Bayesian Analysis (2022) **17**, Number 4, pp. 1153–1191

Finite Mixtures of ERGMs for Modeling Ensembles of Networks[*]

Fan Yin[†], Weining Shen[‡], and Carter T. Butts[†,§,¶]

Abstract. Ensembles of networks arise in many scientific fields, but there are relatively few statistical tools for inferring their generative processes, particularly in the presence of both dyadic dependence and cross-graph heterogeneity. To address this gap, we propose characterizing network ensembles via finite mixtures of exponential family random graph models (ERGMs), a class of parametric statistical models that has been successful in explicitly modeling the complex stochastic processes that govern the structure of edges in a network. Our proposed modeling framework can also be used for applications such as model-based clustering of ensembles of networks and density estimation for complex graph distributions. We develop a joint approach to estimate the number of mixture components and identify cluster-specific parameters simultaneously as well as to obtain an identified model under the Bayesian paradigm. Specifically, we develop a Metropolis-within-Gibbs algorithm to perform Bayesian inference, and estimate the number of mixture components using a strategy of deliberate overfitting with sparse priors that removes excess components during MCMC. As the true ERGM likelihood is generally intractable for model specifications with dyadic dependence terms, we consider two tractable approximations (pseudolikelihood and adjusted pseudolikelihood) to facilitate efficient statistical inference. We run simulation studies to compare the performance of these two approximations with respect to multiple metrics, showing conditions under which both are useful. We demonstrate the utility of the proposed approach using an ensemble of political co-voting networks among U.S. Senators and an ensemble of brain functional connectivity networks.

Keywords: exponential-family random graph models (ERGMs), Bayesian mixture model, MCMC, adjusted pseudolikelihood, political co-voting networks, brain functional connectivity networks.

1 Introduction

Data involving ensembles of networks – that is, multiple independent networks – arise in various scientific fields, including sociology (Slaughter and Koehly, 2016; Stewart et al., 2019), neuroscience (Simpson et al., 2011; Obando and De Vico Fallani, 2017), molecular biology (Unhelkar et al., 2017; Grazioli et al., 2019), and political science (Moody and Mucha, 2013) among others. Typically, ensembles of networks represent the action of multiple generative processes, with different processes being prominent

[*] This work was partially supported by the National Science Foundation (NSF) award DMS-1361425.

[†] Department of Statistics, University of California, Irvine, yinf2@uci.edu

[‡] Department of Statistics, University of California, Irvine, weinings@uci.edu

[§] Department of Sociology, Computer Science, and Electrical Engineering and Computer Science, University of California, Irvine, buttsc@uci.edu

[¶] Corresponding author.

https://doi.org/10.1214/21-BA1298

in different settings. A reasonable starting point for analysis of such data is to posit that this variation can be represented in terms of a discrete set of subpopulations, such that the networks drawn from any given subpopulation tend to be produced by similar generative processes. Given a set of potential generative models, one would then like to identify the subsets of networks drawn from a particular subpopulation, or a probabilistic mixture of multiple subpopulations. It is natural to view this as a hierarchical finite mixture problem, with the base distributions being parametric distributions on graphs. As a plausible approximation to the underlying data generating process, the hierarchical finite mixture framework provides a flexible approach for predictive modeling of ensemble of networks. If one seeks to predict graph structures drawn from a heterogeneous (super)population learned from observed data, one needs to average over the possible generative processes that might end up producing the observation that one wants to predict. Such a view is similar in spirit to model averaging techniques (Hoeting et al., 1999; Hjort and Claeskens, 2003), especially if interpreted in terms of a hierarchical problem in which we seek to predict an outcome of interest (e.g., co-voting prevalence among U.S. senators) by first predicting network structure and then predicting the behavior of a process on that network. In that setting, if it turned out that there were K types of possible network formation processes and we did not know which one ours happened to be, we would certainly want to average across the types.

There is a growing body of literature on the analysis of ensembles of networks. This includes work on discriminative analysis of networks via distance or similarity measures (e.g. Banks and Carley, 1994; Butts and Carley, 2005; Fitzhugh et al., 2015), which can be broadly viewed as mapping the ensemble of interest into some high-dimensional space (e.g., the Hamming space of graphs), and then employing standard multivariate analysis techniques (e.g., hierarchical clustering, multidimensional scaling) to seek an informative low-dimensional approximation. Other approaches work with user-selected graph statistics, either directly (e.g. Pržulj, 2007; Sweet et al., 2019) or by e.g., modeling quantiles of the observed statistics relative to a reference distribution to control for size and density effects (Butts, 2011). As such, these approaches do not attempt to provide generative models for the networks within the ensemble, though they may in some cases provide generative models for summary statistics (e.g., predicting the conditional uniform graph quantile for the transitivity of a new graph drawn from the same ensemble). Another relevant work is Durante et al. (2017), where the authors propose a Bayesian nonparametric approach for modeling the population-level network based on a mixture model framework on the joint distribution of the edges through a latent space model representation.

In the category of generative models for complex networks, a common approach is to employ multilevel models with exponential random graph models (ERGMs, a general family of parametric models for networks (see, e.g. Schweinberger et al., 2020, for a review)), as base distributions. Faust and Skvoretz (2002) introduced both multivariate meta-analysis of ERGM parameters from a common model family (fit to an ensemble of graphs) and predicted conditional edge probabilities from the generative base models as tools for leveraging ERGMs to compare networks. More elaborate meta-analytic procedures and hierarchical models for population of networks were subsequently developed by, among others, Zijlstra et al. (2006); Slaughter and Koehly (2016); McFarland et al. (2014); Butts (2017), and Stewart et al. (2019). In general, those methods have either not posited a generative model for the parameters of the base distribution, as in descriptive meta-analytic approaches (which can be problematic when model interpretation and simulation from the resulting model are of interest), or not suitable for identifying subpopulations from heterogeneous data (as in hierarchical models without mixture structure). Although work such as that of Lehmann et al. (2021) enables the modeling of heterogeneity in brain functional connectivity networks, it requires that subpopulation labels be observed (which is often not the case). Therefore, joint modeling of population-level and network-level parameters where subpopulation memberships are unknown, or where the true generative process otherwise involves a mixture of graph distributions, has remained an open problem to date in the ERGM context.

In this paper, we propose using mixtures of ERGMs to model the generative process of ensembles of networks in which the subpopulation labels are not available, under the general framework of finite mixture models (McLachlan and Basford, 1988; Fraley and Raftery, 2002; Bouveyron et al., 2019). Such a formulation provides a useful probabilistic interpretation of the results and allows for convenient statistical inference; we note in particular that related approaches have proven to be efficacious for modeling structure *within* networks (e.g. Salter-Townshend and Murphy, 2015; Schweinberger and Handcock, 2015; Snijders and Nowicki, 1997). Recent work on using mixtures of network models with conditionally independent dyads (e.g., a priori stochastic blockmodels, the p_1 model) for modeling multiple network observations (Signorelli and Wit, 2020) can encounter difficulties when the observed networks exhibit strong dyadic dependence, which is often the case for real-world networks. We develop a joint approach to simultaneously estimate the number of mixture components and identify cluster-specific parameters, as well as to obtain an identified model under the Bayesian paradigm. Specifically, we propose a Metropolis-within-Gibbs algorithm to perform Bayesian inference, and estimate the number of mixture components using the method of Malsiner-Walli et al. (2016) that employs a deliberately overfitting model with sparsity-inducing priors over a large number of mixture components, with the excess components being "emptied" during MCMC. We also adapt the Bayesian information criterion and Deviance information criterion to enable the comparison of different specifications of ERGMs.

The remainder of this paper is structured as follows. In Section 2, we briefly introduce the exponential-family random graph models (ERGMs) and common estimation techniques. Section 3 describes the idea of finite mixtures of ERGMs, along with the estimation algorithm and the method for selecting the number of mixture components. Section 4 presents simulation studies showing that the proposed method can accurately recover the true subpopulation assignments and model parameters. Sections 5 and 6 show the results of our method applied to a political co-voting data and brain functional connectivity networks, respectively. Section 7 concludes with a discussion.

2 Exponential-Family Random Graph Models (ERGMs)

In recent years, ERGMs have found applications in empirical research in a wide range of scientific fields. Recent examples include the study of large friendship networks (Goodreau, 2007), genetic and metabolic networks (Saul and Filkov, 2007), disease transmission networks (Groendyke et al., 2012), conflict networks in the international system (Cranmer and Desmarais, 2011), the structure of ancient networks in various of archaeological settings (Amati et al., 2019), the structural comparison of protein structure networks (Grazioli et al., 2019), the effects of functional integration and functional segregation in brain functional connectivity networks (Simpson et al., 2011; Sinke et al., 2016; Obando and De Vico Fallani, 2017), and the impact of endogenous network effects on the formation of interhospital patient referral networks (Caimo et al., 2017). While addressing very different problems in different empirical settings, what these studies have in common is a clear methodological commitment to modeling network mechanisms directly via parametric effects, rather than just attempting to "control for" unspecified dependence among the observations (e.g., via latent structure). The ability to provide generative and interpretable models of complex network structure is an important asset of this approach, which we leverage here in the context of graph ensembles.

2.1 Definition and Estimation

Exponential-family random graph models (ERGMs) (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Snijders et al., 2006; Hunter and Handcock, 2006), also known as p-star models (Wasserman and Pattison, 1996), are a family of parametric statistical models developed for explicitly modeling the complex stochastic processes that govern the formation of edges among pairs of nodes in a network. We introduce them first in the single-network case. Consider the set of nodes in the network of interest, $\boldsymbol{V}$, and let $|\boldsymbol{V}| = n$ be its cardinality, i.e. the number of nodes in the network. We represent the network's structure via an order-n random adjacency matrix $\mathbf{Y}$, in which each element takes 1 or 0 representing the presence or absence of a tie between incident nodes. Letting $\mathcal{Y}_n$ be the set of all possible network configurations on n nodes, we write the probability mass function (pmf) of $\mathbf{Y}$ taking a particular configuration $\mathbf{Y}$ in the form of a discrete exponential family as

$$\mathbb{P}_{\boldsymbol{g},h,\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \frac{\exp\big(\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y};\mathbf{X})\big)h(\mathbf{y})}{z_{\boldsymbol{g},h,\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta})}, \quad \mathbf{y} \in \mathcal{Y}_n, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q) \in \mathbb{R}^q$ is a vector of (curved) model parameters, mapped to the natural parameters by $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \cdots, \eta_p(\boldsymbol{\theta})) \in \mathbb{R}^p$. The natural parameters $\boldsymbol{\eta}$ may depend on the sizes of the networks and may be non-linear functions of a parameter vector $\boldsymbol{\theta}$. The user-defined sufficient statistics $\boldsymbol{g} : \mathcal{Y}_n \to \mathbb{R}^p$ may incorporate fixed and known covariates $\mathbf{X}$ that are measured on the nodes or dyads. The sufficient statistics incorporate network features of interest that are believed to be crucial to the mechanisms giving rise to the graph set (see, e.g., Morris et al., 2008). Here, h defines the reference measure for the model family; often chosen to be the counting measure $h(\mathbf{y}) \equiv 1, \forall \mathbf{y} \in \mathcal{Y}_n$ for unvalued graphs with fixed n, other reference measures can make more sense in different settings (see e.g. Schweinberger et al., 2020). Finally, the normalizing factor $z_{\boldsymbol{g},h,\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}_n} \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{g}(\boldsymbol{y}';\mathbf{X})\right\} h(\boldsymbol{y}')$ ensures that (1) sums to 1 over the support $\mathcal{Y}_n$. Relatedly, normalizability requires that $\boldsymbol{\theta}$ only take values for which $z_{\boldsymbol{g},h,\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta})$ is finite.

Exact maximum likelihood estimation (MLE) for ERGM parameters requires direct evaluation of the normalizing factor, which involves integrating an extremely rough (i.e., high variance) function over all possible network configurations ($2^{\binom{n}{2}}$ non-negative terms for an undirected network of size n). This cannot be done by brute force except for trivially small graphs ($n \leqslant 7$), and the roughness of the underlying function precludes simple Monte Carlo strategies; thus, alternative approaches that approximate or avoid this calculation are of substantial interest. To date, the most frequently used approaches include: maximum pseudo-likelihood estimation (MPLE; Besag (1974)) adapted by Strauss and Ikeda (1990); Markov Chain Monte Carlo MLE (MCMC MLE; Geyer and Thompson (1992)) by Handcock (2003); Hunter and Handcock (2006); approximate MLE based on Stochastic approximation (SA; Robbins and Monro (1951); Pflug (1996)) by Snijders (2002); fully Bayesian inference based on exchange algorithm (Caimo and Friel, 2011); approximate Bayesian inference based on adjusted pseudolikelihood (Bouranis et al., 2017); and variational Bayesian inference based on fully adjusted pseudolikelihood (Tan and Friel, 2020). As simulation-based MLE-finding algorithms (e.g., MCMC MLE, SA) rely on *good* initial parameter configuration to seed their simulations, there are also some work on this aspect, including the *partial stepping* technique (Hummel et al., 2012) and the *contrastive divergence* (CD, Hinton (2002))-based techniques adapted to ERGMs by Krivitsky (2017). Despite the computational challenges, these and related strategies have made ERGM inference practical for well-posed model families (see Schweinberger et al., 2020, for a recent review).

