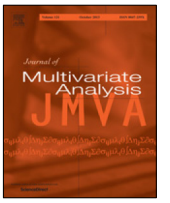
Journal of Multivariate Analysis 175 (2020) 104540

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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



Model-based clustering of time-evolving networks through temporal exponential-family random graph models

Kevin H. Lee [a], Lingzhou Xue [b,*], David R. Hunter [b]

[a] Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA
[b] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 30 December 2018
Received in revised form 19 August 2019
Accepted 19 August 2019
Available online 5 September 2019

AMS 2010 subject classifications:
primary 62H12
secondary 62H30

Keywords:
Minorization–maximization
Model-based clustering
Model selection
Temporal ERGM
Time-evolving network
Variational EM algorithm

ABSTRACT

Dynamic networks are a general language for describing time-evolving complex systems, and discrete time network models provide an emerging statistical technique for various applications. It is a fundamental research question to detect a set of nodes sharing similar connectivity patterns in time-evolving networks. Our work is primarily motivated by detecting groups based on interesting features of the time-evolving networks (e.g., stability). In this work, we propose a model-based clustering framework for time-evolving networks based on discrete time exponential-family random graph models, which simultaneously allows both modeling and detecting group structure. To choose the number of groups, we use the conditional likelihood to construct an effective model selection criterion. Furthermore, we propose an efficient variational expectation–maximization (EM) algorithm to find approximate maximum likelihood estimates of network parameters and mixing proportions. The power of our method is demonstrated in simulation studies and empirical applications to international trade networks and the collaboration networks of a large research university.

1. Introduction

Dynamic networks are a general language for describing time-evolving complex systems, and discrete time network models provide an emerging statistical technique to study biological, business, economic, information, and social systems. For example, time-evolving networks shed light on understanding critical processes such as the study of biological functions using protein–protein interaction networks [14,40], and also contribute to assessing infectious disease epidemiology and time-evolving structures of social networks [6,25,32].

A group can be defined as a set of nodes sharing similar connectivity patterns. In computer science and statistical physics, many node clustering algorithms have been developed. Girvan and Newman [13] propose an algorithm to identify groups based on edge "betweenness". They construct groups by progressively removing the edges that connect groups most from the original network. Newman and Girvan [34] proposed three different measures of "betweenness" and compared the results based on modularity, which measures the quality of a particular division of a network. On the other hand, in statistics, analyzing and clustering networks are often based on statistical mixture models. One idea of model-based clustering in networks comes from Handcock et al. [15], who propose a latent position cluster model that extends the latent space model of Hoff et al. [18] to take account of clustering, using the model-based clustering ideas

* Corresponding author.
E-mail address: lzxue@psu.edu (L. Xue).

https://doi.org/10.1016/j.jmva.2019.104540
0047-259X/

of Fraley and Raftery [12]. In the current literature, there are two very popular statistical modeling frameworks. One is the stochastic blockmodel (SBM), and the other is the exponential-family random graph model (ERGM).

Stochastic blockmodels were first introduced by Holland et al. [19], and they focused on the case of a priori specified blocks, where the memberships are known or assumed, and the goal is to estimate a matrix of edge probabilities. A statistical approach to a posteriori block modeling for networks was introduced by Snijders and Nowicki [39] and Nowicki and Snijders [35], where the objective is to estimate the matrix of edge probabilities and the memberships simultaneously. Airoldi et al. [2] relax the assumption of a single latent role for nodes and develop a mixed membership stochastic blockmodel. Karrer and Newman [22] relax the assumption that a stochastic blockmodel treats all nodes within a group as stochastically equivalent and proposes a degree-corrected stochastic blockmodel. Moreover, in recent years, asymptotic theory of these models has been advanced by Bickel and Chen [8], Choi et al. [9] and Amini et al. [5], and others. A group, or community, in a stochastic blockmodel is defined as a set of nodes with more edges amongst its nodes than between its nodes. The communities are interpreted meaningfully in many research fields. For example, in citation and collaboration networks, such communities can be interpreted as scientific disciplines [21,33]. Communities in food web networks can be interpreted as ecological subsystems [13].

As dynamic network analysis has become an emergent scientific field, there has been a growing number of dynamic network models in the stochastic blockmodel framework. Yang et al. [49] propose a model that captures the evolution of communities with fixed connectivity parameters. Xu and Hero [47] and Matias and Miele [31] propose methods that allow both community memberships and connectivity parameters to vary over time. Xing et al. [46] and Ho et al. [17] propose dynamic extensions of the mixed membership stochastic blockmodel using a state space model. More details about recent statistical methods for dynamic networks with latent variables can be found in a recent survey [23].

Different from the stochastic blockmodel framework, exponential-family random graph models allow researchers to incorporate interesting features of the network into models. Researchers can specify a model capturing those features and cluster nodes based on the specified model. Indeed, the stochastic blockmodel is a special case of a mixture of exponential-family random graph models. Some estimation algorithms for exponential-family random graph models do not scale well computationally to large networks. Vu et al. [42] propose ERGM-based clustering for large-scale cross-sectional networks that solves the scalability issue by assuming dyadic independence conditional on the group memberships of nodes. Agarwal and Xue [1] propose ERGM-based clustering to study large-scale weighted networks. In recent years, several authors also have proposed discrete time network models based on ERGM. Hanneke et al. [16] propose a temporal ERGM (TERGM) to fit the model to a network series and Krivitsky and Handcock [26] propose a separable temporal ERGM (STERGM) that gives more flexibility in modeling time-evolving networks.

Our work is primarily motivated by detecting groups based on interesting features of time-evolving networks, and our results advance existing literature by introducing a promising framework that incorporates model-based clustering while remaining computationally scalable for time-evolving networks. This framework is based on discrete time exponential-family random graph models and inherits the philosophy of finite mixture models, which simultaneously allows both modeling and detecting groups in time-evolving networks. The groups can be defined differently based on how researchers and practitioners incorporate interesting features of the time-evolving networks into the models. For example, in sociology, researchers are interested in whether, say, same-gender friendships are more stable than other friendships or whether there are differences among ethnic categories in forming lasting sexual partnerships over time [24,28]. In this case, stability parameters can be incorporated into the model, resulting in different groups of nodes with different degrees of stability. We also propose an efficient variational expectation–maximization (EM) algorithm that exhibits computational scalability for time-evolving networks by exploiting variational methods and minorization–maximization (MM) techniques. Moreover, we propose a conditional likelihood Bayesian information criterion to solve the model selection problem of determining an appropriate number of groups.

The rest of this paper is organized as follows. In Section 2, we present our model-based clustering method for time-evolving networks based on a finite mixture of discrete time exponential-family random graph models. Section 3 designs an efficient variational expectation–maximization algorithm to find approximate maximum likelihood estimates of network parameters and mixing proportions. Given these estimates, we can infer membership labels and solve the problem of detecting groups for time-evolving networks. In Section 4, we use conditional likelihood to construct an effective model selection criterion. The power of our method is demonstrated by simulation studies in Section 5 and real-world applications to international trade networks and collaboration networks in Section 6. Section 7 includes a few concluding remarks. Proofs and technical details are provided in Appendix and the supplementary file.

2. Methodology

2.1. Model-based clustering of time-evolving networks

We present the model-based clustering for time-evolving networks based on a finite mixture of discrete time exponential-family random graph models. First, we introduce the necessary notation. Let n nodes be fixed over time and indexed by $1, \ldots, n$. Let $\mathbf{Y}_t = \{Y_{t,ij}\}_{1\le i,j\le n} \in \mathcal{Y}$ represent a random binary network at time $t \in \{0, \ldots, T\}$ and denote by $\mathbf{y}_t = \{y_{t,ij}\}_{1\le i,j\le n}$ the corresponding observed network, where $\mathcal{Y}$ is the set of all possible networks. Let $\boldsymbol{\theta} \in \mathbb{R}^p$ be a

vector of p network parameters of interest. Under the kth-order Markov assumption, discrete time exponential-family random graph models are of the form

$$\Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}, \ldots, \mathbf{y}_0) = \exp\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{y}_t, \ldots, \mathbf{y}_{t-k}) - \psi(\boldsymbol{\theta}, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-k})\},$$

where $\psi(\boldsymbol{\theta}, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-k}) = \ln \sum_{\mathbf{y}^* \in \mathcal{Y}} \exp\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{y}^*, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-k})\}$, ensuring that $\Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}, \ldots, \mathbf{y}_0)$ sums to 1. Here, $\mathbf{g}(\mathbf{y}_t, \ldots, \mathbf{y}_{t-k})$ is a p-dimensional vector of sufficient statistics on networks $\mathbf{y}_t, \ldots, \mathbf{y}_{t-k}$.

We now focus on the simplest case of discrete time exponential-family random graph models under the first-order Markov assumption and we write the one-step transition probability from $\mathbf{Y}_{t-1}$ to $\mathbf{Y}_t$ as

$$\Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}) = \exp\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{y}_t, \mathbf{y}_{t-1}) - \psi(\boldsymbol{\theta}, \mathbf{y}_{t-1})\}, \tag{1}$$

where $\psi(\boldsymbol{\theta}, \mathbf{y}_{t-1})$ and $\mathbf{g}(\mathbf{y}_t, \mathbf{y}_{t-1})$ are defined as above.

Remark 1. Given covariates $\mathbf{x}_t$ and a vector $\boldsymbol{\beta} \in \mathbb{R}^q$ of covariate coefficients, we can also write the transition probability from $\mathbf{Y}_{t-1}$ to $\mathbf{Y}_t$ with covariates as

$$\Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{x}_t) = \exp\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{y}_t, \mathbf{y}_{t-1}) + \boldsymbol{\beta}^\top \boldsymbol{h}(\mathbf{y}_t, \mathbf{x}_t) - \psi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}_{t-1})\},$$

where $\boldsymbol{h}(\mathbf{y}_t, \mathbf{x}_t)$ is a q-dimensional vector of statistics and $\psi(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}_{t-1}) = \ln \sum_{\mathbf{y}^* \in \mathcal{Y}} \exp\{\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{y}^*, \mathbf{y}_{t-1}) + \boldsymbol{\beta}^\top \boldsymbol{h}(\mathbf{y}^*, \mathbf{x}_t)\}$.

In general, for some choices of $\mathbf{g}(\mathbf{y}_t, \mathbf{y}_{t-1})$, the model in (1) is not tractable for modeling large networks, since the computing time to evaluate the likelihood function directly grows as $2^{\binom{n}{2}}$ in the case of undirected edges. Here, we restrict our attention to scalable exponential-family models by only choosing statistics that preserve dyadic independence wherein the distribution of $\mathbf{Y}_t$ given $\mathbf{Y}_{t-1}$ factors over the edge states $Y_{t,ij}$ given the previous time point edge states $Y_{t-1,ij}$, i.e.,

$$\Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}) = \prod_{i<j}^{n} \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}). \tag{2}$$

For example, we may consider the following statistics that preserve dyadic independence and capture interesting time-evolving network features in both TERGM and STERGM:

$$g^d(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{i<j}^{n} y_{t,ij}, \tag{3}$$

$$g^s(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{i<j}^{n} \left\{ y_{t,ij} y_{t-1,ij} + (1 - y_{t,ij})(1 - y_{t-1,ij}) \right\}, \tag{4}$$

$$g^f(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{i<j}^{n} \left(y_{t,ij} - y_{t,ij} y_{t-1,ij} \right), \tag{5}$$

$$g^p(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{i<j}^{n} y_{t,ij} y_{t-1,ij}. \tag{6}$$

The subscripted $i < j$ and superscripted n mean that summation should be taken over all pairs (i, j) with $1 \le i < j \le n$; the same is true for products as in Eq. (2). Corresponding to the first and second statistics above are TERGM parameters: θ^d relates to density, or the number of edges in the network at time t, while θ^s relates to stability, or the number of edges maintaining their status from time $t-1$ to time t. Corresponding to the third and fourth statistics above are STERGM parameters: θ^f relates to formation, or the number of edges absent at time $t-1$ but present at time t, while θ^p relates to persistence, or the number of edges existing at time $t-1$ that survive to time t.

