
Article:
Ren, S. orcid.org/0000-0001-9040-249X, Wang, X., Liu, P. orcid.org/0000-0002-0492-0029 et al. (1 more author) (2023) Bayesian nonparametric mixtures of Exponential Random Graph Models for ensembles of networks. Social Networks, 74. pp. 156-165. ISSN 0378-8733

https://doi.org/10.1016/j.socnet.2023.03.005

Social Networks 74 (2023) 156–165

Contents lists available at ScienceDirect

Social Networks

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

Bayesian nonparametric mixtures of Exponential Random Graph Models for ensembles of networks

Sa Ren [a,*], Xue Wang [b], Peng Liu [c], Jian Zhang [c]

[a] School of Health and Related Research, The University of Sheffield, Sheffield, United Kingdom
[b] Walsn Limited, Canterbury, United Kingdom
[c] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, United Kingdom

ARTICLE INFO

Keywords:
Network clustering
Dirichlet process
Markov Chain Monte Carlo
Importance sampling
Adjusted pseudo likelihood

ABSTRACT

Ensembles of networks arise in various fields where multiple independent networks are observed, for example, a collection of student networks from different classes. However, there are few models that describe both the variations and characteristics of networks in an ensemble at the same time. In this manuscript, we propose to model ensembles of networks using a Dirichlet Process Mixture of Exponential Random Graph Models (DPM-ERGMs), which divides an ensemble into different clusters and models each cluster of networks using a separate Exponential Random Graph Model (ERGM). By employing a Dirichlet process mixture, the number of clusters can be determined automatically and changed adaptively with the data provided. Moreover, in order to perform full Bayesian inference for DPM-ERGMs, we develop a Metropolis-within-slice sampling algorithm to address the problem of sampling from the intractable ERGMs on an infinite sample space. We also demonstrate the performance of DPM-ERGMs with both simulated and real datasets.

1. Introduction

Networks, as representations of relational data, are widely used in various scientific fields, such as sociology, neuroscience and biology. They provide valuable insight in understanding the diverse processes behind the complex dependent interactions among different objects. With the recent development of technology, ensembles of networks are increasingly available; they stand for multiple network observations obtained on the same or similar set of nodes across different subjects or time points. Examples of ensembles of networks include a collection of social networks from different schools (Sweet et al., 2019), and a population of brain networks from a number of participants (Simpson et al., 2013). There are high demands for developing the methodology to identify the characteristics common to or unique across individuals by taking advantage of the wealth of data presented in an ensemble.

The statistical modelling of ensembles of networks has also been motivated by the accessibility of abundant network data. Lubbers and Snijders (2007) employed a meta-analysis approach to analyse 102 student networks based on the estimation of each network. Slaughter and Koehly (2016) built a multilevel model using a hierarchical Bayesian approach to study the systematic network patterns within the population and the different structural patterns across networks. Paul and Chen (2020) developed a random effect stochastic block model, where the individual variations from the mean community structure of the population are considered in the model. Similarly, Arroyo et al. (2021) introduced a common subspace independent-edge multiple random graph model that includes both the common invariant submatrix for modelling the shared latent structures and an individual score matrix for describing the individual characteristics. MacDonald et al. (2022) developed an extension to the latent space models that is used to infer the underlying shared structure of multiplex networks. Sweet et al. (2013) also proposed a hierarchical latent space model for ensembles of networks.

Within an ensemble, some networks share common structures, while others exhibit distinct features. Group representation is a powerful tool to capture the similarities and differences of network structures in the same ensemble. Durante and Dunson (2018) introduced a Bayesian method to test the differences between two given groups of networks. Lehmann and White (2021) developed a multilevel network model to compare networks from different groups. In most cases, the underlying group structure is unknown and it is therefore necessary to develop a methodology that identifies the group membership and compares groups of networks simultaneously. Signorelli and Wit (2020) introduced a model-based clustering method based on mixtures of generalised linear models for populations of networks. Yin et al. (2022) proposed a finite mixture of exponential random graph models to model

* Corresponding author.
E-mail address: sarah.ren@sheffield.ac.uk (S. Ren).

https://doi.org/10.1016/j.socnet.2023.03.005

Available online 28 March 2023

the ensemble of networks based on the pseudo likelihood method. Durante et al. (2017) extended the latent space models using a Bayesian nonparametric approach.

In this manuscript, we propose the Dirichlet Process Mixtures of Exponential Random Graph Models (DPM-ERGMs) for ensembles of networks. The Dirichlet process mixture model uses the Dirichlet process as a prior over an infinite mixture model, where the number of mixtures can grow adaptively with the data. This enables the model to determine the group structure of the ensemble automatically, in other words, to compare different networks without prior knowledge of the number of clusters. Moreover, the Dirichlet process provides a large sample space and tractable posterior distributions, facilitating inference on the infinite sample space (Ferguson, 1973). On the other hand, the Exponential Random Graph Model (ERGM), a versatile network model, is employed to model networks for its ability to represent various types of topological features (Schweinberger et al., 2020). Thus, DPM-ERGMs are capable of determining the group structure and describing the group characteristics of an ensemble simultaneously.

Other extensions of ERGMs in combination with mixture models have also been proposed to explore the within-network structures. Salter-Townshend and Murphy (2015) developed a mixture of ERGMs for clustering nodes based on ego-networks. Schweinberger and Handcock (2015) characterised the local dependence in random graph models by taking account of the cluster membership of nodes. Henry et al. (2019) developed a finite mixture of ERGMs to model the unobserved heterogeneity in the effects of nodal covariates and network features. Moreover, ERGMs have also been extended to model the dependence between different layers of networks, such as multilevel ERGMs (Wang et al., 2013) or multilayer ERGMs (Caimo and Gollini, 2020; Krivitsky et al., 2020).

Under the Bayesian nonparametric framework, performing Bayesian inference of the proposed model involves the evaluation and comparison of an infinite number of ERGMs, the intractability of which increases the difficulty of the estimation dramatically. To sample from the infinite sample space, we introduce a latent variable to the model, which helps us to find a finite set of components required to produce the correct Markov chain, borrowing the idea from the slice sampling algorithm (Walker, 2007). Then the inference can be performed by sampling from the full conditional distributions of all variables on a finite space. However, for each network sample, the sampling of the membership variable requires model comparison between different group representations. This is challenging because the current method to estimate ERGM likelihood relies on approximating the intractable normalising constant ratio of two parameters that are close to each other. Parameters from different clusters can lie quite far apart because they represent networks of different characteristics, and parameters of empty groups can be very different from the rest.

One way to sample from the posterior distributions of ERGMs is to use Metropolis Hastings algorithms. Standard Metropolis Hastings algorithms are not applicable since the acceptance probability depends on the intractable normalising constants. To address this issue, Caimo and Friel (2011) applied the exchange algorithm (Murray et al., 2006), where a perfect sampler is employed to facilitate the Metropolis Hastings algorithm, avoiding the calculation of the intractable normalising constant. As the perfect sampler from the ERGM is unavailable in most cases, a sample from the MCMC method is used in practice. Liang and Jin (2013) developed a Monte Carlo Metropolis Hastings algorithm to sample from the intractable posterior distributions. The algorithm is implemented by approximating the unknown normalising constant ratio in the acceptance probability using a Monte Carlo estimate and is proved to converge to the desired target distribution. The exchange algorithm can be seen as a special case of the Monte Carlo Metropolis Hastings algorithm. However, most of the literature on ERGMs only deals with the single network situation. In DPM-ERGMs, networks from the same group are multiple samples from the same ERGM distribution. This requires the Bayesian inference to have the ability to incorporate multiple network samples.

To sample from the posterior distributions of DPM-ERGMs, we develop a Metropolis-within-slice sampling algorithm that employs Metropolis Hastings inside the slice sampling algorithm. Specifically, we propose three different estimation methods to deal with the intractability issue under the Metropolis-within-slice sampling framework. In the first method, we employ the importance sampling technique to estimate the true model. We express the posterior distributions of the membership variable in such a way that a ratio of normalising constants can be approximated with an intermediate importance sampling estimator. Moreover, we extend the Monte Carlo Metropolis Hastings algorithm to incorporate multiple networks from the same group. In the second method, we replace the true likelihood function with the pseudo likelihood function in Metropolis-within-slice scheme. In the third method, we replace the true likelihood function with the adjusted pseudo likelihood function. We will illustrate all three methods in detail later.

