UNIVERSITY OF CALIFORNIA,
IRVINE

Advances in Exponential-family Random Graph Models: Computation, Model Selection, and Methodology

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Fan Yin

Dissertation Committee:
Professor Carter T. Butts, Chair
Professor Michele Guindani
Assistant Professor Weining Shen
Professor Hal S. Stern

2020

DEDICATION

To my parents, Lan Yin and Qian Zhou, whose love is the foundation of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Carter Butts, for all of his advice, encouragement, and wisdom in my journey of doctoral studies. Carter's meticulous nature in all aspects of research has set up a great example for me to follow. Carter has always been willing to talk through the issues and obstacles I encountered, offering hands-on guidance and helping me grow professionally. There are many brilliant advice and bright moments that I shall forever remember, and in particular, I found this one to be extraordinarily enlightening – "It took hundreds of thousands if not millions of years for human beings to develop a civilization as it is today. As a graduate student, you should take nothing for granted, because you have the responsibility and the privilege to help pass on the torch of human civilization." He has made me, and many others, better scientists.

I would also like to thank the members of my dissertation committee, Professor Hal Stern, Professor Michele Guindani, Professor Weining Shen, and one additional member from my advancement committee, Professor Katherine Faust, for their insightful suggestions and advice.

I am grateful and proud to have many great collaborators. Thanks to Professor Rachel Martin, Professor Gianmarc Grazioli, Dr. Domarin Khago for many fruitful discussions on ideal gas law and light scattering experiments. I also want to thank Dr. Nolan Phillips for his remarkable contributions on the held-out predictive evaluations (HoPE) project. Also many thanks to all the members of our SARS-CoV-2 Command Center, Loring Thomas, Peng Huang, Xiaoshuang Luo, Professor Zack Almquist, Professor John Hipp and Professor Carter Butts, together we spared no time to make our contributions to fight SARS-CoV-2 by publishing the paper "Spatial Heterogeneity Can Lead to Substantial Local Variations in COVID-19 Timing and Severity" on PNAS, which can be highly valuable to policy makers.

I also owe a lot of thanks to all members of the Networks, Computation, and Social Dynamics (NCASD) Lab, for all the helpful discussions and fun conversations.

The funds are crucial for allowing me to fully concentrate on my research, and I would like to acknowledge the assistance I have received during my graduate studies. The works presented in this thesis are based on research supported by NSF awards IIS-1526736, DMS-1361425, and IIS-1939237.

On the personal side, I would like to thank my fellow friends, Yiwei Wang, Albert Jiang, Junjie Shen, Nolan Phillips, Luming Chen, Chenzhe Tian, Biao Yang, Yannan Tang, Di Zhang, Lechuan Hu, Mingwei Tang, Tong Shen, Yadong Lu, Rui Xiang, Gongjin Sun, Zhengli Zhao, Zhaohao Zhou, Bingchen Yu, Meng Zhao, Huanqiang Cao, Chen Yang, Zhuoran Zhang, Shiyao Zhang, Jiashu Ren, Yue Wang, Fei Huang, Chengyi Xu, and Francesco Denti. You all make my life at UCI colorful.

Mostly, to my girlfriend, Maozhu Dai, who loved and supported me during my most difficult

times. I cannot imagine a life without you.

Finally, I want to dedicate this thesis to my mom and dad, as well as the rest of my family, who always stand behind me and encourage me to pursue my dreams. With their unconditional love and support, I am able to make it to the end and proudly present my PhD work.

VITA

Fan Yin

EDUCATION

Doctor of Philosophy in Statistics **2015-2020**
University of California, Irvine *Irvine, CA, USA*

Bachelor of Science in Statistics **2010-2014**
University of Science and Technology of China *Hefei, Anhui, China*

RESEARCH EXPERIENCE

Graduate Student Researcher **2016–2020**
University of California, Irvine *Irvine, CA, USA*

TEACHING EXPERIENCE

Teaching Assistant **9/2015-6/2017, 4/2020–6/2020**
University of California, Irvine *Irvine, CA, USA*

INDUSTRY EXPERIENCE

Data Science Summer Intern **6/2019-9/2019**
Uber Technologies, Inc. *San Francisco, CA, USA*

Quantitative Research Summer Intern **6/2017–9/2017**
Moody's Analytics *San Francisco, CA, USA*

PUBLICATIONS and WORKING PAPERS

Yin, Fan; Shen, Weining; and Butts, Carter T. (2019). "Finite Mixtures of ERGMs for Modeling Ensembles of Networks." Revised and resubmitted, *Bayesian Analysis*.

Thomas, Loring J.; Huang, Peng; **Yin, Fan**; Luo, Xiaoshuang I.; Almquist, Zack W.; Hipp, John R. and Butts, Carter T. (2020). "Spatial Heterogeneity Can Lead to Substantial Local Variations in COVID-19 Timing and Severity." Accepted, *Proceedings of the National Academy of Sciences.*

Yin, Fan and Butts, Carter T. (2020). "Kernel-based Approximate Bayesian Inference for Exponential Family Random Graph Models." Under review, *Statistics and Computing*.

Yin, Fan; Phillips, Nolan; and Butts, Carter T. (2019). "Selection of Exponential-Family Random Graph Models via Held-Out Predictive Evaluation (HOPE)." Under revision, *Social Networks*.

Yin, Fan; Khago, Domarin; Martin, Rachel W. and Butts, Carter T. "Bayesian Analysis of Static Light Scattering Data for Globular Proteins." In preparation for *Journal of the Royal Statistical Society, Series C (Applied Statistics)*.

CONFERENCE PRESENTATIONS

Yin, Fan and Butts, Carter T. (5/2017). "Scalable ERGM Inference for Organizational Interaction and Behavioral Simulation Data." 37th Sunbelt Network Conference (INSNA), Beijing, China.

Yin, Fan and Butts, Carter T. (6/2019). "Approximate Bayesian Computation for ERGMs via Copula Model." 39th Sunbelt Network Conference (INSNA), Montreal, Quebec, Canada.

Phillips, Nolan; **Yin, Fan**; and Butts, Carter T. (6/2019). "A New HOPE: Held-Out Predictive Evaluation (HOPE) for Exponential Family Random Graph Models." 39th Sunbelt Network Conference (INSNA), Montreal, Quebec, Canada.

Yin, Fan and Butts, Carter T. (7/2019). "Bayesian Inference for Exponential Random Graph Models via Kernel Bayes Rule." Section on Bayesian Statistical Science, Joint Statistical Meeting, Denver, CO, USA.

Yin, Fan; Shen, Weining; and Butts, Carter T. (8/2020). "Finite Mixtures of ERGMs for Modeling Ensembles of Networks." Section on Bayesian Statistical Science, Joint Statistical Meeting, Joint Statistical Meeting, 2020 Virtual Conference.

SOFTWARE

Programming Languages	**Fluent** in R, Python; **Experienced** in MySQL, MATLAB
Tools	PyTorch, WinBUGS, JAGS, Stan, GitHub, LINUX, QGIS, LaTeX

ABSTRACT OF THE DISSERTATION

Advances in Exponential-family Random Graph Models: Computation, Model Selection, and Methodology

by

Fan Yin

Doctor of Philosophy in Statistics

University of California, Irvine, 2020

Professor Carter T. Butts, Chair

Networks (graphs) are broadly used to represent relations between entities in a wide range of scientific fields. Exponential-family random graph models (ERGMs) provide a highly general way of specifying distributions on graphs, allowing the complex dependence structure of edges in a network to be specified in terms of local structural properties. This thesis addresses problems related to three lines of inquiry for ERGMs: faster Bayesian inference algorithms; comparison of newly proposed and traditional model selection techniques; and methodological innovation for modeling ensembles of networks.

In Chapter 2 of this dissertation, we present a highly parallel algorithm that enables fast Bayesian inference on ERGMs. The impetus for this work comes from the facts that conducting Bayesian inference for ERGMs is challenging because of the intractability of both the likelihood and posterior normalizing factor and auxiliary-variable based Markov Chain Monte Carlo (MCMC) methods for this problem are asymptotically exact but computationally demanding. We propose a kernel-based approximate Bayesian computation algorithm for fitting ERGMs, which is easily parallelizable. Through empirical comparisons against the state-of-the-art approximate exchange algorithm, we show that the proposed algorithm yields comparable accuracy to the state-of-the-art MCMC approach, the approximate ex-

change algorithm (Caimo and Friel, 2011), while cutting the wallclock runtime by half with 5 cores, and by 80% with 30 cores.

In Chapter 3 of this dissertation, we carry out simulation studies to compare newly proposed and traditional model selection techniques. This work is driven by the importance of understanding the strengths and weaknesses of those model selection techniques for ERGMs that are currently available, including Akaike information criterion (Akaike, 1973), Bayesian information criterion (Schwarz, 1978), Held-Out Predictive Evaluation (HOPE) (Yin et al., 2019), Bayes factors (Raftery, 1995) and graphical goodness of fit (Hunter et al., 2008a). In particular, we focus on the first three techniques, as the calculation of Bayes factor for ERGMs relies on reversible jump Markov chain Monte Carlo algorithm extension of the approximate exchange algorithm (Caimo and Friel, 2013), which is hard to implement and tune; the graphical goodness of fit is more suitable for checking whether a model is adequate rather than comparing competing models. The simulation studies are carried out under two scenarios, closed-$\mathcal{M}$ (under which the true model is among the set of candidate models) and open-$\mathcal{M}$ (under which the true model is not among the set of candidate models), and we evaluate the performance of model selection techniques from various aspects covering the model selection accuracy, predictive deviance and prediction accuracy of edge variables.

In Chapter 4 of this dissertation, we propose a novel methodology that can be used for modeling the generative processes of ensembles of networks. The motivation of this work is that ensembles of networks arise in many scientific fields, but there are few statistical tools for inferring their generative processes, particularly in the presence of both dyadic dependence and cross-graph heterogeneity. To fill in this gap, we propose characterizing network ensembles via finite mixtures of exponential family random graph models, a framework for parametric statistical modeling of graphs that has been successful in explicitly modeling the complex stochastic processes that govern the structure of edges in a network. Our proposed methodology 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 Metropolis-within-Gibbs algorithm to conduct fully Bayesian inference and adapt a version of deviance information criterion for missing data models to choose the number of latent heterogeneous generative mechanisms. Simulation studies show that the proposed procedure can recover the true number of latent heterogeneous generative processes and corresponding parameters. We demonstrate the utility of the proposed approach using an ensemble of political co-voting networks among U.S. Senators and an ensemble of advice-seeking networks among school teachers.

Chapter 1

Introduction

Networks are broadly used to represent relations between entities in a wide range of scientific fields. Statistical analysis of network data emerged as early as 1930s, and have continued to offer open problems for current research. The fundamental challenge in statistical modeling of network data is to capture the complex relational dependence (or *dyad* dependence), that is, the existence or strength of a relationship between two entities can affect other relationships in the network in a complex way.

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), emerged as one of the main families of models capable of capturing the complex dependence structure among dyads.

In recent years, ERGMs have found applications in empirical research in many scientific fields. 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., 2019b), 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).

1.1 Network Data Representation

We let $\mathbf{V}$ denote the set of vertices (also referred to as nodes, entities, actors, etc.) in the network (graph) of interest, assumed known and fixed. The cardinality of the node set, $|\mathbf{V}|$, is the number of nodes in the network. A dyad is defined as a pair of actors, ordered if the network of interest is directed, unordered if not. Dyads can take binary values (1 or 0), indicating the presence or absence of a relation between incident nodes; as well as counts and even continuous values indicating the strength of the respective relations. For the purposes of this thesis, we focus on binary relations without loops and multiple edges (i.e., disallow any relations between a node and itself; and disallow two or more relations between the same pair of nodes), also known as, simple graphs.

It is often mathematically convenient to represent the network structure via an adjacency matrix Y. An adjacency matrix for a simple graph of n nodes is a squared matrix of order-n with binary values (1 or 0) on the off-diagonal elements (the diagonal elements do not carry any information and are set at zeros by convention, because no loop is allowed). The adjacency matrices of undirected networks are symmetric while those of directed networks might not be symmetric.

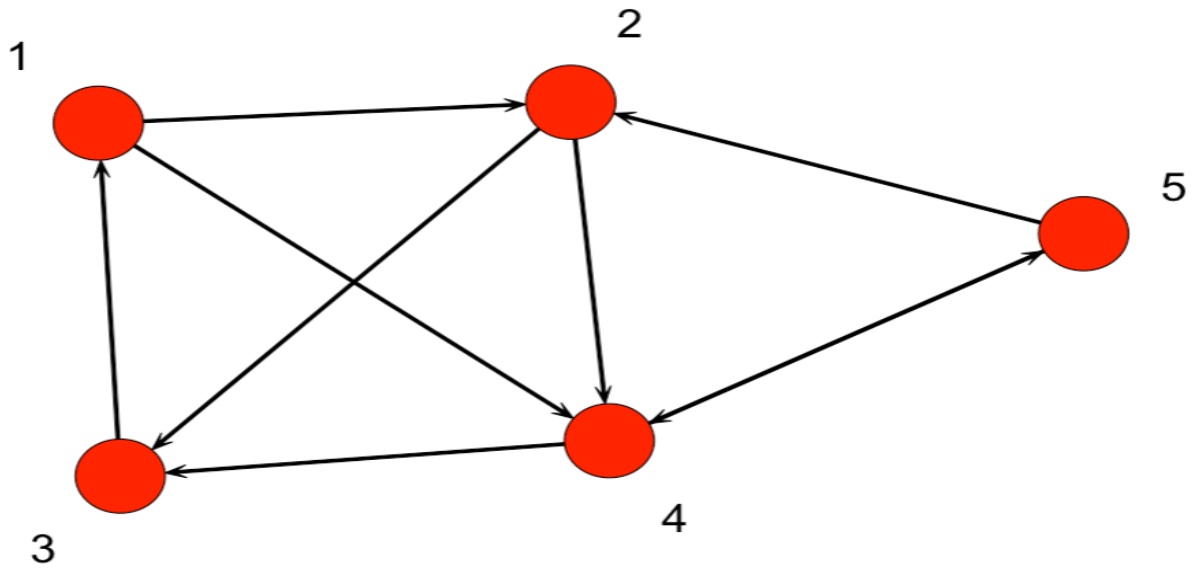

Figure 1.1: An example graph of 5 nodes

Figure 1.1 shows a graphical representation of a directed binary network of 5 nodes with the following adjacency matrix –

$$Y = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

The element Y_{ij} (1 or 0) represents the presence or absence of a tie from node i to node j.

