the Laplacian matrix of each graph (defined as $L = D - A$, where A is the adjacency matrix of the graph and D a diagonal matrix with the degrees of each node as diagonal entries) and then taking the L^1 distance between the Laplacian matrices rather than between the adjacency matrices. Once a distance matrix has been obtained with one of the aforementioned methods, we apply the PAM clustering algorithm with number of clusters equal to M and derive $\mathbf{P}^0$ accordingly.

3.2 Selection of the number of components

In practice, the number of subpopulations M that form the mixture is typically unknown and it needs to be estimated. The estimation of the number of components M can be performed by minimizing model selection criteria such as the Akaike information criteria (AIC) and the Bayesian information criteria (BIC). The choice of the effective sample size to be used for the computation of BIC is particularly crucial for multivariate data (Berger et al., 2014) and it is selected to be equal to K for this purpose. We assess the performance of AIC and BIC on simulated data in Section 4.2.

4 Simulations

In this section, we first evaluate the accuracy of the proposed clustering method with respect to network size (represented by the number of nodes v), to the number of networks K and to the number of subpopulations M on simulated data. Then, we assess the capacity of the selection criteria introduced in Section 3.2 to correctly identify the true number of subpopulations M. We conclude discussing the scalability of the proposed method to large populations of networks and to populations of large networks. The $\mathbb{R}$ code to simulate the data and to perform model-based clustering of populations of networks can be found at `http://www.statmod.org/smij/archive.html`.

4.1 Clustering accuracy

We begin the assessment of the performance of the proposed method with nine simulations (A-I) where we study the clustering accuracy of our method with respect

Table 1 Synthetic overview of simulations A-I. We consider two parsimonious models, the p_1 model (Section 2.2.1) and the stochastic blockmodel (SBM) a priori (Section 2.2.2), and a more general, unconstrained network model (Section 2.2.3) that contains as many parameters as edge pairs. In simulations *A*, *D* and *G* we increase v, keeping K and M fixed. In simulations *B*, *E* and *H* we increase K, keeping v and M fixed. In simulations *C*, *F* and *I* we increase the number of subpopulations M while keeping v fixed; each subpopulation consists of 10 graphs (hence, $K = 10 \cdot M$).

Simulation	Network model	v	K	M
A	p_1 model	from 10 to 40	50	2
B	p_1 model	20	from 12 to 60	2
C	p_1 model	30	$10 \cdot M$	from 2 to 7
D	SBM a priori	from 10 to 40	50	2
E	SBM a priori	20	from 12 to 60	2
F	SBM a priori	30	$10 \cdot M$	from 2 to 7
G	unconstrained	from 10 to 40	50	2
H	unconstrained	15	from 12 to 60	2
I	unconstrained	15	$10 \cdot M$	from 2 to 7

to the three network models introduced in Section 2.2 as v, K or M increases. We focus on how the *purity* (Schütze et al., 2008) of the clusters is affected by these parameters. Purity is a measure of clustering accuracy that attains value 1 if perfect classification is achieved; for M equally sized (true) subpopulations, the worst-case purity value is $1/M$.

Table 1 summarizes the features of the mixtures of networks considered in each simulation. Within each simulation, we compute 50 repetitions for each combination of (v, K, M) considered; we consider 10 different initializations for the EM, 3 of which are obtained as described in Section 3.1 and the remaining 7 are obtained from the previous 3 starting points by randomly replacing the initial probabilities of 30% of the graphs. A more detailed description of the parameters involved in each simulation can be found in Section 1 of the supplementary material.

The distribution of purity across repetitions for the p_1 model is illustrated in Figure 1. Purity quickly increases with respect to the number of nodes present in a graph (panel A); this steep increase is mainly due to the fact that the number of edge pairs increases quadratically with v, making prediction for populations of larger graphs a much easier task. Panel B shows that purity is already fairly high with a small number of graphs, but is highly variable; a larger K results in reduced variability for the purity, which is more concentrated around its median value. Finally, simulation C shows that purity decreases with the number of subpopulations considered; this result is intuitive, since a larger number of subpopulations produce a harder classification problem; nevertheless, even for larger values of M there is an evident improvement over random allocation of graphs to subpopulations (panel C).

Similar observations hold for the simulations (D, E and F) with the stochastic blockmodel a priori (Supplementary Figure 1) and for those (G, H and I) with the unconstrained network model (Supplementary Figure 2).