The rest of manuscript is organised as follows. In Section 2, we describe how the DPM-ERGMs are formulated. Section 3 provides the sampling methodology. Section 4 presents the simulation studies. We summarise the manuscript in Section 5.

2. Model formulation

2.1. Exponential random graph models

ERGMs describe the generating process of networks through exponential family distributions with summary statistics showing various connecting patterns as explanatory variables. A network with n nodes is typically represented by a random adjacency matrix $Y \in \{0,1\}^{n\times n}$, where $Y_{ij} = 1$ indicates an edge between nodes i and j, and $Y_{ij} = 0$ otherwise. The realisation of Y is denoted by y while the set of all possible outcomes of Y is denoted by $\mathcal{Y}$. The covariate information regarding the nodal or network attribute that affects the connections are denoted by $X \in \mathcal{X}$. The network structures of interest are expressed using a summary statistics vector, $S(y, X) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^d$. It represents the characteristics of the network, such as the number of edges, triangles, etc, which are crucial to the formation and dissolution of networks. The general ERGM has the following form,

$$P(Y = y \mid \theta, X) = \frac{\exp\{\theta^\top S(y, X)\}}{k(\theta)}, \tag{1}$$

where $\theta \in \mathbb{R}^d$ is the vector of model parameters, and $S(y, X)$ is the summary statistics (Morris et al., 2008). The normalising constant $k(\theta) = \sum_{y\in\mathcal{Y}} \exp\left\{\theta^\top S(y, X)\right\}$ is the sum over all potential graphs in the sample space, which is usually intractable except for very small networks. Given a realisation of network y, the aim of statistical inference is to find which value of θ provides best description for the data under ERGM framework. The intractability of the normalising constant is a strong barrier to the estimation of ERGMs as the likelihood function can only be specified up to a parameter dependent constant.

Bayesian inference is a natural choice for ERGMs since it allows uncertainty on model parameters. The posterior distribution of ERGMs is

$$f(\theta \mid y, X) = \frac{\pi(\theta)P(Y = y \mid \theta, X)}{P(Y = y \mid X)}, \tag{2}$$

where $\pi(\theta)$ is the prior, $P(Y = y|X) = \int_{R^d} \pi(\theta)P(Y = y|\theta, X)d\theta$. The standard MCMC algorithm is not suitable since the acceptance probability as shown in (3) to move from θ to the new proposal θ' requires evaluation of the intractable constants $k(\theta)$ and $k(\theta')$ at each step of the algorithm

$$\frac{\pi(\theta')h(\theta|\theta')}{\pi(\theta)h(\theta'|\theta)} \cdot \frac{\exp\{\theta'^\top S(y, X)\}}{\exp\{\theta^\top S(y, X)\}} \cdot \frac{k(\theta)}{k(\theta')}. \tag{3}$$

Here, $h(\cdot)$ stands for the proposal distribution. Monte Carlo Metropolis Hastings algorithm (Liang and Jin, 2013) samples from the posterior ERGMs by using an importance sampling estimator to approximate $k(\theta)/k(\theta')$ in the Metropolis Hastings algorithm.

2.2. Dirichlet process mixtures of ERGMs

Ensembles of networks include multiple network observations. In addition to the complex structures within each network, one may also be interested in studying the variations across different networks. Mixture models are a natural approach to describe such a population as they can detect and characterise the subpopulations that share common structures and represent networks that are different using separate distributions. In particular, the infinite mixture model can detect the cluster structures of the population without requiring a pre-specified number of clusters. Here, we propose to model the ensemble of networks through an infinite mixture of ERGMs, each component of which represents a cluster (subpopulation) of networks that share common structures using a cluster-specific ERGM.

An ensemble with N network samples is denoted by $\{Y_i\}_{i=1}^N$, and the corresponding covariate information is $\{X_i\}_{i=1}^N$. In such an ensemble, the single network Y_i is represented using an infinite mixture of ERGMs as follows

$$P_{w,\theta}(Y_i = y_i \mid X_i) = \sum_{j=1}^{\infty} w_j \frac{\exp\{\theta_j^\top S(y_i, X_i)\}}{k(\theta_j)}, \tag{4}$$

where j is the cluster label, w_j is the mixing proportion, θ_j is the cluster specified parameter vector, $S(y_i, X_i)$ is the summary statistics of network y_i, and $k(\theta_j) = \sum_{y\in\mathcal{Y}} \exp\left\{\theta_j^\top S(y, X)\right\}$ is the normalising constant. Without restrictions on the number of clusters, the infinite mixture model is able to provide a wide range of distributions for the data provided.

Alternatively, if we introduce a latent variable Z_i to indicate the membership of network y_i, e.g. $Z_i = k_i$ if y_i belongs to cluster k_i, (4) can also be written as

$$P_\theta(Y_i = y_i \mid X_i, Z_i = k_i) = \frac{\exp\{\theta_{k_i}^\top S(y_i, X_i)\}}{k(\theta_{k_i})}.$$

Therefore, the likelihood of the ensemble of networks can be expressed as

$$P_{w,\theta}(\{Y_i = y_i\}_{i=1}^N \mid \{X_i\}_{i=1}^N) = \prod_{i=1}^{N}\sum_{j=1}^{\infty} w_j \frac{\exp\{\theta_j^\top S(y_i, X_i)\}}{k(\theta_j)},$$

or

$$P_\theta(\{Y_i = y_i\}_{i=1}^N \mid \{X_i, Z_i = k_i\}_{i=1}^N) = \prod_{i=1}^{N} \frac{\exp\{\theta_{k_i}^\top S(y_i, X_i)\}}{k(\theta_{k_i})}.$$

It is informative to consider an infinite mixture model especially when it is not appropriate to have a limit on the number of groups. However, the inference of this model is challenging because the intractable normalising constant has to be evaluated in the infinite sample space.

To perform Bayesian inference on the proposed infinite mixture of ERGMs, we adopt a Dirichlet process prior DP(β, H) (Ferguson, 1973), which is arguably the most commonly used Bayesian nonparametric prior. Under the constructive definition, also known as the stick-breaking representation (Sethuraman, 1994), the mixing proportion w is constructed using a stick-breaking procedure with an auxiliary variable v. A sequence of independent and identically distributed auxiliary variables $v_1, v_2, \dots$ are sampled from a prior distribution Beta$(1, \beta)$, and the mixing proportions are set as $w_1 = v_1$, $w_j = v_j \prod_{l=1}^{j-1}(1 - v_l)$ (for $j > 1$). The membership indicator variable Z follows a multinomial distribution Mult(w) with probability $w = (w_1, w_2, \dots)$. For the prior of ERGM parameter θ_j, we use a multivariate Gaussian distribution $\mathcal{N}(\mu_0, \Sigma_0)$. Given the membership $Z_i = k_i$, the network Y_i is modelled by an ERGM with parameter θ_{k_i}. In the remaining of this manuscript, we will use Dirichlet Process Mixtures of Exponential Random Graph Models (DPM-ERGMs) with the following form,

$$\begin{aligned} &v_j \sim \text{Beta}(1, \beta) \\ &w_1 = v_1, w_j = v_j \prod_{l=1}^{j-1}(1 - v_l) \\ &Z_i | w \sim \text{Mult}(w) \\ &\theta_j | \mu_0, \Sigma_0 \sim \mathcal{N}(\mu_0, \Sigma_0) \\ &y_i | Z_i = k_i, \theta \sim P_{\theta_{k_i}}(Y_i = y_i \mid X_i). \end{aligned} \tag{5}$$

Here, $P_{\theta_{k_i}}(Y_i = y_i \mid X_i) = \exp\{\theta_{k_i}^\top S(y_i, X_i)\}/k(\theta_{k_i})$ is the ERGM with parameter θ_{k_i}.

3. Posterior computation

The statistical inference for the proposed model is very challenging due to the infinite number of mixture components and the intractable ERGM likelihood. In this section, we first develop a Metropolis-within-slice sampling algorithm to address the issue of sampling from the infinite sample space of DPM-ERGMs. Then, we provide details of the algorithms based on a true likelihood method, a pseudo likelihood method and an adjusted pseudo likelihood method separately.