1.2 Exponential-family Random Graph Models (ERGMs)

Letting $\mathcal{Y}_n$ be the set of all possible network configurations on n nodes, we write the probability mass function (pmf) of Y taking a particular configuration y in the form of a discrete

exponential family as

$$\mathbb{P}_{\boldsymbol{\eta}}(Y = y|\mathbf{X};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\intercal}\mathbf{g}(y;\mathbf{X}) - \psi_{\mathbf{g},\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta})\right)h(y), \quad y \in \mathcal{Y}_n, \tag{1.1}$$

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 $\mathbf{g} : \mathcal{Y}_n \to \mathbb{R}^p$ may incorporate fixed and known covariates $\mathbf{X}$ that are measured on the nodes or dyads. The sufficient statistics incorporate network features of interests that are believed to be crucial to the social process that gave rise to it (see, e.g., Morris et al., 2008). Here h defines the reference measure for the model family; often chosen to be the counting measure on $\mathcal{Y}_n$ for unvalued graphs with fixed n, other reference measures can make more sense in different settings. As discussed below, we employ a sparse graph reference that leads to a mean degree that is asymptotically constant in n. Finally, the normalizing factor $\psi_{\mathbf{g},\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta}) = \log \sum_{y' \in \mathcal{Y}_n} \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^{\intercal}\mathbf{g}(y';\mathbf{X})\right\} h(y')$ ensures that (1.1) sums up to 1 over the support $\mathcal{Y}_n$. To make notation simpler, we often assume that $\mathbf{V}$ is implicitly absorbed into $\mathbf{X}$.

Exact evaluation of the normalizing factor involves integrating an extremely rough 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 (see Hunter et al., 2012, for a review).

1.3 Contributions and Outline

We explore three ways to advance the research in ERGMs modeling framework in this thesis. Chapter 2 begins with a review on classical estimation algorithms for ERGMs and then presents a highly parallel algorithm for fast Bayesian inference for ERGMs. Chapter 3 presents a cross-validation-analogue approach, named Held-Out Predictive Evaluation (HOPE), for model selection of ERGMs. Chapter 4 extends the ERGMs for modeling ensembles of networks that arise from heterogeneous graph distributions via finite mixture models. Chapter 5 summarizes my discoveries, points out limitations of some of the methods, and provides future directions that others could follow, to promote the research of ERGMs and network modeling.

Chapter 2

Kernel-based Approximate Bayesian Computation for ERGMs

2.1 Introduction

Despite remarkable success in network modeling, parametric inference for ERGMs with complex dependence has been a historical challenge and continues to offer open problems for current research. The central challenge stems from the normalizing factor of the ERGM likelihood, which involves integrating an extremely rough function over all possible network configurations. While somewhat ad-hoc methods of estimating parameters of network models based on path lengths were explored in a pre-ERGM context by e.g. Rapoport (1957); Fararo and Sunshine (1964); Rapoport (1979), the first work to investigate inference for random graphs with dependence structure in a fully modern sense (and in ERGM form) was the iterative scaling algorithm proposed for the p_1 model (Holland and Leinhardt, 1981), now identified as a sub-class of ERGMs where the dependence is within each dyad (i.e. reciprocity). As an attempt to incorporate higher-order dependence structure, Frank and

Strauss (1986) introduced the *Markov graphs*, where edge variables are dependent only if they share a common node; unfortunately, the accompanying estimation algorithm based on cumulant approximations was not practical for use in typical settings. A major advance was made with Strauss and Ikeda's (1990) adaptation of the maximum pseudolikelihood estimation (MPLE) strategy of Besag (1974)), in which the likelihood is approximated by a product of full conditional distributions, to the estimation of ERGMs. MPLE is still in use to date in some applications, being relatively fast, algorithmically convenient, and able to provide parameter estimates (albeit sometimes innaccurate ones) for even badly-specified models. As an approximation to the MLE, however, the MPLE is often biased with respect to the mean value parameter space (which the MLE is not), less efficient than the MLE, prone to instability, and very poorly calibrated (Van Duijn et al., 2009). Given these issues, most subsequent work has focused on attempting to perform maximum likelihood estimation (MLE). Fitting general ERGMs using maximum likelihood is numerically demanding, as the likelihood can only be specified up to a parameter dependent normalizing constant, making the exact calculation of the MLE extremely difficult except for extremely small graphs (Von et al., 2020) or in cases for which the likelihood function can be analytically simplified (e.g. homogeneous and inhomogeneous Bernoulli graphs). State-of-the-art frequentist estimation approaches for ERGMs thus hinge on simulation-based algorithms to obtain high-quality approximations to the MLE, including

- *Markov chain Monte Carlo maximum likelihood estimation* (MCMC MLE), originally introduced by (Geyer and Thompson, 1992), adapted to ERGMs by Handcock (2003) and Hunter and Handcock (2006).
- *Stochastic approximation* (SA), originally introduced by Robbins and Monro (1951); Pflug (1996), adapted to ERGMs by Snijders (2002).

Bayesian inference for general ERGMs is even more challenging and has been historically

prohibitive, as the parameter dependent normalizing constant in the likelihood does not cancel out when taking posterior ratios (as is required, e.g., for standard MCMC strategies). This produces a target distribution which is termed *doubly intractable* (Murray et al., 2006), given the intractability of both the likelihood and the normalizing constant of the posterior density, and rendering conventional sampling schemes (e.g. Metropolis-Hastings algorithms (Metropolis et al., 1953; Hastings, 1970)) impractical. There have been some recent developments on asymptotic approximations for ERGMs (Pu et al., 2015; He and Zheng, 2015), but they are derived for a very specific set of models employing only permutation invariant subgraph statistics, typically do not converge in the sparse graph regime, and cannot be employed for models with covariate effects or other inhomogeneities.

The development of Bayesian inference has the potential to offer special advantages vis a vis several issues arising in typical ERGM use cases. Per standard theories of exponential-family models (Barndorff-Nielsen, 1978), the MLE for an ERGM's parameters does not exist (i.e., no finite maximizer of the likelihood exists) when the observed statistics for a given model happen to lie on the relative boundary of the convex hull of possible values of the sufficient statistics (Handcock, 2003). This issue is not peculiar to ERGMs, and indeed is present in all discrete exponential families (including trivial cases like the binomial model). However, typical ERGM specifications often include statistics based on sums of small numbers of sparse binary variables, creating a high risk of observing at least one extreme statistic. Though this can be partially "patched" by taking the estimate to be the infinite limit of the parameter in the direction of recession, the resulting model is overconfident (e.g., it may predict that ties between two groups not observed to be in contact are not only rare, but impossible) and lacks well-defined standard errors. By contrast, Bayes estimators under suitably regular priors are still well-defined in such cases, and will shrink estimates away from extreme values. As another example, the standard error of the MLE is currently obtained by employing the inverse of the Fisher information matrix, an approach that is conventionally justified by asymptotics under replication. In typical use cases, however,

models are based on only one observed graph, raising questions about the appropriateness of the underlying theory. While recent results have provided positive justification for using such approaches for certain classes of ERGMs (Schweinberger et al., 2019), and there is empirical evidence showing that the resulting estimates of standard error are similar to that yielded by parametric bootstrap (Fellows and Handcock, 2012), it is attractive to have alternative frameworks for quantifying uncertainty that do not depend on asymptotic assumptions. Bayesian answers regarding uncertainty in parameter estimates are well-defined even in the finite sample case, and hence provide an immediate way of addressing this issue that does not depend on any particular model specification. As noted, fully Bayesian inference for ERGMs is a doubly intractable problem, with both the likelihood and the normalizing constant of the posterior being infeasible to calculate. Early attempts at resolving this issue were based on conventional Metropolis-Hastings algorithms in which the likelihood ratio at each iteration is approximated by a linked importance sampler auxiliary variable algorithm (Koskinen, 2008; Koskinen et al., 2010). Caimo and Friel (2011) attempted to improve performance by proposing an approximate version of the *exchange algorithm* (Murray et al., 2006) to draw posterior samples of model parameters. This approximate exchange algorithm has become the state-of-the-art approach to Bayesian inference for ERGMs, with the potential to yield high-quality posterior draws, but the algorithm can be very expensive to use due to the need to serially draw high-quality ERGM simulations in an auxiliary chain at each iteration. Bouranis et al. (2017) introduce an approximate method to the approximate exchange sampler by calibrating the posterior samples drawn from a "pseudo-posterior" – where the exact ERGM likelihood is replaced by a tractable approximation (e.g. the pseudo-likelihood) – via an affine transformation that requires the existence of the mode of likelihood (i.e. MLE) and a Monte-Carlo approximation of the gradient and curvature around the mode of likelihood (i.e. MCMC-MLE if MLE cannot be solved precisely). Pseudo-posteriors are also employed by Grazioli et al. (2019a), who instead obtain posterior draws using a "Bayesian bootstrap" procedure; though computationally efficient, this approach is limited

to cases in which large numbers of graphs are observed from the same generating process. As these examples suggest, the existing approaches to Bayesian inference for ERGMs are, broadly speaking, either limited to relatively special cases or computationally expensive, non-parallelizable, and difficult to extend to new settings (e.g., ERGMs with endogenous vertex sets (Almquist and Butts, 2014) or inference from non-sufficient statistics) without substantial re-engineering of the underlying algorithms. This hence remains an area of active research, with substantial room for new techniques.

In this chapter, we consider another possible direction for fast Bayesian estimation of ERGM parameters by proposing a parallelizable kernelized approximate Bayesian computation (K-ABC) algorithm. We show that the proposed algorithm can yield comparable estimates to the gold-standard approximate exchange sampler, with significantly reduced computational time when multiple cores are available. We discuss the choice of distance measure, kernel functions, bandwidth selection, and offer some guidance on selecting optimal settings. We also show the inherent connection between the proposed algorithm and Kernel Bayes' rule (KBR) (Fukumizu et al., 2011), offering an interpretation of the resulting estimates from a kernel regression perspective. The KBR interpretation provides a more direct route to obtaining estimates of posterior moments, and also suggests the opportunity to exploit developments in machine learning (e.g., kernelized WLS) to obtain improved posterior approximations.

The outline of the remainder of this chapter is as follows. In Section 2.2, we give an introduction to the exponential random graph models along with some simulation and computational methods that serve as building blocks for the proposed method. In Section 2.3, we present our parallelizable kernelized approximate Bayesian computation (K-ABC) algorithm for fast Bayesian estimation of ERGMs, and provide implementation details. In Section 2.4, we describe the application of our approach in the context of two benchmark social network datasets of varying sizes, showing the accuracy and computational efficiency of our algo-

rithm compared to approximate exchange algorithm, which is the current "gold standard" for Bayesian inference of ERGMs. We discuss possible future extensions of the proposed algorithm in Section 2.5 and conclude in Section 2.6.

2.2 Classical Estimation and Simulation Algorithms for ERGMs

In the following subsections we focus on the technical details of the classical estimation algorithms for ERGMs, some of which serve as building blocks and benchmark for the proposed ABC-based algorithm.

To better illustrate the general idea behind the ERGM-fitting algorithms, we focus on regular ERGMs ($\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$) with counting measure $h(y) \equiv 1$. Therefore, (1.1) becomes

$$p(y|\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^\intercal g(y) - \psi(\boldsymbol{\theta})\right\}, \quad y \in \mathcal{Y}_n \tag{2.1}$$

where $\mathcal{Y}_n$ represents the set of all possible configurations of binary networks of size n and $\psi(\boldsymbol{\theta}) = \log \sum_{y' \in \mathcal{Y}_n} \exp\left\{\boldsymbol{\theta}^\intercal g(y')\right\}$, which involves the summation of $2^{\binom{n}{2}}$ non-negative terms for any value of $\boldsymbol{\theta}$, and the covariates $\mathbf{X}$ are absorbed into y to further simplify the notations.

2.2.1 Maximum Pseudolikelihood Estimation

An approximate approach to maximum likelihood estimation for ERGMs is based on the *pseudolikelihood* function (Strauss and Ikeda, 1990),

$$p(y|\boldsymbol{\theta}) \approx f_{PL}(y|\boldsymbol{\theta}) = \prod_{(i,j)\in\mathcal{D}} p(y_{ij}|y_{-ij},\boldsymbol{\theta}). \tag{2.2}$$

where $\mathcal{D}$ denotes the set of all pairs of dyads on $\mathbf{V}$. For directed networks, $\mathcal{D} = \{(i,j)|i,j \in \mathbf{V}, i \neq j\}$, while for undirected networks, $\mathcal{D} = \{(i,j)|i,j \in \mathbf{V}, i < j\}$. Equation (2.2) is simply the product of full conditional distributions, which has the following form

$$\begin{aligned}
\text{logit}\left\{p(y_{ij}=1|y_{-ij},\boldsymbol{\theta})\right\} &= \log\frac{p(y_{ij}=1|y_{-ij},\boldsymbol{\theta})}{p(y_{ij}=0|y_{-ij},\boldsymbol{\theta})} \\
&= \boldsymbol{\theta}^{\intercal}\left\{g(y_{ij}^{+}) - g(y_{ij}^{-})\right\} \\
&= \boldsymbol{\theta}^{\intercal}\Delta_{i,j}g(y)
\end{aligned} \tag{2.3}$$

where $\Delta_{i,j}g(y) = g(y_{ij}^{+}) - g(y_{ij}^{-})$ 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. Following (2.2), the log pseudo-likelihood can be written as

$$\begin{aligned}
\log f_{PL}(y|\boldsymbol{\theta}) &= \sum_{(i,j)\in\mathcal{D}} [y_{ij}\text{logit}\left\{p(y_{ij}=1|y_{-ij},\boldsymbol{\theta})\right\} + \log\left\{1 - p(y_{ij}=1|y_{-ij},\boldsymbol{\theta})\right\}] \\
&= \sum_{(i,j)\in\mathcal{D}} [y_{ij}\boldsymbol{\theta}^{\intercal}\Delta_{ij}g(y) - \log\left\{1+\exp(\boldsymbol{\theta}^{\intercal}\Delta_{ij}g(y))\right\}].
\end{aligned} \tag{2.4}$$

Note that (2.4) is no different from the likelihood of a logistic regression where y_{ij} are

the responses and $\Delta_{ij}g(y)$ as the corresponding row in the model matrix, facilitating fast estimation. For exponential family distributions with log-likelihood $\ell(\boldsymbol{\theta})$, the estimate of standard error is based on the inverse of Fisher information matrix $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$, where

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla\ell(\boldsymbol{\theta})\nabla\ell(\boldsymbol{\theta})^{\intercal}] = Var_{\boldsymbol{\theta}}[g(Y)]. \tag{2.5}$$

Note that under the framework of pseudolikelihood, we substitute $\log f_{PL}(y|\boldsymbol{\theta})$ for the true log-likelihood $\ell(\boldsymbol{\theta}) \equiv \log p(y|\boldsymbol{\theta})$, where y is omitted for the convenience of notation. In fact, pseudolikelihood is a special form of composite likelihood (Lindsay, 1988), which is a more general class of inference functions used to approximate complex likelihoods (see Varin et al., 2011, for a most recent review). Despite the empirical observations that MPLE generally causes bias and underestimates the standard errors (Van Duijn et al., 2009) (especially for models with strong dyadic dependence), it has been the default choice for the initial value in MCMC MLE. 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 MLE in such cases (Desmarais and Cranmer, 2010).