Remark 2. We can extend the proposed model to handle directed binary networks by considering statistics that preserve dyadic independence and capture directed features of the network. For example, we can consider the TERGM statistics $g^r(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{i,j}^{n} y_{t,ij} y_{t-1,ji}$, which represent reciprocity. The subscripted i, j and superscripted n mean that summation should be taken over all pairs (i, j) with $1 \le i \ne j \le n$

Here, as in Vu et al. [42], we assume that the joint probability mass function, given an initial network $\mathbf{y}_0$, has a K-component mixture form as follows:

$$\begin{aligned} \Pr(\mathbf{y}_1, \ldots, \mathbf{y}_T \mid \mathbf{y}_0) &= \sum_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \Pr(\boldsymbol{Z} = \boldsymbol{z} \mid \mathbf{y}_0) \prod_{t=1}^{T} \Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}, \boldsymbol{z}) \right\} \\ &= \sum_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \Pr(\boldsymbol{Z} = \boldsymbol{z} \mid \mathbf{y}_0) \prod_{t=1}^{T} \prod_{i<j}^{n} \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, \boldsymbol{z}) \right\}, \end{aligned} \tag{7}$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ denotes the membership indicators that follow the following multinomial distributions

$$\boldsymbol{Z}_i \mid \pi_1, \ldots, \pi_K \overset{\text{i.i.d.}}{\sim} \text{Multinomial}(1; \pi_1, \ldots, \pi_K),$$

and $\mathcal{Z}$ denotes the support of the membership indicators $\boldsymbol{Z}$. In the mixture form (7), the assumption of conditional dyadic independence given $\boldsymbol{z}$ strikes a balance between model complexity and parsimony because it allows for marginal dyadic dependence and it means nodes in the same group share the same model parameters. For now, the number of groups K is fixed and known. In Section 4, we will discuss how to choose an optimal number of groups K.

Now, we observe a series of networks, $\mathbf{y}_1, \ldots, \mathbf{y}_T$, given an initial network $\mathbf{y}_0$. The log-likelihood of the observed network series is

$$\ell(\boldsymbol{\pi}, \boldsymbol{\theta}) = \ln \left[\sum_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \Pr(\boldsymbol{Z} = \boldsymbol{z} \mid \mathbf{y}_0) \prod_{t=1}^{T} \Pr(\mathbf{Y}_t = \mathbf{y}_t \mid \mathbf{y}_{t-1}, \boldsymbol{z}) \right\} \right].$$

Our aim is to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ via maximizing the log-likelihood $\ell(\boldsymbol{\pi}, \boldsymbol{\theta})$, i.e.,

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = \underset{(\boldsymbol{\pi}, \boldsymbol{\theta})}{\arg\max}\, \ell(\boldsymbol{\pi}, \boldsymbol{\theta}).$$

However, directly maximizing the log-likelihood function is computationally intractable since it is a sum over all possible latent block structures. Hence, in Section 3, we design a novel variational EM algorithm to efficiently find the approximate maximum likelihood estimates. We shall see that the parameter estimates obtained by this algorithm can provide group membership labels.

Before proceeding, we give specific examples of discrete time exponential-family random graph models with stability parameter(s) that control the rate of evolution of a network.

Example 1. The model in (7) takes the following form when $\boldsymbol{g}(\mathbf{y}_t, \mathbf{y}_{t-1})$ consists only of the stability parameters (4) and when node i belongs to group k and node j belongs to group l:

$$\Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, \boldsymbol{z}) \propto \exp\{(\theta_k^s + \theta_l^s)(y_{t,ij} y_{t-1,ij} + (1 - y_{t,ij})(1 - y_{t-1,ij}))\}.$$

Example 2. The model in (7) takes the following form when $\boldsymbol{g}(\mathbf{y}_t, \mathbf{y}_{t-1})$ consists of both formation parameters (5) and persistence parameters (6) and when node i belongs to group k and node j belongs to group l:

$$\Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, \boldsymbol{z}) \propto \exp\{(\theta_k^f + \theta_l^f)(y_{t,ij} - y_{t,ij} y_{t-1,ij}) + (\theta_k^p + \theta_l^p) y_{t,ij} y_{t-1,ij}\}.$$

Remark 3. When $K = 1$, Example 1 reduces to TERGM with a stability parameter as in Hanneke et al. [16] and the model in Example 2 reduces to STERGM with formation and persistence parameters as in Krivitsky and Handcock [26].

2.2. Parameter identifiability

Parameter identifiability is essential to avoid inconsistent parameter estimation results among different methods. The unique identifiability of the parameters in a broad class of random graph mixture models has been shown by Allman et al. [3,4]. Here we prove generic identifiability for our proposed discrete time exponential-family random graph mixture model where the distribution of $\mathbf{Y}_t$ given $\mathbf{Y}_{t-1}$ is factorized as in Eq. (2). Theorem 1, whose proof is given in Appendix, extends the identifiability result of the stochastic blockmodel of Allman et al. [3,4] to discrete time exponential-family random graph mixture models. In this context, "generically identifiable" means the set of all uniquely identifiable parameters has a complement of Lebesgue measure zero in the full parameter space.

Theorem 1. *Suppose a time-evolving network on n nodes is first-order Markov and has the form given in Eq. (2). The parameters π_k, $1 \le k \le K$, and $\boldsymbol{p}_{kl} = \Pr(Y_{t,ij} = 1 \mid y_{t-1,ij}, Z_{ik} = Z_{jl} = 1)$, $1 \le k \le l \le K$, are generically identifiable from the joint distribution of $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, given an initial network $\mathbf{y}_0$, up to permutations of the subscripts $1, \ldots, K$, if*

$$\begin{cases} \sqrt{n} \ge 4, & \text{for } K = 2; \\ \sqrt{n} \ge K - 1 + \dfrac{(K+2)^2}{4}, & \text{for } K \text{ even}, K \ge 4; \\ \sqrt{n} \ge K - 1 + \dfrac{(K+1)(K+3)}{4}, & \text{for } K \text{ odd}. \end{cases}$$

Moreover, when the model also takes the exponential-family form (1), the network parameters $\boldsymbol{\theta}_k \in \mathbb{R}^p$, $1 \le k \le K$ are generically identifiable, up to permutations of the subscripts $1, \ldots, K$, if the corresponding network statistics are not linearly dependent and $p \le \lfloor (K+1)/2 \rfloor$ where $\lfloor \cdot \rfloor$ is a floor function that maps the given value to the largest integer less than or equal to it.

3. Computation

We present a novel variational EM algorithm to solve model-based clustering for time-evolving networks. Our algorithm is based on the algorithm presented by Vu et al. [42]. The algorithm combines the power of variational methods [43] and minorization–maximization techniques [20] to effectively handle both the computationally intractable log-likelihood function $\ell(\boldsymbol{\pi}, \boldsymbol{\theta})$ and the non-convex optimization problem of the lower bound of the log-likelihood. We introduce an auxiliary distribution $A(\boldsymbol{z}) \equiv \Pr(\boldsymbol{Z} = \boldsymbol{z})$ to derive a tractable lower bound on the intractable log-likelihood function. Using Jensen's inequality, the log-likelihood function may be shown to be bounded from below as follows:

$$\ln \Pr(\mathbf{y}_1, \ldots, \mathbf{y}_T \mid \mathbf{y}_0) = \ln \left\{ \sum_{\boldsymbol{z} \in \mathcal{Z}} \frac{\Pr(\mathbf{y}_1, \ldots, \mathbf{y}_T, \boldsymbol{z} \mid \mathbf{y}_0)}{A(\boldsymbol{z})} A(\boldsymbol{z}) \right\} \geq \sum_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \ln \frac{\Pr(\mathbf{y}_1, \ldots, \mathbf{y}_T, \boldsymbol{z} \mid \mathbf{y}_0)}{A(\boldsymbol{z})} \right\} A(\boldsymbol{z}). \tag{8}$$

We would obtain the best lower bound when $A(\boldsymbol{z})$ is given by $\Pr(\boldsymbol{Z} = \boldsymbol{z} \mid \mathbf{y}_T, \ldots, \mathbf{y}_0)$, where the inequality becomes equality. However, this form of $A(\boldsymbol{z})$ is computationally intractable since it cannot be further factored over nodes. We therefore constrain $A(\boldsymbol{z})$ to a subset of tractable choices and maximize the tractable lower bound to find approximate maximum likelihood estimates.

Here, we constrain $A(\boldsymbol{z})$ to the mean-field variational family where the $\boldsymbol{Z}_i$ are mutually independent, i.e., $A(\boldsymbol{z}) = \prod_{i=1}^{n} \Pr(\boldsymbol{Z}_i = \boldsymbol{z}_i)$. We further specify $\Pr(\boldsymbol{Z}_i = \boldsymbol{z}_i)$ to be Multinomial$(1; \gamma_{i1}, \ldots, \gamma_{iK})$ for $i \in \{1, \ldots, n\}$, where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)$ is the variational parameter. In the estimation phase, whenever it is necessary to assign each node to a particular group, the ith node is assigned to the group with the highest value among $\hat{\gamma}_{i1}, \ldots, \hat{\gamma}_{iK}$.

If we denote the right side of (8) by $\mathrm{LB}(\boldsymbol{\pi}, \boldsymbol{\theta}; \boldsymbol{\Gamma})$, we may write

$$\mathrm{LB}(\boldsymbol{\pi}, \boldsymbol{\theta}; \boldsymbol{\Gamma}) = \sum_{t=1}^{T} \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{ik} \gamma_{jl} \ln \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, Z_{ik} = Z_{jl} = 1) + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} (\ln \pi_k - \ln \gamma_{ik}). \tag{9}$$

Additional details on obtaining the lower bound are presented in the supplementary file. If $\boldsymbol{\pi}^{(\tau)}$, $\boldsymbol{\theta}^{(\tau)}$, and $\boldsymbol{\Gamma}^{(\tau)}$ denote the parameter estimates at the τth iteration of our variational EM algorithm, the algorithm alternates between

- **Idealized Variational E-step:**
  Let $\boldsymbol{\Gamma}^{(\tau+1)} = \arg\max_{\boldsymbol{\Gamma}} \mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma})$.
- **Variational M-step:**
  Let $(\boldsymbol{\pi}^{(\tau+1)}, \boldsymbol{\theta}^{(\tau+1)}) = \arg\max_{(\boldsymbol{\pi}, \boldsymbol{\theta})} \mathrm{LB}(\boldsymbol{\pi}, \boldsymbol{\theta}; \boldsymbol{\Gamma}^{(\tau+1)})$.