The slice sampling algorithm (Walker, 2007; Kalli et al., 2011) provides a way to sample from the infinite mixture components. Similar to the slice sampling, we first introduce a latent variable u to our proposed model to identify the exact number of components that are required to produce a valid Markov chain with the correct stationary distributions. The joint density of (y, u) is written as

$$\begin{aligned} P_{w,\theta}(Y = y, u \mid X, \xi) &= \sum_{j=1}^{\infty} \frac{w_j}{\xi_j} U(u|0, \xi_j)\xi_j P_{\theta_j}(Y = y|X) \\ &= \sum_{j=1}^{\infty} \frac{w_j}{\xi_j} 1(u < \xi_j) P_{\theta_j}(Y = y|X) \end{aligned}$$

Then the inference can be performed by sampling from the clusters that satisfy $\{j : \xi_j > u\}$ instead of an infinite number of clusters, which simplifies the problem dramatically. Here, ξ is a deterministic decreasing sequence used to address the update of u. See Walker (2007) and Kalli et al. (2011) for more details about the introduction of the latent variable u and the choices of ξ.

Furthermore, with the indicator variable Z, the joint density can be expressed as

$$P_{w,\theta}(Y = y, u, Z = k \mid X, \xi) = \frac{w_k}{\xi_k} 1(u < \xi_k) P_{\theta_k}(Y = y|X).$$

Hence, the likelihood for the ensemble $\{Y_i\}_{i=1}^N$ with latent variable u and sequence ξ is

$$\begin{aligned} &l_{w,\theta}(\{Y_i = y_i, Z_i = k_i, u_i\}_{i=1}^N \mid \{X_i\}_{i=1}^N, \xi) \\ &= \prod_{i=1}^{N} \frac{w_{k_i}}{\xi_{k_i}} \mathbf{1}(u_i < \xi_{k_i}) P_{\theta_{k_i}}(Y_i = y_i|X_i). \end{aligned} \tag{6}$$

With the prior distribution specified in (5), the full conditional distributions of all variables (u, w, θ, Z) are available. Then we employ a Metropolis-within-slice sampling algorithm to sample (u, w, θ, Z) from their full conditional distributions in turn.

3.1. True likelihood based algorithm

In order to overcome the intractability issue and perform accurate estimation to the true model, we propose to employ the intermediate importance sampling technique in the Metropolis-within-slice sampling scheme. The sampling procedures of the true likelihood based Metropolis-within-slice sampling algorithm are listed as follows.

Step 1. Sample u_i from a uniform distribution,

$$u_i \sim U(0, \xi_{k_i}) \quad (i = 1, 2, \dots, N), \tag{7}$$

where k_i is the current allocation of network y_i.

Step 2. Sample v_j from a beta posterior distribution,

$$v_j \sim \text{Beta}(1 + a_j, \beta + b_j) \quad (j = 1, 2, \dots, K^*). \tag{8}$$

Here, $a_j = \sum_{i=1}^N \mathbf{1}(k_i = j)$ denotes the number of networks in group j and $b_j = \sum_{i=1}^N \mathbf{1}(k_i > j)$ corresponds to the number of networks in the

groups whose label are bigger than j. K^* denotes the current number of clusters.

Update w_j with

$$w_1 = v_1, w_j = v_j \prod_{l=1}^{j-1}(1 - v_l) \quad (j = 2, \dots, K^*). \tag{9}$$

Step 3. Sample Z_i with the following two steps,

(1) Introduce $k(\theta_c)$ to construct a computable normalising constant ratio and estimate the normalising constant ratio $k(\theta_c)/k(\theta_j)$ $(j = 1, \dots, K^*)$ using an intermediate importance sampling estimator γ_j.

(2) Calculate the conditional probability with the intermediate importance sampling estimator replacement,

$$P(Z_i = k_i | \cdots) \propto \mathbf{1}(\xi_{k_i} > u_i)\frac{w_{k_i}}{\xi_{k_i}} \cdot \exp\{\theta_{k_i}^\top S(y_i, X_i)\} \cdot \gamma_{k_i}, \quad (i = 1, 2, \dots, N). \tag{10}$$

Here, θ_{k_i} is the parameter of group k_i.

Step 4. Sample θ_j $(j = 1, 2, \dots, K^*)$ using the Metropolis–Hastings algorithm with the following procedures,

(1) Draw θ_j' from a proposal distribution $h(\cdot|\theta_j)$.

(2) Estimate the normalising constant ratio $k(\theta_j')/k(\theta_j)$ with an intermediate importance sampling estimator γ.

(3) Accept θ_j' with probability

$$\alpha = \min\left(1, \frac{\pi(\theta_j')h(\theta_j|\theta_j')}{\pi(\theta_j)h(\theta_j'|\theta_j)} \frac{\exp\{(\theta_j' - \theta_j)^\top \sum_{z_i=j} S(y_i, X_i)\}}{\gamma^{\sum_i \mathbf{1}(z_i=j)}}\right). \tag{11}$$

If there are no networks allocated to group j, update θ_j using prior $\pi(\theta_j)$.

Next, we will show the construction of formula (10) in Section 3.1.1 and explain how the Metropolis Hastings algorithm is developed in Section 3.1.2.

3.1.1. Sample Z

The full conditional distribution of Z_i is

$$P(Z_i = k_i \,|\, \cdots) \propto \mathbf{1}(\xi_{k_i} > u_i)\frac{w_{k_i}}{\xi_{k_i}} \cdot \frac{\exp\{\theta_{k_i}^\top S(y_i, X_i)\}}{k(\theta_{k_i})}. \tag{12}$$

The ratio on the right hand side depends on an intractable normalising constant $k(\theta_{k_i})$, which makes the direct sampling infeasible.

Gelman and Meng (1998) provides a way to estimate the normalising constant ratio using the importance sampling technique,

$$\frac{k(\theta_a)}{k(\theta_b)} \approx \frac{1}{m_2}\sum_{s=1}^{m_2} \exp\{(\theta_a - \theta_b)^\top S(z^s)\}, \tag{13}$$

with z^s $(s = 1, 2 \dots, m_2)$ denoting a sequence of m_2 independent auxiliary networks sampled from the ERGM with parameter θ_b. However, the importance sampling estimate will be incorrect if the compared parameters θ_a and θ_b are not close enough (Neal, 2005). This obstacle can be overcome by introducing intermediate distributions between θ_a and θ_b. Specifically, we interpolate m_1 intermediate values, θ_r^{im} $(r = 1, 2, \dots, m_1)$, so that θ_r^{im} and θ_{r+1}^{im} are close enough, and factorise the normalising constant ratio using intermediate values,

$$\frac{k(\theta_a)}{k(\theta_b)} = \prod_{r=0}^{m_1} \frac{k(\theta_{r+1}^{im})}{k(\theta_r^{im})} = \frac{k(\theta_1^{im})}{k(\theta_0^{im})}\frac{k(\theta_2^{im})}{k(\theta_1^{im})} \cdots \frac{k(\theta_{m_1+1}^{im})}{k(\theta_{m_1}^{im})}, \tag{14}$$

where $\theta_0^{im} = \theta_b$ and $\theta_{m_1+1}^{im} = \theta_a$. Then, each factor $k(\theta_{r+1}^{im})/k(\theta_r^{im})$ is estimated using the importance sampling estimator, and $k(\theta_a)/k(\theta_b)$ is approximated by

$$\gamma = \prod_{r=0}^{m_1} \frac{1}{m_2}\sum_{s=1}^{m_2} \exp\{(\theta_{r+1}^{im} - \theta_r^{im})^\top S(z_r^s)\}. \tag{15}$$

where z_r^s $(s = 1, 2, \dots, m_2)$ is a sequence of m_2 independent networks sampled from the ERGM with parameter θ_r^{im}.

If we can construct a normalising constant ratio in the posterior membership probability, we will be able to borrow the strength of intermediate importance sampling to allocate the network samples. To do so, we multiply a constant $k(\theta_c)$ to each term of the posterior probability vector and obtain

$$P(Z_i = k_i \,|\, \cdots) \propto \mathbf{1}(\xi_{k_i} > u_i)\frac{w_{k_i}}{\xi_{k_i}} \cdot \exp\{\theta_{k_i}^\top S(y_i, X_i)\}\frac{k(\theta_c)}{k(\theta_{k_i})},$$

where the constructed normalising constant ratios $k(\theta_c)/k(\theta_{k_i})$ $(k_i = 1, 2, \dots, K^*)$ can be approximated using the intermediate importance sampling estimation as shown in (15). Thus, the posterior probability ratios will not change and the sampling can be performed.