2.2.2 Simulation Methods

More advanced estimation techniques, including simulation-based methods for finding maximum likelihood estimates, as well as Bayesian methods, require sampling from the ERGM distribution. To simulate from $p(y|\boldsymbol{\theta})$, Snijders (2002) propose to use a Metropolis-Hastings sampling procedure: given a proposal y' from density $q(y'|y)$, accept with probability

$$\begin{aligned}\alpha &= \min\left(1, \frac{q(y|y')p(y'|\boldsymbol{\theta})}{q(y'|y)p(y|\boldsymbol{\theta})}\right) \\ &= \min\left(1, \frac{q(y|y')}{q(y'|y)}\exp\left\{\boldsymbol{\theta}^{\intercal}(g(y') - g(y))\right\}\right).\end{aligned} \tag{2.6}$$

Based on (2.3), note that if we restrict $q(y'|y) > 0$ only if $\sum_{(i,j)\in\mathcal{D}} \mathbb{1}(y'_{ij} \neq y_{ij})$ equals one, i.e. only the networks that can be constructed by toggling exactly one dyad from y are allowed to be proposed, then $g(y') - g(y)$ reduces to $\pm\Delta_{ij}g(y)$ with the sign depending on the direction of the toggle, resulting in a Gibbs sampling algorithm. To avoid unnecessary computational cost spent on highly improbable graphs, the starting point $y^{(0)}$ of the sampling is usually set as the observed network y^{obs} (if available). Furthermore, as opposed to the basic MCMC algorithm in which each dyad is selected to be toggled uniformly at random, the adoption of asymmetric proposals have been demonstrated to be more favorable for sparse graphs. For example, the "TNT"(tie-no-tie) sampler implemented as default in the `ergm` package for R (Morris et al., 2008), while the "OTNT"(open triangle-tie-no tie) (Wang and Atchadé, 2014) has been shown to improve performance in clustered networks and the improved fixed density (IFD) sampler has been shown to be a promising tool for large, sparse networks (Byshkin et al., 2016). Specifically, the "TNT" sampler selects an tied dyad (i.e., realized edge) with probability 1/2 (instead of the graph density, which is close to 0 for a sparse graph) at each iteration, which often leads to better mixing in the typical case of ERGMs that concentrate probability mass on sparse graphs. The above MCMC routines produce a sequence of networks $\left\{y^{(0)}, \cdots, y^{(T)}\right\}$, of which the initial part is highly dependent on the starting point (network) and hence is usually discarded as burn-in. These are referred to as the *auxiliary iterations* required before a simulated network can be claimed as a random draw from $p(y|\boldsymbol{\theta})$. Exact sampling from ERGMs is also possible at a higher computational cost (Butts, 2018), and non-MCMC approximate samplers have also been proposed (Butts, 2015).

2.2.3 Approximate MLE via Simulation-based Algorithms

The maximum likelihood estimation for general ERGMs is currently based on simulation-based algorithms, of which the two most prevalent approaches are Markov chain Monte Carlo

maximum likelihood estimation (MCMC MLE) and stochastic approximation (SA).

Given observed network y^{obs}, SA finds MLE by solving the *moment equation*

$$\nabla_{\boldsymbol{\theta}} \log p(y^{obs}|\boldsymbol{\theta}) = g(y^{obs}) - \mathbb{E}_{\boldsymbol{\theta}}[g(Y)] = 0, \tag{2.7}$$

where $\nabla_{\boldsymbol{\theta}} \log p(y|\boldsymbol{\theta})$ denotes the gradient of the log-likelihood function $\log p(y|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Snijders (2002) proposed solving Equation (2.7) using stochastic approximation (Robbins and Monro, 1951; Pflug, 1996). The iteration starts with an initial guess $\boldsymbol{\theta}_0$ (usually MPLE), the stochastic approximation method updates $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$ as follows,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - a_t \hat{D}_t^{-1}(g(Y_{\boldsymbol{\theta}^t}) - g(y^{obs})), \tag{2.8}$$

where $\hat{D}_t$ is an approximation of the Hessian of the log-likelihood function in the neighborhood of $\boldsymbol{\theta}^t$, a_t is a sequence of positive numbers approaching 0 as t increases, and $Y_{\boldsymbol{\theta}^t}$ is a network sample from the ERGM with parameter $\boldsymbol{\theta}^t$ by MCMC methods. Rather than solving the moment equation (2.7), MCMC MLE (Geyer and Thompson, 1992) makes more efficient use of MCMC samples as it targets maximization of the log ratio of the likelihoods, $\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is a fixed parameter value that should ideally be close to the true MLE $\hat{\boldsymbol{\theta}}$. Specifically, Handcock (2003); Hunter and Handcock (2006) developed a version of MCMC MLE for ERGMs as follows

$$\begin{aligned} LR_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) =& (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal g(y^{obs}) - (\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)) \\ =& (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal g(y^{obs}) - \log \mathbb{E}\left[\exp\left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal g(Y) \right\} \right] \\ \approx& (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal g(y^{obs}) - \log \left[\frac{1}{m} \sum_{i=1}^{m} \exp\left\{ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal g(y^{(i)}) \right\} \right], \end{aligned} \tag{2.9}$$

where $\left\{ y^{(1)}, \cdots, y^{(m)} \right\}$ are random draws from $p(y|\boldsymbol{\theta}_0)$ via MCMC as introduced in 2.2.2. It

is worth noting that long burn-in and high thinning factor are required to obtain nearly independent and representative samples and thus a better approximation to $LR_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$. In practice, $\boldsymbol{\theta}_0$ is often taken to be the easy-to-calculate MPLE, while novel approaches for finding a better $\boldsymbol{\theta}_0$ have been proposed in recent years, including the *partial stepping* technique (Hummel et al., 2012) and a *contrastive divergence* (CD,Hinton (2002))-based technique adapted to ERGMs by Krivitsky (2017).

The implementation of the above two algorithms is publicly available in `ergm` package (Hunter et al., 2008b; Handcock et al., 2008) from the Statnet suite of R packages and software `PNet`(Wang et al., 2009).

2.2.4 Bayesian Inference for ERGMs

We consider the Bayesian treatment of ERGM inference as illustrated in Koskinen (2004). Given observed network y^{obs}, and prior distribution $\pi(\boldsymbol{\theta})$ placed on $\boldsymbol{\theta}$, the full posterior distribution of $\boldsymbol{\theta}$ is

$$\begin{aligned}\pi(\boldsymbol{\theta}|y^{obs}) &= \frac{p(y^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(y^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto p(y^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \end{aligned} \tag{2.10}$$

where $\int p(y^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal probability of data, which is often intractable as a (potentially) high-dimensional integral for general models.

Standard MCMC approaches, e.g. the Metropolis-Hastings (MH) algorithm can address intractable normalizing constants of a posterior density as long as the posterior density of interest is known up to a constant. However, the likelihood itself is only known up to a parameter dependent constant $\psi(\boldsymbol{\theta})$, and hence leads to the so-called "doubly intractable" problem, which cannot be dealt with using naive implementation of MH or other conventional

MCMC algorithms designed for models with tractable likelihood functions. This gap has led to the development of a body of MCMC approaches that by design generate samples from doubly intractable posterior densities, most of which rely on augmenting the posterior density so that the augmented posterior probability distribution is easy to sample from.

The exchange algorithm has evolved as a popular approach for tackling problems with intractable likelihood such as the Ising and Potts models (Møller et al., 2006; Murray et al., 2006). The exchange algorithm samples from the augmented distribution,

$$\pi(\boldsymbol{\theta}', y', \boldsymbol{\theta}|y^{obs}) \propto p(y^{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})p(y'|\boldsymbol{\theta}') \tag{2.11}$$

where $p(y^{obs}|\boldsymbol{\theta})$ and $p(y'|\boldsymbol{\theta}')$ correspond to the same distribution but with different parameter values. The distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ is any distribution for augmented variable $\boldsymbol{\theta}'$ which might depend on $\boldsymbol{\theta}$, for example, a random walk centered at $\boldsymbol{\theta}$. Sampling auxiliary variables on an extended state space allows the normalizing constants in likelihood to be canceled in the Metropolis-Hastings acceptance probability,

$$\begin{aligned}\alpha &= \min\left(1, \frac{p(\boldsymbol{\theta}')p(y^{obs}|\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')p(y'|\boldsymbol{\theta})}{p(\boldsymbol{\theta})p(y^{obs}|\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})p(y'|\boldsymbol{\theta}')}\right) \\ &= \min\left(1, \frac{p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \exp\left\{(\boldsymbol{\theta}' - \boldsymbol{\theta})^{\intercal}(g(y^{obs}) - g(y'))\right\}\right),\end{aligned} \tag{2.12}$$

which is tractable and therefore the Metropolis-Hastings type algorithm operating on the augmented state space is applicable to general ERGMs by design. However, the exact exchange algorithm requires exact simulation of the auxiliary variable y' from the likelihood, which is typically infeasible for general ERGMs. The *approximate exchange algorithm* (AEA) of Caimo and Friel (2011) modifies the original exchange algorithm by substituting MCMC-based approximate samples for exact draws. Specifically, the "tie-no-tie" (TNT) sampler (Morris et al., 2008) was advocated as a more efficient approach to simulate from ERGM likelihood at each MCMC iteration, according to the implementation in `bergm` function from

the R package `Bergm` (Caimo and Friel, 2014). The default implementation of approximate exchange algorithm in `Bergm` package uses the idea of adaptive direction sampling (ADS) method (Gilks et al., 1994; ter Braak and Vrugt, 2008) from Population Monte Carlo to propose "parallel ADS" move to improve the mixing, and the default number of chains is set to be twice the number of model parameters.

2.3 Approximate Bayesian Computation for ERGMs

In this paper, we focus on alternatives to exchange sampling based on kernel methods and approximate Bayesian computation. ABC has emerged as a powerful tool for (approximate) Bayesian analysis of complex models for which the likelihood $p(y|\boldsymbol{\theta})$ is unavailable or computationally intractable but where simulation of $Y|\boldsymbol{\theta}$ is feasible (Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003; Sisson and Fan, 2011; Marin et al., 2012; Sisson et al., 2018).

In approximate Bayesian computation (ABC), inference is concerned with the *partial* posterior distribution $\pi(\boldsymbol{\theta}|s^{obs})$ (Doksum and Lo, 1990),

$$\pi(\boldsymbol{\theta}|s^{obs}) = \frac{p(s^{obs})|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(s^{obs})|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.13}$$

where $s^{obs} \equiv S(y^{obs})$ represents a vector of d-dimensional summary statistics computed from the observed data y^{obs}. The classical rejection ABC (R-ABC) (see Algorithm 1) approximates the partial posterior distribution $\pi(\boldsymbol{\theta}|s^{obs})$ by

$$\pi_h^{ABC}(\boldsymbol{\theta}|s^{obs}) \propto \int \mathbb{1}(\left\| s^* - s^{obs} \right\| \leqslant h) p(s^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})ds^*. \tag{2.14}$$

and proceeds by first drawing N values $\boldsymbol{\theta}^{(i)}$, $i = 1, \cdots, N$ from the prior distribution π and then simulating data from the likelihood $p(y|\boldsymbol{\theta}^{(i)})$, retaining those $\boldsymbol{\theta}^{(i)}$ with $\left\|s^{(i)} - s^{obs}\right\| \leqslant h$ ($h > 0$, usually a sufficiently small number to control the precision of the approximation) under some distance metric $\|\cdot\|$. The underlying idea (based on the work of Rubin (1984)) is that $\boldsymbol{\theta}^{(i)}$ is unlikely to have generated the observed data, if $\left\|s^{(i)} - s^{obs}\right\|$ is large. Such algorithms converge to the exact posterior when $h \to 0$ and s contains all sufficient statistics, because the posterior $\pi(\boldsymbol{\theta}|s^{obs})$ can be regarded as a slice of the joint distribution $\pi(\boldsymbol{\theta}, s)$ at $s = s^{obs}$. A somewhat deeper observation (which we will exploit below) is that R-ABC is a form of *kernel method,* in which a uniform kernel with respect to the metric $\|\cdot\|$ with bandwidth h used to perform a simulation-based analog of kernel regression (predicting posterior quantities at s^{obs}). This informal intuition (which can be made precise, see e.g. Fukumizu et al., 2011) suggests a number of potential improvements to the base algorithm, some of which we will leverage here.

In practice, despite the fact that the sampling step of R-ABC is *embarrassingly* parallel, a naive implementation of R-ABC can perform poorly given limited computational resources. In the ERGM context, two immediate problems arise:

- Under a weakly informative prior, an extremely large proportion of sampled parameters may generate graphs nowhere close to the observed graph. For example, a prior such as a multivariate Gaussian centered at zero with large standard deviations places most of its mass in unrealistic regions of the parameter space: e.g. positive values of the parameter associated with the *edges* term in ERGMs are rarely seen (when there is no `edgecov`), as most real-world network data are sparse (Kolaczyk and Krivitsky, 2015); and large positive values of parameters associated with dependence terms such as *k-stars*, *triangles*, or *shared partners* can lead to *degenerate* probability distribution on graphs that are not useful for network modeling (Handcock, 2003; Schweinberger, 2011). R-ABC algorithms can be very inefficient under such prior specifications, as

the feasible region for realistic real-world networks in the parameter space is often very thin and irregularly-shaped (Handcock, 2003; Rinaldo et al., 2009).