In the idealized variational E-step, it is difficult to directly maximize the nonconcave function $\mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma})$ with respect to $\boldsymbol{\Gamma}$. To address this challenge, we use a minorization–maximization technique to construct a tractable minorizing function of $\mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma})$, then maximize this minorizer. Define

$$\begin{aligned} Q(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\Gamma}^{(\tau)}; \boldsymbol{\Gamma}) = & \sum_{t=1}^{T} \sum_{i<j}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \left(\gamma_{ik}^2 \frac{\gamma_{jl}^{(\tau)}}{2\gamma_{ik}^{(\tau)}} + \gamma_{jl}^2 \frac{\gamma_{ik}^{(\tau)}}{2\gamma_{jl}^{(\tau)}} \right) \ln \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, Z_{ik} = Z_{jl} = 1) \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left(\ln \pi_k^{(\tau)} - \ln \gamma_{ik}^{(\tau)} - \frac{\gamma_{ik}}{\gamma_{ik}^{(\tau)}} + 1 \right), \end{aligned} \tag{10}$$

which satisfies the defining characteristics of a minorizing function, namely for all $\boldsymbol{\Gamma}$,

$$Q(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\Gamma}^{(\tau)}; \boldsymbol{\Gamma}) \leq \mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma}) \tag{11}$$

and

$$Q(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\Gamma}^{(\tau)}; \boldsymbol{\Gamma}^{(\tau)}) = \mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma}^{(\tau)}). \tag{12}$$

Additional details on constructing this minorizing function are presented in the supplementary material. Since (10) is concave in $\boldsymbol{\Gamma}$ and separates into functions of the individual γ_{ik} parameters, maximizing (10) is equivalent to solving a sequence of constrained quadratic programming subproblems with respect to $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ respectively, under constraints $\gamma_{i1}, \ldots, \gamma_{iK} \geq 0$ and $\sum_{k=1}^{K} \gamma_{ik} = 1$ for $i \in \{1, \ldots, n\}$.

To maximize $\mathrm{LB}(\boldsymbol{\pi}, \boldsymbol{\theta}; \boldsymbol{\Gamma}^{(\tau+1)})$ in the variational M-step, maximization with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ may be accomplished separately. First, to derive the closed-form update for $\boldsymbol{\pi}$, we introduce a Lagrange multiplier with the constraint $\sum_{k=1}^{K} \pi_k = 1$, which yields

$$\pi_k^{(\tau+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ik}^{(\tau+1)}, \quad k \in \{1, \ldots, K\}.$$

We could obtain $\boldsymbol{\theta}^{(\tau+1)}$ using the Newton–Raphson method, though naive application of Newton–Raphson will not guarantee an increase in (9), which is necessary for the ascent property of the lower bound of the log-likelihood. Since

the Hessian matrix $H(\boldsymbol{\theta}^{(\tau)})$ at the τth iteration is positive definite, it can be shown that if $\boldsymbol{\theta}$ moves in the direction $h^{(\tau)} = -H(\boldsymbol{\theta}^{(\tau)})^{-1}\nabla \mathrm{LB}(\boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma}^{(\tau+1)})$ of the Newton–Raphson method, the LB$(\cdot)$ function is guaranteed to increase initially. In our modified Newton–Raphson method we do not find the successor point $\boldsymbol{\theta}^{(\tau+1)} = \boldsymbol{\theta}^{(\tau)} + h^{(\tau)}$ as in the standard Newton–Raphson method. We instead take $h^{(\tau)}$ as a search direction and perform a line search [7] to find $\lambda^* = \arg\max_{\lambda\in(0,1]} \mathrm{LB}(\boldsymbol{\theta}^{(\tau)} + \lambda h^{(\tau)}; \boldsymbol{\Gamma}^{(\tau+1)})$, then we find the successor point $\boldsymbol{\theta}^{(\tau+1)}$ by

$$\boldsymbol{\theta}^{(\tau+1)} = \boldsymbol{\theta}^{(\tau)} - \lambda^* H(\boldsymbol{\theta}^{(\tau)})^{-1}\nabla \mathrm{LB}(\boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma}^{(\tau+1)}). \tag{13}$$

In Algorithm 1, we summarize the proposed variational EM algorithm.

Algorithm 1 Variational EM algorithm

- Initialize $\boldsymbol{\Gamma}^{(0)}$, $\boldsymbol{\pi}^{(0)}$, and $\boldsymbol{\theta}^{(0)}$.
- Iteratively solve the Variational E-step and M-step with $\tau \in \{0, 1, \ldots\}$ until convergence:
  - **Variational E-step**: Solve $\boldsymbol{\Gamma}^{(\tau+1)}$ from the maximization of $Q(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\Gamma}^{(\tau)}; \boldsymbol{\Gamma})$ under the constraints that $\gamma_{i1}, \ldots, \gamma_{iK} \geq 0$ and $\sum_{k=1}^{K} \gamma_{ik} = 1$ for $i \in \{1, \ldots, n\}$;
  - **Variational M-step**: Compute $\pi_k^{(\tau+1)} = (1/n)\sum_{i=1}^{n} \gamma_{ik}^{(\tau+1)}$ for $k \in \{1, \ldots, K\}$, and solve $\boldsymbol{\theta}^{(\tau+1)}$ using the modified Newton–Raphson method (13) with the gradient and Hessian of (9).

Remark 4. The initial parameters $\gamma_{ik}^{(0)}$ are chosen independently and uniformly on (0, 1), and each $\boldsymbol{\gamma}_i^{(0)}$ is multiplied by a normalizing constant so that $\sum_{k=1}^{K} \gamma_{ik}^{(0)} = 1$ for every i. We start with an M-step to obtain initial $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$.

Remark 5. Using standard arguments that apply to minorization–maximization algorithms [20], we can show that our variational EM algorithm preserves the ascent property of the lower bound of the log-likelihood, namely,

$$\mathrm{LB}(\boldsymbol{\pi}^{(\tau)}, \boldsymbol{\theta}^{(\tau)}; \boldsymbol{\Gamma}^{(\tau)}) \leq \mathrm{LB}(\boldsymbol{\pi}^{(\tau+1)}, \boldsymbol{\theta}^{(\tau+1)}; \boldsymbol{\Gamma}^{(\tau+1)}).$$

4. Model selection

In practice, the number of groups is unknown and should be chosen. Handcock et al. [15] propose a Bayesian method of determining the number of groups by using approximate conditional Bayes factors in a latent position cluster model. Daudin et al. [11] also derive a Bayesian model selection criterion that is based on the integrated classification likelihood (ICL). In this section, we use the conditional likelihood of the network series, given an estimate of the membership vector, to construct an effective model selection criterion.

Let $\hat{z}_{ik} = \mathbb{1}(\hat{\gamma}_{ik} = \max_{1\leq j\leq K} \hat{\gamma}_{ij})$ indicate the assignment of each node to a component once the algorithm has converged. We obtain the conditional log-likelihood of the network series $\mathbf{y}_1, \ldots, \mathbf{y}_T$, given initial network $\mathbf{y}_0$ and estimated component membership vector $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n)$, as

$$cl(\boldsymbol{\theta}, \hat{\boldsymbol{z}}) = \ln \Pr(\mathbf{y}_1, \ldots, \mathbf{y}_T \mid \mathbf{y}_0, \hat{\boldsymbol{z}}),$$

which can be written using conditional dyadic independence in the form

$$cl(\boldsymbol{\theta}, \hat{\boldsymbol{z}}) = \sum_{t=1}^{T}\sum_{i<j}^{n} \ln \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, \hat{\boldsymbol{z}}).$$

We propose the following conditional likelihood Bayesian information criterion to choose the number of groups for our method:

$$\text{CL-BIC}_K = -2cl(\hat{\boldsymbol{\theta}}_{\mathrm{cl}}, \hat{\boldsymbol{z}}) + d_K(\hat{\boldsymbol{\theta}}_{\mathrm{cl}}, \hat{\boldsymbol{z}}) \ln\{Tn(n-1)/2\},$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{cl}}$ is the conditional likelihood estimate given K groups and $d_K(\boldsymbol{\theta}, \hat{\boldsymbol{z}}) = \mathrm{tr}(H_K^{-1}V_K)$ is the model complexity, following Varin and Vidoni [41], based on $H_K = \mathrm{E}\{-\nabla^2_{\boldsymbol{\theta}} cl(\boldsymbol{\theta}, \hat{\boldsymbol{z}})\}$ and $V_K = \mathrm{var}\{\nabla_{\boldsymbol{\theta}} cl(\boldsymbol{\theta}, \hat{\boldsymbol{z}})\}$. We choose the optimal K by minimizing the CL-BIC score. The composite likelihood BIC has been previously studied by Saldana et al. [37] for stochastic blockmodels and by Xue et al. [48] for discrete graphical models.

We may derive the explicit conditional likelihood BIC for Examples 1 and 2. For TERGM with stability parameters $\theta_1^s, \ldots, \theta_K^s$ in Example 1, we obtain

$$cl(\boldsymbol{\theta}^s, \hat{\boldsymbol{z}}) = \sum_{t=1}^{T}\sum_{i<j}^{n}\Big[-\ln\{1+\exp(\theta^s_{\hat{z}_i} + \theta^s_{\hat{z}_j})\} + \{y_{t,ij}y_{t-1,ij} + (1-y_{t,ij})(1-y_{t-1,ij})\}(\theta^s_{\hat{z}_i} + \theta^s_{\hat{z}_j})\Big].$$

For any given K and the corresponding estimate $\hat{\boldsymbol{\theta}}^s_{\text{cl}}$, we derive the explicit estimate of V_K as

$$\hat{V}_K(\hat{\boldsymbol{\theta}}^s_{\text{cl}}) = \sum_{t=1}^{T} \boldsymbol{u}(\hat{\boldsymbol{\theta}}^s_{\text{cl}})\boldsymbol{u}(\hat{\boldsymbol{\theta}}^s_{\text{cl}})^\top,$$

where $\boldsymbol{u}(\hat{\boldsymbol{\theta}}^s_{\text{cl}}) = \left(u(\hat{\theta}^s_{\text{cl},1}), \ldots, u(\hat{\theta}^s_{\text{cl},K})\right)^\top$ and for $k \in \{1, \ldots, K\}$,

$$u(\hat{\theta}^s_{\text{cl},k}) = \sum_{i<j}^{n} \left\{ -\frac{\exp(\hat{\theta}^s_{\text{cl},\hat{z}_i} + \hat{\theta}^s_{\text{cl},\hat{z}_j})}{1+\exp(\hat{\theta}^s_{\text{cl},\hat{z}_i} + \hat{\theta}^s_{\text{cl},\hat{z}_j})} + y_{t,ij}y_{t-1,ij} + (1-y_{t,ij})(1-y_{t-1,ij}) \right\} \left(\hat{z}_{ik} + \hat{z}_{jk}\right).$$