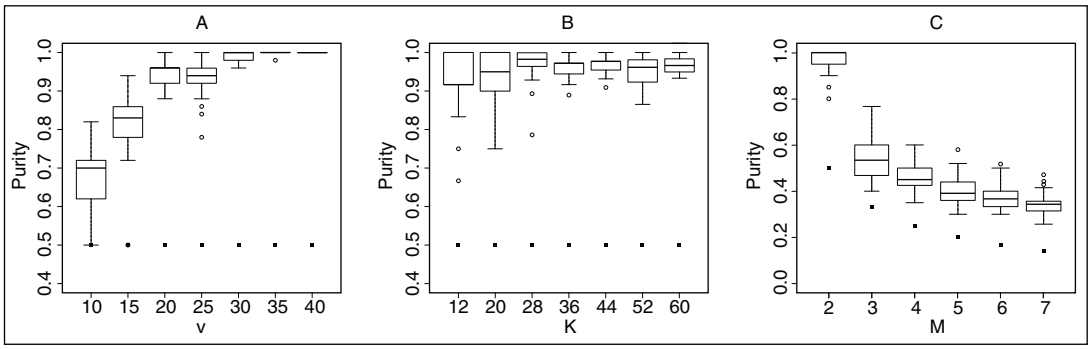

Figure 1 Purity in simulations A, B, C. Each boxplot represents the distribution of purity over 50 repetitions, whereas the squares denote the value of purity that corresponds to a random assignment of graphs to clusters (i.e., $1/M$).

4.2 Selection of the number of subpopulations

In order to assess the performance of the model selection criteria introduced in Section 3.2, in simulation J we repeatedly sample K networks from a mixture of unconstrained network models (defined in Section 2.2.3) with $M = 3$ subpopulations of equal size (more details can be found in Section 1 of the supplementary material). We fix $v = 20$ and let $K \in \{30, 90, 180, 300\}$. We repeat each simulation 100 times, computing the MLEs of the mixture model parameters for $M \in \{1, 2, 3, 4, 5\}$. Then, we compute AIC and BIC and derive the optimal number of subpopulations according to each criterion.

Figure 2 shows the distribution of the optimal number of subpopulations obtained with AIC and BIC for the different values of K considered. We note that as expected, both AIC and BIC can accurately select the correct number of subpopulations ($M = 3$) when a sufficiently large number of graphs K is available. When K is small, however, BIC tends to systematically underestimate M and it is thus outperformed by AIC.

4.3 Scalability of method to large and many graphs

In Section 4.1, we have considered simulation scenarios with a relatively small number of graphs of moderate size. This has allowed us to show how the proposed approach can achieve a good accuracy in allocating graphs to their correct subpopulation already in problems where v or K are relatively small. Here, we consider two simulations with larger v and K to illustrate the scalability of our approach, focusing on how the computing time is affected by the number of networks K as well as by the size of the networks. In general, we see that the computing time increases linearly with K and M, and super-linearly with v.

In simulation K, we simulate data from a mixture of stochastic blockmodels a priori with five blocks, setting $v = 50$, $M = 2$, and we let K increase from 100 to 1 000. Figure 3 shows that the median computing time is linear in the number of

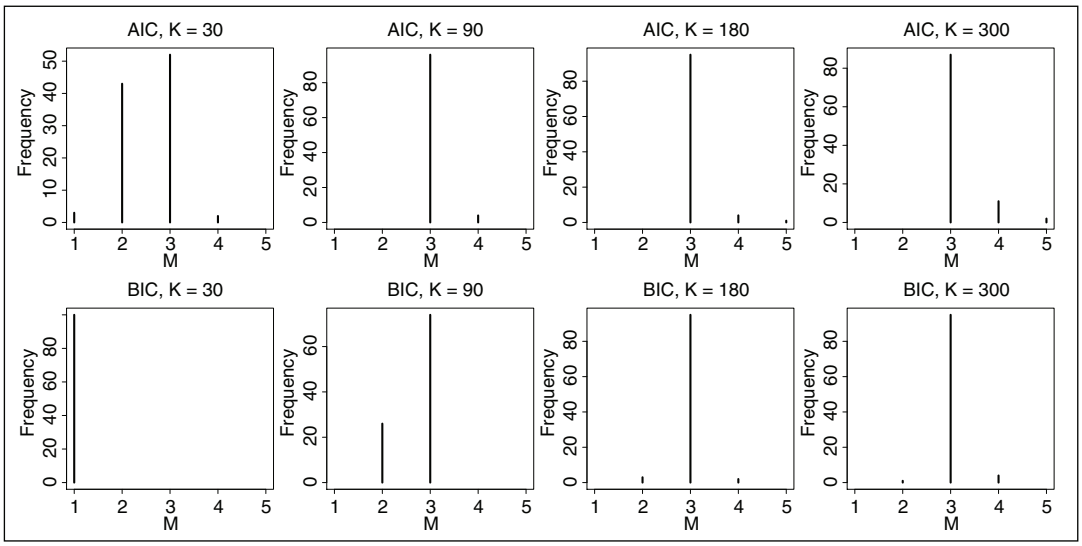

Figure 2 Distribution of the optimal number of subpopulations in simulation J according to AIC and BIC, for different values of K. The true number of populations is equal to 3.