- The distance metric $\|\cdot\|$ and the rejection threshold h are determined based on the so-called "reference table" (simulated parameter-data pairs obtained from a pilot run). The former is usually chosen as a version of weighted Euclidean distance with the weights being selected to normalize the summary statistics so that they vary over roughly the same scale, preventing the distance being dominated by the most variable statistic. The threshold h controls the trade-off between runtime and approximation accuracy, and for R-ABC it is usually selected using the 1% quantile of the distance computed based on the pilot run. However, the relatively high cost of ERGM simulation can make such algorithm tuning fairly expensive, especially where sampling must be based on an imprecise prior (which, as described above, will lead in most cases to degenerate or otherwise non-viable graph distributions).

Algorithm 1 Rejection-ABC (R-ABC) algorithm (Pritchard et al., 1999)

Require: Observed summary statistics $s^{obs} = S(y^{obs})$, data generating mechanism $p(y|\boldsymbol{\theta})$, prior $\pi(\boldsymbol{\theta})$

Input: Summary statistics $s = S(y)$
A desired sample size $N > 0$.
A distance metric $\|\cdot\|$
A threshold parameter $h > 0$.
Burn-in for MCMC-based simulation for the likelihood (default $B = 2n^2$, where n is the network size)
Compute observed summary statistics $s^{obs} = S(y^{obs})$

1: **while** $i \leqslant N$ **do**
2: $\quad\boldsymbol{\theta}' \sim \pi(\boldsymbol{\theta})$
3: $\quad y' \sim p(y|\boldsymbol{\theta}')$ (burn-in for the MCMC-based simulation B)
4: $\quad s' = S(y')$
5: $\quad$ **if** $\left\|s' - s^{obs}\right\| \leqslant h$ **then**
6: $\quad\quad$ Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'$, $i = i + 1$
7: $\quad$ **end if**
8: **end while**

Output: A set of parameter values $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{N}$ with equal weights, drawn from $\pi_h^{ABC}(\boldsymbol{\theta}|s^{obs})$

Given these issues, a number of adaptations are required to make ABC feasible for ERGM inference, Here, we provide a strategy for approximate Bayesian inference for ERGM parameters under two different scenarios. We first consider the most common scenario, in which we either observe the full network or the set of sufficient statistics is from it, developing a highly parallel algorithm for fast Bayesian inference. In the second scenario, only a subset of the sufficient statistics can be observed (potentially alongside other, proxy statistics), which is typical for sampled, incompletely reported, or obfuscated network data. While conventional estimation schemes are difficult to apply in these cases without extensive re-engineering, we show that our ABC approach easily accommodates them. In both cases, we propose a version of a kernelized ABC-MCMC algorithm for posterior simulation, though we also discuss KBR-style approaches for efficient posterior moment estimation.

2.3.1 Kernel ABC Importance Sampling Algorithm

To improve the sampling efficiency when only weakly informative priors are available, we propose to sample from an importance density rather than the prior. We also consider alternative kernels to the standard uniform kernel employed in the default R-ABC algorithm. This leads to a *kernel ABC importance sampling algorithm* (K-ABC-IS), shown here as Algorithm 2. The easy-to-calculate MPLE is a natural choice for constructing an initial proposal distribution when the full network is available. Specifically, we employ a location-scale family centered at the MPLE, with a scale matrix based on the Hessian of the log pseudolikelihood. Because the curvature of the pseudolikelihood about the MPLE generally underestimates the variability of the parameters, we use an "inflated" multivariate Student's t proposal $\mathcal{T}_\nu$ with a relatively small degree of freedom parameter (e.g. $\nu = 4$) whose scale matrix is the inverse Hessian matrix of the negative log pseudolikelihood at the MPLE, multiplied by a scaling factor $\omega > 1$ to ensure that the sampled parameters are not confined to an overly narrow region near the MPLE. With the parameters sampled from the importance density,

we are more likely to generate graphs that are more similar to the observed graph, hence improving sampling efficiency. (Equivalently, we expect that–so long as the data is reasonably informative, and under reasonable choices of priors–substantial posterior mass will be concentrated in the vicinity of the MLE, and hence typically the MPLE. Use of a heavy-tailed proposal "hedges" this expectation against the possibility that the MPLE is a poor initial guess, and ensures adequate weight to majorize the tails of the posterior distribution.) Algorithm 2 presents the kernel ABC importance sampling algorithm. Intuitively, the key idea is to "doubly re-weight" the sampled parameters by both their importance ratio versus the prior (i.e., how likely the draw would be to arise under the prior versus the proposal) and their likelihood of generating graphs that are similar to the observed graph in terms of the summary statistics. As a metric on the space of statistics we employ the *Mahalanobis* distance; it serves as a natural choice because it takes both the variability and correlation of the summary statistics into consideration (an important factor, since many typical ERGM statistics are highly correlated). We here suggest a Gaussian kernel due to the facts that it is a non-compact kernel and yields fairly efficient estimator for smooth distributions. The bandwidth, h, is here chosen based on a simple heuristic for kernel density estimation (Silverman, 1986) applied to the the computed Mahalanobis distance $d^{(i)}, i = 1, \cdots, N$ distribution. It should be noted that we do not reject samples in this algorithm, instead assigning them different weights according to both importance ratio and kernel weights. We have found that the proposed Algorithm 2 is an improvement over ABC with smooth rejection (Beaumont et al., 2002) in the ERGM setting. It is worth noting that more sophisticated approaches for bandwidth selection exist but can increase the computational cost; as discussed in detail in later sections, our experience has suggested that the heuristic bandwith provides comparable performance to more elaborate schemes with much greater computational efficiency.

Note that the sampling step in Algorithm 2 is *embarrassingly* parallel; since this accounts for the overwhelming majority of the algorithm's computational cost, dramatic performance enhancements are possible on multi-core hardware. By contrast, the approximate exchange

algorithm must be run serially, and cannot take advantage of this level of parallelism. On the other hand, the exchange algorithm has the advantage of exploring the parameter space in a more controlled manner, guided by the likelihood ratio and prior ratio defined in (2.12) at each iteration, and in our experiments has proved to be slightly more efficient than K-ABC-IS when the latter is run on a single core. When multiple cores are available, K-ABC-IS can be substantially faster.

Algorithm 2 K-ABC importance sampling algorithm (K-ABC-IS)

Require: Observed summary statistics $s^{obs} = S(y^{obs})$, data generating mechanism $p(y|\boldsymbol{\theta})$, prior $\pi(\boldsymbol{\theta})$

Input: A desired sample size $N > 0$.

A parametric family for proposal distribution (e.g.multivariate Student's t proposal distribution, $\mathcal{T}_\nu$, degree of freedom ν).

A scale factor ω

Burn-in for MCMC-based simulation for the likelihood (default $B = 2n^2$, where n is the network size)

A distance metric $\|\cdot\|$ (e.g. mahalanobis distance)

Smoothing kernel $K_h(\cdot)$ and scale parameter $h > 0$.

(Optional) $\hat{\boldsymbol{\theta}}_{MPLE}$, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_{MPLE})$

Initialization: $\hat{\mu}$, $\hat{\Sigma}$ (default $\hat{\mu} = \hat{\boldsymbol{\theta}}_{MPLE}$, $\hat{\Sigma} = \omega \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}_{MPLE})$), set $f(\boldsymbol{\theta}) \equiv \mathcal{T}_4(\hat{\mu}, \hat{\Sigma})$

for $i = 1, 2, \cdots, N$ **do**

  $\boldsymbol{\theta}^{(i)} \sim f(\boldsymbol{\theta})$, (unnormalized) importance weight $w_I^{(i)} = \frac{\pi(\boldsymbol{\theta}^{(i)})}{f(\boldsymbol{\theta}^{(i)})}$

  $y^{(i)} \sim p(y|\boldsymbol{\theta}^{(i)})$ (burn-in of the MCMC-based simulation B)

  $s^{(i)} = S(y^{(i)})$

end for

$W = \frac{1}{N}\sum_{i=1}^{N}(s^{(i)} - \bar{s})(s^{(i)} - \bar{s})^\intercal$, where $\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s^{(i)}$

$d^{(i)} = (s^{(i)} - s^{obs})^\intercal W^{-1}(s^{(i)} - s^{obs})$ for $i = 1, 2, \cdots, N$

Perform univariate kernel density estimate on $d^{(i)}, i = 1, 2, \cdots, N$ to obtain a (heuristic) bandwidth h, and fix h as the scale parameter for the smoothing kernel $K_h(\cdot)$, kernel weight $w_K^{(i)} \propto K_h(d^{(i)})$.

Assign weight to $\boldsymbol{\theta}^{(i)}$ as $\tilde{w}^{(i)} \propto w_I^{(i)} w_K^{(i)}$, and the corresponding normalized $w^{(i)}$

Output: A set of parameter values $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{N}$ with weights $w^{(i)}$, drawn from $\pi_h^{ABC}(\boldsymbol{\theta}|s^{obs})$

Based on the weighted parameters returned by Algorithm 2, we estimate the partial posterior

mean of any scalar function of model parameters, $a(\boldsymbol{\theta})$, $\mathbb{E}[a(\boldsymbol{\theta})|s^{obs}]$ by the kernel estimate

$$\begin{aligned} m^{0,a} =& \sum_{i=1}^{N} w^{(i)} a(\boldsymbol{\theta}^{(i)}) \\ =& \frac{\sum_{i=1}^{N} a(\boldsymbol{\theta}^{(i)}) w_I^{(i)} w_K^{(i)}}{\sum_{i=1}^{N} w_I^{(i)} w_K^{(i)}} \\ =& \frac{\sum_{i=1}^{N} a(\boldsymbol{\theta}^{(i)}) w_I^{(i)} K_h(d^{(i)})}{\sum_{i=1}^{N} w_I^{(i)} K_h(d^{(i)})}. \end{aligned} \tag{2.15}$$

Note that (2.15) is similar to the Nadaraya-Watson type estimator (Nadaraya, 1964; Watson, 1964), which can be found by minimizing the weighted sum of squared residuals

$$WSSR_0 = \sum_{i=1}^{N} \left\{ a(\boldsymbol{\theta}^{(i)}) - \alpha \right\}^2 w^{(i)} \tag{2.16}$$

By letting $a(\boldsymbol{\theta}) = \boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_i \boldsymbol{\theta}_j, i, j = 1, \cdots, p$, we can obtain the estimate for posterior mean and posterior second moments, hence yielding a natural estimate of posterior variance based on the identity $\text{Var}[\boldsymbol{\theta}_j|s^{obs}] = \mathbb{E}[\boldsymbol{\theta}_j^2|s^{obs}] - (\mathbb{E}[\boldsymbol{\theta}_j|s^{obs}])^2$. From a non-parametric regression perspective, the proposed estimator in (2.15) corresponds to a locally constant estimate, and more intricate estimation of the posterior moments might be achieved by using locally linear or polynomial estimators (Blum, 2010) or any state-of-the-art machine learning techniques as long as the optimization is with respect to the squared error loss (e.g., kernelized WLS). Note that when the sufficient statistics $g(y)$ are a subset of the selected summary statistics, the resulting estimator targets the true posterior mean and standard deviation.

The construction of posterior intervals is straightforward given weighted samples $\left\{ (\boldsymbol{\theta}^{(i)}, w^{(i)}) \right\}_{i=1}^{N}$. For $j = 1, \cdots, p$, the general procedure is as follows:

1. Find the empirical cumulative distribution function (ECDF) as $\hat{F}_j(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\boldsymbol{\theta}_j^{(i)} \leqslant x)$.

2. Obtain a smooth approximation $\tilde{F}_j(x)$ of the ECDF $\hat{F}_j(x)$ using monotonic spline (e.g. `splinefun` function in R, with method set as "monoH.FC").

3. Find the q-th quantile of $\tilde{F}_j(x)$ by minimizing the squared error.

To obtain independent samples from the joint posterior with equal weights, it is possible to use sampling-importance resampling (SIR) techniques (Rubin, 1987, 1988). The resampling step can be conducted with or without replacement, but the latter should be favored when only a few large weights and many small weights are present (Gelman et al., 1995). Improved SIR with faster convergence rates and bias-reduced SIR were proposed and studied by Skare et al. (2003). The SIR-based techniques have proved to be useful for ABC algorithms producing weighted samples (see, e.g. Mengersen et al., 2013; Zhu et al., 2016).

The theoretical validity of the proposed ABC algorithm can be justified from an approximate likelihood perspective (Karabatsos and Leisen, 2018), as it implicitly works with a kernel density estimate of the likelihood, i.e.

$$\pi_h^{ABC}(\boldsymbol{\theta}|s^{obs}) \propto \int K_h(\left\|s^* - s^{obs}\right\|)p(s^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})ds^*. \tag{2.17}$$

In particular, in the typical ERGM case s corresponds to the sufficient statistics of the proposed model, and hence π_h^{ABC} can closely approximate the true posterior for appropriate choice of K_h. As we show below, a Gaussian kernel appears to work well for the cases studied here.

2.3.2 Kernel ABC Adaptive Importance Sampling Algorithm

We consider the idea of adaptive importance sampling (AIS) (see,e.g. Ortiz and Kaelbling, 2000; Liu, 2008; Pennanen and Koivu, 2006; Rubinstein and Kroese, 2004) where the initial proposal distribution is not good enough. This idea can be particularly useful when MPLE is suspected to be heavily biased (e.g. network size is small and with strong dyadic dependence) or even not available (e.g. fitting egocentrically sampled data with terms involving counts of triangles or higher-order cycles). Algorithm 3 describes the K-ABC adaptive importance sampling (K-ABC-AIS) algorithm, in which both the proposal distribution and distance function are updated iteratively based on the points sampled in most recent step. Prangle (2017) gave some theoretical support for similar algorithms with compact smoothing kernels.