2.2 Size-Adjusted Parameterizations

It is worth noting that the behavior of (1) across n is highly dependent on the choice of reference measure, h. In particular, the counting measure – while a mathematically convenient choice – implicitly sets the base distribution of the network to be the uniform distribution on $\mathcal{Y}_n$, and has the side effect of generating graphs whose densities are *ceteris paribus* constant in n. When network size varies, this is not always realistic: in many networks, mean degree is approximately constant in n, implying that density must scale as n^{-1}. To correct for this, Krivitsky et al. (2011) propose the reference measure $h(\mathbf{y}) = n^{-M(\mathbf{y})}$, where M is the edge count. This is equivalent to adding a size-dependent offset of $-\log n$ to the natural parameter associated with the edge count, i.e., $\eta_1(\boldsymbol{\theta}) = \theta_1 - \log n$, where $\theta_1 \in \mathbb{R}$ is a parameter that does not depend on the network size. In the present work, we employ the *Krivitsky reference measure* as above for the networks of varying sizes, though other size-adjusted parameterizations are also possible (e.g., Butts and Almquist, 2015; Kolaczyk and Krivitsky, 2015).

3 Finite Mixtures of ERGMs

We assume a population of networks $(\mathbf{Y}^{(1)}, \boldsymbol{V}^{(1)}, \mathbf{X}^{(1)}), \ldots, (\mathbf{Y}^{(m)}, \boldsymbol{V}^{(m)}, \mathbf{X}^{(m)})$, where $\mathbf{Y}^{(i)}$ is a graph structure on vertex set $\boldsymbol{V}^{(i)}$ with covariate set $\mathbf{X}^{(i)}$. Our interest is in modeling $\underline{\mathbf{Y}} = \left(\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(m)}\right)$ given $(\boldsymbol{V}^{(1)}, \mathbf{X}^{(1)}), \cdots, (\boldsymbol{V}^{(m)}, \mathbf{X}^{(m)})$, where it will be assumed that the respective graph structures are conditionally independent given the respective subpopulation membership, vertex sets, and covariates. To simplify notation, $\boldsymbol{V}$ is implicitly absorbed into $\mathbf{X}$ for the remainder of this paper.

We formulate the data generating mechanism as a hierarchical model and we note that Bayesian inference is a natural choice here, since (1) it is more robust to initialization and less prone to converge to local minima than maximum likelihood; (2) interval estimation is straightforward and does not rely on the assumption of approximate normality; and (3) it provides principled answers in fixed-n, m settings.

With a prespecified upper bound for the number of subpopulations (or equivalently, mixture components) K, we have the following hierarchical representation of the model (see Figure 1 for a graphical representation)

$$\begin{aligned}
&\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K) \sim \mathrm{Dir}_K(\alpha, \ldots, \alpha),\\
&P(Z_i = j) = \tau_j \quad \text{for every} \quad i = 1, \ldots, m, \text{ and } j = 1, \ldots, K, \quad \text{given } \boldsymbol{\tau},\\
&\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim \mathcal{N}_p(\boldsymbol{\mu}, \Psi) \text{ independently given } \boldsymbol{\mu}, \Psi,\\
&Y^{(i)} \sim \mathbb{P}_{\boldsymbol{g},h,\boldsymbol{\eta}}(\cdot|\mathbf{X}_i; \boldsymbol{\theta}_{Z_i}) \text{ independently for } i = 1, \ldots, m,\\
&\text{given } \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K,\ Z_1, \ldots, Z_m, \text{and } \mathbf{X}_1, \ldots, \mathbf{X}_m,
\end{aligned} \tag{2}$$

where $Z_i, \ldots, Z_m$ are indicator variables taking values in $\{1, \ldots, K\}$ for $Y_i, \ldots, Y_m$, and $\boldsymbol{\mu}, \Psi$ are hyper-parameters in the multivariate normal prior distributions for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. Without the Krivitsky reference measure, a multivariate normal prior with mean 0 for all elements reflects the prior belief that the graph density is around 0.5, which exceeds the density for most known real-world graphs. Therefore we let $\boldsymbol{\mu} = (-1, \underbrace{0, \ldots, 0}_{p-1})$ and $\Psi = 25 \times I_p$ to specify a flat multivariate normal prior for $\boldsymbol{\theta}_k$'s, which is a standard choice in the ERGM literature (Caimo and Friel, 2014; Bouranis et al., 2018) to reflect our belief that the real-world graph is more likely to have a density smaller than 0.5. Other mean values can be used if there is additional prior knowledge about the graph density. With the Krivitsky reference measure, a multivariate normal prior with mean 0 for all elements reflects the prior belief that expected baseline mean degree (i.e., discounting other effects) is around 1, which is more reasonable for typical sparse networks. Again, other mean values can be used if there is additional prior knowledge about the mean degree of the graphs of interests. To identify the number of mixture components, we deliberately choose a large value for K and a small value for α (e.g., 0.01 or 0.001, as demonstrated in Malsiner-Walli et al. (2016)), so that the superfluous components can be eliminated during MCMC, and hence a straightforward estimator for the true number of components is given by the most frequent number of non-empty components visited. In our numerical examples, we choose α from a set of pre-specified small values that achieves the most accurate estimate of the cluster number. It is of interest to develop efficient marginal likelihood computing methods for our proposed method in a future work and use them to guide the choice of α using an empirical Bayesian approach.

From this point onward, we focus on ERGM model specifications in canonical form (i.e., $\boldsymbol{\eta}(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}$) as it is a rich model family comprising most commonly-used model specifications and known to be computationally stable. For networks of equal sizes, it

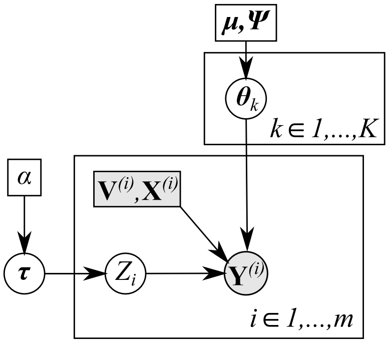

Figure 1: Structure of the graph mixture model. Random quantities are depicted within circles, fixed quantities within rectangles; observables are shaded.

is natural to use counting measure $h(\mathbf{y}) \equiv 1$, therefore (1) becomes

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y};\mathbf{X})\right)}{z(\boldsymbol{\theta})}, \quad \mathbf{y} \in \mathcal{Y}_n, \tag{3}$$

where $z(\boldsymbol{\theta}) = \sum_{\boldsymbol{y'} \in \mathcal{Y}_n} \exp\left(\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y'};\mathbf{X})\right)$. And for networks of varying sizes, with Krivitsky reference measure, (1) becomes

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y};\mathbf{X}) - \log(n)\sum_{i,j} y_{ij}\right)}{z_n(\boldsymbol{\theta})}, \quad \mathbf{y} \in \mathcal{Y}_n, \tag{4}$$

where the normalizing factor $z_n(\boldsymbol{\theta}) = \sum_{\boldsymbol{y'} \in \mathcal{Y}_n} \exp\left(\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y'};\mathbf{X}) - \log(n)\sum_{i,j} y'_{ij}\right)$.

3.1 Bayesian Estimation

As noted, we perform posterior inference via MCMC. Our algorithm iterates over the model parameters $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\tau}) = (\underline{\boldsymbol{\theta}}, \boldsymbol{\tau})$ with the priors given above, and the latent variables $\boldsymbol{Z} = (Z_1, \cdots, Z_m)$. Where possible, we sample from the full conditional posterior distributions; otherwise we use Metropolis-Hastings steps.

The proposal distribution $q(\cdot|\boldsymbol{\theta})$ in the Metropolis step is set by the user to achieve good performance of the algorithm. Following literature on Bayesian ERGMs (Caimo and Friel, 2011; Bouranis et al., 2017), we use a symmetric multivariate Gaussian proposal for updating $\boldsymbol{\theta}_k, k = 1, 2, \ldots, K$. As suggested by Frühwirth-Schnatter (2001), an additional permutation step is added to randomly permute the current labeling of the components at the end of each MCMC iteration, which ensures that the sampler explores all $K!$ modes of the full posterior distribution and prevents the sampler from being trapped around a single posterior mode (Geweke, 2007).

Algorithm 1 Metropolis-within-Gibbs sampler for the ERGM mixture model.

```
1:  Initialization: Set $\boldsymbol{\tau}^0$, $\underline{\boldsymbol{\theta}}^0$ and $\boldsymbol{Z}^0$ to random initial values.
2:  for $t = 1, 2, \cdots, T$ do
3:      Generate $Z_i^t$ $(i = 1, \cdots, m,\ k = 1, \cdots, K)$ from
          $\mathbb{P}(Z_i^t = k | \eta_k^{t-1}, \boldsymbol{\theta}_k^{t-1}, \mathbf{y}^{(i)}) \propto \eta_k^{t-1} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{t-1})$
4:      Compute $\nu_k^t = \sum_{i=1}^m \mathbb{1}_{Z_i^t = k}$; $k = 1, \cdots, K$
5:      Generate $\boldsymbol{\tau}^t$ from Dirichlet$(\alpha_1 + \nu_1^t, \cdots, \alpha_K + \nu_K^t)$
6:      for $k = 1, \cdots, K$ do
7:          Propose $\boldsymbol{\theta}_k^{'} \sim q(\cdot | \boldsymbol{\theta}_k^{t-1})$
8:          Accept $\boldsymbol{\theta}_k^{'}$ as $\boldsymbol{\theta}_k^t$ with probability equal to
            $\frac{\pi(\boldsymbol{\theta}_k^{'}) \prod_{Z_i^t = k} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{'}) q(\boldsymbol{\theta}_k^{t-1} | \boldsymbol{\theta}_k^{'})}{\pi(\boldsymbol{\theta}_k^{t-1}) \prod_{Z_i^t = k} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{t-1}) q(\boldsymbol{\theta}_k^{'} | \boldsymbol{\theta}_k^{t-1})}$
9:      end for
10:     Random permutation of the labeling: select randomly one permutation $\rho$ of $K!$
    possible permutations of $\{1, \ldots, K\}$ and substitute
            $\boldsymbol{\tau}^t = \boldsymbol{\tau}^t_{\rho(1,\ldots,K)},$
            $\underline{\boldsymbol{\theta}^t} = \underline{\boldsymbol{\theta}^t}_{\rho(1,\ldots,K)},$
            $\boldsymbol{Z}^t = \rho(\boldsymbol{Z}^t).$        (5)
11: end for
```

3.2 Approximations to Intractable Likelihoods

The likelihood $\mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta})$ is intractable for general ERGMs with dyadic dependent terms, which poses two challenges to the proposed Algorithm – (1) updating $\boldsymbol{\theta}_k$'s via Metropolis-Hastings ratio (8) is a *doubly-intractable* problem (Murray et al., 2006); (2) updating latent indicator variables Z_i requires a tractable likelihood. Although the exchange algorithm (Caimo and Friel, 2011) can be used to tackle the former challenge by cancelling out the intractable normalizing factors based on auxiliary variables, it falls short of the latter due to its inability to provide an approximation to the ERGM likelihood. Importance-sampling based approximation (Koskinen, 2004, 2008) appears to be a straightforward solution, but it can be too expensive to be practical when the cost of simulating from ERGMs is considerable, as such simulations are needed for each network observation at each MCMC iteration. Therefore, we propose to consider tractable approximations to the true ERGM likelihood (1) when fitting the proposed model.

Pseudolikelihood

The *pseudolikelihood* approximates the true ERGM likelihood by a product of full conditional distributions of edge variables

$$f_{PL}(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta}) = \prod_{(i,j) \in \mathcal{D}} \mathbb{P}(y_{ij} | y_{-ij}; \mathbf{X}; \boldsymbol{\theta}) = \prod_{(i,j) \in \mathcal{D}} \frac{1}{1 + \exp\{-\boldsymbol{\theta}^\top \Delta_{i,j} \boldsymbol{g}(\mathbf{y}; \mathbf{X})\}}, \tag{6}$$

where $\Delta_{i,j}\boldsymbol{g}(\mathbf{y};\mathbf{X}) = \boldsymbol{g}(y_{ij}^+;\mathbf{X}) - \boldsymbol{g}(y_{ij}^-;\mathbf{X})$ are the so-called *change statistics* associated with the dyad (i,j), representing the change in sufficient statistics when y_{ij} is toggled from 0 (y_{ij}^-) to 1 (y_{ij}^+) with the rest of the network remaining unchanged; $\mathcal{D}$ denotes the set of all pairs of dyads. For directed networks, $\mathcal{D} = \{(i,j)|i,j \in \mathcal{N}, i \neq j\}$, while for undirected networks, $\mathcal{D} = \{(i,j)|i,j \in \mathcal{N}, i < j\}$. In the frequentist paradigm, maximizing (6) gives the so-called MPLE, which is relatively fast, algorithmically convenient, and able to provide approximate parameter estimates for even badly-specified models.

While empirical observations show that MPLE can cause bias and underestimate standard errors (Van Duijn et al., 2009) for models with strong dyadic dependence, it has been the default choice for initialization of MCMC-MLE algorithms. There is also promising work on using bootstrapped MPLE to construct confidence intervals (Schmid and Desmarais, 2017) for large and sparse networks, as the MPLE is usually close to the MLE in such cases (Desmarais and Cranmer, 2010). Similar logic has motivated the use of Bayesian bootstrap estimation based on "pseudo maximum a posteriori (MAP)" estimates using the PL approximation to the likelihood (Grazioli et al., 2019).