The choice of θ_c is important to the accuracy of the intermediate importance sampling estimation. The estimation will be incorrect if the parameters to be compared, θ_c and θ_j, are not close enough. As each group has a unique θ_j, it is impossible to find one θ_c close to all θ_j at the same time. Simple importance sampling is not applicable here and multiple intermediate values must be used to ensure the quality of estimation. Also, some θ_{k_i} can be quite difficult to sample from, especially when it is representing an empty group.

3.1.2. Sample θ

The posterior distribution of group parameter θ_j is proportional to the product of prior $\pi(\theta_j)$ and the joint likelihood of the networks in group j, which is

$$f(\theta_j | \cdots) \propto \pi(\theta_j) \prod_{Z_i=j} \frac{\exp\{\theta_j^\top S(y_i, X_i)\}}{k(\theta_j)}. \tag{16}$$

Sampling from such a posterior distribution is challenging as it depends on the product of multiple intractable likelihood functions.

The use of MCMC algorithm to sample from this posterior distribution of θ_j involves the calculation of $k(\theta_j')/k(\theta_j)$. As the product of the multiple normalising constant ratios has to be calculated, it is necessary to have a more accurate estimation for each $k(\theta_j')/k(\theta_j)$. To achieve this, we use the intermediate importance sampling estimator γ as in (15) to substitute $k(\theta_j')/k(\theta_j)$.

Therefore, to sample from (16) using Metropolis Hastings algorithm, we propose θ_j' from $h(\cdot|\theta_j)$, and accept θ_j' with probability

$$\frac{\pi(\theta_j')h(\theta_j|\theta_j')}{\pi(\theta_j)h(\theta_j'|\theta_j)} \frac{\exp\{(\theta_j' - \theta_j)^\top \sum_{z_i=j} S(y_i, X_i)\}}{\gamma^{\sum_i \mathbf{1}(z_i=j)}}. \tag{17}$$

With the approximation to the normalising constant ratio available, the acceptance ratio is calculable and thus the posterior sampling is feasible. Compared with importance sampling, the use of intermediate values increases the quality of estimation by introducing intermediate distributions. Similar techniques like annealed importance sampling and linked importance sampling (Neal, 2005) can be used as well.

3.2. Pseudo likelihood based algorithm

In addition to the true likelihood approach in Section 3.1, we also propose a fast estimation method based on the pseudo likelihood (Strauss and Ikeda, 1990), which is an approximation to the true likelihood. To be specific, the algorithm is developed by employing a pseudo likelihood approximation in the Metropolis-within-slice sampling algorithm. In the pseudo likelihood based algorithm, (u, w) are sampled in the same way as in the true likelihood based algorithm, and (θ, Z) are updated with pseudo likelihood replacement.

The pseudo likelihood method approximates the true likelihood using the product of conditional probabilities of all edges in a network,

$$PL(Y = y \mid X, \theta) = \prod_{r \neq s} P\left(y_{rs} = 1 \mid y_{-rs}, X, \theta\right)^{y_{rs}} \times \left\{1 - P\left(y_{rs} = 1 \mid y_{-rs}, X, \theta\right)\right\}^{1 - y_{rs}},$$

where $y_{-rs} = \{y_{kl}, (k,l) \neq (r,s)\}$ denotes all the dyads of the graph excluding y_{rs}. Here, y_{rs} follows a Bernoulli distribution with probability defined by change statistics, $\Delta S_{rs} = S(y_{rs} = 1, y_{-rs}, X) - S(y_{rs} = 0, y_{-rs}, X)$, which indicates the changes of y_{rs} on the summary statistics,

$$P\left(y_{rs} = 1 | y_{-rs}, X, \theta\right) = \frac{\exp(\theta^\top \Delta S_{rs})}{1 + \exp(\theta^\top \Delta S_{rs})}.$$

If we replace the true likelihood with pseudo likelihood, then the acceptance ratio for sampling θ_j using Metropolis Hastings algorithm is

$$\frac{\pi(\theta_j')h(\theta_j|\theta_j')}{\pi(\theta_j)h(\theta_j'|\theta_j)} \cdot \frac{\prod_{z_i=j} PL\left(Y_i = y_i | X_i, \theta_j'\right)}{\prod_{z_i=j} PL\left(Y_i = y_i | X_i, \theta_j\right)}, \tag{18}$$

and the posterior probability of cluster membership Z_i is proportional to

$$\mathbf{1}(\xi_j > u_i)\frac{w_j}{\xi_j} \cdot PL\left(Y_i = y_i | X_i, \theta_j\right). \tag{19}$$

Thus, the sampling of θ, Z is possible with the pseudo likelihood replacement.

Pseudo likelihood based algorithm is faster than true likelihood based algorithm, but it is less accurate. The major issue is that it may underestimate the endogenous network formation process, since pseudo likelihood only uses local information within a whole graph (van Duijn et al., 2009). Moreover, when the model is near-degenerate, posterior samples from pseudo likelihood method may fall into the degenerate region (Caimo and Friel, 2011).

3.3. Adjusted pseudo likelihood based algorithm

The adjusted pseudo likelihood (Bouranis et al., 2017, 2018) is proposed as an improvement to the pseudo likelihood. The adjusted pseudo likelihood for a single network Y is

$$APL(Y = y|X, \theta) = C \cdot PL(Y = y|X, \psi(\theta)), \tag{20}$$

where $C > 0$ is the magnitude adjustment constant. ψ is a model-specific invertible and differentiable mapping that adjusts the mode and curvature of the pseudo likelihood function, defined as

$$\psi(\theta) = \hat{\theta}_{MPLE} + W \cdot (\theta - \hat{\theta}_{MLE}). \tag{21}$$

Here, W is a transformation matrix, $\hat{\theta}_{MPLE}$ is the maximum pseudo likelihood estimate, and $\hat{\theta}_{MLE}$ is the maximum likelihood estimate. The transformation matrix W aims to match the gradient and the Hessian of the log-pseudo likelihood and the log-likelihood. The details about the estimation of magnitude adjustment constant C and the transformation matrix W can be found in Bouranis et al. (2017, 2018).

Similarly, we use the adjusted pseudo likelihood as the replacement to the true likelihood and get the adjusted pseudo likelihood based Metropolis-within-slice sampling algorithm. Variables (u, w) are sampled in the same way as in the true likelihood based Metropolis-within-slice sampling algorithm, described in Section 3.1. θ_j is sampled using Metropolis Hastings algorithm with the acceptance ratio

$$\frac{\pi(\theta_j')h(\theta_j|\theta_j')}{\pi(\theta_j)h(\theta_j'|\theta_j)} \cdot \frac{\prod_{z_i=j} APL\left(Y_i = y_i | X_i, \theta_j'\right)}{\prod_{z_i=j} APL\left(Y_i = y_i | X_i, \theta_j\right)}, \tag{22}$$

and the posterior probability of cluster membership Z_i is proportional to

$$\mathbf{1}(\xi_j > u_i)\frac{w_j}{\xi_j} \cdot APL\left(Y_i = y_i | X_i, \theta_j\right). \tag{23}$$

The label switching problem is a common issue in the Bayesian analysis of mixture models, where the posterior distributions remain invariant to the permutation of clusters. For the proposed Metropolis-within-slice sampling algorithm, we handle the label switching problem by post-processing the output from the algorithm. Specifically, we first obtain the new labels by performing K-centroids cluster analysis (Malsiner-Walli et al., 2016; Leisch, 2006) on the cluster parameters θ of the non-empty groups after burn in and thinning. Then we relabel the cluster components using the new labels. More details can be found in Yin et al. (2022).

4. Empirical results

In this section, we show the performance of the proposed methods using simulation studies. We first apply the DPM-ERGMs on synthetic network samples so that we can compare the model results with the true parameter values. Then we demonstrate how to find the underlying cluster structure of the ensemble of real networks using Krackhardt's advice networks as an example. The synthetic network samples are generated using R package ergm (Hunter et al., 2008) and the R code for the simulation studies is available at GitHub.[1]

4.1. Synthetic networks

We use synthetic networks to compare the clustering accuracy of different methods. Firstly, we choose three most commonly used network sufficient statistics for the ERGM distribution,

- $S^1(y_i) = \sum_{r \neq s} y_{rs,i}$, the total number of edges in the network.
- $S^2(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} EP_k(y_i)$, $\phi = 0.25$, geometrically weighted edgewise shared partner, GWESP, a representation for transitivity. $EP_k(y_i)$ is the number of connected pairs that have k common neighbours.
- $S^3(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r = X_s)$, the total number of connections between individuals with the same covariate.