2.3.3 Proposal Distributions for Importance Sampling

Similar to the importance of proposal distributions in MCMC (Roberts et al., 1997; Rosenthal, 2011), our proposed algorithms can greatly benefit from a well-chosen proposal distribution. Our focus here is on probability densities constructed from a common yet flexible distributional family, the multivariate Student's t distribution, $\mathcal{T}_\nu(\mu, \Sigma)$. The easy-to-calculate MPLE is a natural choice for μ, as it is typically not very far from the high density region (or, at minimum, likely to be closer than the prior mean). To mitigate the potential issue caused by an overly confident estimate of uncertainty, we consider a relatively small degree of freedom $\nu = 4$ and use a scale factor $\omega = 4$ to inflate the nominal covariance matrix given by MPLE. When using adaptive importance sampling, we advocate the use of a sequence of gradually decreasing scaling factors $\omega_1, \cdots, \omega_T$ so as to avoid wasting too many draws in low-density regions. Similarly, increasing the degree of freedom in later rounds of importance sampling is also an option. One potential pitfall for K-ABC-AIS is to split the fixed computational budget into multiple, very thin portions, which can in turn lead to an even

Algorithm 3 K-ABC adaptive importance sampling algorithm (K-ABC-AIS)

Require: Observed summary statistics $s^{obs} = S(y^{obs})$, data generating mechanism $p(y|\boldsymbol{\theta})$, prior $\pi(\boldsymbol{\theta})$

Input: $K_h(\cdot)$ ($h > 0$), B, distance metric $\|\cdot\|$ (e.g. mahalanobis distance), $\mathcal{T}_\nu$
Number of rounds of importance sampling T.
Desired sample sizes $N_t > 0, t = 1, \cdots, T$.
Scale factors $\omega_t > 0, t = 1, \cdots, T$
(Optional) $\hat{\boldsymbol{\theta}}_{MPLE}$, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}_{MPLE})$

Initialization: Let $t = 1$ and find $\hat{\mu}_1$, $\hat{\Sigma}_1$ (default $\hat{\mu}_1 = \hat{\boldsymbol{\theta}}_{MPLE}$, $\hat{\Sigma}_1 = \omega_1 \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}_{MPLE})$), set $f_1(\boldsymbol{\theta}) \equiv \mathcal{T}_\nu(\hat{\mu}_1, \hat{\Sigma}_1)$
for $t = 1, \cdots, T$ **do**
  for $i = 1, \cdots, N$ **do**
    $\boldsymbol{\theta}_t^{(i)} \sim f_t(\boldsymbol{\theta})$, (unnormalized) importance weight $w_{I,t}^{(i)} = \frac{\pi(\boldsymbol{\theta}_t^{(i)})}{f(\boldsymbol{\theta}_t^{(i)})}$
    $y_t^{(i)} \sim p(y|\boldsymbol{\theta}_t^{(i)})$ (burn-in of the MCMC-based simulation B)
    $s_t^{(i)} = S(y_t^{(i)})$
  end for
  $W_t = \frac{1}{N}\sum_{i=1}^{N}(s_t^{(i)} - \bar{s})(s_t^{(i)} - \bar{s})^\intercal$, where $\bar{s}_t = \frac{1}{N}\sum_{i=1}^{N} s_t^{(i)}$
  $d_t^{(i)} = (s_t^{(i)} - s^{obs})^\intercal W_t^{-1}(s_t^{(i)} - s^{obs})$ for $i = 1, \cdots, N$
  Perform univariate kernel density estimate on $d_t^{(i)}, i = 1, 2, \cdots, N$ to obtain a (heuristic) bandwidth h_t, and fix h_t as the scale parameter for the smoothing kernel $K_{h_t}(\cdot)$
  Assign weights to $\boldsymbol{\theta}_t^{(i)}$ as $\tilde{w}_t^{(i)} \propto w_{I,t}^{(i)} w_{K,t}^{(i)}$, and the corresponding normalized $w_t^{(i)}$
  $\hat{\mu}_t = \sum_{i=1}^{N_t} w_t^{(i)} \boldsymbol{\theta}_t^{(i)}$, $\hat{\Sigma}_t = \omega_t \sum_{i=1}^{N_t} w_t^{(i)} (\boldsymbol{\theta}_t^{(i)} - \hat{\mu}_t)(\boldsymbol{\theta}_t^{(i)} - \hat{\mu}_t)^\intercal$.
  if $t<T$ **then**
    $f_{t+1}(\boldsymbol{\theta}) \equiv \mathcal{T}_4(\hat{\mu}_t, \hat{\Sigma}_t)$
  end if
end for

Output: A set of $i = 1, \cdots, N_T$ samples $\left\{\boldsymbol{\theta}^{(i)}\right\}_{i=1}^{N_T}$ with weights $w_T^{(i)}$, drawn from $\pi_h^{ABC}(\boldsymbol{\theta}|s^{obs})$

worse proposal distribution than the one initially suggested (e.g., if the first round yields an estimate that is worse than the MPLE itself due to insufficient sampling). We note that samples from previous rounds can be retained in subsequent calculations, provided that their importance weights are handled appropriately; otherwise, however, our observation has been that using more than two to three rounds of refinement yields little benefit, and hence it is more efficient to split a fixed sampling budget into two (or at most three) waves of sampling than in a larger number. We describe the results of a simulation experiment investigating the impact of sample size below.

2.3.4 Bandwidth Selection

A key parameter determining the accuracy of inference via ABC is the bandwidth, h. If s is sufficient for $\boldsymbol{\theta}$, and $h \to 0$ we can obtain an arbitrarily good approximation to the true posterior; however, this insight is of relatively little practical use, since exact matching of simulated to observed statistics is an event of vanishingly small probability. A good working bandwidth thus strikes a balance between accuracy in approximating the target distribution (or at least its first several moments) and computational efficiency. As introduced in 2.3.1, we find that a simple bandwidth heuristic calculated on the distribution of the simulated Mahalanobis distances can yield satisfactory and stable performance at very low cost. To achieve a higher accuracy on the estimation of posterior moments, alternative approaches that might better serve the purpose in principle include *cross validation* (CV), or k-nearest neighbor CV (kNN CV), given that the goal is to estimate $\mathbb{E}[a(\boldsymbol{\theta}_j)|s^{obs}], j = 1, \cdots, p$ at the observed summary statistics s^{obs}. Bandwidths can also be chosen for each dimension, albeit with modification of the kernel and distance metric. However, preliminary experiments using these methods suggested that bandwidth selection using these approaches was often unstable, and did not yield systematic improvement on the heuristic option. At the same time, these methods were substantially more computationally expensive than the heuristic, increasing

estimation time. Per-statistic bandwidth selection methods also create challenges for tasks requiring the same weight to be applied to all elements of a draw (e.g., posterior sampling, as opposed, e.g., to estimation of marginal posterior moments), as orthogonal kernels lose the advantage of the Mahalanobis distance in accounting for correlations among statistics (and thereby efficiency) and correlated kernels are difficult to calibrate. Because we found more sophisticated bandwidth selection schemes to add cost without improving performance, we do not pursue them further here. However, it is plausible that better procedures are possible, and we regard this as an open problem.

2.4 Applications

We apply our approach to two benchmark social network datasets of varying sizes. Approximate exchange algorithm (AEA) was considered to be the current "gold-standard" for Bayesian inference of ERGMs. Hence, to illustrate the accuracy and computational efficiency of our approach, we compare the proposed algorithm with AEA. All computations in this paper are implemented in **R** (R Core Team, 2020) on a computing server (96GB RAM, with 4 Intel Xeon E5-2690v2 processors, operating at 3.00GHz, with 10 processing cores in each) – we use software suite `statnet` (Handcock et al., 2008) to simulate networks from ERGMs, and we implement AEA using the R package `Bergm` (Caimo and Friel, 2014). (Note that `Bergm` also uses `statnet` for MCMC simulation, and hence its implementation and those of K-ABC methods employed in this paper are directly comparable.) The **R** code to implement the algorithm and the data are available from `https://github.com/fyin-stats/K_ABC_ERGMs`.

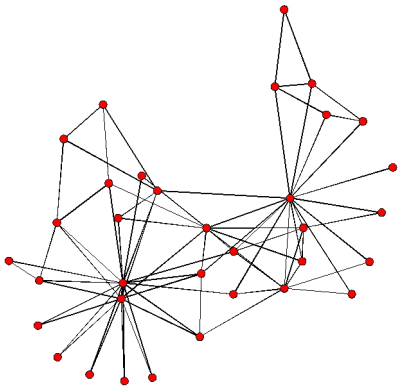

Figure 2.1: Karate club friendship network

2.4.1 Karate Club Network

The Karate club data (Zachary, 1977) represents a friendship network between 34 members in a US university karate club in the 1970's. This network consists of 78 undirected edges as presented in Figure 2.1, and the interest lies on the effect of triad closure. We consider the optimal model specification identified in Bouranis et al. (2018), which is $g(y) = (g_1(y), v(y, \phi))$. Specifically, $g_1(y) = \sum_{i<j} y_{ij}$ is the total number of edges in the network and $v(y, \phi)$ is the *geometrically weighted edgewise shared partner*(GWESP) statistic (Hunter and Handcock, 2006) defined as

$$v(y, \phi) = e^{\phi} \sum_{k=1}^{n-2} \left\{ 1 - (1 - e^{-\phi})^k \right\} EP_k(y)$$

where $EP_k(y)$ is the number of connected pairs that have exactly k common neighbors and parameter ϕ controls the decreasing rates of weights placed on higher order terms. The GWESP statistic is a common choice for modeling the tendency of forming local clusters in a network, and it has intuitive interpretation as there is diminishing positive return on the odds of an edge for each additional shared partner (e.g. one more common friend in the context of friendship network).

We obtain the "ground truth" based on a long AEA run consisting of 4 population chains (burn-in period 2500, 12500 main iterations for each chain) with a "conservative" burn-in for MCMC-based simulation, e.g., 1×10^5, which takes 4373.7 seconds (1.215 hours) to fit. While acknowledging that a holistic comparison between K-ABC and AEA sampler cannot be easily conducted due to algorithmic differences, we here provide a more limited comparison of AEA versus the K-ABC approach for typcal desiderata within a basic test case. When the total size of proposed samples is fixed, prior theory leads us to expect the AEA to outperform K-ABC. However, K-ABC can base inference on sample sizes that scale with the number of available cores given fixed computational time, as the sampling step of K-ABC is *embarrassingly* parallel. (Equivalently, given a fixed number of simulated graphs, wallclock time can be reduced under K-ABC by employing a larger number of cores.) Taking these facts into consideration, we consider the following settings,

- K-ABC-IS : One round of importance sampling, sample size: 32000, degree of freedom $\nu = 4$, scale factor $\omega = 4$, burn-in for MCMC-based simulation $B = 10^4$.
- K-ABC-AIS : Two rounds of importance sampling, sample size: (8000,24000), degree of freedom $\nu_1 = 4, \nu_2 = 4$, scale factor $\omega_1 = 4, \omega_2 = 2$, burn-in for MCMC-based simulation $B = 10^4$.
- AEA : 4 population chains, each chain with burn in = 500, main iters = 1500, auxiliary burn-in (i.e. burn-in for MCMC-based simulation) = 10^4

Note that we allow the K-ABC-IS and K-ABC-AIS to draw a total of 32000 samples, which is 4 times the total sample size in AEA. To ensure the simulated networks used in these three algorithms are of the same quality, we fix the burn-in period of MCMC-based simulation to be 10000. Under the above settings, all the algorithms are run 20 times, and their results are summarized in Table 2.1. Given the stochastic nature of these algorithms, the results differ from run to run, but overall they are very close to the ground truth. K-ABC-AIS and AEA

Table 2.1: Comparison between K-ABC and AEA. K-ABC algorithms (K-ABC-IS, K-ABC-AIS) are run on 30 cores. Wallclock runtime reported is the average across 20 runs. The ground truth value is obtained based on a long AEA run.

	Ground truth	MAE	RMSE	Runtime (secs)
K-ABC-IS (θ_1)	-3.25	0.09	0.11	14.2
K-ABC-IS (θ_2)	1.10	0.07	0.08	14.2
K-ABC-AIS (θ_1)	-3.25	0.03	0.03	14.8
K-ABC-AIS (θ_2)	1.10	0.02	0.03	14.8
AEA (θ_1)	-3.25	0.02	0.03	94.4
AEA (θ_2)	1.10	0.02	0.02	94.4

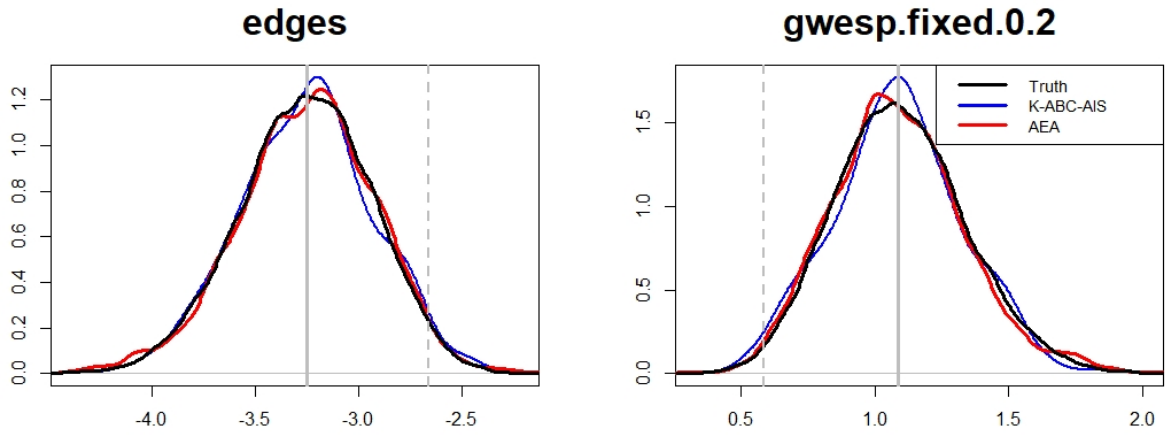

Figure 2.2: Estimated marginal posterior distribution of θ. The grey line and grey dotted line represent the MLE and MPLE, respectively.

yield essentially identical performance with respect to the estimation of posterior means, but the runtime of K-ABC-AIS is almost one seventh of that for AEA. We also note that the K-ABC-AIS performs better than K-ABC-IS, as the former gives more accurate posterior mean estimates, indicating that the adaptive scheme is indeed helpful for producing a better proposal distribution.

Figure 2.2 shows that K-ABC matches closely to each marginal posterior distribution from the ground truth. It is worth mentioning that the posterior marginal density for K-ABC is constructed based on unweighted samples obtained using sampling-importance resampling (SIR) with replacement.