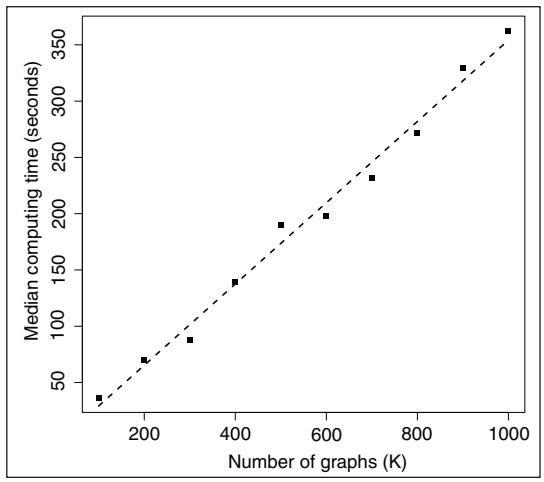

Figure 3 Relationship between number of graphs and median computing time in simulation K. It can be observed that computing time is approximately linear in the number of graphs. Computations were performed using a processor with 2.3 GHz CPU.

graphs, and it increases from 36.5 seconds when $K = 100$ up to 363 seconds when $K = 1\,000$.

In simulation L, we simulate data from a mixture of stochastic blockmodels a priori with five blocks, setting $K = 50$, $M = 2$, and we let v increase from 100 to 1 000. Figure 4 shows that the median computing time is quadratic in the number

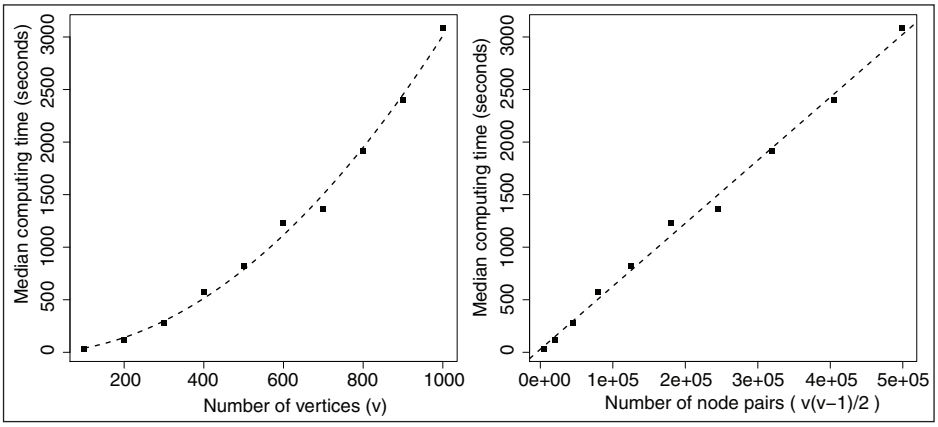

Figure 4 Relationship between graph size and median computing time in simulation L. It can be observed that computing time is approximately quadratic in the number of vertices v, and linear in the number of edge pairs $v(v-1)/2$. Computations were performed using a processor with 2.3 GHz CPU.

of vertices v and linear in the number of node pairs $v(v-1)/2$, increasing from 35.1 seconds when $v = 100$ to 3 089 seconds when $v = 1\,000$.

5 Application

In this section, we illustrate an application of our model-based network clustering method to a population of networks on advice relationships within a small business collected by Krackhardt (1987), whose aim was to find ways to summarize the reconstructions of an unobserved social network reported by different perceivers. In this study, 21 employees of a high-tech US company were asked to fill in a questionnaire where, among other questions, each employee was requested to reconstruct advice relationships between the 21 employees. From the answers to this questionnaire, Krackhardt (1987) derived a population of $K = 21$ directed advice networks, wherein each network is the reconstruction of advice relationships according to a different employee. Given the difficulty to analyse data within the resulting three-dimensional array, which he called *cognitive social structure*, Krackhardt (1987) proposed three simple aggregation techniques to reduce the dimensionality of the problem and simplify interpretation. Alternatively, we show here how a suitably defined mixture of network models may be employed to highlight important patterns in Krackhardt's population of networks.