X is a binary covariate with half of nodes taking value 0 and the other half taking 1. Then we simulate networks from mixtures of ERGM distributions under four different scenarios. In the first scenario, we consider an ensemble with $N = 30$ networks and $K = 3$ balanced groups. The number of nodes in each network is $n = 40$. In the second scenario, we keep the same ensemble size but increase the network size to $n = 100$. In the third scenario, we consider a larger ensemble with $N = 80$ networks and $K = 4$ groups. The number of networks in each group is 25, 25, 25 and 5. The network size of scenario 3 is $n = 40$. Similarly, the scenario 4 has $N = 80$ networks and $K = 4$ groups with a larger network size $n = 100$. The parameter values θ for each group under different scenarios are specified as follows,

$$\theta^{(40,30)} = \begin{pmatrix} -0.85 & -0.10 & -0.10 \\ -3.45 & 0.75 & 2 \\ -5.10 & 2.5 & 0.5 \end{pmatrix}, \quad \theta^{(100,30)} = \begin{pmatrix} -2.03 & -0.10 & -0.10 \\ -4.15 & 0.75 & 2 \\ -5.85 & 2.5 & 0.5 \end{pmatrix},$$

$$\theta^{(40,80)} = \begin{pmatrix} -0.85 & -0.10 & -0.10 \\ -3.45 & 0.75 & 2 \\ -5.10 & 2.5 & 0.5 \\ -2.00 & 0.20 & 1.0 \end{pmatrix}, \quad \theta^{(100,80)} = \begin{pmatrix} -2.03 & -0.10 & -0.10 \\ -4.15 & 0.75 & 2 \\ -5.85 & 2.5 & 0.5 \\ -3.00 & 0.20 & 1.0 \end{pmatrix}.$$

Next, we apply the proposed infinite mixture models to the synthetic ensembles. We run the simulations for 10,000 iterations starting with all networks in one group using the true likelihood based methods. The prior distribution of variable v is a beta distribution Beta$(1, 0.1)$, and the prior of ERGM parameters θ is selected to be a multivariate normal distribution $\mathcal{N}(\mu_0, \Sigma_0)$ with $\mu_0 = (-3, 0, 0)$, $\Sigma_0 = 4^2 I_p$, where I_p is a p dimension diagonal matrix, with p denoting the number of sufficient statistics. The proposal distribution in the Metropolis–Hastings algorithm is $\mathcal{N}(0, \Sigma_p)$, $\Sigma_p = 0.05^2 I_p$. For sequence $\xi_1, \xi_2, \ldots$, we use an exponential decreasing sequence, $\xi_i = e^{-i}$. The number of components that satisfies $\{j : \xi_j > u_i\}$, K_i, is also the smallest integer that satisfies $\{e^{-K_i} > u_i\}$, thus $K_i = \lfloor -log(u_i) \rfloor$. Also, we choose

[1] https://github.com/SRenStats/DPM-ERGMs.

Table 1
Estimation accuracy of K across 50 replicates for various experimental conditions with five different estimation methods.

(n, N)	K	Accuracy of $\hat{K}$					Average $\hat{K}$				
		IF-TL	IF-APL	IF-PL	F-TL	F-PL	IF-TL	IF-APL	IF-PL	F-TL	F-PL
(40, 30)	3	1	0.96	0.56	0.92	0.70	3	3.04	3.40	2.92	3.32
(100, 30)	3	1	1	0.54	1	0.84	3	3	3.52	3	3.14
(40, 80)	4	1	1	0.26	0.26	0.42	4	4	5.26	3.2	4.06
(100, 80)	4	1	1	0.46	0.86	0.56	4	4	4.60	3.86	3.76

Table 2
Average ARI and RI across 50 replicates for various experimental conditions with five different estimation methods.

(n, N)	K	Average ARI					Average RI				
		IF-TL	IF-APL	IF-PL	F-TL	F-PL	IF-TL	IF-APL	IF-PL	F-TL	F-PL
(40, 30)	3	1	0.992	0.926	0.962	0.948	1	0.997	0.968	0.981	0.979
(100, 30)	3	1	1	0.944	1	0.974	1	1	0.977	1	0.988
(40, 80)	4	0.992	0.990	0.841	0.903	0.886	0.997	0.996	0.939	0.957	0.954
(100, 80)	4	1	0.999	0.928	0.987	0.931	1	0.999	0.972	0.994	0.970

$m_1 = 2, m_2 = 10$ in the Metropolis Hastings step and $m_1 = 5, m_2 = 10$ for the sampling of membership variable. Details on the choices of m_1, m_2 can be found in the supplementary materials.

For comparison, we apply the finite mixture model of Yin et al. (2022) to the same synthetic ensembles with pseudo likelihood approximation as well as the true likelihood estimation method we developed. For each scenario, we repeat the same process for 50 times. For simplicity, we use IF-TL to stand for the infinite mixture model with the true likelihood method, IF-APL for the infinite mixture model with the adjusted pseudo likelihood method, IF-PL for the infinite mixture model with the pseudo likelihood method, F-TL for the finite mixture model with the true likelihood method and F-PL for the finite mixture model with the pseudo likelihood method. For the pseudo likelihood based methods, we run the simulation for 100,000 iterations and discard the first 60% samples as burn in and set the thinning parameter as 50, to be consistent with (Yin et al., 2022).

After running simulations, we present the accuracy of the estimated number of clusters of each method in Table 1. From Table 1, we can see that the two infinite methods, IF-TL and IF-APL, perform very well with respect to the estimation accuracy for the number of clusters K, with an average accuracy of almost 100%. F-TL performs better than F-PL on three scenarios, other than the third scenario, where F-TL gives a lower estimate on K. The estimation accuracy of IF-PL is low. This is because the variances of the pseudo likelihood estimates are often underestimated (Bouranis et al., 2018), leading to a high and narrow posterior distributions. Compared to the true posterior distribution, one pseudo posterior distribution can only represent a smaller number of samples. Thus, in a mixture model, more distributions are required to represent the whole population, and the estimates of average $\hat{K}$ are increased. As it is also shown in Table 1, the average $\hat{K}$ of IF-PL and F-PL is higher than the true K. This issue is worse for the infinite mixture model than the mixture model. Regarding the convergence rate, the true likelihood based methods converge faster than the pseudo likelihood based methods.

Furthermore, we also evaluate the accuracy of the cluster memberships using Rand index (RI) (Rand, 1971) and adjusted Rand index (ARI). The RI takes values between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 stands for perfect match. The ARI is the adjusted-for-chance version of the RI. Random labellings have an ARI close to 0 and 1 indicates that the data clusterings are exactly the same. We compare the clustering results of each iteration with the true cluster membership and calculate the average ARI and RI across 50 replicates. The average ARI and average RI are shown in Table 2.

From Table 2, we can see that all the methods perform well in regard to the accuracy of the cluster memberships. IF-TL, IF-APL and F-TL have higher values of ARI and RI, compared to the two pseudo likelihood based methods, IF-PL and F-PL. The average ARI for scenario 3 with FL-TL is 0.903, higher than F-PL, 0.886, regardless of the lower accuracy of $\hat{K}$. Although the IF-PL provides a higher estimation for K, it only divides the original group into more than one groups. As it does not mix networks from different groups, ARI values from IF-PL are still satisfactory despite the high estimates on $\hat{K}$.

Table 3
Overall computation time (in hours) of each method under different experimental conditions.

(n, N)	K	IF-TL	IF-APL	IF-PL	F-TL	F-PL
(40, 30)	3	6.13	0.72	0.68	6.61	0.61
(100, 30)	3	9.84	1.36	1.21	6.60	1.36
(40, 80)	4	7.74	1.55	1.59	6.11	1.66
(100, 80)	4	9.61	3.71	3.32	9.26	3.97

We also display the overall computation time of each method in Table 3. The computing is performed at a single core (Intel Core i5-11500 @ 2.70 GHz). Compared to the pseudo likelihood based methods, the true likelihood based methods, IF-TL and F-TL, take the longest time. They aim to get the exact estimation to the model and thus require estimating the intractable normalising constants at each iteration of the algorithms. This significantly increases the computation time. The time differences between IF-TL and F-TL are because of the differences on the number of non-empty clusters at each iteration. IF-TL has a flexible number of clusters at each iteration while F-TL uses a over-clustering method with a pre-specified number of clusters. If the pre-specified number is higher than the average cluster number in the infinite mixture model, then the finite mixture model takes longer time. The computation time of the pseudo likelihood based methods are similar. The adjusted pseudo likelihood methods need extra time to estimate the adjusted pseudo likelihood function for each network, while the pseudo likelihood methods have to compare a higher number of clusters during iteration.