2.4.2 Faux Mesa High School Network

The Faux Mesa High school network represents a total of 203 undirected friendship relations in a synthetic high school of 205 students based on an observed high school in the western U.S. (Handcock et al., 2008), and it is widely used as a realistic test network for statistical procedures. Figure 2.3 shows that the network is sparse and a large proportion of edges are formed between students in the same grade, suggesting a strong *homophily* effect on grade. The presence of some local clusters is indicative of the bias towards the formation of triangles (i.e. transitivity effect). Bearing the observed facts in mind, we consider a model with the following 3 statistics:

$$g_1(y) = \sum_{i<j} y_{ij} \qquad g_2(y) = \sum_{i<j} y_{ij} \mathbb{1}(x_i = x_j)$$

$$g_3(y) = v(y, \phi)$$

where x_i represents the grade of i-th individual and $\mathbb{1}(\cdot)$ is the indicator function, hence $g_2(y)$ counts the total number of edges connecting individuals from the same grade. $v(y, \phi)$ is the *geometrically weighted edgewise shared partner* (GWESP) statistic

$$w(y, \phi) = e^{\phi} \sum_{k=1}^{n-2} \left\{ 1 - (1 - e^{-\phi})^k \right\} EP_k(y)$$

where the decay parameter is fixed at 0.5 here, as suggested in the model proposed in Hunter et al. (2008b).

As large friendship networks are usually sparse with a high degree of homophily and transitivity (Goodreau, 2007), we consider a multivariate Gaussian prior centered at $\mu_0 =$

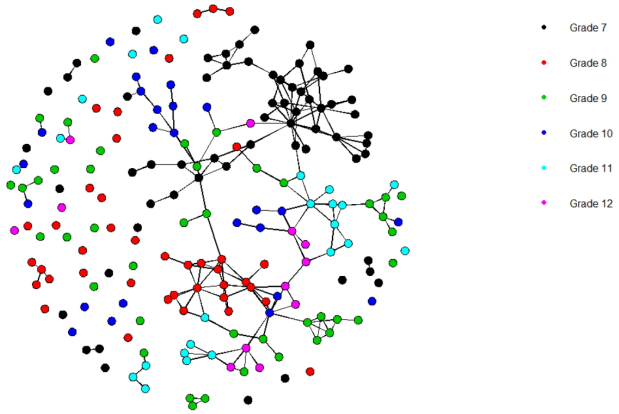

Figure 2.3: Faux Mesa High School friendship network. Colours indicate the grade.

$(-2, 0.5, 0.5)$ and covariance matrix $\Sigma_0 = 5\boldsymbol{I}_3$; this corresponds to an *a priori* belief that the coefficient associated with the edges term $g_1(y)$ is likely to be negative and those associated with the effect of grade homophily and transitivity are likely to be positive. The relatively large standard deviations ensure the statistical inference cannot be dominated by the prior belief. Also, given the sparsity of the observed data (network density ≈ 0.01), the observed edge-count based sufficient statistics might be a lot closer to their lower bound (0) than to their upper bound (total number of edges, $\frac{n(n-1)}{2}$) and their distribution might be right-skewed, hence the proposed weighting scheme will unfairly favor the sparse graph. As a remedy, we consider a monotonic power-law transformation, $T(u) = \sqrt{u+1}$ on the sufficient statistics when implementing K-ABC procedure.

The "ground truth" is again obtained based on a long AEA run, where we choose a "conservative" burn-in for MCMC-based simulation, e.g. 5×10^5, and run AEA for sufficiently long – 6 population chains, each with burn-in = 4000, main-iters = 16000 to ensure that the resulting samples can provide an adequate approximation to the "true" target (17.7 hours, 63669.7 seconds). We compare the K-ABC-AIS algorithm with the AEA under the following settings –

- K-ABC-AIS : Two rounds of importance sampling, sample size: (24000,96000); degree

Table 2.2: Comparison between K-ABC and AEA. K-ABC algorithm (K-ABC-AIS) is run on 30 cores. Wallclock runtime reported is the average across 20 runs.

	Ground truth	MAE	RMSE	Runtime (secs)
K-ABC-AIS (θ_1)	-6.20	0.04	0.05	283.6
K-ABC-AIS (θ_2)	1.97	0.01	0.01	283.6
K-ABC-AIS (θ_3)	1.24	0.07	0.07	283.6
AEA (θ_1)	-6.20	0.03	0.03	1078.3
AEA (θ_2)	1.97	0.01	0.01	1078.3
AEA (θ_3)	1.24	0.06	0.06	1078.3

of freedom $\nu_1 = 4, \nu_2 = 4$; scale factor $\omega_1 = 4, \omega_2 = 2$, burn-in for MCMC-based simulation $B = 5 \times 10^4$.

- AEA : 6 population chains, each chain with burn-in = 1000, main iters = 4000, auxiliary burn-in (i.e. burn-in for MCMC-based simulation) $= 5 \times 10^4$

Note that in this case we also allow the K-ABC-AIS to draw a total of 120000 samples, which is 4 times the total sample size in AEA. Table 2.2 shows that the point estimates given by K-ABC-AIS and AEA show virtually identical performance. Figure 2.4 shows that the estimated marginal distributions are similar, but we also notice that there is discrepancy between the marginal posterior distribution of the GWESP parameter estimated based on the K-ABC-AIS, AEA and the "ground truth." Such behavior suggests that a sufficiently long burn-in period for simulating from ERGMs can play a crucial role in both AEA and K-ABC type algorithms.

2.4.3 Computational Efficiency of K-ABC with Parallel Computing

We further investigate the computational efficiency of the proposed K-ABC approach. The presented results suggest that - (1) K-ABC seems to be able to produce comparable results to AEA when the total sample size is 4 times that of AEA; (2) K-ABC-AIS can produce

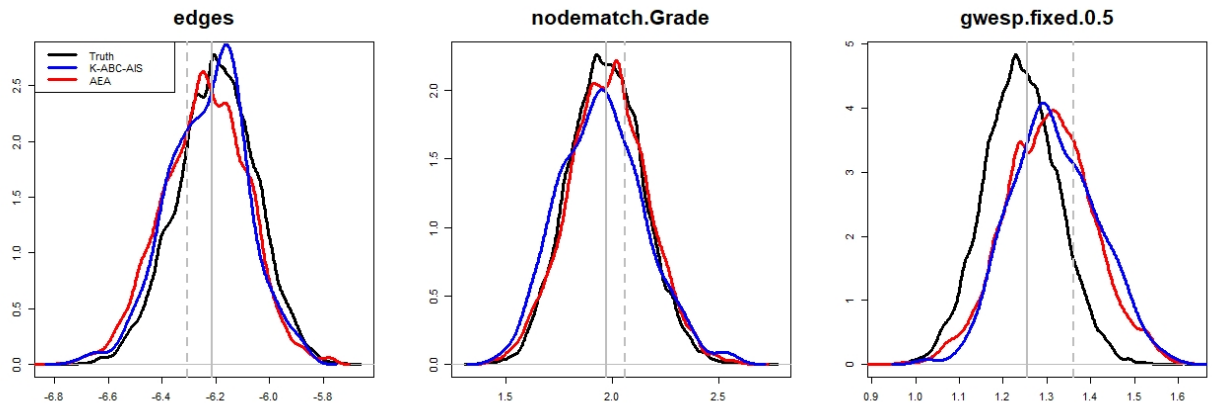

Figure 2.4: Estimated marginal posterior distribution of θ. The grey line and grey dotted line represent the MLE and MPLE, respectively.

more accurate estimations than K-ABC-IS, and it is advisable to allocate one fifth of the total sample size to the first round of importance sampling step. Therefore, we compare the computational efficiency between K-ABC-AIS and AEA under the settings which give similar level of estimation accuracy. Figure 2.5 illustrates the relative computing time of the K-ABC-AIS algorithm and AEA for the two networks Karate club (34 nodes), Faux Mesa High (205 nodes) for an increasing number of computing cores.

The relative computing time is defined as the ratio of K-ABC-AIS time to AEA time, and thus a relative computing time greater than 1 indicates that the AEA computing time is shorter, while a relative computing time smaller than 1 indicates that the K-ABC-AIS provides faster results.

Figure 2.5 demonstrates that both networks only require five cores for the K-ABC-AIS to outperform the computing time of the AEA and that the computing time can be further reduced if more computing cores are available as we can get five-fold reduction on the computing time when 30 computing cores are used. We expect further reduction on the computing time as the serial part of the K-ABC-AIS algorithm only takes a small portion of the total runtime.

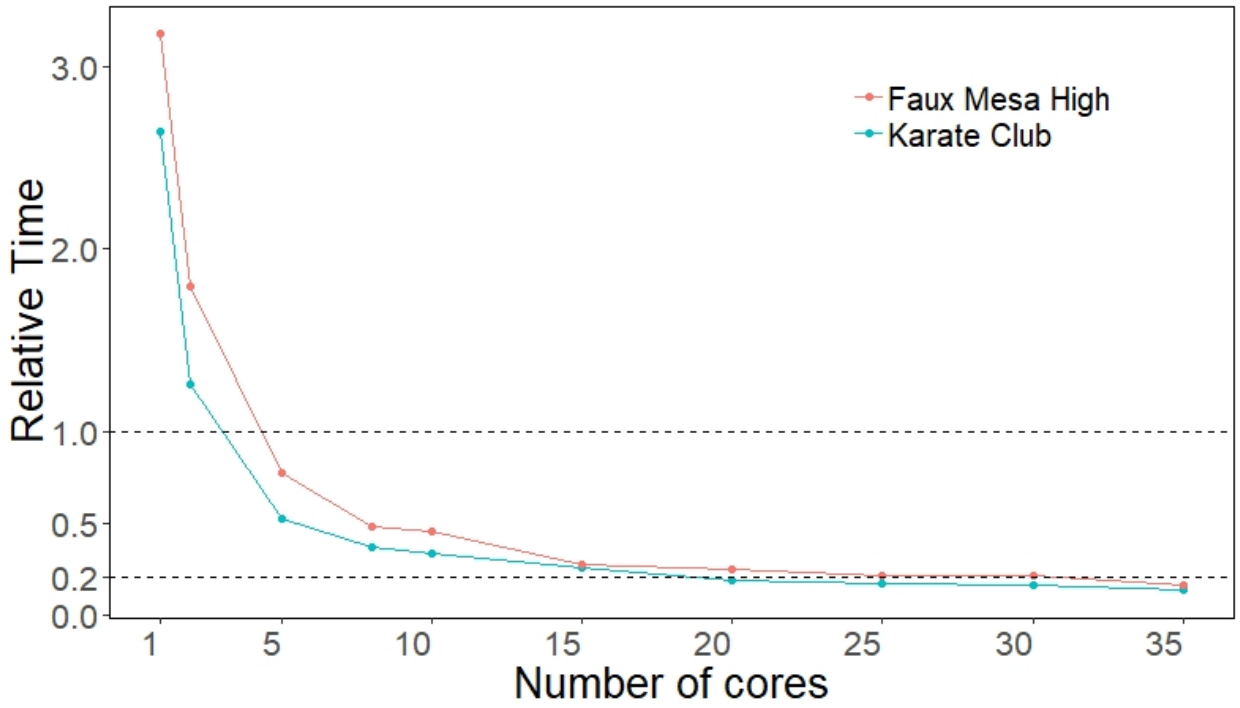

Figure 2.5: The y-axis gives the ratio of the K-ABC-AIS time to that of the AEA time. Values below 1 indicate that the K-ABC-AIS requires a shorter computing time.

2.5 Further Extensions

The proposed K-ABC approach has a wide range of connections to existing Bayesian computation techniques, including regression-adjustment ABC, Bayes Linear Analysis, and Kernel Bayes' rule. Techniques and extensions developed in these literatures could naturally be applied our case, without requiring extensive modification of our approach.

It is particularly worth noting the connection between K-ABC and Kernel Bayes' rule (KBR) (Fukumizu et al., 2011). Both of them provide posterior estimates in the form of a kernel mean, but the fundamental goal of K-ABC is obtaining samples from an approximation to the posterior distribution, while KBR can generate empirical estimates via the kernel approaches that converge to the true posterior mean embedding in the limit of infinite sample size (Fukumizu et al., 2013). The connection with KBR also makes plain the extent to which estimation of posterior moments (and hence quantiles) is fundamentally a nonparametric

regression problem, where we seek to estimate $\mathbb{E}[a(\boldsymbol{\theta})|y^{obs}]$ (for some function $a(\cdot)$) from a superpopulation defined by the joint distribution of y and $\boldsymbol{\theta}$. Because this regression is for us a purely computational device, any scheme that performs well and is computationally efficient is potentially useful. While we here use an approach that is equivalent to classical kernel regression, kernelized weighted least squares would be another natural choice, as might more exotic alternatives such as random forests or neural networks. The primary advantage of such methods is their flexibility in fitting complex functions with minimal user input, an asset that is of obvious relevance in this application. On the other hand, methods that require expensive training procedures to calibrate nuisance parameters may not improve performance sufficiently to justify the increased cost. Further work will need to be done to determine which techniques, including linear adjustments (Beaumont et al., 2002; Blum et al., 2013), non-linear adjustments (Blum and François, 2010), yield net performance gains.

Another possible direction might be approximating the posterior by a multivariate normal distribution, based on the classic Bernstein-von Mises theorem (Van der Vaart, 2000). There is recent work on variational Bayesian inference for ERGMs (Tan and Friel, 2020) based on the adjusted pseudo-likelihood (Bouranis et al., 2018), in which the posterior density $\pi(\boldsymbol{\theta}|y^{obs})$ is approximated by a Gaussian distribution $q_\lambda(\boldsymbol{\theta})$, and the parameters $\lambda = \{\mu, \Sigma\}$ are found by minimizing the Kullback-Leibler divergence (or equivalently maximizing the *evidence lower bound*). Provided the Gaussian approximation is valid, the proposed approach enables the construction of Gaussian distribution based on estimated posterior mean and posterior second moments. In case the resulting covariance matrix is not positive definite, post correction methods can be adopted (e.g., Løland et al., 2013). To date, general variational inference for ERGMs without the knowledge of MLE has not proven successful outside of demonstration models, and it is unclear whether its limitations can be overcome. However, variational approximations may be useful as additional tools for seeding ABC proposal distributions, especially in the dense graph regime where typical MCMC algorithms are often slow.

Finally, we note that extensions of the methods considered here to temporal ERGMs (TERGMs), ERGMs with latent variables, or other more complex cases are fairly straightforward given the ability to simulate from the data generating process. In particular, the modular structure of the K-ABC algorithm makes it relatively easy to accommodate such extensions within a single computational framework, so long as a simulation algorithm for the extended model is available. This is in contrast with existing strategies for ERGM inference, where are generally specialized for fairly narrow classes of models. This feature makes K-ABC a promising foundation for building ERGM-based modeling tools that are substantially more flexible than those currently in use.