Because each employee attempted to reconstruct the actual network of advice relationships within the firm, which is unobserved, we may expect that not only different perceivers would reconstruct advice relationships in a different manner but also some perceivers may provide substantially similar reconstructions of the advice network. In other words, it seems reasonable to hypothesize the presence

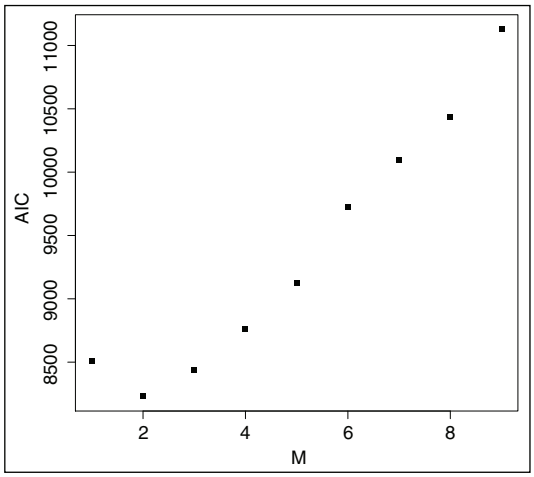

Figure 5 Value of the AIC for mixtures of unconstrained network models with increasing number of subpopulations M. The minimum AIC is attained when $M = 2$.

of different clusters of networks, corresponding to groups of employees who have similar perceptions of advice relationships within the firm.

We begin our analysis in an exploratory manner by considering at first a mixture of M unconstrained network models: we do not make any assumption on how each arrow is formed, thus leaving the probabilities to observe an arrow from node i to node j ($i \neq j$) unconstrained. The aim of this first analysis is twofold: first, we want to find an appropriate number of clusters, unconfounded by a too restrictive network model; secondly, we aim to exploit patterns in the estimated probabilities of observing an arrow in each subpopulation to further refine the analysis. Later in this section, we will use this information to define a more parsimonious network model where we will let these probabilities depend on a set of covariates.

The first model that we consider simply assumes that $y_{ij}^k | z_k \sim \text{Bern}\,(\pi_{ij}^{z_k})$, with $z_k \in \{1, \ldots, M\}$ and $i \neq j$. We estimate the optimal number of subpopulations $\hat{M}$ following the approach outlined in Section 3.2, using AIC as model selection criterion based on the results presented in Section 4.2. As Figure 5 shows, AIC attains a minimum when $M = 2$, so we set $\hat{M} = 2$. Estimation of the mixture model with $M = 2$ components leads to the detection of a first cluster that comprises six perceivers, namely employees 1, 3, 4, 5, 10 and 21, and of a further cluster comprising the other 15 employees. In Figure 6 we show the predicted probabilities of observing advice relationships from sender i to receiver j in each subpopulation, that is, $\hat{\pi}_{ij}^m$, with $m \in \{1, 2\}$ and $i \neq j$.

Graphical inspection of Figure 6 clearly reveals that graphs in the first subpopulation are denser than graphs in the second subpopulation; moreover, in both subpopulations we can intuitively observe that department affiliation seems to have a strong influence on the predicted probabilities of advice relationships. However, the