We also derive the explicit estimate of H_K denoted by $\hat{H}_K(\hat{\boldsymbol{\theta}}^s_{\text{cl}})$ as follows:

$$\begin{bmatrix} T\sum_{i<j}^{n}\left[\frac{4\exp(\hat{\theta}^s_{\text{cl},1}+\hat{\theta}^s_{\text{cl},1})}{\{1+\exp(\hat{\theta}^s_{\text{cl},1}+\hat{\theta}^s_{\text{cl},1})\}^2}\right] I^{1}_{i,j} & \cdots & T\sum_{i<j}^{n}\left[\frac{\exp(\hat{\theta}^s_{\text{cl},1}+\hat{\theta}^s_{\text{cl},K})}{\{1+\exp(\hat{\theta}^s_{\text{cl},1}+\hat{\theta}^s_{\text{cl},K})\}^2}\right] I^{1,K}_{i,j} \\ \vdots & \ddots & \vdots \\ T\sum_{i<j}^{n}\left[\frac{\exp(\hat{\theta}^s_{\text{cl},K}+\hat{\theta}^s_{\text{cl},1})}{\{1+\exp(\hat{\theta}^s_{\text{cl},K}+\hat{\theta}^s_{\text{cl},1})\}^2}\right] I^{K,1}_{i,j} & \cdots & T\sum_{i<j}^{n}\left[\frac{4\exp(\hat{\theta}^s_{\text{cl},K}+\hat{\theta}^s_{\text{cl},K})}{\{1+\exp(\hat{\theta}^s_{\text{cl},K}+\hat{\theta}^s_{\text{cl},K})\}^2}\right] I^{K}_{i,j} \end{bmatrix},$$

where $I^k_{i,j} = \hat{z}_{ik}\hat{z}_{jk}$ and $I^{k,l}_{i,j} = \hat{z}_{ik}\hat{z}_{jl} + \hat{z}_{il}\hat{z}_{jk}$ for $k, l \in \{1, \ldots, K\}$. We now obtain the estimate of d_K as $\hat{d}_K = \text{tr}(\hat{H}_K^{-1}\hat{V}_K)$. Finally, for clustering time-evolving networks through TERGM with a stability parameter, we determine the optimal number of groups from

$$\hat{K} = \arg\min_K \widehat{\text{CL-BIC}}_K = \arg\min_K \; -2cl(\hat{\boldsymbol{\theta}}^s_{\text{cl}}, \hat{\boldsymbol{z}}) + \hat{d}_K(\hat{\boldsymbol{\theta}}^s_{\text{cl}}, \hat{\boldsymbol{z}}) \ln\{Tn(n-1)/2\},$$

where $\hat{\boldsymbol{\theta}}^s_{\text{cl}}$ and $\hat{\boldsymbol{z}}$ are the estimates of $\boldsymbol{\theta}^s$ and $\boldsymbol{z}$ corresponding to a given K. The details for STERGM with formation and persistence parameters in Example 2 are presented in the Appendix.

We also introduce an alternative model selection criterion based on modified integrated classification likelihood. Again for the TERGM with stability parameters, the modified ICL can be written as

$$\text{ICL}_K = \sum_{t=1}^{T}\sum_{i<j}^{n} \ln \Pr(Y_{t,ij} = y_{t,ij} \mid y_{t-1,ij}, \hat{\boldsymbol{z}}) - K \ln\{Tn(n-1)/2\},$$

where the second term penalizes the K stability parameters and we choose the optimal number of groups as

$$\hat{K} = \arg\max_K \text{ICL}_K.$$

Model selection results using both conditional likelihood BIC and modified ICL in the simulation studies are presented in Section 5.

5. Simulation studies

5.1. Simulation studies for mixture of TERGMs and STERGMs

Firstly simulation studies were conducted when the dataset was generated from a mixture of TERGMs and STERGMs. To simulate time-evolving networks from the K-component mixture of TERGMs with stability parameters, i.e., Example 1, we first specify network structure by choosing randomly the categories of the nodes according to the fixed mixing proportions and by defining initial densities for each group. We then obtain the initial network $\mathbf{y}_0$ by simulating all the edges based on the probabilities specified by the density parameters and the categories of the nodes. Next, we set different stability parameters for each group and simulate the edges in networks $\mathbf{y}_1, \ldots, \mathbf{y}_T$ sequentially, based on the probabilities determined by the parameters, the preceding networks, and the categories of the nodes. Similarly, we simulate time-evolving networks from the K-component mixture of STERGMs with formation and persistence parameters, i.e., Example 2. For each of the four model settings listed in Table 1, we use 100 nodes and 10 discrete time points.

To check the performance of the algorithm at identifying the correct number of groups, we count the frequencies of min CL-BIC and max ICL over 100 repetitions. To assess the clustering performance, we calculate the average value of the Rand Index (RI) and the average value of Normalized Mutual Information (NMI) over the 100 repetitions. The NMI can be used to compare the performance of clusterings with different numbers of clusters since it is a normalized metric. To assess the estimation performance of the algorithm, we calculate the average root squared error for estimated mixing proportions and network parameters over the 100 repetitions: $\text{RSE}_{\boldsymbol{\pi}} = (\sum_{k=1}^{K}(\hat{\pi}_k - \pi_k)^2)^{1/2}$ and $\text{RSE}_{\boldsymbol{\theta}} = (\sum_{k=1}^{K}(\hat{\theta}_k - \theta_k)^2)^{1/2}$.

Table 1
Models 1–2 are two different simulation settings for TERGM with stability parameters (Example 1) and Models 3–4 are two different simulation settings for STERGM with formation and persistence parameters (Example 2).

	Model 1		Model 2		
	G_1	G_2	G_1	G_2	G_3
Mixing proportion π_k	0.5	0.5	0.33	0.33	0.33
Stability parameter θ_k^s	−0.5	0.5	−1	0	1
Initial network density parameter θ_k^d	−0.5	0.5	−1	0	1
	Model 3		Model 4		
	G_1	G_2	G_1	G_2	G_3
Mixing proportion π_k	0.5	0.5	0.33	0.33	0.33
Formation parameter θ_k^f	−1.5	1.5	−1.5	0	1.5
Persistence parameter θ_k^p	−1	1	−1	0	1
Initial network density parameter θ_k^d	−0.5	0.5	−1	0	1

Table 2
Comparison of the model selection performance using the frequencies of min CL-BIC and max ICL over 100 repetitions. The details about Models 1–4 can be found in Table 1.

	Model 1 ($K_0 = 2$)				Model 2 ($K_0 = 3$)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
min CL-BIC	0	99	0	1	0	0	97	3
max ICL	0	100	0	0	0	0	96	4
	Model 3 ($K_0 = 2$)				Model 4 ($K_0 = 3$)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
min CL-BIC	0	99	1	0	0	3	93	4
max ICL	0	100	0	0	0	0	99	1

Table 3
Comparison of the clustering performance using the average Rand Index (RI) and Normalized Mutual Information (NMI) for 100 repetitions for various models and values of K, with sample standard deviations in parentheses. The details about Models 1–4 can be found in Table 1.

	Model 1 ($K_0 = 2$)		
	$K = 2$	$K = 3$	$K = 4$
RI	1.000 (0.000)	0.874 (0.033)	0.773 (0.030)
NMI	1.000 (0.000)	0.674 (0.043)	0.521 (0.025)
	Model 2 ($K_0 = 3$)		
	$K = 2$	$K = 3$	$K = 4$
RI	0.751 (0.044)	0.996 (0.031)	0.943 (0.025)
NMI	0.482 (0.075)	0.990 (0.064)	0.827 (0.045)
	Model 3 ($K_0 = 2$)		
	$K = 2$	$K = 3$	$K = 4$
RI	1.000 (0.000)	0.874 (0.033)	0.774 (0.031)
NMI	1.000 (0.000)	0.675 (0.046)	0.521 (0.026)
	Model 4 ($K_0 = 3$)		
	$K = 2$	$K = 3$	$K = 4$
RI	0.761 (0.045)	0.998 (0.021)	0.945 (0.018)
NMI	0.511 (0.084)	0.995 (0.049)	0.836 (0.033)

We check the performance of our criterion functions in choosing the correct number of groups, with K_0 denoting the true number of groups. As in Table 2, both CL-BIC and modified ICL perform well. The average values of Rand Index and Normalized Mutual Information are reported in Table 3.

Table 4 summarizes the estimation performance of our algorithm using RSE_{π}, RSE_{θ^s}, RSE_{θ^f}, and RSE_{θ^p}. The results of Tables 2–4 together tell us that our algorithm performs convincingly on this set of test datasets. We also found that our proposed algorithm performs convincingly in an unbalanced setting where the cluster proportions are unequal, and these results are provided in the supplementary material.

Table 4
Comparison of the estimation performance for mixing proportions and network parameters over 100 repetitions with standard deviations shown in parentheses. The details about Models 1–4 and the root squared error can be found in Table 1 and Section 5.1.

Model 1 ($K_0 = 2$)		Model 2 ($K_0 = 3$)	
RSE_π	RSE_{θ^s}	RSE_π	RSE_{θ^s}
0.056 (0.043)	0.013 (0.008)	0.073 (0.043)	0.038 (0.098)

Model 3 ($K_0 = 2$)			Model 4 ($K_0 = 3$)		
RSE_π	RSE_{θ^f}	RSE_{θ^p}	RSE_π	RSE_{θ^f}	RSE_{θ^p}
0.057 (0.043)	0.030 (0.020)	0.023 (0.015)	0.072 (0.040)	0.045 (0.094)	0.039 (0.071)

Table 5
Models 5–6 are two different simulation settings when the time-evolving networks are not simulated from the true model.

	Model 5		Model 6		
	G_1	G_2	G_1	G_2	G_3
Mixing proportion	0.4	0.6	0.3	0.4	0.3
Mean relational duration	5	2.5	7.5	5	2.5
Average network density	0.15	0.1	0.1	0.25	0.3

Table 6
Frequencies of min CL-BIC over 100 repetitions. The details about Models 5–6 can be found in Table 5.

	Model 5 ($K_0 = 2$)				Model 6 ($K_0 = 3$)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
TERGM	0	87	12	1	0	1	96	3
STERGM	0	93	2	5	0	8	91	1

Table 7
Frequencies of max ICL over 100 repetitions. The details about Models 5–6 can be found in Table 5.

	Model 5 ($K_0 = 2$)				Model 6 ($K_0 = 3$)			
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
TERGM	0	49	23	28	0	0	65	35
STERGM	0	53	32	15	0	0	59	41

5.2. Simulation studies for robustness of model selection

We conduct additional simulation studies to check which model selection criterion is more robust in choosing the correct number of groups when the time-evolving networks are not simulated from the true model. This time we use the 'simulate.stergm' function in the `tergm` package [27] in `R` [36] to simulate K time-evolving networks [26] and combine them into a larger time-evolving network in which each of the original time-evolving networks is the subnetwork containing nodes from one of the K groups. First, we specify each network structure by choosing randomly the categories of the nodes according to the fixed mixing proportions and by defining the network densities. Next, we set different mean relational durations, which represent different degrees of stability, and simulate each time-evolving network to have the average network density we defined over the time points. Finally, we combine the K time-evolving networks into a single time-evolving network by adding a fixed number of edges between randomly chosen pairs of individuals in different groups. For each of the two model settings listed in Table 5, we use 100 nodes, 10 discrete time points, and 10 edges added between randomly chosen pairs of nodes in different groups.

As shown in Table 6, in the new simulation setting where the time-evolving networks are not simulated from the true model, CL-BIC still performs well in choosing the correct number of groups using either TERGM with a stability parameter or STERGM with formation and persistence parameters. However, as shown in Table 7, modified ICL fails to choose the correct number of groups. Tables 6 and 7 show that the performance of our proposed CL-BIC in choosing the correct number of groups seems more robust than modified ICL when the model assumptions are violated.

The average RI and NMI results are reported in Tables 8 and 9. In all models, TERGM with a stability parameter and STERGM with formation and persistence parameters achieve a high average RI and NMI for the correct number of mixtures. Moreover, we see a fairly high average RI and NMI with the selected (via minimum CL-BIC) number of groups $\hat{K}$. The results of Tables 6, 8, and 9 together tell us that our algorithm based on CL-BIC can perform convincingly in choosing the correct number of groups and assigning nodes to groups even when the time-evolving networks are not generated from the true model.

Table 8
Comparison of the clustering performance using average Rand Index with standard deviations shown in parentheses. The details about Models 5–6 can be found in Table 5.