2.6 Conclusion

In this paper, we introduced a kernelized approximate Bayesian computation (K-ABC) procedure for ERGMs, exploiting the algorithm's parallelizability to show substantial performance gains versus standard methods when multiple cores are available. In typical cases, the availability of sufficient statistics facilitates inference using this approach, as does the availability of relatively inexpensive crude initial estimates that can be used to construct effective proposal distributions for importance sampling. Further enhancements in performance can be obtained by iterative refinement of initial estimates, though our simulation studies suggest that (given a fixed budget) a small number of larger samples is usually preferred to many waves of small samples. Comparing our approach with the current state-of-the-art (the approximate exchange algorithm), we find that K-ABC adaptive importance sampling algorithm (K-ABC-AIS) is able to produce estimates of comparable quality at greatly reduced wallclock time so long as multiple cores are available. In a serial setting (i.e., when only one core can be used), the more refined sampling scheme of AEA is more efficient than the ABC techniques explored here, and we would recommend it as the preferred approach in this case.

AEA also has the advantage of providing very high-quality posterior approximations when run with sufficiently rigorous settings (albeit at very high cost). The two approaches thus have distinct advantages and disadvantages. One potentially useful asset of the proposed K-ABC algorithm is that it is immediately extensible to non-standard cases (such as inference from proxy statistics) that are difficult to handle using other techniques. It is also far easier to implement than AEA. This makes K-ABC a natural choice when flexibility or ease of implementation are considerations, especially if speed is of the essence.

Though parsimoniously modeling dependencies of scientific interest in networks is the primary objective for ERGMs, the development of efficient Bayesian inference on higher-dimensional ERGMs is favorable. With the recent development on high dimensional ABC algorithms ((Nott et al., 2014; Li et al., 2017)), we envision ABC as a promising framework.

Finally, we note that there are many variations on the specific implementation decisions pursued here; though we investigated the consequences of several such decisions via two case studies, there are far more possibilities for expansion and modification of the base algorithm than can be considered in any one study. We are thus optimistic for the potential for further enhancement of this very promising approach to ERGM inference.

Chapter 3

Comparisons of Model Selection Methods for ERGMs

3.1 Introduction

The exponential family random graph modeling (ERGM) framework (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Snijders et al., 2006; Hunter and Handcock, 2006) (known in older work by the term p^* (Wasserman and Pattison, 1996)) has emerged as an important approach to the statistical analysis of social network data, providing a highly general way of specifying distributions on graphs and allowing the complex dependence structure of edges in a network to be specified in terms of local structural properties (Robins et al., 2007). A wide variety of features have been proposed as potential instantiations of the different types of driving forces governing the formation of social networks (Morris et al., 2008), with the potential to accommodate an increasingly rich range of network types. At the same time, however, poor specifications can lead to unrealistic model behavior (Handcock, 2003; Schweinberger, 2011), and in practice considerable domain expertise can be required to select

terms that implement the correct dependence structure for a specific setting (Lusher et al., 2013).

Choosing among competing model specifications can be viewed as a model selection problem, making information criteria such as Akaike Information Criterion (AIC) (Akaike, 1973) and Bayesian Information Criterion (BIC)(Schwarz, 1978) natural choices for adjudication. However, these criteria rely on a number of theoretical assumptions that are frequently problematic in a network modeling context. First, edge variables in typical network models are non-independent, making it difficult to determine the effective sample size needed for size-corrected AIC and BIC calculations (Hunter et al., 2008a); indeed, at this time the theoretical justification for these criteria is unclear in the case of models for single networks with dyadic dependence (though see Kolaczyk and Krivitsky, 2015; Schweinberger et al., 2019, for some possible directions). Second, likelihood calculations for complex ERGMs rely on stochastic approximations (e.g., bridge sampling) that become expensive for large networks and where high precision is needed. This is not a barrier to parameter estimation (which typically relies on quantities such as likelihood ratios between identically specified models with similar parameters that can be more precisely computed), but makes fine distinctions between the likelihoods of similarly performing but differently specified models difficult.

The theoretical (if not computational) challenges noted above can be avoided in a Bayesian context by performing model selection via Bayes factors (Raftery, 1995). ERGM applications to date require fairly expensive, high-quality posterior simulation using methods such as the *Reversible Jump Exchange Algorithm* (Caimo and Friel, 2013), a tailored version of the conventional *Reversible Jump MCMC* algorithm (Green, 1995) combined with the *exchange algorithm* (Murray et al., 2006) by Caimo and Friel (2011) to deal with the double intractability of the posterior density, as the ERGM likelihood cannot be computed analytically in general. This approach has the advantage of being theoretically principled in the setting of fixed sample size, but the need to obtain a high-quality approximation of the Bayes

factor can be computationally demanding. Promising directions in this area include efforts like those of Bouranis et al. (2018) to reduce the computational cost by adjusting the posterior samples yielded by an analytically tractable approximation of the ERGM likelihood, e.g. the pseudolikelihood. However, its magnitude adjustment step requires approximating the *true* ERGM likelihood evaluated at MLE, which raises the same issue encountered in the calculation of AIC/BIC; thus, an efficient and easily used method for obtaining Bayes factors for general ERGMs remains elusive. Moreover, the Bayes factor itself is not always an ideal tool for model selection. The Bayes factor (and its multi-model generalizations) provides an answer to the question, "which of a set of proposed models is more likely to be the true data generating process?" assuming that the models being evaluated are *a priori* equally probable and that one is correct. While the equal probability assumption (i.e., uniform model priors) can be adjusted, the hidden assumption that one proposed model is correct – or, at least, that among incorrect models the model "more likely" to be correct is also "better" – is not entirely innocent (Bernardo and Smith, 1994; Spiegelhalter et al., 2002). When no available model is correct, the model preferred by the Bayes factor may or may not have other desirable properties (e.g., better predictive performance for some task of interest), and indeed the Bayes factor may heavily weight aspects of model performance that are not in practice those most valued by the analyst.

Relatedly, the Bayes factor can be sensitive to model features, such as the tail weight of the parameter priors, that are often chosen on semi-arbitrary grounds (and that in practice often have little impact on estimation). This creates the risk that model selection will be unduly influenced by choices of the analyst that are difficult to constrain and that are otherwise of minimal substantive importance. Moreover, the approach is only applicable to Bayesian inference, which is not at present widely used for ERGMs due to computational challanges. Thus, while the Bayes factor can be an important tool in the network analyst's arsenal, it also poses considerable difficulty in practice.

As an attempt to compensate for the difficulties of likelihood-based criteria, Goodreau (2007) and Hunter et al. (2008a) introduced graph-level "goodness of fit" (GOF) plots to assess the fit of ERGMs, in which several graph-level statistics (e.g., degree distributions, edgewise shared partner distributions) of observed networks are compared against those of simulated networks from the estimated model. The underlying idea is that draws from a fitted ERGM should have structural properties similar to the observed one, and, in particular, that network properties not explicitly used to fit the model ("out of model" properties) should be reproduced by those that were used ("in model" properties). The properties used to assess a model are usually chosen on substantive grounds, though some efforts have been made to suggest relatively "generic" statistics of broad utility (e.g. Hunter et al., 2008a; Wang et al., 2013a,b; Shore and Lubin, 2015). While this approach has proven useful in practice, it is properly a *model adequacy checking* strategy rather than a *model selection* strategy: it provides ways to identify performance deficiencies in a chosen model, but it does not provide a general rubric for choosing among competing models. Likewise, the GOF approach is not designed to provide strong information regarding the predictive performance of a fitted model. Rather, it only answers the question of how well networks drawn from a model fit to a specific data set reproduce other features of that data set. This is useful for detecting when a fitted model is *incapable* of producing realistic behavior, but it does not establish that the model will predict well (either in the context of extrapolation to new structures or interpolation of held-out data).

While it could perhaps be argued that predictive performance is not always a major concern for ERGMs, lack of predictive power at the very least suggests limitations of a model that should be borne in mind when using it. Moreover, predictive performance is clearly a consideration in many applications. For example, studies of international conflict (Hoff and Ward, 2004; Maoz et al., 2006) or bill cosponsorship (Fowler, 2006; Cranmer and Desmarais, 2011) are concerned with consequential relations for which predictions are of significant interest. These could include e.g. the ability to forecast future network states from past network

states (or merely from covariates), or conditional prediction of edge states given covariates and/or information on other edges. In the latter case, performance involving specific edge states (as opposed e.g. to unlabeled properties such as the degree variance) is of obvious importance: it is important to know whether a predicted conflict is between e.g. China and United States versus Bulgaria and Croatia. A model that successfully reproduced the degree distribution for a given year's conflict network might not perform well at predicting the degree distribution for the next year's network, nor at predicting who will be in a conflict with whom. Such a model might be judged satisfactory via conventional adequacy checks, but its value for understanding global conflict would be questionable at best.

The above suggest more explicit predictive metrics as potentially useful tools for model selection (as well as model assessment). In many fields, cross-validation (CV) techniques have been fruitfully used in this role, allowing one to assess how well the predictions from a model generalize to a new data set; flexible, easily understood, and able to be linked directly with performance outcomes of substantive interest, CV methods are well-adapted to model selection (see Arlot and Celisse, 2010, for a recent review). Typically, CV divides the data set into a *training set* (to which the model is fit) and a *test set* (against which the fitted model's predictions are evaluated under pre-specified loss functions), selecting the model with the smallest estimated loss. Many variants of this procedure exist (e.g., leave-one-out CV, k-fold CV, bootstrap CV, etc.), but all share the common feature of assessing predictive performance on a data subset that was held out during parameter estimation. Classical CV for regression models with independent and identically distributed (i.i.d.) data was proposed as early as Geisser (1975), and CV procedures have been tailored for the purpose of performance evaluation in latent variable modeling of relational data (e.g., Hoff, 2008; Dabbs and Junker, 2016; Li et al., 2020; Chen and Lei, 2018), where edge variables are conditionally independent given latent variables and hence can be straightforwardly held out. Likewise, there is work on applying CV to ERGMs where multiple networks from the same population model are available, and entire networks can be held out (Stewart et al.,

2019). Such work suggests considerable potential utility in applying CV to the more typical setting of general ERGMs on single graphs realizations.

A difficulty with standard CV techniques in a general ERGM setting is the direct dependence of edge variables, which makes it impossible to simply omit edge variables without changing the underlying probability model.[1] Intuitively, the *presence* of an edge variable in such a model is itself informative, and this information must be retained for predictions to be meaningful. This same issue arises in the context of ERGM estimation from networks with missing edge data, where the presence of edge variables must still be accounted for even when their states are unknown. Handcock and Gile (2010) introduced an estimation scheme for handling such data in the case of ignorable missingness, which resolves this difficulty by integrating over the unknown states of the missing edge variables (and thereby preserving the impact of their interactions with the variables whose states are observed). The missing data case suggests the key to obtaining CV-like procedures for ERGMs: while one cannot meaningfully hold out edge variables, one can hold out the *states* (i.e., whether a given edge variable contains an edge or a null) of edge variables, retaining their presence but treating them as missing data. Building on this intuition, Wang et al. (2016) proposed a held-out evaluation scheme for evaluating model-based imputation that was analogous to CV, which they dubbed Held-Out Predictive Evaluation (HOPE). Unlike CV, the edge variables in the validation set under HOPE are only marked as missing (i.e. NA) in the model training phase, instead of being completely eliminated. Thus, the trained model accounts for the presence of the edge variables, but it is not given information on their states. Testing is then performed by conditional prediction of the held-out edge states from the fitted model conditional on the edge variables that are not held-out. Since the core techniques needed to perform this procedure (estimation with missing edge data and conditional procedure) are supported in standard ERGM software, it can be used without the need for custom software

[1] This is related to the inconsistency of dependence models under naive subsampling, when the presence of unmeasured vertices is not accounted for (Shalizi and Rinaldo, 2013); procedures that do allow consistent estimation are discussed by Schweinberger et al. (2019).

implementations or other special considerations.

While HOPE was originally introduced in the context of imputation assessment, it is a general CV-analogue and the idea of holding out a portion of data as missing can be used for a wide range of evaluation tasks. For instance, Koskinen et al. (2018) proposed a model-based approach for the identification of influential nodes in a network by assessing the sensitivity of estimated parameters when all edge variables associated with the corresponding node is held out. In this paper, we introduce the use of HOPE to perform model selection for ERGMs, with an emphasis on simple metrics and procedures that are applicable to a wide range of network data. Using HOPE, researchers can gain information on how well the model is able to predict held-out portions of the data from other edge observations; where the model performs poorly based on held-out data, which can point to weaknesses in parameterization; and how one model's predictive performance compares to others. Because such predictive assessments automatically correct for overfitting (which by definition improves in-sample performance while harming out-of-sample performance), they can be quantitatively compared across specifications in a way that some other metrics cannot. Taken together, these assessments can assist researchers in improving models and facilitate the comparison of fit across multiple models.

With the development of HOPE as a promising model selection technique for ERGMs based on cross-validation, a well-established model selection framework for general statistical models, a natural and critical question that remains to be answered is, how well does HOPE perform compared to other methods? As a first attempt to answer this question, we propose to conduct simulation studies and evaluate the performance of various model selection techniques using the following metrics: model selection accuracy; predictive deviance on an independent test data; and prediction accuracy of edge variables on independent test data. These three criteria exemplify three potential desiderata in model selection: the ability to select the true model specification among a set of candidate models (where the true generating

process is present); the ability to select a model that can yield the smallest deviance on new data generated from the same data generating process (whether or not the predictive model reflects that generative process); and the ability to produce accurate edgewise predictions on the new data generated from the same data generating process (again, without necessarily assuming access to a "true" model). Details of the simulation study is given in subsequent sections.

The remainder of this chapter is structured as follows. In Section 3.2, we review the conventional methods for comparing / selecting ERGMs. In Section 3.3, we introduce Held-Out Predictive Evaluation (HOPE), a cross-validation analogue method, for selecting competing ERGM specifications. Section 3.4 provides several comprehensive simulation studies to compare the conventional and the novel model-selection methods. Finally, we conclude with a discussion in Section 3.5.

3.2 Review on Traditional Model Selection Methods

In this section, we review the classic methods for comparing / selecting ERGMs. In particular, we shall focus on the information-criterion based methods, including the Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978) and graphical goodness of fit (Hunter et al., 2008a). As BIC approximates the Bayes factor, and more exact calculation of Bayes factor is very challenging for ERGMs (Caimo and Friel, 2013), we do not include the Bayes factor in the set of candidate model selection techniques in this simulation study.