As the full conditionals of edge variables are tractable (see (6)), sampling from ERGMs can be implemented through a Gibbs sampling procedure. To improve the mixing in the typical case of ERGMs that concentrate probability mass on sparse graphs, the default simulation algorithm in `statnet` implements a "tie no tie"(TNT) sampler, in which the dyad is randomly selected with equal probability from the set of dyads with ties or the dyads without ties.

Adjusted Pseudolikelihood

Bouranis et al. (2018) proposed an *adjusted pseudolikelihood* for correcting the mode, curvature and magnitude of the pseudolikelihood, and their simulation studies show that the adjusted pseudolikelihood can provide an accurate approximation to the true likelihood in the presence of strong dyadic dependence (where pseudolikelihood falls short). Building upon adjusted pseudolikelihood approximation, Tan and Friel (2020) developed a variety of variational methods for Gaussian approximation of the posterior density and model selection, which is shown to yield comparable performance to that of the exchange algorithm. The adjusted pseudolikelihood is defined as follows,

$$\tilde{f}(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = C \cdot f_{PL}(\mathbf{y}|\mathbf{X};\phi(\boldsymbol{\theta})), \tag{7}$$

where the constant $C > 0$ adjusts the magnitude and $\phi : \mathbb{R}^p \to \mathbb{R}^p$ is an invertible affine transformation that adjusts the mode and curvature to match the true ERGM likelihood.

$$\phi(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_{PL} + W(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ML}). \tag{8}$$

In (8), W is a $p \times p$ upper triangular matrix, $\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$ is the maximum likelihood estimate (MLE), and $\hat{\boldsymbol{\theta}}_{PL} = \arg\max_{\boldsymbol{\theta}} f_{PL}(\mathbf{y}|\mathbf{X};\phi(\boldsymbol{\theta}))$ is the maximum pseudolikelihood estimate (MPLE). The estimation of matrix W and scaling constant C require simulating from ERGMs at the MLE to approximate the Hessian of $\log\mathbb{P}(\mathbf{y}|\mathbf{X};\boldsymbol{\theta})$ and at a sequence of points (i.e., *temperatures*, using the terms for path sampling (Gelman and Meng, 1998)) from the coordinate origin to the MLE to approximate the normalizing factor $z(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})$ (see supplementary material (Yin et al., 2021) for more details).

3.3 Identifying the Number of Clusters

We derive the posterior distribution of the number of non-empty components, K_0, from the MCMC output following the approach of Ishwaran et al. (2001) and Nobile et al. (2004). Letting $N_k^{(t)}$ denote the number of observations allocated to component k at iteration $t = 1, 2, \ldots, \tilde{T}$ (after discarding the burn-in and thinning) and $I(\cdot)$ denote the indicator function, we represent the number of non-empty components at iteration t as $K_0^{(t)} = K - \sum_{k=1}^{K} I(N_k^{(t)} = 0)$, and therefore estimate the posterior $\mathbb{P}(K_0 = h|\underline{\mathbf{y}})$ for $h = 1, 2, \ldots, K$ by the respective relative frequency. A natural point estimator for the number of clusters is then the posterior mode, $\hat{K}_0 = \arg\max_h \mathbb{P}(K_0 = h|\underline{\mathbf{y}})$, which is optimal under a zero-one loss.

3.4 Post MCMC Inference

Identification of the finite mixture model requires handling the label switching problem caused by the invariance of representation with respect to ordering the components. The resulting multimodality and symmetry of the posterior distribution for symmetric priors makes it difficult to conduct inference on component-specific parameters. Following Malsiner-Walli et al. (2016), we perform K-centroids cluster analysis with Mahalanobis distance (Leisch, 2006) on the point process representation of the MCMC draws to post-process the MCMC output and obtain a unique labeling as follows

1. Estimate the number of mixture components ($\hat{K}_0$) based on the retained MCMC iterations $t = 1, 2, \ldots, \tilde{T}$ after discarding the burn-in and thinning using the method described in Section 3.3, that is, $\hat{K}_0 = \text{mode}(K_0^{(t)}), t = 1, 2, \ldots, \tilde{T}$.

2. For $\tilde{T}$ retained MCMC iterations, extract only the subsequence $T_0 (\leqslant \tilde{T})$ where the number of non-empty components is exactly equal to $\hat{K}_0$.

3. For all T_0 iterations in the subsequence, remove the draws from empty components and arrange the remaining draws of the different components in a matrix with $\hat{K}_0 \times T_0$ rows and p columns.

4. Perform K-centroid cluster analysis based on the Mahalanobis distance on $\hat{K}_0 \times T_0$ draws to group them into $\hat{K}_0$ clusters.

5. Examine the clustering results and only keep those draws where the resulting cluster membership indicator vector forms a permutation of $1, 2, \ldots, \hat{K}_0$, which gives a subsequence of $\tilde{T}_0 (\leqslant T_0)$ draws.

6. For the remaining $\tilde{T}_0$ draws, a unique labeling is achieved by resorting the draws according to the cluster membership from the K-centroid cluster analysis in step 4 and rescaling the elements by their sum in each $\boldsymbol{\tau}_t$ such that they sum to 1.

3.5 Choosing Between Competing Model Specifications

When multiple model specifications $\mathcal{M}_j, j = 1, \dots, J$ are under consideration, it is important to have a fast and convenient model selection index. In particular, we illustrate two popular information criterion based choices here, the *Bayesian information criterion* (BIC) (Schwarz, 1978) and a version of the observed *deviance information criteria* (DIC) introduced by Celeux et al. (2006), which is an extension of the original DIC (Spiegelhalter et al., 2002) to latent variable models. Consider the likelihood of network $\mathbf{Y}^{(i)}$ taking value $\mathbf{y}^{(i)}$ under the mixture model

$$\mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\tau}, \underline{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \tau_k \mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)} | \mathbf{X}; \boldsymbol{\theta}_k) \tag{9}$$

therefore the BIC is defined as follows

$$BIC(\mathcal{M}_j) = 2 \times \sum_{i=1}^{m} \log \mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\tau}, \hat{\underline{\boldsymbol{\theta}}}_{\mathcal{M}_j}) - \left(p_{\mathcal{M}_j} \hat{K}_0(\mathcal{M}_j) + p_{\mathcal{M}_j} - 1\right) \log(m), \tag{10}$$

where $\hat{\underline{\boldsymbol{\theta}}}_{\mathcal{M}_j}$, $\hat{K}_0(\mathcal{M}_j)$ and $p_{\mathcal{M}_j}$ corresponds to the posterior mode of $\underline{\boldsymbol{\theta}}$, the estimated number of mixture components, and the number of parameters in each ERGM component, respectively. Following Celeux et al. (2006), given remaining posterior draws $\boldsymbol{\tau}^t = (\tau_1^t, \dots, \tau_{\hat{K}_0}^t), \underline{\boldsymbol{\theta}}^t = (\boldsymbol{\theta}_1^t, \dots, \boldsymbol{\theta}_{\hat{K}_0}^t)$ and observed ensemble of networks $\underline{\mathbf{y}} = (\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(m)})$, we have observed DIC for model specification $\mathcal{M}_j$ as

$$DIC(\mathcal{M}_j) = -4\mathbb{E}_{\underline{\boldsymbol{\theta}}}[\log \mathbb{P}(\underline{\mathbf{y}} | \underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}) | \underline{\mathbf{y}}] + 2 \log \hat{\mathbb{P}}(\underline{\mathbf{y}} | \underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}), \tag{11}$$

where

$$\hat{\mathbb{P}}(\underline{\mathbf{y}} | \underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} \hat{\mathbb{P}}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} \left(\frac{1}{\tilde{T}_0} \sum_{t=1}^{\tilde{T}_0} \mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\tau}, \underline{\boldsymbol{\theta}}^{(t)}) \right),$$

and

$$\mathbb{E}_{\underline{\boldsymbol{\theta}}}[\log \mathbb{P}(\underline{\mathbf{y}} | \underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}) | \underline{\mathbf{y}}] = \frac{1}{\tilde{T}_0} \sum_{t=1}^{\tilde{T}_0} \sum_{i=1}^{m} \log \left(\mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\tau}, \underline{\boldsymbol{\theta}}) \right).$$

3.6 Posterior Probability of Cluster Membership

We estimate the posterior marginal probability of individuals belonging to each cluster (alternately: graphs having been generated by a particular process) via Monte Carlo:

$$\mathbb{P}(Z_i = k | \mathbf{y}^{(i)}) = \int \frac{\tau_k \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \tau_k \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k)} \pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\tau} | \underline{\mathbf{y}}) d\underline{\boldsymbol{\theta}} d\boldsymbol{\tau}, \tag{12}$$

where $\pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\tau} | \underline{\mathbf{y}})$ is the posterior distribution of $\underline{\boldsymbol{\theta}}, \boldsymbol{\tau}$. The integral (12) is computationally intractable. Hence we use posterior samples $\underline{\boldsymbol{\theta}}^1, \cdots, \underline{\boldsymbol{\theta}}^L$ and $\boldsymbol{\tau}^1, \cdots, \boldsymbol{\tau}^L$ to obtain its

Monte-Carlo approximation,

$$\hat{\mathbb{P}}(Z_i = k|\mathbf{y}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} \frac{\tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)};\boldsymbol{\theta}_k^l)}{\sum_{k=1}^{K} \tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)};\boldsymbol{\theta}_k^l)}. \tag{13}$$

It is also possible to obtain the joint probabilities for the partitions and use decision theoretic approaches (Fritsch and Ickstadt, 2009; Wade and Ghahramani, 2018) to assign cluster memberships. These are promising future directions to pursue if further enhancements of performance are needed. In our work, we focus on the marginal probability method, as we have found it to perform well in practice.

4 Simulation Studies

We conduct simulation studies to evaluate the performance of the proposed algorithm with pseudolikelihood (PL) and adjusted pseudolikelihood (APL) on the following metrics: identification of mixture components; parameter estimation; and posterior predictive distribution of key network characteristics. We also compare the proposed method to K-means clustering in terms of their cluster membership recovery performance based on (i) network sufficient statistics (SS), (ii) maximum likelihood estimates, and (iii) maximum pseudolikelihood estimates. In all of the simulation studies, we consider ensembles of networks simulated from mixtures of ERGMs with three components of equal sizes.

4.1 Performance Evaluation

The identification of mixture components consist of two aspects, namely, the identification of true number of components and the recovery of true cluster memberships. To examine if the proposed algorithm can consistently identify the true number of components, we consider the proportion of times the true number of components is identified across all replicates under each experiment setting. To evaluate how accurately the true cluster memberships can be recovered, we calculate the average value of Rand index (RI) (Rand, 1971) and the average value of variation of information (VI) (Meilă, 2007) over 50 replicates. The RI measures the proportion of concordant pairs between two data clusterings, taking value 1 when two clusterings are identical and 0 when two clusterings do not agree on any pair of points. The VI, a true metric on clusterings, yields a value of 0 for identical clusterings with higher values indicating more disparate clustering assignments. We also compare the summary statistics for post MCMC inference, including T_0, $\tilde{T}_0$, $1 - T_0/\tilde{T}$, and $1 - \tilde{T}_0/T_0$ as defined in Section 3.4 (see supplementary material for more details).

For parameter estimation, we compute the relative bias of posterior mean and frequentist coverage rates of 95% posterior credible intervals. As it is possible that the estimated number of clusters is different from the true number of components, we focus on those runs in which the true number of components is successfully identified when evaluating the parameter estimation accuracy.

We examine the posterior predictive performance of the resulting mixture models by checking the discrepancy between the distributions of several critical network summary measures calculated on the observed networks and those of simulated networks.

As there is a computational overhead associated with the calculation of adjusted pseudolikelihood, we also include a comparison of computation time.

4.2 Simulation Design

In this simulation study, we consider ensembles of networks simulated from mixtures of ERGM distributions defined on three most commonly used network sufficient statistics

- $g_1(\mathbf{y}) = \sum_{i<j} y_{ij}$, total number of edges.
- $g_2(\mathbf{y}) = e^{\phi} \sum_{k=1}^{n-2} \left\{1 - (1 - e^{-\phi})^k\right\} EP_k(\mathbf{y})$, geometrically weighted edgewise shared partners (GWESP). Here $EP_k(\mathbf{y})$ is the number of connected pairs that have exactly k common neighbors, which measures local clustering in a network. The decay parameter ϕ controls the relative contribution of $EP_k(\mathbf{y})$ to the GWESP statistic, and it is fixed at 0.25 in this case.
- $g_3(\mathbf{y}; \mathbf{X}) = \sum_{i<j} y_{ij} \mathbb{1}_{\{\mathbf{X}_i = \mathbf{X}_j\}}$, the total number of edges with endpoints sharing the same value on nodal covariate $\mathbf{X}$, often known as a nodematch term.

We fix nodal covariate $\mathbf{X}$ to be a binary variable, and let one half of nodes take value 0, while the other half take value 1 on $\mathbf{X}$. To examine the performance of the proposed approach across a range of different conditions, we run a full-factorial experiment on the following two treatments

- Network size: 40, 100.
- Cluster size: 10, 30.