	Model 5 ($K_0 = 2$)			
	$K = 2$	$K = 3$	$K = 4$	$K = \hat{K}$
TERGM	0.976 (0.032)	0.797 (0.045)	0.716 (0.036)	0.948 (0.080)
STERGM	0.979 (0.025)	0.798 (0.054)	0.727 (0.041)	0.966 (0.056)
	Model 6 ($K_0 = 3$)			
	$K = 2$	$K = 3$	$K = 4$	$K = \hat{K}$
TERGM	0.753 (0.054)	0.976 (0.033)	0.935 (0.023)	0.975 (0.037)
STERGM	0.756 (0.055)	0.972 (0.036)	0.931 (0.024)	0.961 (0.052)

Table 9
Comparison of the clustering performance using average Normalized Mutual Information with standard deviations shown in parentheses. The details about Models 5–6 can be found in Table 5.

	Model 5 ($K_0 = 2$)			
	$K = 2$	$K = 3$	$K = 4$	$K = \hat{K}$
TERGM	0.922 (0.093)	0.535 (0.074)	0.403 (0.073)	0.863 (0.179)
STERGM	0.928 (0.076)	0.532 (0.084)	0.410 (0.071)	0.901 (0.131)
	Model 6 ($K_0 = 3$)			
	$K = 2$	$K = 3$	$K = 4$	$K = \hat{K}$
TERGM	0.508 (0.094)	0.936 (0.064)	0.795 (0.044)	0.933 (0.070)
STERGM	0.500 (0.092)	0.926 (0.069)	0.778 (0.050)	0.901 (0.111)

6. Applications to real-world time-evolving networks

Here, we apply our proposed model-based clustering methods to detect groups in two time-evolving network datasets: International trade networks of 58 countries from 1981 to 2000, and collaboration networks of 151 researchers at a large American research university from 2004 to 2013. In particular, we are interested in analyzing the rate of evolution of these time-evolving networks as in Knecht [24], Snijders et al. [38], and Krivitsky and Handcock [26]. Before proceeding, we introduce metrics to measure the instability of edges in the estimated groups $\mathrm{G}_1, \ldots, \mathrm{G}_{\hat{K}}$. For $k, l \in \{1, \ldots, \hat{K}\}$ and $t \in \{1, \ldots, T\}$, define

- the "$1 \to 0$" instability of edges between G_k and G_l at the time t:

$$\mathcal{S}^{kl}_{1\to 0}(t) = \frac{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} y_{t-1,ij}(1 - y_{t,ij})}{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} y_{t-1,ij} y_{t,ij}};$$

- the "$0 \to 1$" instability of edges between G_k and G_l at the time t:

$$\mathcal{S}^{kl}_{0\to 1}(t) = \frac{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} (1 - y_{t-1,ij}) y_{t,ij}}{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} (1 - y_{t-1,ij})(1 - y_{t,ij})};$$

- the total instability of edges between G_k and G_l at the time t:

$$\mathcal{S}^{kl}_{\mathrm{tot}}(t) = \frac{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} \left\{ y_{t-1,ij}(1 - y_{t,ij}) + (1 - y_{t-1,ij}) y_{t,ij} \right\}}{\sum_{i\in \mathrm{G}_k, j\in \mathrm{G}_l} \left\{ y_{t-1,ij} y_{t,ij} + (1 - y_{t-1,ij})(1 - y_{t,ij}) \right\}}.$$

The three instability statistics defined above evaluate the within-group instability when $k = l$ and the between-group instability when $k \neq l$. Next, we define $\mathcal{AS}^{kl}_{1\to 0}$, $\mathcal{AS}^{kl}_{0\to 1}$ and $\mathcal{AS}^{kl}_{tot}$ as the averages over all t of $\mathcal{S}^{kl}_{1\to 0}(t)$, $\mathcal{S}^{kl}_{0\to 1}(t)$, and $\mathcal{S}^{kl}_{tot}(t)$, respectively. Here, a larger value of $\mathcal{AS}^{kl}_{1\to 0}$ indicates that the network is more likely to dissolve ties, a larger value of $\mathcal{AS}^{kl}_{0\to 1}$ implies that the network is more likely to form ties, and a larger value of $\mathcal{AS}^{kl}_{tot}$ implies that the network is less stable overall.

6.1. International trade networks

We first consider finding groups for the yearly international trade networks of $n = 58$ countries studied by Ward and Hoff [44]. We follow Westveld and Hoff [45] and Saldana et al. [37] to define networks $\mathbf{y}_{1981}, \ldots, \mathbf{y}_{2000}$ as follows: for any $t \in \{1981, \ldots, 2000\}$, $y_{t,ij} = 1$ if the bilateral trade between country i and country j in year t exceeds the median bilateral trade in year t, and $y_{t,ij} = 0$ otherwise. By definition, this setup results in networks in which the edge density is roughly one half. Here, we employ model-based clustering through TERGM with a stability parameter, i.e., Example 1.

Table 10
Summary of network statistics, parameter estimates, and average memberships for each group. Groups G_1, G_2, and G_3 correspond to the medium stability, low stability, and high stability groups, respectively.

	G_1	G_2	G_3
Total # of nodes	24	21	13
Average # of edges per node	17.11	34.28	42.42
Average # of triangles per node	150.17	443.95	613.15
Estimated mixing proportion $\hat{\pi}$	0.3920	0.3677	0.2403
Estimated stability parameter $\hat{\theta}^s$	1.6323	1.3168	2.0712
Average membership of $\hat{\gamma}_{\cdot 1}$	0.8147	0.1158	0.0579
Average membership of $\hat{\gamma}_{\cdot 2}$	0.1060	0.8262	0.1101
Average membership of $\hat{\gamma}_{\cdot 3}$	0.0794	0.0579	0.8320

Table 11
Summary of within-group and between-group instability statistics for the proposed model-based clustering group assignments for the international trade network dataset. $\mathcal{AS}^{kl}_{1\to 0}$, $\mathcal{AS}^{kl}_{0\to 1}$, and $\mathcal{AS}^{kl}_{tot}$ measure the average over all t of $\mathcal{S}^{kl}_{1\to 0}(t)$, $\mathcal{S}^{kl}_{0\to 1}(t)$, and $\mathcal{S}^{kl}_{tot}(t)$, respectively, with standard deviations shown in parentheses.

$\mathcal{AS}^{11}_{1\to 0}$	$\mathcal{AS}^{11}_{0\to 1}$	$\mathcal{AS}^{11}_{\text{tot}}$	$\mathcal{AS}^{22}_{1\to 0}$	$\mathcal{AS}^{22}_{0\to 1}$	$\mathcal{AS}^{22}_{\text{tot}}$	$\mathcal{AS}^{33}_{1\to 0}$	$\mathcal{AS}^{33}_{0\to 1}$	$\mathcal{AS}^{33}_{\text{tot}}$
0.088	0.014	0.023	0.048	0.248	0.082	0.002	0.014	0.004
(0.094)	(0.009)	(0.013)	(0.041)	(0.132)	(0.047)	(0.006)	(0.034)	(0.006)
$\mathcal{AS}^{12}_{1\to 0}$	$\mathcal{AS}^{12}_{0\to 1}$	$\mathcal{AS}^{12}_{\text{tot}}$	$\mathcal{AS}^{13}_{1\to 0}$	$\mathcal{AS}^{13}_{0\to 1}$	$\mathcal{AS}^{13}_{\text{tot}}$	$\mathcal{AS}^{23}_{1\to 0}$	$\mathcal{AS}^{23}_{0\to 1}$	$\mathcal{AS}^{23}_{\text{tot}}$
0.098	0.042	0.058	0.041	0.048	0.043	0.011	0.049	0.016
(0.030)	(0.014)	(0.015)	(0.021)	(0.026)	(0.008)	(0.017)	(0.088)	(0.019)

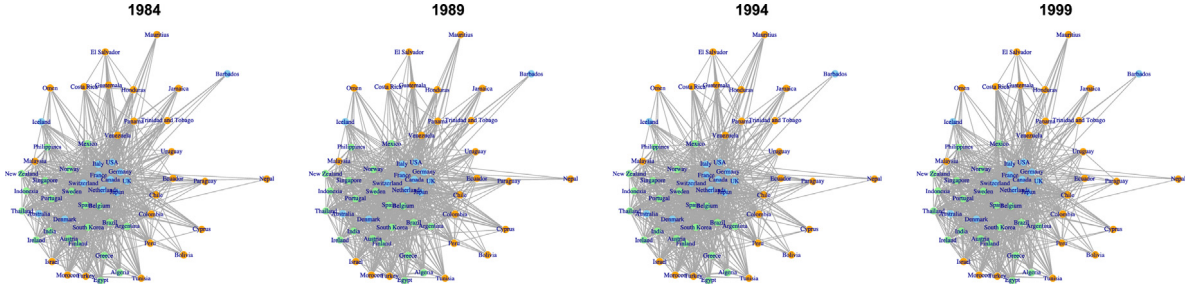


Fig. 1. International trade networks with estimated groups in four different years. Nodes assigned to G_1, G_2, and G_3 are colored orange, light green, and light blue, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

First, we use our proposed CL-BIC to determine the number of groups and $\hat{K} = 3$ is chosen. Hence, we shall identify three groups of countries based on different degrees of stability.

We summarize the characteristics of the three estimated groups, including some basic network statistics and parameter estimates, in Table 10. We also calculate the within-group and between-group instability measures. As shown in Table 11, group G_3 has the smallest total instability $\mathcal{AS}^{33}_{\text{tot}}$, which implies that those countries in G_3 consistently maintain their trading countries. In addition, the "$1 \to 0$" and "$0 \to 1$" instability measures show that countries in group G_2 change their trading countries more actively than countries in G_1 and G_3. To summarize, groups G_1, G_2, and G_3 correspond to the medium stability, low stability, and high stability groups, respectively.

In Fig. 1, we plot the international trade networks with estimated groups represented by orange for G_1, light green for G_2, and light blue for G_3 to illustrate our model-based clustering result. To illustrate how networks change over time for countries in each of the three groups, in Figs. 2–4 we isolate one representative country from each group: Israel from G_1, Thailand from G_2, and the United States from G_3.

6.2. Collaboration networks

We next find groups for the yearly collaboration networks at a large research university from 2004 to 2013. There are $n = 151$ researchers from various academic units in this dataset. We define networks $\mathbf{y}_{2004}, \ldots, \mathbf{y}_{2013}$ as follows: for any $t \in \{2004, \ldots, 2013\}$, $y_{t,ij} = 1$ if researcher i and researcher j have an active research grant together during year t, and $y_{t,ij} = 0$ otherwise. Here, we employ model-based clustering through STERGM with formation and persistence parameters, i.e., Example 2.

Again, we use our proposed CL-BIC and $\hat{K} = 2$ is chosen as an optimal number of groups. Therefore, we shall identify two groups of researchers based on different degrees of formation and persistence.