In this simulation study, we compare the performance of different methods for different network ensembles. We find that the infinite mixture model with the true likelihood function, IF-TL, is most accurate, followed by the infinite mixture model with the adjusted pseudo likelihood function, IF-APL. With respect to the computation time, the true likelihood based methods require more time than the pseudo likelihood based methods for both infinite and finite mixture models.

4.2. Krackhardt's advice networks

We next apply the proposed DPM-ERGMs to an advice network ensemble. David Krackhardt (Krackhardt, 1987) studied a sequence of 21 networks about 21 employees in a high-tech machine manufacturing firm. The networks are constructed based on the data collected from a survey on the query "Who does X go to for advice and help with work?"

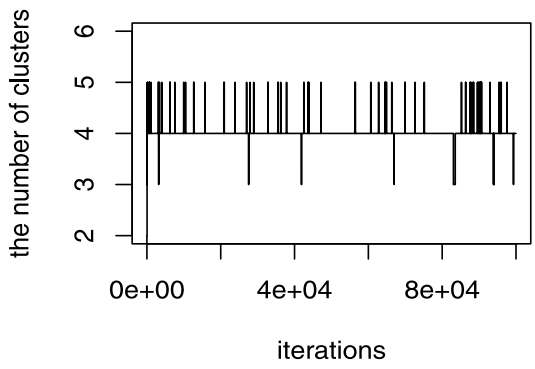

Fig. 1. The number of non-empty groups at each iteration.

Everyone is asked not only the advice relationship of themselves but also other people. Therefore, a collection of 21 perception networks $y_i (i = 1, 2, \ldots, 21)$ is built where each network represents an individual's perspective about the advice relationships among the 21 individuals. $y_{rs,i} = 1$ indicates that in the opinion of individual i, r asks help from s. The covariate information of each individual is represented by a vector X. The original paper focuses on exploring the differences of perception networks through node centrality scores to measure the importance of the nodes. Here, we are interested in learning the differences and similarities of the perception networks using the mixture of ERGMs. In this way, the generating mechanism of the perception networks can be analysed. This helps us to better understand the perception network relationships. For the structure statistics, we choose the following,

- $S^1(y_i) = \sum_{r \neq s} y_{rs,i}$, the total number of edges in the network. This reflects on the communication strength.
- $S^2(y_i) = \sum_{r \neq s} y_{rs,i} \mathbf{1}(X_r = X_s)$, the total number of connections between individuals in the same level. The positive coefficient indicates that people tend to ask for help from people of the same level, while the negative coefficient means that more help is sought from others in a different level.
- $S^3(y_i) = e^{\phi} \sum_{k=1}^{n-2} \{1 - (1 - e^{-\phi})^k\} DP_k(y_i)$, $\phi = 0.25$, geometrically weighted dyad-wise shared partner, GWDSP, a good representation for local clustering property, where $DP_k(y_i)$ represents the number of dyads with k shared partners in the network y_i.

We first estimate the model using the infinite true likelihood method, IF-TL. The hyperparameter are specified as follows. A multivariate Gaussian distribution with mean $\mu_0 = (-3, 0, 0)$ and covariance $\Sigma_0 = 4^2 I_3$ is chosen as the prior distribution for ERGM parameters. The proposal variance in the Metropolis Hastings algorithm is set as $\Sigma_q = 0.05^2 I_3$. A beta prior Beta$(1, 0.1)$ is used for the mixing proportion. $\theta_0 = (-2, 0, 0)$ is the initial value for ERGM parameter. In the intermediate importance sampling procedure, we use $m_1 = 2$ intermediate distributions and $m_2 = 10$ auxiliary networks for Metropolis Hastings algorithm and $m_1 = 5, m_2 = 10$ in the allocation step.

The number of clusters at each iteration from the true likelihood based method IF-TL are shown at Fig. 1. We can see that 4 groups are clustered with networks 15, 20 in the first group, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 14, 18, 19, 21 in the second group, 6, 13, 16, 17 in the third group, and network 1 in the fourth group. The acceptance probability in the Metropolis Hastings algorithm for 4 groups are 0.43, 0.16, 0.49, 0.38 respectively. To learn about the characteristics of each group, we display the posterior density plots from the true likelihood based method in Fig. 2. Group 1 has the smallest coefficient for edges but the biggest for GWDSP. This means that networks 15 and 20 have strong local clustering property, which is consistent with the fact that networks 15 and 20 have hub structures where fewer nodes have most of the connections. The advice relationships they nominate are centred around themselves. Group 2 has a big coefficient for edges and negative coefficient for level effect, indicating that networks are dense in this group and there are more advice between employees of different levels than of same levels. Group 3 has the smallest negative level effect, meaning that the advice relationships they observed are most across employees of different levels. Network 1 individually forms group 4. The level effect of network 1 is around 0, suggesting that individual level does not play a big role in network 1.

Our results are supported by the findings of Krackhardt (1987). Next, we compare our results with the centrality calculated in Krackhardt (1987). Betweenness centrality reflects on the influence of a node has over the flow of information. Group 1 consists of networks 15 and 20, which have unique performances on betweenness centrality. The betweenness centrality of nodes 15, 20 is 81.15 and 65.35, which are much bigger than the rest of nodes. Both of them mentioned a lot of advice relationships they are involved in. This is consistent with our finding of local clustering phenomenon implied by high GWDSP coefficient. The networks in group 3 are distinct from the rest of individuals in terms of low indegree and betweenness centrality. The indegree of individuals 6, 13, 16, 17 is all 0, indicating that they are not asked for advice by anybody. Also, the betweenness centrality of them is 0, 0.2, 0.11, 0.28, smaller than the rest of nodes in the locally aggregated networks. Moreover, employee 1 has high indegree centrality 18, but low betweenness centrality 2.81. It is asked advice often, but rarely asks advice from other people. Of all the 18 edges individual 1 claimed, only 1 relationship is confirmed by others. The speciality of individual 1 explains why the network 1 formed a group of its own.

Posterior assessments can be done by comparing the observed network statistics with simulated network statistics sampled from ERGM with estimation as parameters. Specifically, we generate 500 networks using the posterior mean as parameters and draw the density plots of the simulated network statistics in Fig. 3. As we can see, the simulated network statistics are close to the observed network statistics, suggesting that the true likelihood based method IF-TL fits the data well. Note that network 1 located on the right end of the plot is far from other networks regarding the number of total edges and the number of edges within the same level. This is another reason that we think network 1 is better to be in a separate group.

Furthermore, we apply the adjusted pseudo likelihood based method, IF-APL, to the advice ensemble. After 100,000 iterations, 4 stable groups are detected. The networks in groups 1 and 3 from the IF-APL method are the same as from the IF-TL method. Networks 3, 7, 11, 12, 18 form the group 2 and networks 1, 2, 4, 5, 8, 9, 10, 14, 19, 21 form the group 4. The acceptance probability of the Metropolis Hastings algorithm for 4 groups are 0.52, 0.38, 0.49, 0.38 respectively. As for the maximum likelihood estimation $\hat{\theta}_{MLE}$ required for the adjusted pseudo likelihood, we use the Monte-Carlo contrastive divergence estimate (Krivitsky, 2017) since the convergence of some models with the standard MCMC MLE method (Hunter and Handcock, 2006) is very slow.

We also apply the pseudo likelihood based method, IF-PL, to the advice ensemble. After 100,000 iterations, 6 stable groups are detected. The networks in groups 1, 3, 4 from the IF-PL method are the same as from the IF-TL method. The group 2 from the IF-TL method is divided further into 3 groups, where networks 2, 4, 5, 8, 9, 10, 14, 19, 21 form the new second group, 3, 7, 12, 18 make the new fifth group, and 11 is in the sixth group. The acceptance probability of the Metropolis Hastings algorithm for 6 groups are 0.36, 0.23, 0.40, 0.36, 0.29, 0.50 respectively.