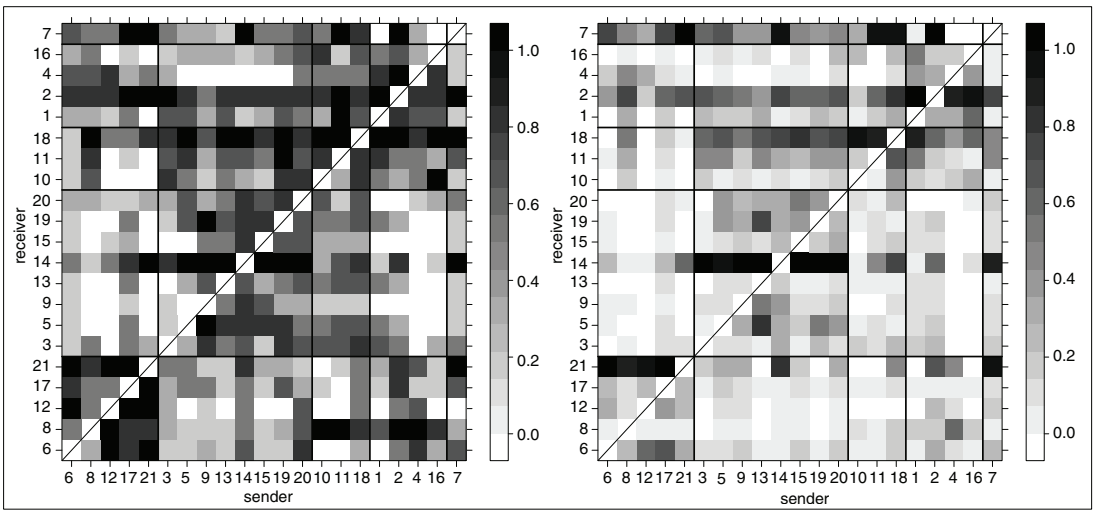

Figure 6 Predicted probabilities to observe an arrow from individual i (x-axis) to individual j (y-axis) in the two subpopulations. On both axes, nodes are ordered by department, and horizontal and vertical lines separate employees into the four departments the firm is divided into (so, e.g., employees 6, 8, 12, 17 and 21 belong to department 1; note that employee 7, the CEO, doesn't belong to any department). It is apparent that graphs within the first subpopulation are denser, and that in both subpopulations department affiliation induces a community structure wherein advice relationships are typically more frequent within the same department than between different departments.

large number of parameters (840) employed by the mixture of unconstrained models makes it difficult to draw any further conclusion on similarities and differences between the two subpopulations, and to relate those to any other known feature of the employees. Therefore, we now consider a more parsimonious model where we try to relate the presence of an arrow to such features. Krackhardt (1987) collected the following additional information about the employees:

- age and length of service (tenure) of each employee;
- position occupied by each employee in the firm; one employee is the CEO, two are vice-presidents and the remaining 18 have supervision roles; here, we consider a binary distinction between CEO and vice-presidents on the one hand, and the other 18 employees on the other;
- the department that each employee belonged to; in total, the firm comprises four departments.

We incorporate these covariates into the analysis by considering a network model where we combine features of the p_2 model (van Duijn et al., 2004) and of the stochastic blockmodel a priori (Holland et al., 1983), and we furthermore let arrows depend on the available set of monadic covariates (Signorelli and Wit, 2018). Such a model can be regarded as a degree-corrected stochastic blockmodel a priori with

Table 2 MLEs and standard errors for β^m, $m \in \{1, 2\}$ in model (5.1), and MLEs of σ^m and τ^m. Asterisks (*) denote regression coefficients that are significantly different from 0 ($H_0 : \beta_j^m = 0$) at $\alpha = 5\%$ level. The last column contains the *p*-value of the test for equality of each parameter in the two subpopulations ($H_0 : \beta_j^1 = \beta_j^2$)

Parameter	$\hat{\theta}^1$	$\hat{\theta}^2$	$SE(\hat{\theta}^1)$	$SE(\hat{\theta}^2)$	p-value ($\theta^1 = \theta^2$)
β_0	0.809	−1.997*	0.671	0.439	0.000
β_1 (age sender)	−0.014	−0.006	0.012	0.010	0.972
β_2 (tenure sender)	−0.035*	−0.016	0.017	0.009	0.930
β_3 (sender in lead pos.)	0.014	0.008	0.016	0.013	0.977
β_4 (perceiver = sender)	1.128*	1.020*	0.231	0.146	0.675
β_5 (age receiver)	0.034	0.044*	0.022	0.012	0.964
β_6 (tenure receiver)	0.543*	0.582*	0.214	0.170	0.876
β_7 (receiver in lead pos.)	1.407*	2.058*	0.287	0.150	0.017
β_8 (perceiver = receiver)	1.353*	1.354*	0.231	0.149	0.998
σ (rand. int. sender)	0.329	0.267			
τ (rand. int. receiver)	0.472	0.232			

covariates, where the blocks are given by the four departments in the company, the (in- and out-) degree-correction is carried out using random effects and where we furthermore account for the effect of several monadic covariates. Let A_i and T_i denote the age and tenure of node i, let L_i be a binary variable distinguishing individuals in leadership positions that is 1 if i is either the CEO or a vice-president and 0 otherwise, and let $I(i = k)$ and $I(j = k)$ be binary variables that are 1 if, respectively, the perceiver (k) is sender (i) or receiver (j), and 0 otherwise. Furthermore, let $D_i \in \{1, 2, 3, 4\}$ denote the department that individual i is affiliated to. We consider the following mixture model: $y_{ij}^k | \left(z_k, u_i^{z_k}, v_i^{z_k}\right) \sim \text{Bern}\left(\pi_{ij}^{z_k}\right)$, where

$$\begin{aligned} \text{logit}(\pi_{ij}^{z_k}) = {} & \beta_0^{z_k} + u_i^{z_k} + v_j^{z_k} + \beta_1^{z_k} A_i + \beta_2^{z_k} T_i + \beta_3^{z_k} L_i + \beta_4^{z_k} I(i = k) \\ & + \beta_5^{z_k} A_j + \beta_6^{z_k} T_j + \beta_7^{z_k} L_j + \beta_8^{z_k} I(j = k) \\ & + \sum_{r=1}^{4} \gamma_r^{z_k} I[D_i = r] + \sum_{s=1}^{4} \delta_s^{z_k} I[D_j = s] + \sum_{r=1}^{4} \sum_{s=1}^{4} \xi_{rs}^{z_k} I[D_i = r] I[D_j = s], \end{aligned} \tag{5.1}$$