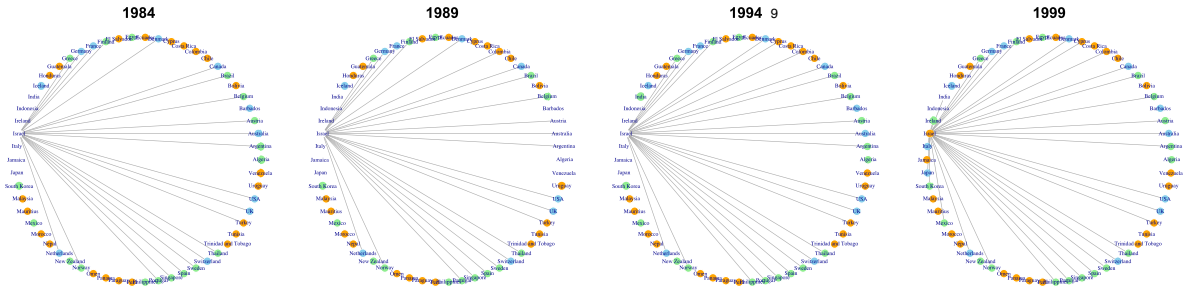


Fig. 2. Trade network of Israel, with estimated membership vector $\hat{\boldsymbol{y}} = (0.711, 0.072, 0.218)$. Although primarily classified as medium stability, Israel has stable trade with high-level GDP countries in G_3, which explains its 21.8% membership in G_3.

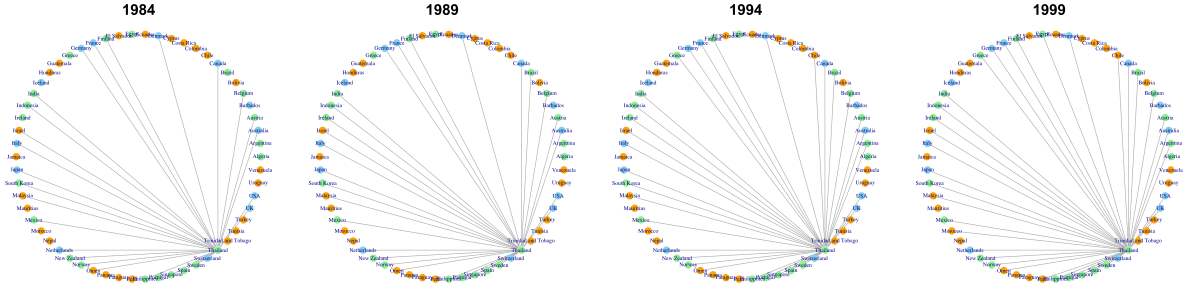


Fig. 3. Trade network of Thailand, with estimated membership vector $\hat{\boldsymbol{y}} = (0.049, 0.910, 0.041)$. Primarily classified in the low-stability group, Thailand's economic boom from 1987 to 1996 was triggered by its improved foreign trade and influx of foreign investment, which is consistent with its rapidly increasing number of trading partners.

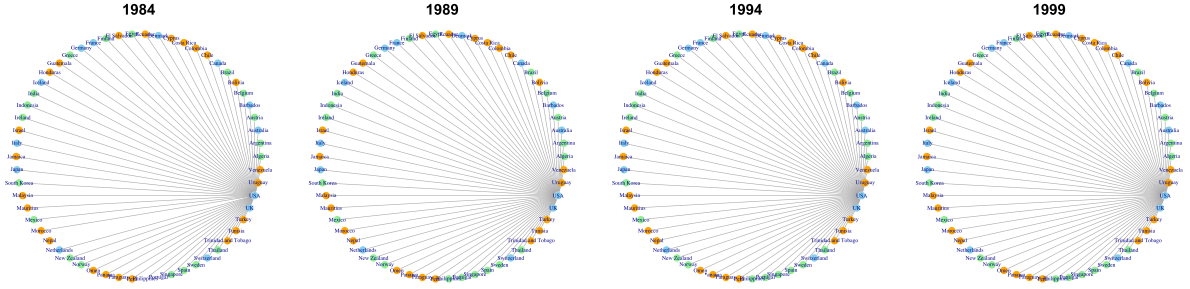


Fig. 4. Trade network of the United States, with estimated membership vector $\hat{\boldsymbol{y}} = (0.004, 0.000, 0.996)$. The United States is solidly in the high-stability group.

Table 12 summarizes their basic network statistics and parameter estimates, while Table 13 displays the within-group and between-group instability measures. As these tables show, G_1 has higher "$1 \to 0$", "$0 \to 1$", and total instability than G_2. Thus, the researchers in G_1 tend to have fewer stable collaborations and work with more collaborators than those in G_2.

Compared to TERGM with a stability parameter, STERGM with formation and persistence parameters provide more detailed insights about time-evolving networks. Based on the parameter estimates in Table 12, in each group, the stability is more explained by the persistence parameter than the formation parameter. In view of this fact, we further calculate the mean relational duration using the persistence parameter estimates for each estimated group. We obtain the mean relational durations of 2.18 years for G_1 and 2.85 years for G_2.

Fig. 5 presents the collaboration networks with estimated groups represented by orange for G_1 and light blue for G_2. We also plot the networks of several representative individual researchers, anonymized by the assignment of four-digit identification numbers, in Figs. 6–8.

Table 12
Summary of network statistics, parameter estimates, and average memberships for each group. Groups G_1 and G_2 correspond to the low stability and high stability groups.

	G_1	G_2
Total # of nodes	34	117
Average # of edges per node	2.92	1.90
Average # of triangles per node	0.8088	0.3957
Estimated mixing proportions $\hat{\pi}$	0.2464	0.7536
Estimated formation parameter $\hat{\theta}^f$	−2.2677	−2.9634
Estimated persistence parameter $\hat{\theta}^p$	0.1647	0.6156
Average membership of $\hat{\gamma}_{\cdot 1}$	0.7706	0.0941
Average membership of $\hat{\gamma}_{\cdot 2}$	0.2294	0.9059

Table 13
Summary of within-group and between-group instability statistics for the proposed model-based clustering group assignments for the collaboration network dataset. $\mathcal{AS}^{kl}_{1\to 0}$, $\mathcal{AS}^{kl}_{0\to 1}$, and $\mathcal{AS}^{kl}_{tot}$ measure the average over all t of $\mathcal{S}^{kl}_{1\to 0}(t)$, $\mathcal{S}^{kl}_{0\to 1}(t)$, and $\mathcal{S}^{kl}_{tot}(t)$, respectively, with standard deviations shown in parentheses.

$\mathcal{AS}^{11}_{1\to 0}$	$\mathcal{AS}^{11}_{0\to 1}$	$\mathcal{AS}^{11}_{\text{tot}}$	$\mathcal{AS}^{22}_{1\to 0}$	$\mathcal{AS}^{22}_{0\to 1}$	$\mathcal{AS}^{22}_{\text{tot}}$	$\mathcal{AS}^{12}_{1\to 0}$	$\mathcal{AS}^{12}_{0\to 1}$	$\mathcal{AS}^{12}_{\text{tot}}$
0.672	0.014	0.029	0.272	0.003	0.005	0.540	0.005	0.010
(0.359)	(0.006)	(0.007)	(0.059)	(0.001)	(0.001)	(0.189)	(0.001)	(0.002)

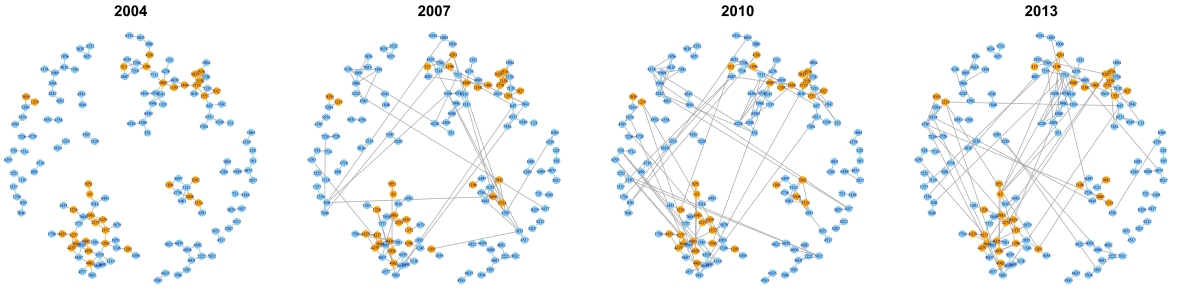


Fig. 5. Collaboration networks with estimated groups in four different years. Nodes assigned to G_1 and G_2 are colored orange and light blue, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

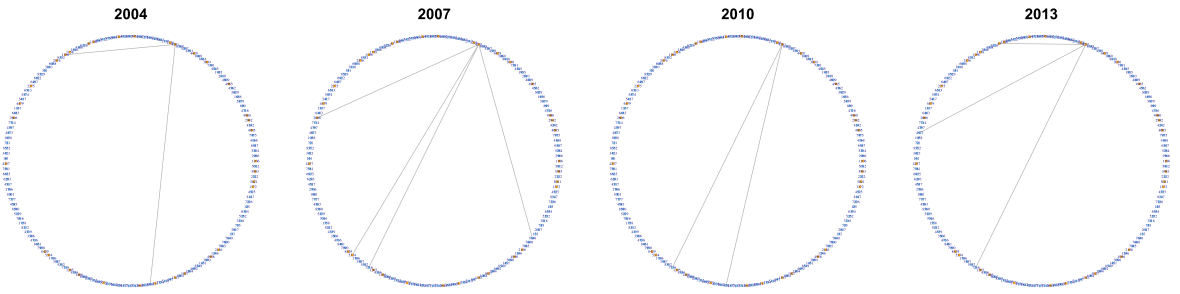


Fig. 6. Collaboration network of researcher #4201, with estimated membership vector $\hat{\boldsymbol{\gamma}} = (0.8506, 0.1494)$.

7. Conclusions

We propose a novel model-based clustering framework for time-evolving networks based on discrete time exponential-family random graph models, which simultaneously allows modeling and detecting group structure. Our model-based framework achieves different goals than a dynamic stochastic blockmodel framework. For example, the communities found via the latter framework are by assumption sets of nodes with higher edge probability within the sets than between the sets. By contrast, our model-based clustering framework allows for groups to be differentiated by other time-evolving features of interest.

Our model can be extended as follows. First, the network parameters can be modeled as time-varying functions (e.g., [30]) instead of constants. For example, Corneli et al. [10] propose a model for dynamic networks based on conditional non-homogeneous Poisson point processes where intensity functions are dependent on the node memberships and assumed to be stepwise constant, with common discontinuity points. Secondly, to effectively capture the time-evolving

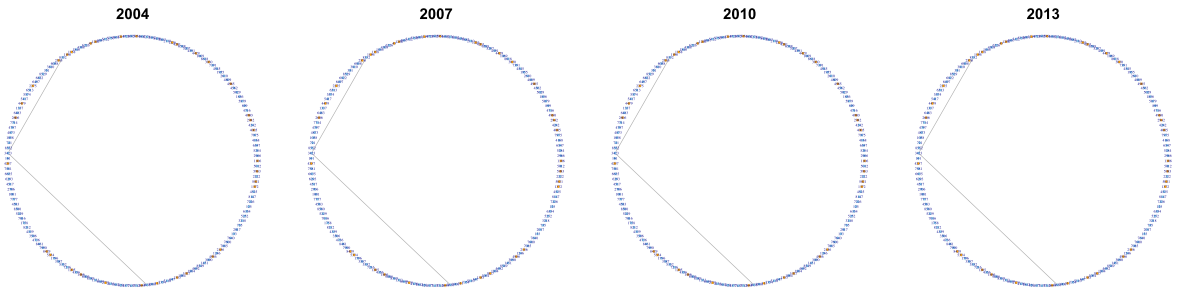


Fig. 7. Collaboration network of researcher #3423, with estimated membership vector $\hat{\boldsymbol{y}} = (0.0175, 0.9825)$.