For comparison, we show the density plots of the simulated network statistics for group 4 on Fig. 4. Simulated network statistics from IF-TL are centred around the observed statistics on the top row, while simulated statistics from IF-APL and IF-PL are distant from the observed statistics. This is because the model for group 4 is near-degenerate. For a near-degenerate model, the underlying parameter values are close to a degenerate region, which increases the difficulty for estimation. This

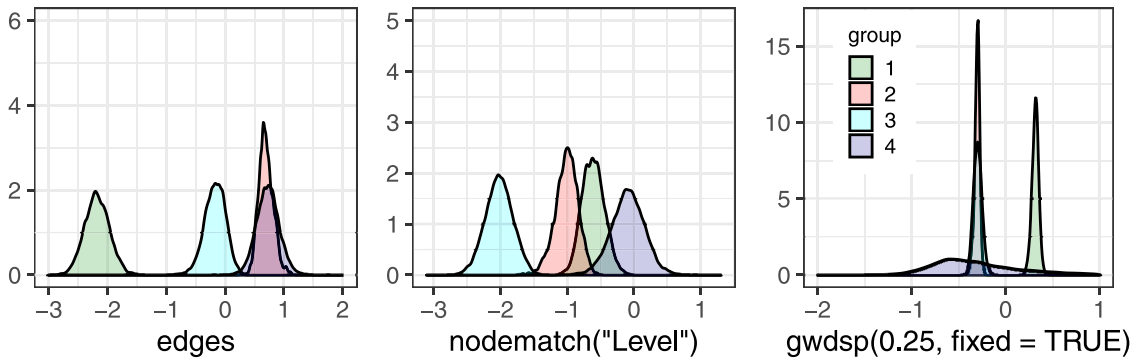

Fig. 2. Density plots of the ERGM parameters for each group using IF-TL.

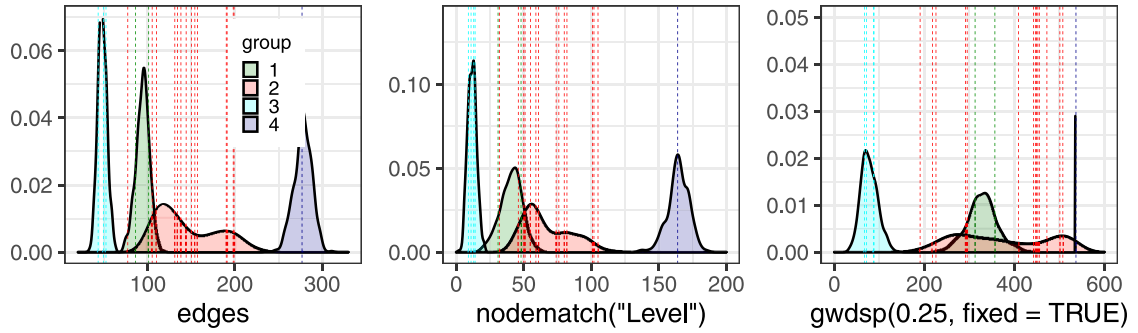

Fig. 3. Density plots of network statistics based on networks simulated from posterior mean. The vertical lines stand for the value of structure statistics of observed networks.

can happen quite often when we fit a ERGM with complicated statistics to real datasets. The pseudo likelihood based methods do not work for the near-degenerate model (Caimo and Friel, 2011). In this case, we can only use true likelihood method. More details about the cluster results as well as the posterior assessments for all three methods can be found in the supplementary materials.

In this case study, we found stable and meaningful clusters with all the three methods developed for DPM-ERGMs. Although pseudo likelihood based methods managed to divide the ensemble into reasonable clusters, they failed to represent the features of networks because they are not suitable for estimating the near-degenerate model in this example. While using DPM-ERGMs for real network ensembles, we suggest to use IF-APL for a quick preliminary analysis, but for the accurate estimation of network structure, especially for networks with complicated dependent interactions, we recommend using IF-TL.

5. Discussion

In this manuscript, we proposed to model the ensemble of networks using a Dirichlet process mixture of ERGMs. Through such a framework, the subpopulations consisting of similar networks can be detected and compared automatically without requiring a fixed number of clusters in advance. On the other hand, multiple networks with similar characteristics are described by the same ERGM, namely, the cluster-specific ERGM, which is better than a single network ERGM, because information from all networks in the same cluster are gathered together on the cluster-specific ERGMs. Moreover, we also developed a novel Metropolis-within-slice sampling algorithm for the posterior inference of the DPM-ERGMs. To handle the intractability issue of the ERGM likelihood in the infinite mixture model, we presented three different estimation methods, the true likelihood based method (IF-TL), the adjusted pseudo likelihood based method (IF-APL) and the pseudo likelihood based method (IF-PL). Simulation studies have shown that all three methods perform well in recovering the clustering memberships of networks. Regarding the estimation accuracy of the true number of clusters, IF-TL and IF-APL perform very well.

Although the IF-TL method provides accurate estimation to the proposed models, it is also time consuming as auxiliary networks are sampled using MCMC technique at each iteration of the algorithm to approximate the normalising constant ratio. For networks with different sizes or covariates the time is even longer. The IF-APL method provides good estimation but care is needed especially when estimating the MLE for ERGMs. Pseudo likelihood based approximations are fast but should be treated with caution because they can lead to an unreasonable inference. Despite the fact that the computational burden can be reduced with the use of multiple computer cores, it is still worthwhile to explore more accurate and faster estimation methods, especially at the context of ensembles of networks.

The flexibility of both ERGMs and the Bayesian nonparametric mixture models also offers us many promising future directions. It would be interesting to incorporate more diverse network structures as well as individual properties into the current framework, such as an extension to the conditional ERGMs for single networks (Nasini et al., 2017). Within each network, there exists a multilevel structures at different scales: micro, meso and macro, respectively (Mursa et al., 2021). It might be illuminating to look at the multilevel within-network structures of clusters of networks.

Recently, there is also work on the concentration and consistency results for ERGMs (Schweinberger and Stewart, 2020) as well as for pseudo likelihood based M-estimators (Stewart and Schweinberger, 2020) in a single network case. It would be interesting to derive the posterior consistency of DPM-ERGMs on top of these results. Besides, although the Monte Carlo Metropolis Hastings algorithm (Liang and Jin, 2013) has been proved to converge to the desired target distribution for the single network, the theoretical properties of such an algorithm with multiple network samples still remain unknown. Also, it is interesting to