$u_i^{z_k} \sim N\left[0, (\sigma^{z_k})^2\right]$ and $v_j^{z_k} \sim N\left[0, (\tau^{z_k})^2\right]$ are random intercepts that allow to model parsimoniously the in- and out-degree distributions, and γ_r^m, δ_s^m and ξ_{rs}^m are blockmodel main effects and interactions subject to the constraints that $\sum_{r=1}^4 \gamma_r^m = 0$, $\sum_{s=1}^4 \delta_s^m = 0$ and $\sum_{r=1}^4 \sum_{s=1}^4 \xi_{rs}^m = 0$ for every $m \in \{1, 2\}$.

We remark that not only model (5.1) is considerably thriftier than the unconstrained mixture model previously considered (the former comprises 54 parameters, the latter 840), but it is also more interpretable as it enables us to study how the advice relationships reconstructed by the employees could have been affected by individual (age and tenure) and organizational (department and leading roles) features of the employees and firm.

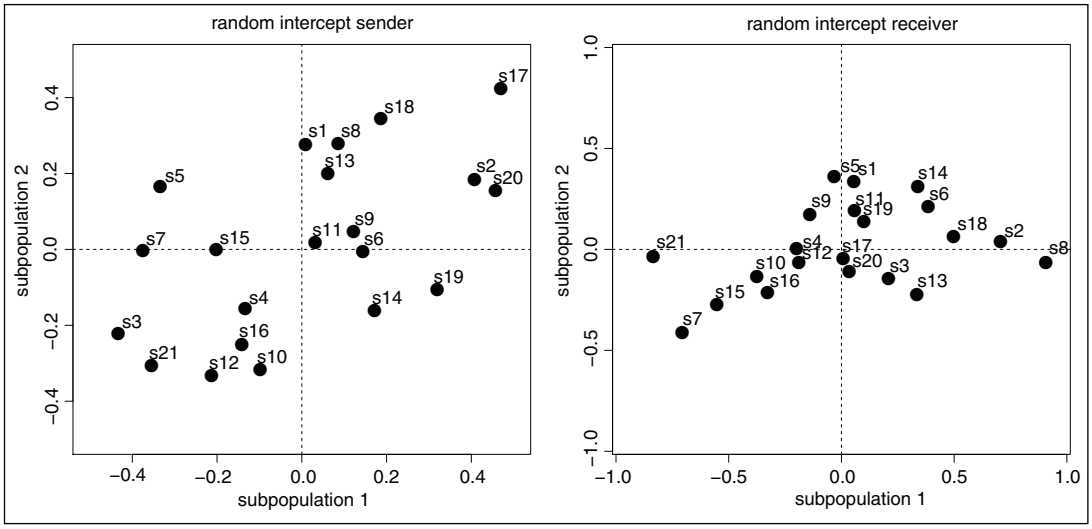

Figure 7 Predicted random intercepts for sender ($\hat{u}_i$) and receiver ($\hat{v}_j$) in subpopulations 1 (x-axis) and 2 (y-axis)

Table 2 contains the MLEs of the fixed effects $\beta_0^m, \ldots, \beta_8^m,\ m \in \{1, 2\}$ and of the standard deviation of the random effects in model (5.1). In both subpopulations we observe that the perceiver tends to report more ingoing and outgoing relationships that involve him ($\hat{\beta}_4 > 0$ and $\hat{\beta}_8 > 0$). Moreover, there is a common tendency to seek advice from employees with longer tenure within the firm ($\hat{\beta}_6 > 0$) and from the CEO and the vice-presidents ($\hat{\beta}_7 > 0$). As concerns differences between the two subpopulations, not only it is apparent that graphs in the first subpopulation are significantly denser than those in the second ($\hat{\beta}_0^1 > \hat{\beta}_0^2$), but we also observe that in the second subpopulation the tendency to seek advice from the CEO and vice-presidents is significantly stronger than in the first one ($\hat{\beta}_7^2 > \hat{\beta}_7^1$). Furthermore, $\hat{\tau}^1 > \hat{\tau}^2$ indicates a less heterogeneous distribution of out-degrees in the second subpopulation (i.e., in subpopulation 1 advice requests tend to be more concentrated on fewer employees).