We thus have a total of 4 experimental conditions, each of which is run under pseudolikelihood (PL) and adjusted pseudolikelihood (APL) (hence there is a total of $8 (= 4 \times 2)$ simulation settings) for 50 repetitions. The true cluster-specific parameters are specified as

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

to ensure that the simulated networks (i) have similar mean degree (~ 9.9, that is, networks of size 100 have density ~ 0.10) across different mixture components and network sizes (Figure 2); and (ii) represent three common and intuitive patterns in real-world networks. Parameter settings in the first row correspond to the case in which there is a weak triadic closure effect (edge variables are nearly independent Bernoulli draws), and parameter settings in the second row correspond to the case in which there is a strong homophily effect but an intermediate triadic closure effect, while the

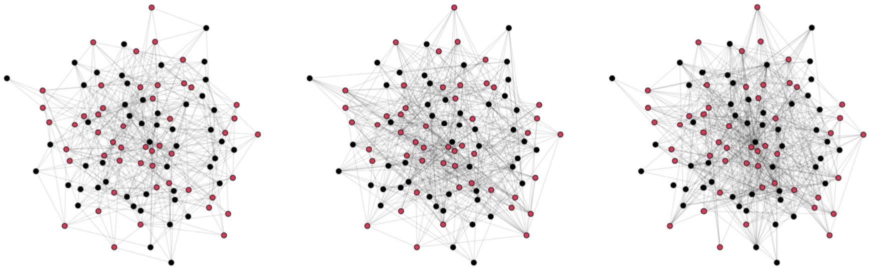

Figure 2: Representative networks from clusters 1 (left), 2 (middle), and 3 (right). Network size: 100. Color indicates nodal covariate value: 0 (black), 1 (red).

parameter settings in the third row correspond to the case in which there is a strong triadic closure effect but an intermediate homophily effect. To maintain this pattern, we fix the values of coefficients associated with GWESP and nodematch terms across settings with different network sizes, and only modify the coefficient of edges term to keep the mean degree value as desired.

We apply the proposed Algorithm 1 to analyze the synthetic data sets with the initial number of clusters $K = 6$. All MCMC chains are run for 60000 iterations with the first 60% of draws discarded as burn-in and a thinning interval of size 50 to ensure convergence. We draw random values from a discrete uniform distribution taking values in $\{1, \dots, K\}$ to initialize the latent membership indicators $\boldsymbol{Z}_i^0$, and draw random values from the Dirichlet prior to initialize the value of $\boldsymbol{\tau}^0$. The initial values of $\underline{\boldsymbol{\theta}}$ are drawn independently from a uniform distribution $\mathcal{U}(-0.1, 0.1)$. We fix $\alpha = 0.001$ on the basis of some experimentation (more details are available in the Supplementary Material). All computations in this paper are implemented in **R** (R Core Team, 2019), and we use software suite `statnet` (Handcock et al., 2008) to generate networks from ERGMs where the burn-in for the first simulated network (network size: n) from each mixture component under each replicate is set as $20n^2$ with a thinning interval of $4n^2$ iterations.

4.3 Identification of Mixture Components

To check the performance of the proposed inferential procedure at identifying the correct number of mixture components, we first count the frequencies of $\hat{K}_0$ over 50 repetitions across different experimental conditions and likelihood approximation methods. Table 1 shows the proportion of times the true number of mixture components is identified. We observe that the performance of PL at identifying the true number of mixture components appears to be slightly better than that of APL when the network size is large ($n = 100$), but the performance of adjusted pseudolikelihood (APL) is more robust across all 4 experimental conditions (accuracy $> 70\%$) and it is apparently superior to PL when the network size is small ($n = 40$).

With respect to the clustering performance measured by average RI and average VI, we note that the proposed mixture model yields satisfactory performance at absolute

Table 1: Estimation accuracy of $\hat{K}_0$, and average RI across 50 replicates for four experimental conditions (network size, cluster size), under the proposed mixture model (bayesERGMmix), and K-means clustering on MPLE, MLE and sufficient statistics (SS). The number of clusters in K-means clustering are based on the $\hat{K}_0$ estimated from bayesERGMmix (PL or APL, whichever yields higher accuracy of $\hat{K}_0$ in the respective experimental condition).

	Accuracy of $\hat{K}_0$		Average RI					Average VI				
	bayesERGMmix		bayesERGMmix		Kmeans			bayesERGMmix		Kmeans		
Condition	PL	APL	PL	APL	MPLE	MLE	SS	PL	APL	MPLE	MLE	SS
(40, 10)	0.70	0.78	0.975	0.961	0.909	0.923	0.828	0.115	0.123	0.467	0.360	0.956
(40, 30)	0.32	0.84	0.954	0.989	0.919	0.882	0.794	0.289	0.041	0.514	0.521	1.10
(100, 10)	0.86	0.80	0.983	0.954	0.982	0.954	0.969	0.060	0.133	0.064	0.133	0.145
(100, 30)	0.86	0.72	0.986	0.937	0.986	0.738	0.983	0.059	0.187	0.059	0.890	0.082

scale (average RI $>$ 95% for all settings except for APL under network size $=$ 100, cluster size $=$ 30). For K-means clustering (the number of clusters in K-means is set to the estimated number of mixture components from our mixture model in the corresponding replicate to facilitate a fair comparison), we find that K-means clustering on the sufficient statistics is uniformly inferior to K-means clustering on the parameter estimates (MPLE or MLE), and it is worth noting that the K-means clustering based on MPLE is better than that of MLE when the network size is large but worse when the network size is small. Our proposed mixture model has a clear advantage over K-means when the network size is small ($n = 40$) and this advantage becomes less significant as the sample size and network size increases. All these results confirm the numerical advantage of our proposed mixture model and also suggest that K-means may provide a useful initialization for our method.

4.4 Parameter Estimation

Figure 3 summarizes the relative bias of posterior mean estimate under different experimental conditions and likelihood approximation methods (PL and APL). We observe that the relative biases are small in general for both likelihood approximations, and decrease as sample size increases given the same network size. We also note that the relative biases are even smaller for large networks, which confirm the current understanding of PL as discussed in Section 3.2. Although it appears that the relative bias for the parameters associated with GWESP term estimated from APL is quite substantial at relative scale (about 20%) for the cluster "W" (i.e., cluster with weak dyadic dependent effects), it is very small at absolute scale as the true parameter value is -0.10.

Figure 4 shows the frequentist coverage of 95% posterior credible intervals under different experimental conditions and likelihood approximation methods. We notice that the posterior intervals estimated from APL are superior to those estimated from PL in terms of the coverage rates, especially for the parameters associated with the dyadic dependent term (i.e., GWESP, in this context). Also, we observed a fairly general pattern that the frequentist coverage of the posterior credible intervals yielded by APL approaches 95% as the sample size (cluster size) increases regardless of the network size. Therefore we have empirical evidence showing that the curvature correction step for APL functions well to combat the over-confidence issue caused by PL for the interval estimation of parameters associated with the dyadic dependent terms reported in Van Duijn et al. (2009).

In sum, PL can lead to good performance in terms of the point estimation but is lacking in the uncertainty quantification when the dyadic dependent effect is strong, which can be substantially mitigated, if not fully taken care of, by the use of APL.

4.5 Posterior Predictive Assessments

We consider four widely used metrics that characterize different aspects of graph structure and compare the mean values of these metrics calculated from posterior predictive distribution to those from the observed data. In practice, other metrics (such as

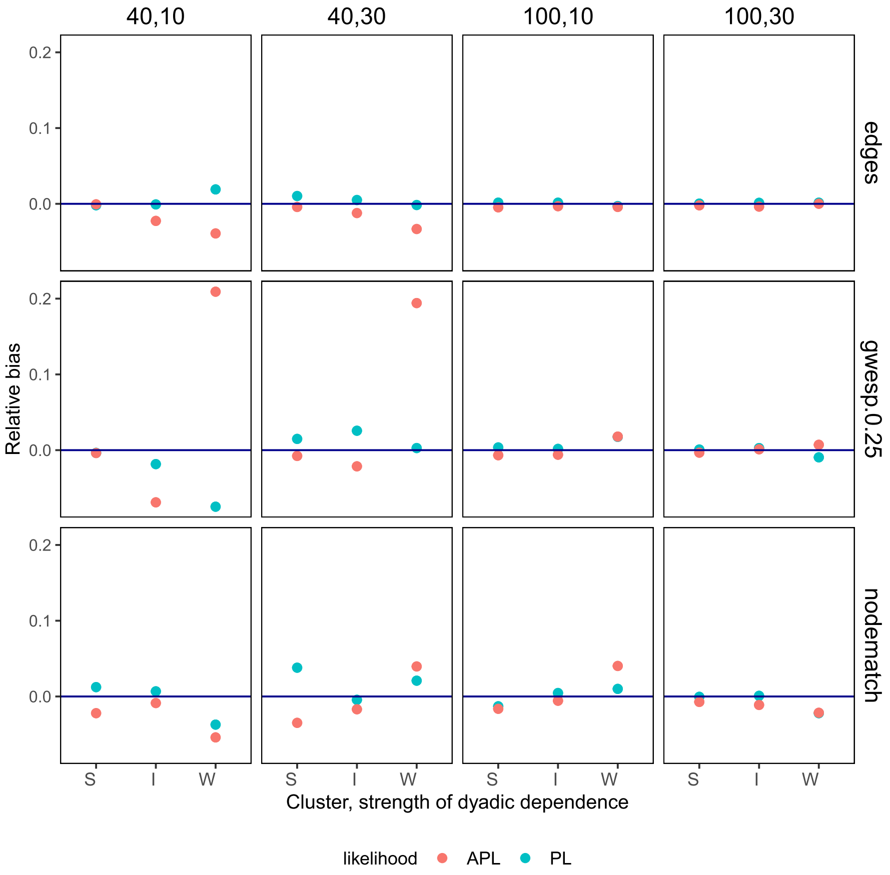

Figure 3: Relative bias. Three clusters are represented by the strength of dyadic dependent effects (S: Strong; I: Intermediate; W: Weak), which is GWESP with $\phi = 0.25$ in this case. The solid blue horizontal line corresponds to the target level: 0.

homophily and posterior predictive p-values) can also be used based on practitioner's domain knowledge and preference. A utility function (e.g. a weighted average of metrics) may also be used to summarize the posterior predictive checks.

Specifically, our selected graph-level metrics are

- Mean eigenvector centrality: the eigenvector centrality (EC) is a node-level metric that measures the degree of membership of a given node in the largest core/periphery structure in the graph, and we take mean eigenvector centrality among all nodes in the graph to convert it to a graph-level metric.[1] The eigenvector centrality

[1] Except in very rare cases for which the graph adjacency matrix lacks a principal eigenvalue. In such circumstances, eigenvector centrality is a signed indicator of membership in the two largest core/periphery structures (positive versus negative).

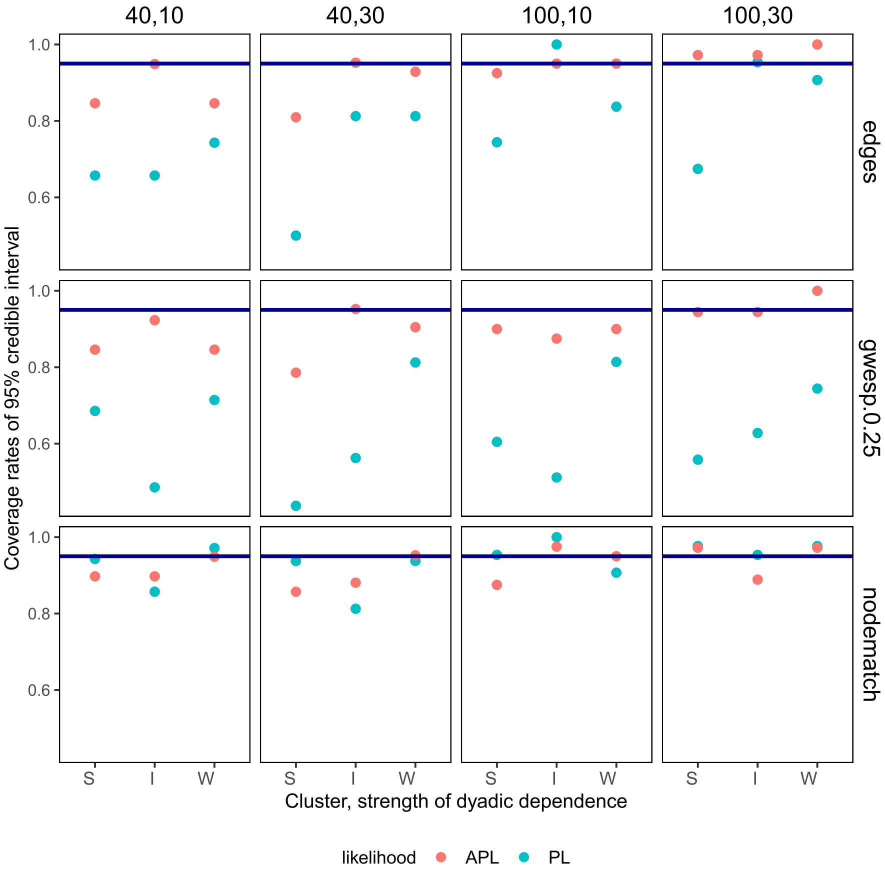

Figure 4: Frequentist coverage of 95% posterior credible intervals. Three clusters are represented by the strength of dyadic dependent effects (S: Strong; I: Intermediate; W: Weak), which is GWESP with $\phi = 0.25$ in this case. The solid blue horizontal line corresponds to the target level: 0.95.

is also the best one-dimensional approximation of the graph structure (in a least-squares sense), and accuracy in reproducing it indicates the extent to which the model is able to recover the broadest structural features of the graph.

- Transitivity: a standard measure of triadic closure in network analysis (Wasserman and Faust, 1994), defined as the ratio of complete triangles to all potentially complete triangles.
- Standard deviation of degree distribution: a measure of the level of heterogeneity in degree distribution.
- Mean of inverse path length: also known as the mean of geodesic distances, a measure of the overall closeness between nodes in a graph.