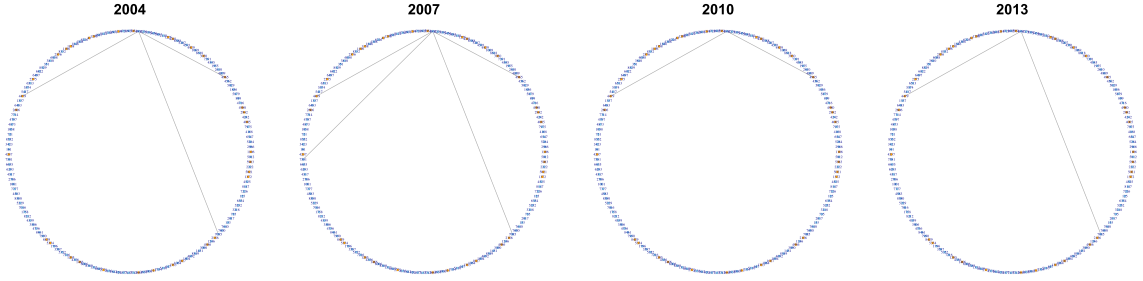


Fig. 8. Collaboration network of researcher #4607, with estimated membership vector $\hat{\boldsymbol{y}} = (0.4000, 0.6000)$.

group structure of discrete time exponential-family random graph models, we can incorporate a hidden Markov structure into our model.

Acknowledgments

We would like to thank the Editor, Associate Editor and referees for their helpful comments and suggestions. Lingzhou Xue's research is supported in part by the National Science Foundation DMS-1811552 and the National Institutes of Health NIDA-P50DA039838.

Appendix

We first present the complete proof of Theorem 1 and then include the details for the proposed conditional likelihood BIC for STERGM with formation and persistence parameters.

Proof of Theorem 1. The proof consists of three parts. First, we show that the mixing proportions π_k, $k \in \{1 \ldots, K\}$, and the conditional probability of observing an edge $\boldsymbol{p}_{kl} = \Pr(Y_{t,ij} = 1 \mid y_{t-1,ij}, Z_{ik} = Z_{jl} = 1)$, $k, l \in \{1, \ldots, K\}$, $k \leq l$, are generically identifiable. Next, we recover the network parameters by showing there exist unique closed-form expressions of $\boldsymbol{\theta}_k$, $k \in \{1, \ldots, K\}$ in terms of the conditional probability of observing an edge. We assume that there are no linear dependencies among the network statistics. Finally, we derive the condition on the number of different network parameters considered in the model to maintain the identifiability.

Allman et al. [4] show in Theorem 2 that it is possible to generically identify the mixing proportions and the conditional probability of observing an edge in the random graph mixture model with binary edge state variables. In our proposed model, the conditional probability of observing an edge is separated into two cases when $y_{t-1,ij} = 0$ and $y_{t-1,ij} = 1$, and we are able to obtain similar matrix A as in the proof of Theorem 2 in Allman et al. [4] containing the probabilities of observing the network time series, conditional on the node states. Then as in Allman et al. [3], Lemma 16, we partition the nodes to obtain three sequences of pairwise edge-disjoint subnetworks with corresponding matrix of probabilities and apply Kruskal's Theorem [29, Theorem 4a]. Finally, with an appropriate marginalization as in the proof of Theorem 2 in Allman et al. [4], we are able to recover the mixing proportions and the conditional probability of observing an edge.

Now to recover the network parameters, we show there exist unique closed-form expressions of $\boldsymbol{\theta}_k$, $k \in \{1, \ldots, K\}$ in terms of $\boldsymbol{p}_{kl}$, $k, l \in \{1, \ldots, K\}$, $k \leq l$ and these expressions can be obtained by exploiting the convex duality of exponential

families. Let

$$\psi^*(\boldsymbol{\mu}, y_{t-1,ij}) = \sup_{\boldsymbol{\theta}}\{\boldsymbol{\theta}^\top \boldsymbol{\mu} - \psi(\boldsymbol{\theta}, y_{t-1,ij})\}$$

be the Legendre–Fenchel transform of $\psi(\boldsymbol{\theta}, y_{t-1,ij})$, where the subscripts k and l have been dropped and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{g}(\mathbf{Y}_t, \mathbf{Y}_{t-1})\}$ is the mean value parameter vector whose entries are the expectations of the corresponding network statistics when $\boldsymbol{\theta}$ is the canonical parameter. Here, $\psi(\boldsymbol{\theta}, y_{t-1,ij})$ can be written as

$$\psi(\boldsymbol{\theta}, y_{t-1,ij}) = \sup_{\boldsymbol{\mu}}\{\boldsymbol{\theta}^\top \boldsymbol{\mu} - \psi^*(\boldsymbol{\mu}, y_{t-1,ij})\} = \sup_{\boldsymbol{p}}[\boldsymbol{\theta}^\top \boldsymbol{\mu}(\boldsymbol{p}) - \psi^*\{\boldsymbol{\mu}(\boldsymbol{p}), y_{t-1,ij}\}],$$

where $\boldsymbol{\mu}(\boldsymbol{p}) = \sum_{y^*\in\{0,1\}} \boldsymbol{g}(y^*, y_{t-1,ij}) \Pr(Y_{t,ij} = y^* \mid y_{t-1,ij}, \boldsymbol{z})$ and $\psi^*\{\boldsymbol{\mu}(\boldsymbol{p}), y_{t-1,ij}\} = \sum_{y^*\in\{0,1\}} \Pr(Y_{t,ij} = y^* \mid y_{t-1,ij}, \boldsymbol{z}) \ln \Pr(Y_{t,ij} = y^* \mid y_{t-1,ij}, \boldsymbol{z})$ since the Legendre–Fenchel transform of $\psi(\boldsymbol{\theta}, y_{t-1,ij})$ is self-inverse [43].

Hence, closed-form expressions of $\boldsymbol{\theta}$ in terms of $\boldsymbol{p}$ can be found by maximizing $\boldsymbol{\theta}^\top \boldsymbol{\mu}(\boldsymbol{p}) - \psi^*\{\boldsymbol{\mu}(\boldsymbol{p}), y_{t-1,ij}\}$ with respect to $\boldsymbol{p}$ and it is unique since the natural parameter space is convex for an exponential family and $\boldsymbol{\theta}^\top \boldsymbol{\mu}(\boldsymbol{p}) - \psi^*\{\boldsymbol{\mu}(\boldsymbol{p}), y_{t-1,ij}\}$ is strictly concave in $\boldsymbol{p}$.

To show $\boldsymbol{\theta}^\top \boldsymbol{\mu}(\boldsymbol{p}) - \psi^*\{\boldsymbol{\mu}(\boldsymbol{p}), y_{t-1,ij}\}$ is strictly concave, we show the Hessian is negative definite. Let c be the cardinality of $|\boldsymbol{p}|$, then the Hessian H_c is a $c \times c$ diagonal matrix with diagonal elements $(-1/p_{y^*})_{y^*\in\{0,1\}}$. Here, $-H_c$ is positive definite because the determinants of all upper-left sub-matrices are positive and therefore, H_c is negative definite.

Lastly, we further need conditions on the choice and number of different network parameters considered in the model. First, the network parameters should be chosen carefully in the model such that we avoid linear dependencies among the corresponding network statistics. Secondly, let p be the number of different network parameters considered in the model. For fixed K, we have Kp total network parameters and this value cannot be greater than $K + \{K(K-1)\}/2$, which is the number of identifiable conditional probabilities of observing an edge between two nodes given edge state of previous time point and their membership labels. Hence, the maximum number of different network parameters identifiable in the model is $\lfloor (K+1)/2 \rfloor$ where $\lfloor \cdot \rfloor$ is a floor function.

Conditional likelihood BIC for STERGM with formation and persistence parameters

Let $\hat{k}_i$ denote the component assigned to node i by the $\hat{\boldsymbol{z}}_i$ vector, i.e., the value of k such that $\hat{z}_{ik} = 1$. The conditional log-likelihood of STERGM with formation and persistence parameters can be written as

$$\mathrm{cl}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^p, \hat{\boldsymbol{z}}) = \sum_{t=1}^{T}\sum_{i<j}^{n} \left\{ -\ln \mathcal{C}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^p) + (y_{t,ij} - y_{t,ij}y_{t-1,ij})(\theta^f_{\hat{k}_i} + \theta^f_{\hat{k}_j}) + (y_{t,ij}y_{t-1,ij})(\theta^p_{\hat{k}_i} + \theta^p_{\hat{k}_j}) \right\},$$

where $\mathcal{C}(\boldsymbol{\theta}^f, \boldsymbol{\theta}^p) = \{1 + \exp(\theta^p_{\hat{k}_i} + \theta^p_{\hat{k}_j})\}^{y_{t-1,ij}} \{1 + \exp(\theta^f_{\hat{k}_i} + \theta^f_{\hat{k}_j})\}^{1-y_{t-1,ij}}$.

For any given K and the corresponding estimates $\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}$ and $\hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}$, we derive the explicit estimate of V_K as

$$\hat{V}_K(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}) = \sum_{t=1}^{T} \boldsymbol{u}(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}) \boldsymbol{u}(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}})^\top,$$

where $\boldsymbol{u}(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}) = \left(u(\hat{\theta}^f_{\mathrm{cl},1}), \ldots, u(\hat{\theta}^f_{\mathrm{cl},K}), u(\hat{\theta}^p_{\mathrm{cl},K+1}), \ldots, u(\hat{\theta}^p_{\mathrm{cl},2K})\right)^\top$ and for $k \in \{1, \ldots, K\}$,

$$u(\hat{\theta}^f_{\mathrm{cl},k}) = \sum_{i<j}^{n} \left\{ -\frac{(1 - y_{t-1,ij}) \exp(\hat{\theta}^f_{\mathrm{cl},\hat{z}_i} + \hat{\theta}^f_{\mathrm{cl},\hat{z}_j})}{1 + \exp(\hat{\theta}^f_{\mathrm{cl},\hat{z}_i} + \hat{\theta}^f_{\mathrm{cl},\hat{z}_j})} + y_{t,ij} - y_{t,ij}y_{t-1,ij} \right\} \left(\hat{z}_{ik} + \hat{z}_{jk}\right),$$

and for $l \in \{K+1, \ldots, 2K\}$,

$$u(\hat{\theta}^p_{\mathrm{cl},l}) = \sum_{i<j}^{n} \left\{ -\frac{y_{t-1,ij} \exp(\hat{\theta}^p_{\mathrm{cl},\hat{z}_i} + \hat{\theta}^p_{\mathrm{cl},\hat{z}_j})}{1 + \exp(\hat{\theta}^p_{\mathrm{cl},\hat{z}_i} + \hat{\theta}^p_{\mathrm{cl},\hat{z}_j})} + y_{t,ij}y_{t-1,ij} \right\} \left(\hat{z}_{il} + \hat{z}_{jl}\right).$$