Figure 7 shows the distribution of the predicted random effects for sender and receiver in model (5.1). In the left-hand-side plot, which displays the sender effect, most points fall in the first and third quadrant; this is an indication that perceivers in the two subpopulations have similar ideas on how many colleagues a certain individual seeks advice from. For example, individual 17 has the highest in-degree correction $\hat{u}_i$ in both subpopulations. Similar observations can be made for the right-hand-side plot; moreover, here we clearly see the different magnitude of the out-degree correction in the two subpopulations, which we already inferred from Table 2.

Table 3 summarizes the significance and sign of the estimated block-interaction parameters $\xi_{rs}^{z_k}$ in model (5.1) (more details on $\gamma_r^{z_k}$, $\delta_s^{z_k}$ and $\xi_{rs}^{z_k}$ are provided in Table 1

Table 3 Sign and significance of the block-interaction parameters $\xi_{rs}^{z_k}$ in cluster 1 (left) and cluster 2 (right). $\oplus$ and $\ominus$ denote parameters significantly different from 0 ($p < 0.05$), + and − parameters with $p > 0.05$. The value and significance of all $\gamma_r^{z_k}$, $\delta_s^{z_k}$ and $\xi_{rs}^{z_k}$ can be found in Table 1 of the supplementary material.

Subpopulation 1					Subpopulation 2				
	Dept. receiver					Dept. receiver			
Dept. sender	1	2	3	4	Dept. sender	1	2	3	4
1	$\oplus$	$\ominus$	$\ominus$	−	1	$\oplus$	$\ominus$	$\ominus$	−
2	$\ominus$	$\oplus$	+	$\ominus$	2	$\ominus$	$\oplus$	+	$\ominus$
3	$\ominus$	$\oplus$	$\oplus$	+	3	$\ominus$	+	$\oplus$	$\ominus$
4	−	$\ominus$	−	$\oplus$	4	−	$\ominus$	−	$\oplus$

of the supplementary material). In both clusters we find evidence of a rather strong community structure induced by department affiliation, which results in employees seeking advice from members of the same department more frequently ($\hat{\xi}_{rr}^{m} > 0 \; \forall r \in \{1, 2, 3, 4\} \wedge \forall m \in \{1, 2\}$). All the other block-interactions are typically negative or non-significant, with the exception of advice relationships from department 3 to department 2 in cluster 1 ($\hat{\xi}_{32}^{1} > 0$). Overall, the two subpopulations appear to have a similar, but not identical, view of the intensity of advice relationships occurring between members from different departments.

6 Discussion

We have developed a model-based clustering approach for populations of networks that specifies a joint statistical model for all graphs in the population and that is capable of identifying subpopulations of graphs which share a similar generative model, but which may still look like quite different networks in edge-space. Building on the fact that GLMs and GLMMs represent a flexible and efficient tool for modelling and estimating a wide variety of generative processes, we have proposed to employ mixtures of GLMs or GLMMs to perform model-based clustering of networks. Estimation of the proposed mixtures of network models can be efficiently carried out by an EM algorithm. The identification of the number of subpopulations that form the mixture has been performed with standard model selection criteria.

Evaluation of the proposed method on simulated data shows that the accuracy of the clustering method strongly depends on the size of the graphs and on the number of clusters, and much less on the number of graphs in the population. In particular, the accuracy increases quickly with the number of vertices and it decreases, as expected, with the number of clusters. The estimation of the number of subpopulations M can be based on the minimization of model selection criteria such as AIC and BIC. As illustrated in Section 4.2, the performance of AIC and BIC appears to be similar when a relatively large number of graphs is available; however, for small K, BIC tends to

systematically underestimate M, so we recommend the use of AIC when dealing with a small number of networks.

The approach presented in this article is able to consider mixtures of network models that make conditional independence assumptions on the probability of existence of edges. Examples of such models include the p_1 model of Holland and Leinhardt (1981), the p_2 model of van Duijn et al. (2004), different types of stochastic blockmodels a priori (Holland et al., 1983; Wang and Wong, 1987; Signorelli and Wit, 2018), the loglinear models proposed by Wolfe and Olhede (2013), the unconstrained model illustrated in Section 2.2.3 and any feasible combination of these models, like the one that we have employed in Equation (5.1). We note that ERGMs (Frank and Strauss, 1986) fall outside this class of models as they violate the conditional independence assumption, although quasi-likelihood estimation via a GLM is possible (van Duijn et al., 2009). We have made an attempt to implement this, but the results were not cleared and therefore we do not recommend it in general.