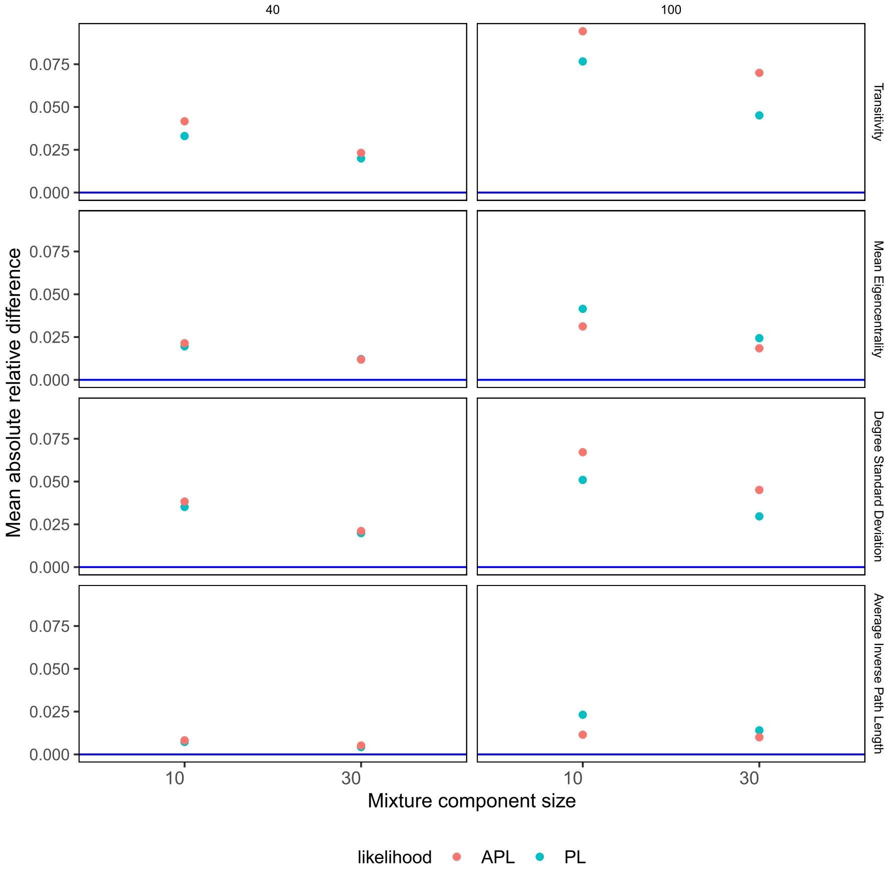

Figure 5: Mean absolute relative difference across 50 repetitions under each simulation setting.

As the above-listed graph-level metric is defined for each graph, we construct an ensemble-level metric out of each one of them by taking the mean, which yields a total of four ensemble-level metrics. When doing posterior predictive simulations, we first select 100 draws from the posterior samples and simulate an ensemble of networks that is of the same size as the synthetic data for each selected draw, and then we compare the posterior predictive data to the distribution giving rise to the synthetic data by calculating the mean absolute relative difference for those four ensemble-level metrics across 50 repetitions under each simulation setting (see Figure 5). We note that PL and APL yield very similar posterior predictive performance when the network size is small ($n = 40$) but their difference becomes considerable when the network size is large ($n = 100$) where we note that the posterior samples estimated from APL have an advantage in producing networks that are closer to the target distribution with respect to mean eigenvector centrality and average inverse path length, whereas the posterior samples estimated from PL have an advantage in producing networks that are closer

		Elapsed time			
Condition	Lik	APL calculation	MCMC	Total	Relative time
(40,10)	PL	0	43.13	43.13	1
(40,10)	APL	146.50 (5.48)	47.20	193.70 (52.68)	4.49 (1.22)
(40,30)	PL	0	116.73	116.73	1
(40,30)	APL	341.72 (4.75)	113.13	454.85 (117.88)	3.90 (1.01)
(100,10)	PL	0	45.73	45.73	1
(100,10)	APL	408.53 (16.23)	45.13	453.66 (61.36)	9.92 (1.34)
(100,30)	PL	0	126.85	126.85	1
(100,30)	APL	1082.412 (13.18)	119.28	1201.69 (132.46)	9.47 (1.04)

Table 2: Elapsed time for four experimental conditions (network size, cluster size) measured in minutes with PL and APL as approximations to true ERGM likelihoods. The "Relative time" column shows the computation time relative to that of PL under the same MCMC setting for the respective simulation setting. The numbers in parentheses in the "APL calculation" column are $\max_{i=1,\ldots,m} T_i^{APL}$ under the respective setting. The numbers in the parentheses in the "Total" and "Relative time" columns are calculated assuming the APL calculation is run in parallel on a machine with m computing cores.

to the target distribution with respect to transitivity and degree standard deviation. More importantly, regardless of the likelihood approximation used, there is a clear trend showing that the recovery of those ensemble-level metrics are becoming better as sample size (i.e., cluster size) increases for each fixed network size.

4.6 Computation Time

Table 2 shows a comparison on the computation time for PL and APL under four different experimental conditions with MCMC settings specified in Section 4.2 (60000 total iterations, initial number of mixture components $K = 6$, random initial values) using a single computing core (Intel Xeon E5-2699 v4 CPU @ 2.20GHz). As both PL and APL provide tractable approximations to true ERGM likelihoods, they are comparable in terms of the MCMC sampling time $T_{MCMC}^{APL} \approx T_{MCMC}^{PL}$ under the same MCMC settings. As the implementation of APL requires a computational overhead upfront as detailed in Section 3.2, the overall computation time under APL can be substantially larger than that under PL if only one computing core is available, but we note that multi-core architectures can be easily exploited to reduce the overall computation time under APL, as the calculations needed to obtain the values required for APL are independent for all networks. Therefore, given the APL calculation time as T_i^{APL} for the i-th network, $i = 1, \ldots, m$, and MCMC sampling time T_{MCMC}^{APL}, the total computation time under APL can be reduced from $\sum_{i=1}^{m} T_i^{APL} + T_{MCMC}^{APL}$ to $\max_{i=1,\ldots,m} T_i^{APL} + T_{MCMC}^{APL}$, if m computing cores are available. Therefore, the relative computation time can also be greatly reduced (as shown by the numbers in parentheses in Table 2). If more comput-

	edges	nodematch("D-D")	nodemix("D-R")	gwesp, $\phi = 0.25$
$\mathcal{M}_1$	✓	✓	✓	
$\mathcal{M}_2$	✓	✓	✓	✓, diff = F

Table 3: List of candidate model specifications ($\mathcal{M}_1$, $\mathcal{M}_2$) for co-voting networks. ✓ indicates the corresponding term is included in the respective model.

ing cores are available, we note that the computation time under APL can be further reduced by exploiting the fact that network samples needed at different temperatures (see Equation (1) in the supplementary material) can also be drawn in a completely independent manner.

5 Application to Political Co-Voting Networks

We apply the proposed method to cluster the co-voting networks among U.S. Senators from 1867 (start year of Congress 40) to 2014 (end year of Congress 113), which were first analyzed by Moody and Mucha (2013) using modularity and role-based blockmodels. The co-voting networks are constructed based on the roll call voting data from http://voteview.com, which contains the voting decision of each Senator (yay, nay, or abstain) for every bill brought to Congress.[2]

in the co-voting network represent Senators and an edge is placed between two nodes if the corresponding Senators vote concurrently (both yay or both nay) on at least 75% of the bills to which they were both present. Here we aim at identifying subgroups of networks that appear to have similar generating characteristics within the group but different characteristics across groups.

5.1 Model Specification and Estimation

Figure 6 shows that the co-voting networks vary in structure on different years, and the party-affiliation appears to be a key factor affecting the co-voting patterns among Senators. Therefore we consider two competing ERGM specifications in Table 3, where the statistic nodemix("D-R") counts the total number of edges between Democrats and Republicans $\sum_{i<j} y_{ij} \mathbb{1}_{\{\mathbf{x}_i = R, \mathbf{x}_j = D\}}$. We note that the interpretation of nodematch("D-D") and nodemix("D-R") should be relative to the baseline propensity of forming edges between Republicans under our model specifications.

We note that these networks vary in size (range: 69–112[3]) and thus include an offset term to adjust for network size, e.g., the Krivitsky reference measure discussed in Section 2.2, which provides a parameterization with constant baseline expected degree. We run long MCMC chains (150000 iterations with first 60% draw discarded as burn-in and a thinning interval of size 50) with $K = 6$ and random initial values using

[2] The data is publicly available online in the R package `VCERRGM`, https://github.com/jihuilee/VCERGM.

[3] Senators can leave the Congress in the middle of their interim and these seats can be filled by newly appointed/elected Senators, and hence causing the network size to be larger than 100 at times.

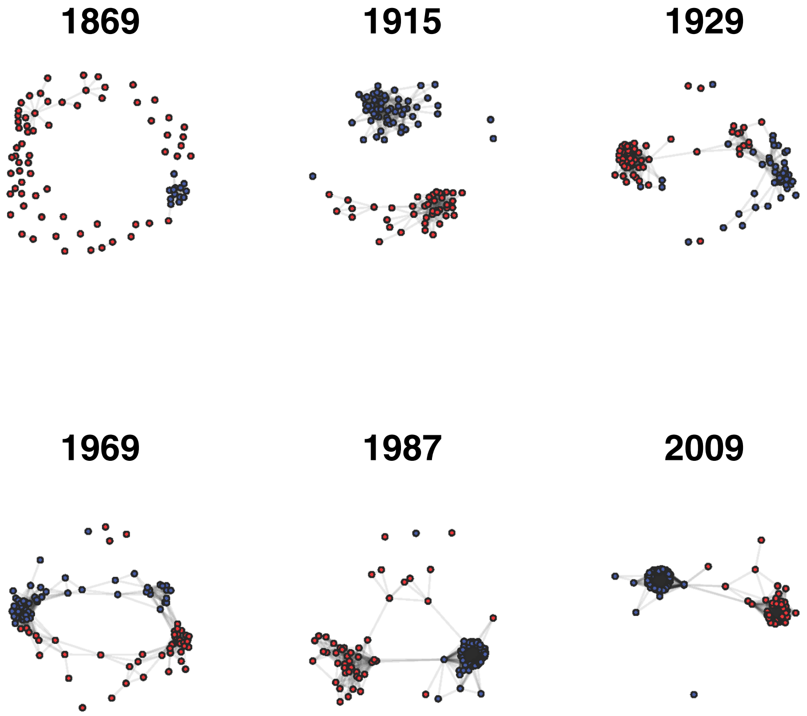

Figure 6: Co-voting networks of 41st, 64th, 71st, 91st, 100th and 111th Congress, which were formed in the year of 1869, 1915, 1929, 1969, 1987 and 2009, respectively. Colors indicate Senators' party affiliations, blue = Democrats(D), red = Republican(R).

the Dirichlet prior for $\boldsymbol{\tau}$ and multivariate normal priors for $\boldsymbol{\theta}_k$'s. We set the mean of the multivariate normal prior $\boldsymbol{\mu}$ to be $(1, \underbrace{0, \ldots, 0}_{p-1})$ to reflect the prior belief that the expected mean degree is larger than 1. Because the total number of cross-party edges is zero for some networks, we use the pseudolikelihood approximation for statistical inference as there is no finite maximizer of the likelihood when the observed statistics happen to lie on the relative boundary of the convex hull of possible values of the sufficient statistics (Handcock, 2003), causing issues for approximating the normalizing factor.

We calculate the DIC and BIC values for the candidate model specifications, which yields $DIC(\mathcal{M}_1) = 220240$, $DIC(\mathcal{M}_2) = 188177$; $BIC(\mathcal{M}_1) = 219869$, $BIC(\mathcal{M}_2) = 187806$; since both model selection criteria identify $\mathcal{M}_2$ as the better model specification, we conduct cluster-specific analysis based on it. Such a huge gap in DIC values is indicative of the substantial importance of incorporating triadic closure effects in modeling co-voting networks. The estimated number of non-empty mixture components $\hat{K}_0$ is 4 for $\mathcal{M}_2$ and we summarize the posterior mean and 95% credible intervals for cluster-specific parameters under $\mathcal{M}_2$ in Table 4.

	Cluster			
	1	2	3	4
edges	0.92 (0.27, 1.08)	1.91 (1.86, 1.97)	2.04 (1.99, 2.08)	2.47 (2.35, 2.58)
D-D	2.08 (1.89, 3.27)	−0.35 (−0.41, −0.23)	−0.12 (−0.15, −0.09)	0.92 (0.86, 0.99)
D-R	−1.91 (−2.29, −1.00)	−2.68 (−2.80, −2.57)	−3.09 (−3.17, −3.00)	−4.47 (−4.63, −4.30)
gwesp.0.25	0.99 (0.91, 1.24)	1.46 (1.37, 1.50)	2.14 (2.11, 2.18)	2.63 (2.54, 2.72)

Table 4: Posterior mean (95% credible interval) of cluster-specific parameters under $\mathcal{M}_2$.

We note that the size-invariant parameters for edge term (first column) can be interpreted as the log of the baseline mean degree (rather than the logit of the baseline density, as in the case of the counting measure), suggesting expected degrees varying from approximately $\exp(0.92) \approx 2.5$ to $\exp(2.47) \approx 11.8$ across clusters prior to consideration of other effects.