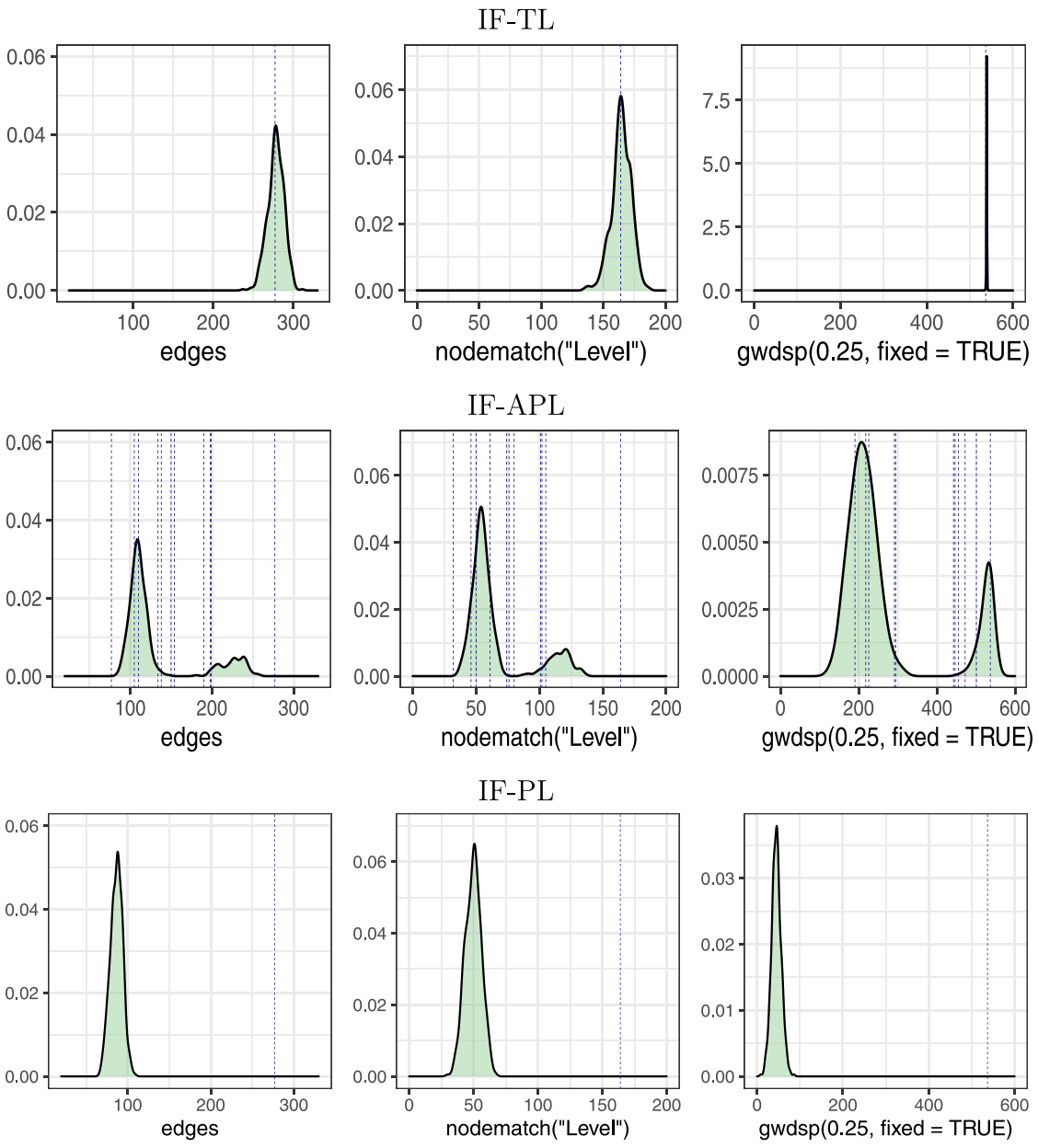

Fig. 4. Simulated network statistics from the IF-TL method for group 4 (or network 1) is in the first row. Simulated network statistics from the IF-APL method for group 4 (or network 1) is in the second row. Simulated network statistics from the IF-PL method for group 4 is in the third row.

explore how the importance sampling technique in sampling the cluster membership variables affects the posterior consistency of the estimates.

Acknowledgements

The authors would like to thank the editor and the reviewers for their constructive comments and suggestions that have helped to improve the manuscript. The authors are grateful to Dr. Fan Yin for sharing his code on GitHub. The authors also appreciate Yanjun Pu for helping with the computing resources for the simulation studies in this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socnet.2023.03.005.

References

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C.E., Vogelstein, J.T., 2021. Inference for multiple heterogeneous networks with a common invariant subspace. J. Mach. Learn. Res. 22 (142), 1–49.

Bouranis, L., Friel, N., Maire, F., 2017. Efficient Bayesian inference for exponential random graph models by correcting the pseudo-posterior distribution. Social Networks 50, 98–108.

Bouranis, L., Friel, N., Maire, F., 2018. Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. J. Comput. Graph. Statist. 27 (3), 516–528.

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

Caimo, A., Gollini, I., 2020. A multilayer exponential random graph modelling approach for weighted networks. Comput. Statist. Data Anal. 142, 106825.

van Duijn, M.A., Gile, K.J., Handcock, M.S., 2009. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. Social Networks 31 (1), 52–62.

Durante, D., Dunson, D.B., 2018. Bayesian inference and testing of group differences in brain networks. Bayesian Anal. 13 (1), 29–58.

Durante, D., Dunson, D.B., Vogelstein, J.T., 2017. Nonparametric Bayes modeling of populations of networks. J. Amer. Statist. Assoc. 112 (520), 1516–1530.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1, 209–230.
Gelman, A., Meng, X.-L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Statist. Sci. 13 (2), 163–185.
Henry, T.R., Gates, K.M., Prinstein, M.J., Steinley, D., 2019. Modeling heterogeneous peer assortment effects using finite mixture exponential random graph models. Psychometrika.
Hunter, D.R., Handcock, M.S., 2006. Inference in curved exponential family models for networks. J. Comput. Graph. Statist. 15 (3), 565–583.
Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. Ergm: A package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. 24 (3), 1–29.
Kalli, M., Griffin, J.E., Walker, S.G., 2011. Slice sampling mixture models. Stat. Comput. 21, 93–105.
Krackhardt, D., 1987. Cognitive social structures. Social Networks 9 (2), 109–134.
Krivitsky, P.N., 2017. Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. Comput. Statist. Data Anal. 107, 149–161.
Krivitsky, P.N., Koehly, L.M., Marcum, C.S., 2020. Exponential-family random graph models for multi-layer networks. Psychometrika 85 (3), 630–659.
Lehmann, B., White, S., 2021. Bayesian exponential random graph models for populations of networks. arxiv:2104.05110.
Leisch, F., 2006. A toolbox for k-centroids cluster analysis. Comput. Statist. Data Anal. 51 (2), 526–544.
Liang, F., Jin, I.-H., 2013. A Monte Carlo Metropolis-Hastings algorithm for sampling from distributions with intractable normalizing constants. Neural Comput. 25 (8), 2199–2234.
Lubbers, M.J., Snijders, T.A., 2007. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. Social Networks 29 (4), 489–507.
MacDonald, P.W., Levina, E., Zhu, J., 2022. Latent space models for multiplex networks with shared structure. Biometrika 1–24.
Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B., 2016. Model-based clustering based on sparse finite Gaussian mixtures. Stat. Comput. 26 (1), 303–324.
Morris, M., Handcock, M.S., Hunter, D.R., 2008. Specification of exponential-family random graph models: terms and computational aspects. J. Stat. Softw. 24 (4), 1548.
Murray, I., Ghahramani, Z., MacKay, D.J.C., 2006. MCMC for doubly-intractable distributions. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence. pp. 359–366.
Mursa, B.-E.-M., Dioşan, L., Andreica, A., 2021. Network motifs: A key variable in the equation of dynamic flow between macro and micro layers in complex networks. Knowl.-Based Syst. 213, 106648.
Nasini, S., Martínez-de Albéniz, V., Dehdarirad, T., 2017. Conditionally exponential random models for individual properties and network structures: Method and application. Social Networks 48, 202–212.
Neal, R.M., 2005. Estimating Ratios of Normalizing Constants Using Linked Importance Sampling. Technical Report No. 0511, Department of Statistics, University of Toronto.
Paul, S., Chen, Y., 2020. A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. Ann. Appl. Stat. 14 (2), 993–1029.
Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66 (336), 846–850.
Salter-Townshend, M., Murphy, T.B., 2015. Role analysis in networks using mixtures of exponential random graph models. J. Comput. Graph. Statist. 24 (2), 520–538.
Schweinberger, M., Handcock, M.S., 2015. Local dependence in random graph models: characterization, properties and statistical inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 77 (3), 647–676.
Schweinberger, M., Krivitsky, P.N., Butts, C.T., Stewart, J.R., 2020. Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios. Statist. Sci. 35 (4), 627–662.
Schweinberger, M., Stewart, J., 2020. Concentration and consistency results for canonical and curved exponential-family models of random graphs. Ann. Statist. 48 (1), 374–396.
Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statist. Sinica 4 (2), 639–650.
Signorelli, M., Wit, E.C., 2020. Model-based clustering for populations of networks. Stat. Model. 20 (1), 9–29.
Simpson, S.L., Lyday, R.G., Hayasaka, S., Marsh, A.P., Laurienti, P.J., 2013. A permutation testing framework to compare groups of brain networks. Front. Comput. Neurosci. 7 (171), 1–13.
Slaughter, A.J., Koehly, L.M., 2016. Multilevel models for social networks: Hierarchical Bayesian approaches to exponential random graph modeling. Social Networks 44, 334–345.
Stewart, J.R., Schweinberger, M., 2020. Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension. arXiv preprint arXiv:2012.07167.
Strauss, D., Ikeda, M., 1990. Pseudolikelihood estimation for social networks. J. Amer. Statist. Assoc. 85 (409), 204–212.
Sweet, T.M., Flynt, A., Choi, D., 2019. Clustering ensembles of social networks. Netw. Sci. 7 (2), 141–159.
Sweet, T.M., Thomas, A.C., Junker, B.W., 2013. Hierarchical network models for education research: Hierarchical latent space models. J. Educ. Behav. Stat. 38 (3), 295–318.
Walker, S.G., 2007. Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. 36, 45–54.
Wang, P., Robins, G., Pattison, P., Lazega, E., 2013. Exponential random graph models for multilevel networks. Social Networks 35 (1), 96–115.
Yin, F., Shen, W., Butts, C.T., 2022. Finite mixtures of ERGMs for modeling ensembles of networks. Bayesian Anal. 1–39.