Acknowledgements

We thank two anonymous reviewers, whose useful suggestions and remarks have contributed to improve this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors gratefully acknowledge funding from the COST Action *European Cooperation for Statistics of Network Data Science* (CA15109).

Supplementary material

Supplementary material for this article is available online at the link: `http://www.statmod.org/smij/archive.html`

References

Abegaz F and Wit E (2013) Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, **14**, 586–99.

Berger J, Bayarri M, and Pericchi L (2014) The effective sample size. *Econometric Reviews*, **33**, 197–217.

Breslow NE and Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**, 1–38.

Durante D, Dunson DB and Vogelstein JT (2017) Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*, **112**, 1516–30.

Frank O and Strauss D (1986) Markov graphs. *Journal of the American Statistical Association*, **81**, 832–42.

Friedman J, Hastie T and Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–41.

Grün B and Leisch F (2008) Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas.*, edited by Shalabh and C. Heumann. pages 205–30. Heidelberg: Available at: https://link.springer.com/content/pdf/10.1007/978-3-7908-2064-5.pdf).

Hanneke S, Fu W and Xing EP (2010) Discrete temporal models of social networks. *Electronic Journal of Statistics*, **4**, 585–605.

Hoff PD, Raftery AE and Handcock MS (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**, 1090–98.

Holland PW, Laskey KB and Leinhardt S (1983) Stochastic blockmodels: First steps. *Social Networks*, **5**, 109–37.

Holland PW and Leinhardt S (1981) An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33–50.

Jaccard P (1912) The distribution of the ora in the alpine zone:. 1. *New Phytologist*, **11**, 37–50.

Krackhardt D (1987) Cognitive social structures. *Social Networks*, **9**, 109–34.

Matias C and Miele V (2017) Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B*. doi: 10.1111/rssb.12200

McCullagh P and Nelder JA (1989) *Generalized Linear Models*. Vol. 37. Boca Raton, FL: CRC press.

Mukherjee SS, Sarkar P and Lin L (2017) On clustering network-valued data. In *Advances in neural information processing systems*, edited by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, pages 7071–81. Available at: https://papers.nips.cc/book/advances-in- neural-information-processing-systems-30-2017

Paul S and Chen Y (2016) Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, **10**, 3807–70.

Perry PO and Wolfe PJ (2012) Null models for network data. arXiv preprint arXiv:1201.5871.

Reyes P and Rodriguez A (2016) Stochastic blockmodels for exchangeable collections of networks. arXiv preprint arXiv: 1606.05277.

Reynolds AP, Richards G, de la Iglesia B and Rayward-Smith VJ (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, **5**, 475–504.

Schutze H, Manning CD and Raghavan P (2008) *Introduction to Information Retrieval*. Vol 39. Cambridge: Cambridge University Press.

Sewell DK and Chen Y (2015) Latent space models for dynamic networks. *Journal of the American Statistical Association*, **110**, 1646–57.

Signorelli M (2017) Variable selection for (realistic) stochastic blockmodels. In *SIS 2017. Statistics and Data Science: New Challenges, New Generations*, edited by A. Petrucci and R. Verde. pages 927–34. Florence: Firenze University Press.

Signorelli M and Wit EC (2018) A penalized inference approach to stochastic block modelling of community structure in the

Italian Parliament. *Journal of the Royal Statistical Society: Series C*, **67**, 355–69.

Snijders TA (2001) The statistical evaluation of social network dynamics. *Sociological Methodology*, **31**, 361–95.

Snijders TA (2011) Statistical models for social networks. *Annual Review of Sociology*, **37**, 131–53.

Snijders TA and Nowicki K (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75–100.

Stanley N, Shai S, Taylor D and Mucha PJ (2016) Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering*, **3**, 95–105.

Sweet TM, Thomas AC and Junker BW (2014) Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In *Handbook on Mixed Membership Models and Their Applications*, edited by Edoardo M. Airoldi, David Blei, Elena A. Erosheva, Stephen E. Fienberg. pages 463–88. Boca Raton, FL: Chapman and Hall/CRC.

Taya F, de Souza J, Thakor NV and Bezerianos A (2016) Comparison method for community detection on brain networks from neuroimaging data. *Applied Network Science*, **1**, 8.

van Duijn MA, Gile KJ and Handcock MS (2009) A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, **31**, 52–62.

van Duijn MA, Snijders TA and Zijlstra BJ (2004) p2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, **58**, 234–54.

Vujacic I, Abbruzzo A and Wit E (2015) A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *Journal of Statistical Computation and Simulation*, **85**, 3628–40.

Wang YJ and Wong GY (1987) Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8–19.

Wolfe PJ and Olhede SC (2013) Nonparametric graphon estimation. arXiv preprint arXiv:1309.5936.