Based on these estimates, we see inhibition of cross-party ties (D-R row in Table 4) except for the first cluster and strong triadic closure (see gwesp.0.25 row in Table 4), as well as apparent qualitative distinctions between clusters. To probe the impact of the four behavioral regimes inferred from the co-voting data, we take a simulation-based approach which exploits a critical advantage of working with a fully generative model: the ability to perform "what-if" analyses that separate effects due to observed covariates from differences in structure arising from differences in generative processes. Specifically, we consider how the entire ensemble of Congressional networks would be expected to have been different, *if* each respective regime had governed the U.S. Congress for the entire study period. Such an analysis proceeds as follows. First, we simulate a set of posterior predictive networks for each Congress during the study period, with parameters drawn from the posterior distribution of each respective cluster. Each collection of networks can be thought of as a simulated "alternate history," in which the size and composition of each Congress were held to their real-world values but the behavioral tendencies that shaped the co-voting networks throughout the period were reflective of only one of the three clusters. Systematic differences in network structure across sets thus provide insight into the potential impact of behavioral regime, controlling for size and composition.

One important property that can be probed in this way is the expected incidence of voting coalitions, which play an important role in party politics. Here, we focus on minimal coalitions, defined as sets of three legislators who consistently vote together (i.e., triangles). Within-party coalitions can be sources of party cohesion, although they also act as blocks that can sometimes resist (and must be negotiated with by) party leaders; cross-party coalitions, by contrast, pose significant challenges to party cohesion, but can also serve as foci for sponsorship and promotion of bipartisan legislation. Both are hence significant, with distinct implications for the political landscape. To examine the coalition structures that would have been expected to occur under our three behavioral regimes, we simulate ten "alternate histories" from the posterior distributions of each cluster, calculating the realized proportions of intra-Democratic, intra-Republican, and inter-Party triangles. (That is, the counts of fully connected triads with all three members as Democrats, all three members as Republicans, or members from both parties,

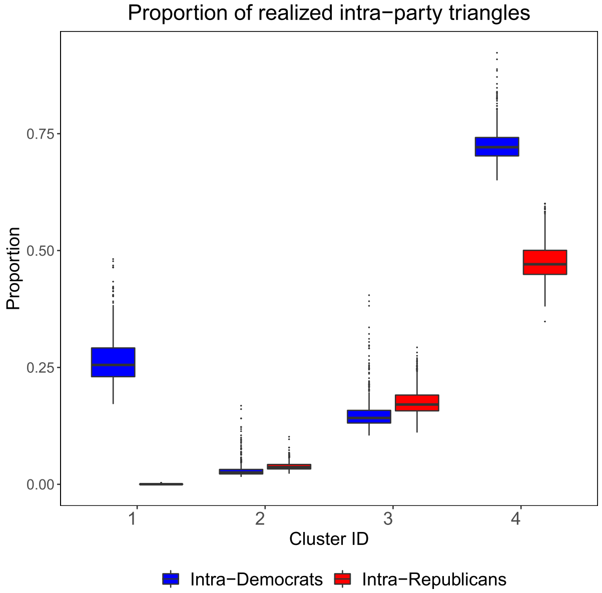

Figure 7: Proportion of realized intra-party triangles in simulated networks. Colors indicate the party affiliation (blue = Democrat (D), red = Republican (R)).

scaled by their maximum possible values.) Using proportions rather than raw counts ensures these metrics are normalized for network size and the distribution of party affiliations in each Congress; substantively, this choice of scaling tells us how close each party (or the cross-party cut) is to forming a perfect coalition, in which all members vote in concert. Figures 7 and 8 respectively show the realized proportion of intra-party and inter-party triangles in the simulated networks. Both figures show substantial differences in coalition structure, implying that the behavioral regimes associated with the three inferred clusters would be expected to have a meaningful impact on the political process. Specifically, we note the following:

- The regime of cluster 1 is marked by the extremely strong coalition of Democrats compared to Republicans.
- The key symbol of the regime of cluster 2 is marked by the formation of very few voting coalitions, either within party or between party.
- By contrast, the regimes of cluster 3 show a much higher incidence of intra-party coalition formation, with roughly 10–20% of the potential intra-party coalitions being present. Coalition incidence differs by party, with Republicans forming more coalitions in percentage versus Democrats. Interestingly, the regime 3 also shows the highest rate of cross-party coalition formation; while the rate is very low overall, it is considerably higher than all other clusters.

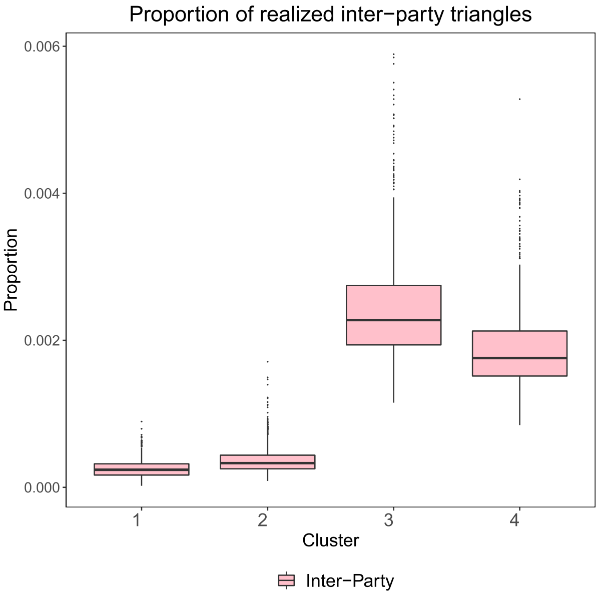

Figure 8: Proportion of realized inter-party triangles in simulated networks.

- Finally, the regime of cluster 4 favors extremely high levels of intra-party cohesion, with rates approaching 50% of the maximum possible for Republicans and 70% for Democrats. As this implies, the resulting networks are also highly asymmetric, with the Democratic party expected to generate a much more cohesive coalition structure than the Republicans. Interestingly, this strong intra-party coalition formation does not exist entirely at the expense of cross-party coalitions: we find an expected rate of cross-party coalition formation that is only slightly less than that expected for networks arising under cluster 3. That said, the much higher incidence of intra-party coalition formation under cluster 4 leads inter-party coalitions to be a smaller fraction of the total coalition set than under cluster 3, potentially making them less critical to the legislative process.

Taken together, these observations suggest that the cluster 2 regime tends to generate *uniformly loose* voting networks with very few coalitions of any kind. These networks may resist polarization, but their high level of fragmentation may make it more difficult to assemble the sorts of alliances needed to push through controversial legislation. By contrast, the regime of cluster 3 tends to produce *uniformly clustered* networks with moderately high levels of coalition formation in both parties coupled with relatively high numbers of cross-party coalitions. These networks may pose particular challenges for party leaders, as they contain a mix of multiple local coalitions that must be courted for votes, “lone wolves” outside of coalitions who must be approached individually, and likely defectors whose cross-party coalitions provide a bulwark against within-party

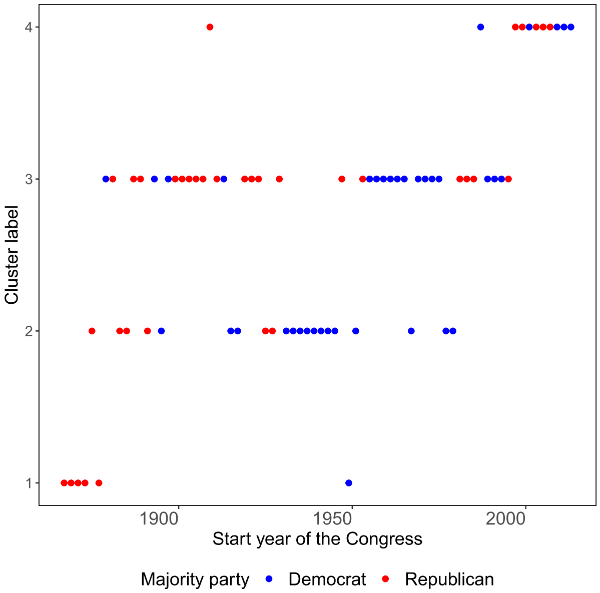

Figure 9: Maximum probability cluster assignments over study period. Colors indicate the majority party determined by the numbers of Senators in the dataset in the corresponding Congress (blue = Democratic (D), red = Republican (R)). Regimes of voting behavior are visibly correlated over time.

influence. The regime of cluster 4 tends to produce *party-cohesive* networks dominated by dense intra-party coalitions on both sides of the aisle (but with substantially higher levels of cohesion among Democratic legislators). This regime offers party leaders the greatest chance of being able to mobilize members in support of legislation, at the cost of potential legislative deadlock during periods of high inter-party conflict. Finally, the extreme behavior of Democrats in early years under the regime of cluster 1 might be partially attributed to the fact that the percentage share of Democratic legislators in the Congress is too low to allow them to be divisive (17% in 1867, 16% in 1869, 24% in 1871, 27% in 1873).

Figure 9 shows maximum probability cluster assignments over the study period, which enables us to connect the co-voting behaviors to different eras. We observe that the co-voting culture represented by cluster 1 is seen mostly at the beginning of the study period (1867, 1869, 1871, 1873 and 1877), the culture represented by cluster 2 and cluster 3 alternate in the late nineteenth and almost the entire twentieth centuries, while the culture represented by asymmetric polarization represented by cluster 4 becomes dominant after the late 1990's.

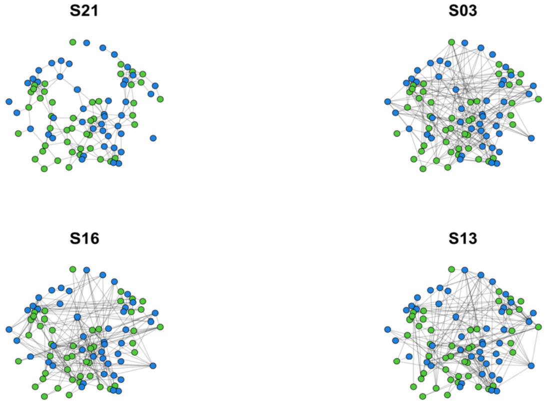

Figure 10: Brain connectivity networks of Subject No. 021, 003, 016, 013. Colours (green, blue) indicate the different hemispheres (left, right); coordinates of nodes are solely determined by the built-in algorithm of R package `network` (Butts, 2008) for the purpose of demonstration, without taking the actual 3D spatial structure into consideration.

6 Application to Brain Functional Connectivity Networks

In this section, we apply the proposed method to model an ensemble of brain functional connectivity networks that was previously studied in Simpson et al. (2011, 2012). The data were collected among 10 normal subjects (5 female, average age: 27.7 years old, standard deviation: 4.7 years) who were part (Subject No. 002, 003, 005, 008, 009, 010, 012, 013, 016, 021) of a larger functional magnetic resonance imaging (fMRI) studies of age-related changes in cross-modal deactivations (Peiffer et al., 2009). The nodes in brain functional connectivity networks correspond to 90 distinct anatomical regions of interest (ROI) defined by the Automated Anatomical Labeling atlas (AAL) (Tzourio-Mazoyer et al., 2002) and these 90 regions are divided symmetrically across left and right hemispheres (see Figure 10). The edges in these functional connectivity networks were constructed by thresholding the temporal correlation coefficient matrix adjusted for motion and physiological noise (see Hayasaka and Laurienti, 2010; Simpson et al., 2011, for further details) to discard the weak and non-significant links that may represent spurious connections. The thresholds were selected at subject level to make $\frac{\log(n)}{\log(\bar{d})} \approx 2.8$, or equivalently $\bar{d} \approx n^{1/2.8} \approx 5$ for each network, where $n = 90$ is the total number of nodes and $\bar{d}$ is the average degree.

6.1 Model Specification

With the edge statistic controlling the baseline propensity of forming edges, Simpson et al. (2011) proposed to use geometrically weighted edgewise shared partner (GWESP)

	edges	gwesp, $\phi = 0.25$	gwnsp, $\phi = 0.25$	nodematch(“Hemisphere”)
$\mathcal{M}_1$	✓	✓	✓	
$\mathcal{M}_2$	✓	✓	✓	✓, diff = F

Table 5: List of candidate model specifications ($\mathcal{M}_1$, $\mathcal{M}_2$) for brain functional connectivity networks. ✓ indicates the corresponding term is included in the respective model.

statistic and geometrically weighted non-edgewise shared partner (GWNSP) statistic[4] to capture the two most important structural features in brain functional connectivity networks, namely, functional segregation and functional integration (Rubinov and Sporns, 2010), where the functional segregation in the brain refers to the ability for specialized processing to occur within densely interconnected groups of brain regions, and the functional integration in the brain means the ability to rapidly combine specialized information from distributed brain regions. In the context of network modeling perspective, the former favors the graphs with more local clusters (local efficiency), whereas the latter requires the presence of bridging edges connecting the local clusters to make different regions in the whole brain coordinate more efficiently (global efficiency). Building upon this seminal work, Sinke et al. (2016) added a nodematch term with respect to hemisphere (i.e., total number of edges connecting nodes in the same hemisphere) to capture the additional propensity of forming edges between nodes in the same hemisphere. Following the choices in literature, we consider two competing model specifications (Table 5).