We also derive the explicit estimate of H_K as

$$\hat{H}_K(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}) = \begin{bmatrix} \hat{H}_{11} & \ldots & \hat{H}_{1K} & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{H}_{K1} & \ldots & \hat{H}_{KK} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \hat{H}_{(K+1)(K+1)} & \ldots & \hat{H}_{(K+1)(2K)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \hat{H}_{(2K)(K+1)} & \ldots & \hat{H}_{(2K)(2K)} \end{bmatrix},$$

where

$$\hat{H}_{11} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{4(1-y_{t-1,ij})\exp(\hat{\theta}^f_{\mathrm{cl},1}+\hat{\theta}^f_{\mathrm{cl},1})}{\{1+\exp(\hat{\theta}^f_{\mathrm{cl},1}+\hat{\theta}^f_{\mathrm{cl},1})\}^2}\right] I^1_{i,j}, \quad \hat{H}_{1K} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{(1-y_{t-1,ij})\exp(\hat{\theta}^f_{\mathrm{cl},1}+\hat{\theta}^f_{\mathrm{cl},K})}{\{1+\exp(\hat{\theta}^f_{\mathrm{cl},1}+\hat{\theta}^f_{\mathrm{cl},K})\}^2}\right] I^{1,K}_{i,j},$$

$$\hat{H}_{K1} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{(1-y_{t-1,ij})\exp(\hat{\theta}^f_{\mathrm{cl},K}+\hat{\theta}^f_{\mathrm{cl},1})}{\{1+\exp(\hat{\theta}^f_{\mathrm{cl},K}+\hat{\theta}^f_{\mathrm{cl},1})\}^2}\right] I^{K,1}_{i,j}, \quad \hat{H}_{KK} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{4(1-y_{t-1,ij})\exp(\hat{\theta}^f_{\mathrm{cl},K}+\hat{\theta}^f_{\mathrm{cl},K})}{\{1+\exp(\hat{\theta}^f_{\mathrm{cl},K}+\hat{\theta}^f_{\mathrm{cl},K})\}^2}\right] I^{K}_{i,j},$$

$$\hat{H}_{(K+1)(K+1)} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{4y_{t-1,ij}\exp(\hat{\theta}^p_{\mathrm{cl},1}+\hat{\theta}^p_{\mathrm{cl},1})}{\{1+\exp(\hat{\theta}^p_{\mathrm{cl},1}+\hat{\theta}^p_{\mathrm{cl},1})\}^2}\right] I^1_{i,j}, \quad \hat{H}_{(K+1)(2K)} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{y_{t-1,ij}\exp(\hat{\theta}^p_{\mathrm{cl},1}+\hat{\theta}^p_{\mathrm{cl},K})}{\{1+\exp(\hat{\theta}^p_{\mathrm{cl},1}+\hat{\theta}^p_{\mathrm{cl},K})\}^2}\right] I^{1,K}_{i,j},$$

$$\hat{H}_{(2K)(K+1)} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{y_{t-1,ij}\exp(\hat{\theta}^p_{\mathrm{cl},K}+\hat{\theta}^p_{\mathrm{cl},1})}{\{1+\exp(\hat{\theta}^p_{\mathrm{cl},K}+\hat{\theta}^p_{\mathrm{cl},1})\}^2}\right] I^{K,1}_{i,j}, \quad \hat{H}_{(2K)(2K)} = \sum_{t=1}^{T}\sum_{i<j}^{n}\left[\frac{4y_{t-1,ij}\exp(\hat{\theta}^p_{\mathrm{cl},K}+\hat{\theta}^p_{\mathrm{cl},K})}{\{1+\exp(\hat{\theta}^p_{\mathrm{cl},K}+\hat{\theta}^p_{\mathrm{cl},K})\}^2}\right] I^{K}_{i,j}.$$

In the equations above, $I^k_{i,j} = \hat{z}_{ik}\hat{z}_{jk}$ and $I^{k,l}_{i,j} = \hat{z}_{ik}\hat{z}_{jl} + \hat{z}_{il}\hat{z}_{jk}$ for $k, l \in \{1, \ldots, K\}$.

We now obtain the estimate of d_K as $\hat{d}_K = \mathrm{tr}(\hat{H}_K^{-1}\hat{V}_K)$. Finally, for clustering time-evolving networks through STERGM with formation and persistence parameters, we determine the optimal number of groups from

$$\hat{K} = \arg\min_K \widehat{\text{CL-BIC}}_K = \arg\min_K \; -2\mathrm{cl}(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}, \hat{\boldsymbol{z}}) + \hat{d}_K(\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}, \hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}, \hat{\boldsymbol{z}})\ln\{Tn(n-1)/2\},$$

where $\hat{\boldsymbol{\theta}}^f_{\mathrm{cl}}$, $\hat{\boldsymbol{\theta}}^p_{\mathrm{cl}}$ and $\hat{\boldsymbol{z}}$ are the estimates of $\boldsymbol{\theta}^f$, $\boldsymbol{\theta}^p$ and $\boldsymbol{z}$ for a given K.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.104540.

References

[1] A. Agarwal, L. Xue, Model-based clustering of nonparametric weighted networks with application to water pollution analysis, Technometrics (2019) http://dx.doi.org/10.1080/00401706.2019.1623076.
[2] E.M. Airoldi, D.M. Blei, S.E. Fienberg, E.P. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008) 1981–2014.
[3] E.S. Allman, C. Matias, J.A. Rhodes, Identifiability of parameters in latent structure models with many observed variables, Ann. Statist. 37 (6A) (2009) 3099–3132.
[4] E.S. Allman, C. Matias, J.A. Rhodes, Parameter identifiability in a class of random graph mixture models, J. Statist. Plann. Inference 141 (5) (2011) 1719–1736.
[5] A.A. Amini, A. Chen, P.J. Bickel, E. Levina, Pseudo-likelihood methods for community detection in large sparse networks, Ann. Statist. 41 (4) (2013) 2097–2122.
[6] P.S. Bearman, J. Moody, K. Stovel, Chains of affection: The structure of adolescent romantic and sexual networks, Am. J. Sociol. 110 (1) (2004) 44–91.
[7] D. Bertsimas, 15.093J Optimization Methods, Massachusetts Institute of Technology: MIT OpenCourseWare. URL: https://ocw.mit.edu, (2009).
[8] P.J. Bickel, A. Chen, A nonparametric view of network models and Newman–Girvan and other modularities, Proc. Natl. Acad. Sci. 106 (50) (2009) 21068–21073.
[9] D.S. Choi, P.J. Wolfe, E.M. Airoldi, Stochastic blockmodels with a growing number of classes, Biometrika 99 (2) (2012) 273–284.
[10] M. Corneli, P. Latouche, F. Rossi, Multiple change points detection and clustering in dynamic networks, Stat. Comput. 28 (5) (2018) 989–1007.
[11] J.-J. Daudin, F. Picard, S. Robin, A mixture model for random graphs, Stat. Comput. 18 (2008) 173–283.
[12] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (458) (2002) 611–631.
[13] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.
[14] J.-D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, M. Vidal, Evidence for dynamically organized modularity in the yeast protein–protein interaction network, Nature 430 (6995) (2004) 88–93.
[15] M.S. Handcock, A.E. Raftery, J.M. Tantrum, Model-based clustering for social networks, J. R. Stat. Soc. A 170 (2) (2007) 301–354.
[16] S. Hanneke, W. Fu, E. Xing, Discrete temporal models of social networks, Electron. J. Stat. 4 (2010) 585–605.
[17] Q. Ho, L. Song, E. Xing, Evolving cluster mixed-membership blockmodel for time-varying networks, J. Mach. Learn. Res.: Workshop Conf. Proc. 15 (2011) 342–350.
[18] P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent space approaches to social network analysis, J. Amer. Statist. Assoc. 97 (460) (2002) 1090–1098.
[19] P.W. Holland, K.B. Laskey, S. Leinhardt, Stochastic blockmodels: First steps, Social Networks 5 (2) (1983) 109–137.
[20] D.R. Hunter, K. Lange, A tutorial on MM algorithms, Amer. Statist. 58 (1) (2004) 30–37.
[21] P. Ji, J. Jin., Coauthorship and citation networks for statisticians, Ann. Appl. Stat. 10 (4) (2016) 1779–1812.
[22] B. Karrer, M.E. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E 83 (1) (2011) 016107.
[23] B. Kim, K. Lee, L. Xue, X. Niu, A review of dynamic network models with latent variables, Stat. Surv. 12 (2018) 105–135.
[24] A.B. Knecht, Friendship Selection and Friends' Influence. Dynamics of Networks and Actor Attributes in Early Adolescence (Dissertation), Utrecht University, 2008.
[25] G. Kossinets, D.J. Watts, Empirical analysis of an evolving social network, Science 311 (5757) (2006) 88–90.
[26] P.N. Krivitsky, M.S. Handcock, A separable model for dynamic networks, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1) (2014) 29–46.
[27] P.N. Krivitsky, M.S. Handcock, `tergm`: Fit, Simulate and Diagnose Models for Network Evolution Based on Exponential-Family Random Graph Models, The Statnet Project (http://www.statnet.org), R package version 3.4.0, (2016).

[28] P.N. Krivitsky, M.S. Handcock, M. Morris, Adjusting for network size and composition effects in exponential-family random graph models, Stat. Methodol. 8 (4) (2011) 319–339.
[29] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra Appl. 18 (2) (1977) 95–138.
[30] K.H. Lee, L. Xue, Nonparametric finite mixture of Gaussian graphical models, Technometrics 60 (4) (2018) 511–521.
[31] C. Matias, V. Miele, Statistical clustering of temporal networks through a dynamic stochastic block model, J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (4) (2017) 1119–1141.
[32] M. Morris, M. Kretzschmar, Concurrent partnerships and transmission dynamics in networks, Social Networks 17 (3–4) (1995) 299–318.
[33] M.E. Newman, Coauthorship networks and patterns of scientific collaboration, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5200–5205.
[34] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[35] K. Nowicki, T.A. Snijders, Estimation and prediction for stochastic blockstructures, J. Amer. Statist. Assoc. 96 (455) (2001) 1077–1087.
[36] R Core Team, R Core Team R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, (2016).
[37] D.F. Saldana, Y. Yu, Y. Feng, How many communities are there? J. Comput. Graph. Statist. 26 (1) (2017) 171–181.
[38] T.A. Snijders, G.G. Van de Bunt, C.E. Steglich, Introduction to stochastic actor-based models for network dynamics, Social Networks 32 (1) (2010) 44–60.
[39] T.A. Snijders, K. Nowicki, Estimation and prediction for stochastic blockmodels for graphs with latent block structure, J. Classification 14 (1) (1997) 75–100.
[40] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, J.L. Wrana, Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nature Biotechnol. 27 (2) (2009) 199–204.
[41] C. Varin, P. Vidoni, A note on composite likelihood inference and model selection, Biometrika 92 (3) (2005) 519–528.
[42] D.Q. Vu, D.R. Hunter, M. Schweinberger, Model-based clustering of large networks, Ann. Appl. Stat. 7 (2) (2013) 1010–1039.
[43] M.J. Wainwright, M.I. Jordan, Graphical models, exponential families, and variational inference, Found. Trends Mach. Learn. 1 (1–2) (2008) 1–305.
[44] M.D. Ward, P.D. Hoff, Persistent patterns of international commerce, J. Peace Res. 44 (2) (2007) 157–175.
[45] A.H. Westveld, P.D. Hoff, A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict, Ann. Appl. Stat. 5 (2A) (2011) 843–872.
[46] E.P. Xing, W. Fu, L. Song, A state-space mixed membership blockmodel for dynamic network tomography, Ann. Appl. Stat. 4 (2) (2010) 535–566.
[47] K.S. Xu, A.O. Hero, Dynamic stochastic blockmodels for time-evolving social networks, IEEE J. Sel. Top. Sign. Proces. 8 (4) (2014) 552–562.
[48] L. Xue, H. Zou, T. Cai, Nonconcave penalized composite conditional likelihood estimation of sparse ising models, Ann. Statist. 40 (3) (2012) 1403–1429.
[49] T. Yang, Y. Chi, S. Zhu, Y. Gong, R. Jin, Detecting communities and their evolutions in dynamic social networks-a Bayesian approach, Mach. Learn. 82 (2) (2011) 157–189.