For each candidate model specification, we use adjusted pseudolikelihood and run long MCMC chains (100000 total iterations where first 60% of draws are discarded as burn-in and a thinning interval of size 100) with $K = 4$ and prior specified in Section 3. The estimated number of mixture components $\hat{K}_0$ is 2 for both model specifications. We calculate the DIC and BIC values for candidate model specifications and find $DIC(\mathcal{M}_1) = 25997.5$, $DIC(\mathcal{M}_2) = 26873.2$; $BIC(\mathcal{M}_1) = 25999.2$, $BIC(\mathcal{M}_2) = 26835.2$, therefore we select $\mathcal{M}_1$ for subsequent analysis (see online supplementary material for the results of posterior predictive checks with respect to two key characteristics of primary interests, global efficiency and local efficiency). We note that our mixture model provides a richer framework for assessing the heterogeneity in brain functional connectivity networks than was feasible in prior work such as Simpson et al. (2011) and Sinke et al. (2016), in which each individual network was fitted to ERGMs separately followed by a direct comparison on the coefficient values across prespecified subgroups (e.g., age).

[4] $g(\mathbf{y}) = e^{\phi} \sum_{k=1}^{n-2} \left\{1 - (1 - e^{-\phi})^k\right\} NP_k(\mathbf{y})$, geometrically weighted non-edgewise shared partner (GWNSP). Here $NP_k(\mathbf{y})$ is the number of non-connected pairs that have exactly k common neighbors. The decay parameter ϕ controls the relative contribution of $NP_k(\mathbf{y})$ to the GWNSP statistic, and it is fixed at 0.25 in this case.

	Cluster 1	Cluster 2
edges	-2.53 (-2.74, -2.35)	-1.47 (-1.79, -1.18)
gwesp, $\phi = 0.25$	1.39 (1.27, 1.54)	0.90 (0.70, 1.11)
gwnsp, $\phi = 0.25$	-0.29 (-0.31, -0.28)	-0.41 (-0.43, -0.38)

Table 6: Posterior mean (95% credible interval) of cluster-specific parameters under $\mathcal{M}_1$.

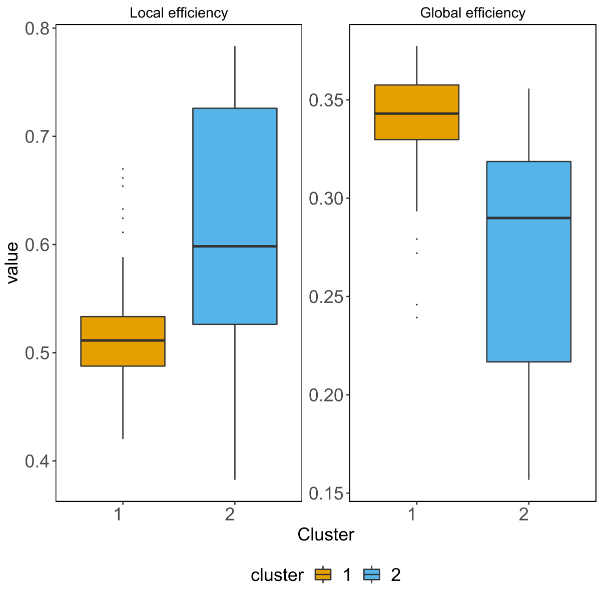

Figure 11: Global efficiency and local efficiency for networks simulated from cluster 1 and cluster 2.

6.2 Results

Table 6 presents the posterior means and 95% credible intervals for cluster-specific parameters and we run simulations to compare the difference in local efficiency and global efficiency between the networks represented by these two clusters. Specifically, we select 100 draws from the retained posterior samples for each cluster and simulate a network for each selected draw, which yields 100 representative networks for each cluster. We calculate the global efficiency and local efficiency (see Rubinov and Sporns, 2010, for details on these metrics) of these simulated networks using function `efficiency` in R package `brainGraph` (Watson, 2020). Figure 11 shows that the networks represented by cluster 1 exhibits higher level of global efficiency (functional aggregation) while those represented by cluster 2 shows higher level of local efficiency (functional segregation).

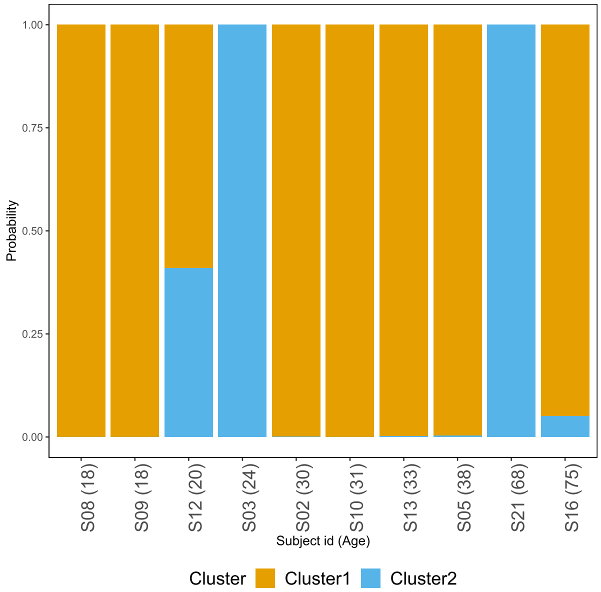

Figure 12: Posterior probability of cluster membership. The x-axis gives the subject id (age) order by age.

Figure 12 shows the posterior probability of cluster membership with subject id ordered by age. Despite the fact that several previous published analyses (Simpson et al., 2011; Sinke et al., 2016) were concerned with comparing the brain functional connectivity networks between younger and older adults, we do not a observe consistent pattern between cluster membership and age from our clustering results; this suggests that, while individual differences are present, they may not correspond to *a priori* obvious characteristics. The ability to discover such patterns (rather than to have to specify them *ex ante*) is an advantage of this approach.

7 Conclusion

This paper develops a mixture of ERGMs approach for modeling the generative process leading to heterogeneous network ensembles. We propose a Metropolis-within-Gibbs algorithm to perform Bayesian inference, and estimate the number of mixture components based on a deliberately overfitting model with sparse priors that allow us to eliminate excess components during MCMC. To deal with the intractable likelihoods of ERGMs with complex dependence, we consider two tractable approximations (pseudolikelihood (PL) and adjusted pseudolikelihood (APL)) to facilitate efficient statistical inference. Simulation studies show that both PL and APL are able to provide accurate point estimates of cluster-specific parameters but the latter provides better quantification of the uncertainties. With respect to clustering performance, the APL appears to be more robust in terms of identifying the true number of clusters across different network sizes and sample sizes, while the PL improves substantially when the network size is large. Although the computation time of the APL is substantially larger than that of PL if only one computing core is available, this disadvantage can be greatly reduced when multi-core architectures are available.

In our model, we choose to use small values of α to promote a small number of clusters for parsimonious estimation and to maintain the desired theoretical properties of the resulting posterior distributions. This idea has been previously explored in Malsiner-Walli et al. (2016) and a theoretical justification is given in Rousseau and Mengersen (2011). However, those two references mainly focus on regular likelihood function and it remains unclear how those results can be extended to PL and APL for complex models such as ERGM. Investigating the impact of α on the posterior distribution under the use of PL and APL is an interesting future working direction.

We apply the proposed approach to study the political co-voting networks among U.S. Senators with size-adjusted parameterizations being used to account for the difference in the size of the observed networks, and identify four clusters that represent very different co-voting patterns. After matching the clusters with temporal information, we observe one extremely asymmetric co-voting pattern under which the Democrats are extremely cohesive, one roughly symmetric co-voting pattern with very few voting coalitions and one mildly asymmetric co-voting patterns under which the Republicans are more cohesive that alternated in late nineteenth and almost the entire twentieth century, and an abrupt shift in the co-voting pattern towards the direction of political party polarization in last two decades. As the regimes of voting behavior are visibly correlated over time according to our analysis, further analysis of this data can leverage dynamic network models, which is a whole area onto itself, to capture the temporal dependency at the node level (i.e., same Senator incumbent for multiple Congresses) as well as at the graph level (i.e., neighboring Congresses are more likely to hold similar co-voting culture).

The application to brain functional connectivity networks showcases that the proposed approach can offer a more principled alternative to existing methods, which fit each network to the ERGM separately, for modeling the heterogeneity in brain functional connectivity networks. Compared to other methods in the literature, our method allows principled statistical inference for the generative processes of heterogeneous ensembles of networks.

In closing, we comment on four important directions for future research that could prove beneficial to the modeling of ensembles of networks. It is worth noting that the sizes of the US congresses between 1867 and 2014 range from 69 to 112, non-identical but broadly similar. More importantly, these size changes occur within a social system whose basic structure remains fairly similar throughout the time period. In other cases, however, large size differences may be accompanied by increasingly complex internal barriers to interaction or other additional exogenous structure that must be accounted for to obtain realistic predictions. While this additional structure is not available in the form of covariates, more sophisticated size-adjusted parameterizations may be required; reference measures or other tools facilitating "automatic" correction of such effects would facilitate mixture modeling in such scenarios. With respect to likelihood calculation, it is encouraging that we obtain favorable results in terms of clustering performance when the network size is large and point estimation in general in our simulation study using the easily computed pseudo-likelihood approximation, and adjusted pseudolikelihood can provide us with better uncertainty estimates and more robust clustering performance in general but at higher computational cost. Although the additional computational costs accompanied by the use of adjusted pseudolikelihood (APL) can be reduced with the use of multi-core architectures, it is of critical interest to identify the *optimal* settings beyond which there is diminishing return for the path sampling step involved in the APL calculation. A natural further extension of the finite mixture modeling framework could be Dirichlet Process mixtures (DPM) or mixture of finite mixtures (MFM) of ERGMs where the number of mixture components can vary depending on the incoming data size. Although computationally challenging, such an extension can provide a highly flexible-yet-interpretable density estimation framework for complex graph distributions. Last but not least, the developments of more scalable inference algorithms (e.g., variational inference) are favorable for handling data of large size.

Supplementary Material

Web-based Supplementary File for "Finite Mixtures of ERGMs for Modeling Ensembles of Networks" (DOI: 10.1214/21-BA1298SUPP; .pdf). The supplementary material contains details about computational details for adjusted pseudolikelihood, selecting the prior hyperparameters, summary of the post MCMC inference for simulation studies and posterior predictive assessments for data applications.

References

Amati, V., Mol, A., Shafie, T., Hofman, C., and Brandes, U. (2019). "A Framework for Reconstructing Archaeological Networks Using Exponential Random Graph Models." *Journal of Archaeological Method and Theory*, 1–28. 1156

Banks, D. and Carley, K. M. (1994). "Metric Inference for Social Networks." *Journal of Classification*, 11(1): 121–149. 1154

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236. 1157

Bouranis, L., Friel, N., and Maire, F. (2017). "Efficient Bayesian inference for exponential random graph models by correcting the pseudo-posterior distribution." *Social Networks*, 50: 98–108. 1157, 1159

Bouranis, L., Friel, N., and Maire, F. (2018). "Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods." *Journal of Computational and Graphical Statistics*, 27(3): 516–528. MR3863754. doi: https://doi.org/10.1080/10618600.2018.1448832. 1158, 1161

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*, volume 50. Cambridge University Press. MR3967046. doi: https://doi.org/10.1017/9781108644181. 1155

Butts, C. T. (2008). "network: a Package for Managing Relational Data in R." *Journal of Statistical Software*, 24(2): 1–36. 1179

Butts, C. T. (2011). "Bayesian meta-analysis of social network data via conditional uniform graph quantiles." *Sociological Methodology*, 41(1): 257–298. 1154

Butts, C. T. (2017). "Baseline Mixture Models for Social Networks." *arXiv preprint* arXiv:1710.02773. 1154

Butts, C. T. and Almquist, Z. W. (2015). "A flexible parameterization for baseline mean degree in multiple-network ERGMs." *The Journal of Mathematical Sociology*, 39(3): 163–167. MR3367715. doi: https://doi.org/10.1080/0022250X.2014.967851. 1157

Butts, C. T. and Carley, K. M. (2005). "Some Simple Algorithms for Structural Comparison." *Computational and Mathematical Organization Theory*, 11(4): 291–305. 1154

Caimo, A. and Friel, N. (2011). "Bayesian inference for exponential random graph models." *Social Networks*, 33(1): 41–55. MR2873466. 1157, 1159, 1160

Caimo, A. and Friel, N. (2014). "Bergm: Bayesian Exponential Random Graphs in R." *Journal of Statistical Software, Articles*, 61(2): 1–25. URL https://www.jstatsoft.org/v061/i02 1158

Caimo, A., Pallotti, F., and Lomi, A. (2017). "Bayesian exponential random graph modelling of interhospital patient referral networks." *Statistics in Medicine*, 36(18): 2902–2920. MR3670398. doi: https://doi.org/10.1002/sim.7301. 1156

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1(4): 651–673. MR2282197. doi: https://doi.org/10.1214/06-BA122. 1163

Cranmer, S. J. and Desmarais, B. A. (2011). "Inferential network analysis with exponential random graph models." *Political Analysis*, 19(1): 66–86. 1156

Desmarais, B. A. and Cranmer, S. J. (2010). "Consistent confidence intervals for maximum pseudolikelihood estimators." In *Proceedings of the Neural Information Processing Systems 2010 Workshop on Computational Social Science and the Wisdom of Crowds*. Citeseer. 1161

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2017). "Nonparametric Bayes modeling of populations of networks." *Journal of the American Statistical Association*, 112(520): 1516–1530. MR3750873. doi: https://doi.org/10.1080/01621459.2016.1219260. 1154

Faust, K. and Skvoretz, J. (2002). "Comparing networks across space and time, size and species." *Sociological Methodology*, 32(1): 267–299. 1154

Fitzhugh, S. M., Pixley, J. E., and Butts, C. T. (2015). "A Life History Graph Approach to the Analysis and Comparison of Life Histories." *Advances in Life Course Research*, 25: 16–34. 1154

Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97(458): 611–631. MR1951635. doi: https://doi.org/10.1198/016214502760047131. 1155

Frank, O. and Strauss, D. (1986). "Markov graphs." *Journal of the American Statistical Association*, 81(395): 832–842. MR0860518. 1156

Fritsch, A. and Ickstadt, K. (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, 4(2): 367–391. MR2507368. doi: https://doi.org/10.1214/09-BA414. 1164

Frühwirth-Schnatter, S. (2001). "Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models." *Journal of the American Statistical Association*, 96(453): 194–209. MR1952732. doi: https://doi.org/10.1198/016214501750333063. 1159

Gelman, A. and Meng, X.-L. (1998). "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." *Statistical Science*, 163–185. MR1647507. doi: https://doi.org/10.1214/ss/1028905934. 1162

Geweke, J. (2007). "Interpretation and inference in mixture models: Simple MCMC works." *Computational Statistics & Data Analysis*, 51(7): 3529–3550. MR2367818. doi: https://doi.org/10.1016/j.csda.2006.11.026. 1159

Geyer, C. J. and Thompson, E. A. (1992). "Constrained Monte Carlo maximum likelihood for dependent data." *Journal of the Royal Statistical Society. Series B (Methodological)*, 657–699. MR1185217. 1157

Goodreau, S. M. (2007). "Advances in exponential random graph (p*) models applied to a large social network." *Social Networks*, 29(2): 231–248. 1155

Grazioli, G., Martin, R. W., and Butts, C. T. (2019). "Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods." *Frontiers in Molecular Biosciences*, 6: 42. 1153, 1156, 1161

Groendyke, C., Welch, D., and Hunter, D. R. (2012). "A network-based analysis of the 1861 Hagelloch measles data." *Biometrics*, 68(3): 755–765. MR3055180. doi: https://doi.org/10.1111/j.1541-0420.2012.01748.x. 1156

Handcock, M. S. (2003). "Assessing Degeneracy in Statistical Models of Social Networks." Technical report, Center for Statistics and Social Sciences, University of Washington. URL https://www.csss.washington.edu/Papers/wp39.pdf. 1157, 1174

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). "statnet: Software tools for the representation, visualization, analysis and simulation of network data." *Journal of Statistical Software*, 24(1): 1548. 1166

Hayasaka, S. and Laurienti, P. J. (2010). "Comparison of characteristics between region- and voxel-based network analyses in resting-state fMRI data." *NeuroImage*, 50(2): 499–508. 1179

Hinton, G. E. (2002). "Training products of experts by minimizing contrastive divergence." *Neural Computation*, 14(8): 1771–1800. 1157

Hjort, N. L. and Claeskens, G. (2003). "Frequentist model average estimators." *Journal of the American Statistical Association*, 98(464): 879–899. MR2041481. doi: https://doi.org/10.1198/016214503000000828. 1154

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical Science*, 382–401. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 1154

Holland, P. W. and Leinhardt, S. (1981). "An exponential family of probability distributions for directed graphs." *Journal of the American Statistical Association*, 76(373): 33–50. MR0608176. 1156

Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012). "Improving simulation-based algorithms for fitting ERGMs." *Journal of Computational and Graphical Statistics*, 21(4): 920–939. MR3005804. doi: https://doi.org/10.1080/10618600.2012.679224. 1157

Hunter, D. R. and Handcock, M. S. (2006). "Inference in curved exponential family models for networks." *Journal of Computational and Graphical Statistics*, 15(3): 565–583. MR2291264. doi: https://doi.org/10.1198/106186006X133069. 1156, 1157

Ishwaran, H., James, L. F., and Sun, J. (2001). "Bayesian model selection in finite mixtures by marginal density decompositions." *Journal of the American Statistical Association*, 96(456): 1316–1332. MR1946579. doi: https://doi.org/10.1198/016214501753382255. 1162

Kolaczyk, E. D. and Krivitsky, P. N. (2015). "On the question of effective sample size in network modeling: an asymptotic inquiry." *Statistical Science*, 30(2): 184. MR3353102. doi: https://doi.org/10.1214/14-STS502. 1157

Koskinen, J. (2004). "Bayesian analysis of exponential random graphs-estimation of parameters and model selection." Technical report, Research Report 2004: 2, Department of Statistics, Stockholm University. 1160

Koskinen, J. H. (2008). "The linked importance sampler auxiliary variable Metropolis Hastings algorithm for distributions with intractable normalising constants." *MelNet Social Networks Laboratory Technical Report*, 08–01. 1160

Krivitsky, P. N. (2017). "Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models." *Computational Statistics & Data Analysis*, 107: 149–161. MR3575065. doi: https://doi.org/10.1016/j.csda.2016.10.015. 1157

Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011). "Adjusting for network size and composition effects in exponential-family random graph models." *Statistical Methodology*, 8(4): 319–339. MR2800354. doi: https://doi.org/10.1016/j.stamet.2011.01.005. 1157

Lehmann, B., Henson, R., Geerligs, L., White, S., et al. (2021). "Characterising group-level brain connectivity: a framework using Bayesian exponential random graph models." *NeuroImage*, 225: 117480. 1155

Leisch, F. (2006). "A toolbox for K-centroids cluster analysis." *Computational Statistics & Data Analysis*, 51(2): 526–544. MR2297469. doi: https://doi.org/10.1016/j.csda.2005.10.006. 1162

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and Computing*, 26(1-2): 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 1155, 1158, 1162, 1183

McFarland, D. A., Moody, J., Diehl, D., Smith, J. A., and Thomas, R. J. (2014). "Network Ecology and Adolescent Social Structure." *American Sociological Review*, 79(6): 1088–1121. 1154

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. M. Dekker New York. MR0926484. 1155

Meilă, M. (2007). "Comparing clusterings—an information based distance." *Journal of Multivariate Analysis*, 98(5): 873–895. MR2325412. doi: https://doi.org/10.1016/j.jmva.2006.11.013. 1164

Moody, J. and Mucha, P. J. (2013). "Portrait of political party polarization." *Network Science*, 1(1): 119–121. 1153, 1173

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). "Specification of exponential-family random graph models: terms and computational aspects." *Journal of Statistical Software*, 24(4): 1548. 1156

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for Doubly-intractable Distributions." In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, 359–366. Arlington, Virginia, United States: AUAI Press. URL http://dl.acm.org/citation.cfm?id=3020419.3020463 1160

Nobile, A. et al. (2004). "On the posterior distribution of the number of components in a finite mixture." *The Annals of Statistics*, 32(5): 2044–2073. MR2102502. doi: https://doi.org/10.1214/009053604000000788. 1162

Obando, C. and De Vico Fallani, F. (2017). "A statistical model for brain networks inferred from large-scale electrophysiological signals." *Journal of The Royal Society Interface*, 14(128): 20160940. 1153, 1156

Peiffer, A. M., Hugenschmidt, C. E., Maldjian, J. A., Casanova, R., Srikanth, R., Hayasaka, S., Burdette, J. H., Kraft, R. A., and Laurienti, P. J. (2009). "Aging and the interaction of sensory cortical function and structure." *Human Brain Mapping*, 30(1): 228–240. 1179

Pflug, G. C. (1996). *Optimization of Stochastic Models. The Interface Between Simulation and Optimization*. Boston: Kluwer Academic. MR1492446. doi: https://doi.org/10.1007/978-1-4613-1449-3. 1157

Pržulj, N. (2007). "Biological network comparison using graphlet degree distribution." *Bioinformatics*, 23(2): e177–e183. doi: https://doi.org/10.1093/bioinformatics/btl301. 1154

R Core Team (2019). *R: A Language and Environment for Statistical Computing.*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ 1166

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association*, 66(336): 846–850. 1164

Robbins, H. and Monro, S. (1951). "A stochastic approximation method." *The Annals of Mathematical Statistics*, 400–407. MR0042668. doi: https://doi.org/10.1214/aoms/1177729586. 1157

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 1183

Rubinov, M. and Sporns, O. (2010). "Complex network measures of brain connectivity: uses and interpretations." *NeuroImage*, 52(3): 1059–1069. 1180, 1181

Salter-Townshend, M. and Murphy, T. B. (2015). "Role Analysis in Networks using Mixtures of Exponential Random Graph Models." *Journal of Computational and Graphical Statistics*, 24(2): 520–538. MR3357393. doi: https://doi.org/10.1080/10618600.2014.923777. 1155

Saul, Z. M. and Filkov, V. (2007). "Exploring biological network structure using exponential random graph models." *Bioinformatics*, 23(19): 2604–2611. 1155

Schmid, C. S. and Desmarais, B. A. (2017). "Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap." In *2017 IEEE International Conference on Big Data (Big Data)*, 116–121. IEEE. 1161

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6(2): 461–464. MR0468014. 1163

Schweinberger, M. and Handcock, M. S. (2015). "Local dependence in random graph models: characterization, properties and statistical inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3): 647–676. MR3351449. doi: https://doi.org/10.1111/rssb.12081. 1155

Schweinberger, M., Krivitsky, P. N., Butts, C. T., Stewart, J. R., et al. (2020). "Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios." *Statistical Science*, 35(4): 627–662. MR4175389. doi: https://doi.org/10.1214/19-STS743. 1154, 1156, 1157

Signorelli, M. and Wit, E. C. (2020). "Model-based clustering for populations of networks." *Statistical Modelling*, 20(1): 9–29. MR4052400. doi: https://doi.org/10.1177/1471082X19871128. 1155

Simpson, S. L., Hayasaka, S., and Laurienti, P. J. (2011). "Exponential random graph modeling for complex brain networks." *PLoS ONE*, 6(5): e20039. 1153, 1156, 1179, 1180, 1182

Simpson, S. L., Moussa, M. N., and Laurienti, P. J. (2012). "An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks." *NeuroImage*, 60(2): 1117–1126. 1179

Sinke, M. R., Dijkhuizen, R. M., Caimo, A., Stam, C. J., and Otte, W. M. (2016). "Bayesian exponential random graph modeling of whole-brain structural networks across lifespan." *NeuroImage*, 135: 79–91. 1156, 1180, 1182

Slaughter, A. J. and Koehly, L. M. (2016). "Multilevel models for social networks: hierarchical Bayesian approaches to exponential random graph modeling." *Social Networks*, 44: 334–345. 1153, 1154

Snijders, T. A. (2002). "Markov chain Monte Carlo estimation of exponential random graph models." *Journal of Social Structure*, 3(2): 1–40. 1157

Snijders, T. A. and Nowicki, K. (1997). "Estimation and prediction for stochastic blockmodels for graphs with latent block structure." *Journal of classification*, 14(1): 75–100. MR1449742. doi: https://doi.org/10.1007/s003579900004. 1155

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). "New specifications for exponential random graph models." *Sociological Methodology*, 36(1): 99–153. 1156

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 1163

Stewart, J., Schweinberger, M., Bojanowski, M., and Morris, M. (2019). "Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms." *Social Networks*, 59: 98–119. 1153, 1154

Strauss, D. and Ikeda, M. (1990). "Pseudolikelihood estimation for social networks." *Journal of the American Statistical Association*, 85(409): 204–212. MR1137368. 1157

Sweet, T. M., Flynt, A., and Choi, D. (2019). "Clustering ensembles of social networks." *Network Science*, 1–19. 1154

Tan, L. S. and Friel, N. (2020). "Bayesian variational inference for exponential random graph models." *Journal of Computational and Graphical Statistics*, 1–19. MR4191251. doi: https://doi.org/10.1080/10618600.2020.1740714. 1157, 1161

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain." *NeuroImage*, 15(1): 273–289. 1179

Unhelkar, M. H., Duong, V. T., Enendu, K. N., Kelly, J. E., Tahir, S., Butts, C. T., and Martin, R. W. (2017). "Structure prediction and network analysis of chitinases from the Cape Sundew, Drosera capensis." *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(3): 636–643. 1153

Van Duijn, M. A., Gile, K. J., and 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. 1161, 1168

Wade, S. and Ghahramani, Z. (2018). "Bayesian cluster analysis: Point estimation and credible balls (with discussion)." *Bayesian Analysis*, 13(2): 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 1164

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press. 1170

Wasserman, S. and Pattison, P. (1996). "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*." *Psychometrika*, 61(3): 401–425. MR1424909. doi: https://doi.org/10.1007/BF02294547. 1156

Watson, C. G. (2020). *brainGraph: Graph Theory Analysis of Brain MRI Data*. R package version 3.0.0. URL https://CRAN.R-project.org/package=brainGraph 1181

Yin, F., Shen, W., and Butts, C. T. (2021). "Web-based Supplementary File for "Finite Mixtures of ERGMs for Modeling Ensembles of Networks"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1298SUPP. 1162

Zijlstra, B. J., Van Duijn, M. A., and Snijders, T. A. (2006). "The multilevel p2 model." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1): 42–47. 1154

Acknowledgments

The authors thank the editor, the associate editor, and anonymous reviewers for their valuable comments and suggestions that considerably improved the paper. The authors also thank Dr. Jihui Lee for making the co-voting network data easily accessible via R package `VCERGM` on GitHub. The authors are grateful to Dr. Sean L. Simpson for sharing brain functional connectivity networks data.