3.2.1 AIC and BIC

Recall the general definition of AIC and BIC for a statistical model $\mathcal{M}$

$$AIC(\mathcal{M}) = 2(\#\text{of parameters in } \mathcal{M}) - 2(\text{maximized log-likelihood under } \mathcal{M}) \quad (3.1)$$

$$BIC(\mathcal{M}) = \log(N)(\#\text{of parameters in } \mathcal{M}) - 2(\text{maximized log-likelihood under } \mathcal{M}) \quad (3.2)$$

where N denotes the sample size. The goal is to search for $\mathcal{M}$ that minimizes $AIC(\mathcal{M})$ or $BIC(\mathcal{M})$. As the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{MLE}$ corresponds to the point at which the log-likelihood is maximized given observed data, the maximized log-likelihood for model $\mathcal{M}$ is

$$\log p(y^{obs}|\hat{\boldsymbol{\theta}}_{MLE}) = \hat{\boldsymbol{\theta}}_{MLE}^{\intercal} g(y^{obs}) - \psi(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}) \quad (3.3)$$

where $\psi(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})$ is the log-partition function evaluated at $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$, which is analytically intractable in general and has to be approximated through Monte Carlo simulations. As discussed in Section 2.2, exact MLE is unavailable for general ERGMs, and hence $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$ is usually approximated by the MCMC MLE.

Consider the partition function (i.e., exponentiated log-partition function), or equivalently the normalizing factor,

$$\kappa(\boldsymbol{\theta}) = \exp(\psi(\boldsymbol{\theta})) = \exp(\log \sum_{y' \in \mathcal{Y}_n} \exp\left\{\boldsymbol{\theta}^{\intercal} g(y')\right\}) = \sum_{y' \in \mathcal{Y}_n} \exp\left\{\boldsymbol{\theta}^{\intercal} g(y')\right\}$$

and note that $\kappa(\mathbf{0}) = 2^{\binom{n}{2}}$ for undirected networks and $2^{n(n-1)}$ for directed networks. The $\kappa(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})$ can be approximated by an unbiased estimator, following Gelman and Meng (1998),

Hunter and Handcock (2006) and Friel (2013)

$$\frac{\kappa(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{\kappa(\mathbf{0})} = \frac{\kappa(u_T\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{\kappa(u_0\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})} = \prod_{l=0}^{T-1} \frac{\kappa(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{\kappa(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})} \tag{3.4}$$

where $0 = u_0 < u_1 < u_2 < \cdots < u_T = 1$ are equally spaced within the range. Importance sampling is used to estimate the ratio of normalizing factors in (3.4), we note

$$\begin{aligned}
\frac{\kappa(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{\kappa(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})} &= \frac{\sum_{y'\in\mathcal{Y}_n}\exp(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))}{\kappa(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})} \\
&= \sum_{y'\in\mathcal{Y}_n}\frac{\exp(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))}{\exp(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))/p(y'|u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})} \\
&= \sum_{y'\in\mathcal{Y}_n} p(y'|u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})\frac{\exp(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))}{\exp(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))} \\
&= \mathbb{E}_{y'\sim p(y'|u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}\big[\frac{\exp(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))}{\exp(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y'))}\big]
\end{aligned} \tag{3.5}$$

and hence a natural unbiased estimate of this expectation is

$$\widehat{\frac{\kappa(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}{\kappa(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\exp(u_{l+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y_k^{(l)}))}{\exp(u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}^{\intercal}g(y_k^{(l)}))}$$

where $y_1^{(l)}, \ldots, y_K^{(l)}$ are i.i.d draws from $p(y|u_l\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})$. Larger Monte Carlo sample size K and the *temperature* T can lead to more precise estimates, but with higher computational costs. We chose $K = 1000$ and $T = 20$ in this work as empirical results show diminishing returns when going beyond these values. With the estimated normalizing factor $\widehat{\kappa(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}$,

the maximized log-likelihood (3.3) is approximated as

$$\log p(y^{obs}|\hat{\boldsymbol{\theta}}_{MLE}) = \hat{\boldsymbol{\theta}}_{MLE}^{\intercal} g(y^{obs}) - \exp(\widehat{\kappa(\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}})}) \tag{3.6}$$

An additional challenge for BIC is that the seemingly straightforward sample size N is no longer straightforward for ERGMs with dyadic-dependent terms. Instead, we only know that the true value of N is within the range $[1, \binom{n}{2}]$ for undirected networks, and $[1, n(n-1)]$ for directed networks, and believed to be closer to 1 when dyadic-dependent terms play more substantial roles in the network formation process. The **R** package `ergm` conservatively uses the maximum of the range to estimated BIC. The current understanding is that both AIC and BIC can be problematic for selecting models, due to the approximations involved in their calculations, which we shall investigate via simulation studies.

3.2.2 Graphical Goodness of Fit

Graphical goodness of fit (GOF) is a simulation-based method that is commonly used to assess the model performance. The basic idea is that a fitted ERGM should recapitulate key structural features similar to the observed observed network (Hunter et al., 2008a), which has its root in posterior predictive assessment (Gelman et al., 1996). The choice of the set of structural features for constructing these GOF procedures depends on both empirical and theoretical questions. An implementation of GOF for ERGMs with several commonly-used high-level graph statistics is provided in the package `ergm` through the function `gof.ergm`.

Figure 3.1 displays an example visualization of the output of `gof.ergm`. The boxplots are generated based on 100 simulated draws, and the bold black lines represent the observed statistics. The top-left, top-right and bottom-left subfigures in Figure 3.1 correspond to

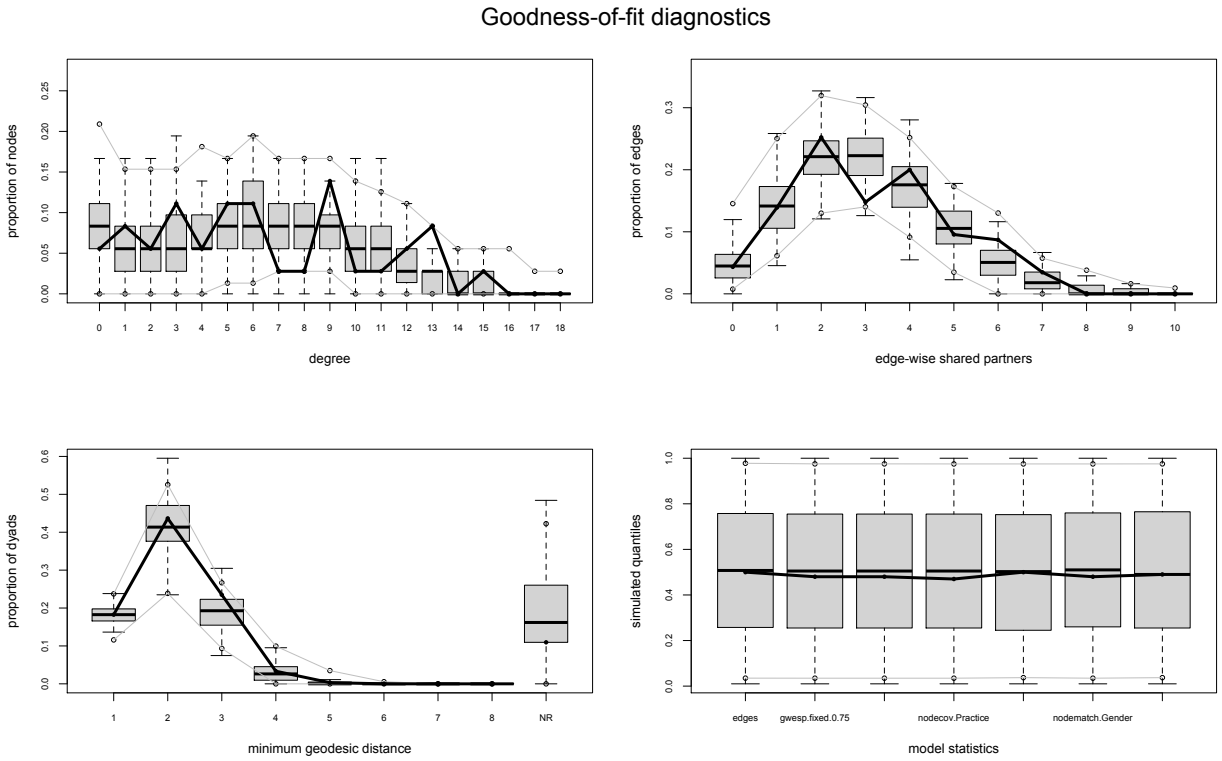

Figure 3.1: Sample graphical goodness of fit plots for ERGMs.

the distributions of degree counts, edgewise shared partner counts, and minimum geodesic distances, respectively. These statistics offer key structural information about networks and hence are often of substantial interest to network researchers (Hunter et al., 2008a). Typically either none of these statistics are included in the model, or only a small subset of them are. The bottom-right subfigure give the quantiles of observed model statistics (bold black lines) under the distribution of simulated model statistics. These quantiles are expected to be not too far away from 0.5 because the true MLE should yield expected statistics equal to observed statistics for exponential-family distributions (Barndorff-Nielsen, 1978). In the example above, the bold black lines lie within the range of boxplots for most of the statistics, which indicates that the observed data is plausible under the fitted model.

This graphical assessment method offers ways to identify performance deficiencies in models, and is particularly useful for detecting when a fitted model is *incapable* of producing realistic behavior. However, it is clearly lacking as a model selection method, because it neither

provides quantification of how poorly the model fits the data nor establishes that the model will predict well (either in the context of extrapolation to new structures or interpolation of held-out data).

3.3 Held-Out Predictive Evaluation (HOPE)

The Held-Out Predictive Evaluation (HOPE) was introduced by Wang et al. (2016) as a model-based technique for imputing missing edge data in social networks, based upon the ability to fit and simulate from ERGMs in the presence of missing data (Handcock and Gile, 2010). We extend HOPE for the purpose of selecting between competing ERGMs.

3.3.1 Likelihood Inference for ERGMs in the Presence of Missing Data

We consider the binary random adjacency matrix Y, with the support $\mathcal{Y}_n$. In the general case where Y may be only partially observed, we introduce the indicator matrix W that is of the same size as Y

$$W_{ij} = \begin{cases} 1, & \text{edge variable } Y_{ij} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

and denote the observed part of Y by $Y_{obs} = \{Y_{ij} : W_{ij} = 1\}$ and the unobserved part by $Y_{mis} = \{Y_{ij} : W_{ij} = 0\}$; then $Y = Y_{obs} \cup Y_{mis}$. The *complete data*, $\{Y_{obs}, Y_{mis}, W\}$, are not fully observed, and the *observed data*, are $\{Y_{obs}, W\}$. Letting lower-case symbol represent

the realized values of random variables, we define $\mathcal{Y}_n(y_{obs}) = \{v : v \cup y_{obs} \in \mathcal{Y}_n\}$, which is the set of possible values of Y_{mis}, under the constraint that the observed part is equal to y_{obs}. Therefore $y_{obs} \cup \mathcal{Y}_n(y_{obs})$ is the subset of $\mathcal{Y}_n$ with the observed part equal to y_{obs}.

Under the assumption that the missing data mechanism is missing at random (MAR) (Rubin, 1976), that is,

$$P(W = w|Y = y, \psi) = P(W = w|Y_{obs} = y_{obs}, \psi) \qquad \forall y \in y_{obs} \cup \mathcal{Y}_n(y_{obs}) \tag{3.8}$$

where ψ denote the parameters that govern the missing data mechanism, and are distinct from the ERGM parameters $\boldsymbol{\theta}$. As (3.8) implies that missing edge variables Y_{mis} do not contain any information about ψ, we have the joint likelihood for ψ and $\boldsymbol{\theta}$ given observed data y_{obs} and observed missing pattern w_{obs}

$$L[\boldsymbol{\theta}, \psi|Y_{obs} = y_{obs}, W = w_{obs}] \propto P(W = w_{obs}|Y_{obs} = y_{obs}, \psi)P(Y_{obs} = y_{obs}|\boldsymbol{\theta})$$

Thus likelihood-based inference for $\boldsymbol{\theta}$ from $L[\boldsymbol{\theta}, \psi|Y_{obs} = y_{obs}, W = w_{obs}]$ will be the same as likelihood-based inference for $\boldsymbol{\theta}$ using the (so-called) *face-value likelihood* based solely on Y_{obs} (Handcock and Gile, 2010)

$$P(Y_{obs} = y_{obs}|\boldsymbol{\theta}) \propto \sum_{v \in \mathcal{Y}_n(y_{obs})} P(Y = y_{obs} \cup v|\boldsymbol{\theta}) \tag{3.9}$$

We note that the above derivation holds for any parametric models for social networks. Starting from this point, we assume the random behaviour of Y is characterized by an

ERGM. To better illustrate the idea, we focus on linear ERGMs with the counting measure specified in (2.1). Combining (2.1) and (3.9), we have the conditional probability of Y_{mis} given $Y_{obs} = y_{obs}$

$$\begin{aligned} P(Y_{mis} = y_{mis}|Y_{obs} = y_{obs}, \boldsymbol{\theta}) &= \frac{P(Y_{obs} = y_{obs}, Y_{mis} = y_{mis}|\boldsymbol{\theta})}{P(Y_{obs} = y_{obs}|\boldsymbol{\theta})} \\ &= \frac{P(Y = y|\boldsymbol{\theta})}{P(Y_{obs} = y_{obs}|\boldsymbol{\theta})} \\ &= \frac{P(Y = y|\boldsymbol{\theta})}{\sum_{v \in \mathcal{Y}_n(y_{obs})} P(Y = y_{obs} \cup v|\boldsymbol{\theta})} \\ &= \exp[\boldsymbol{\theta}^\intercal g(y_{obs} \cup y_{mis}) - \psi(\boldsymbol{\theta}|y_{obs})], \quad y_{mis} \in \mathcal{Y}_n(y_{obs}) \end{aligned} \tag{3.10}$$

where $\psi(\boldsymbol{\theta}|y_{obs}) = \log \sum_{v \in \mathcal{Y}_n(y_{obs})} \exp[\boldsymbol{\theta}^\intercal g(y_{obs} \cup v)]$. This result gives a simple way to simulate from the conditional distribution and hence produce multiple imputations for missing edge variables. Also note that (3.11) becomes

$$P(Y_{obs} = y_{obs}|\boldsymbol{\theta}) \propto \sum_{v \in \mathcal{Y}_n(y_{obs})} P(Y = y_{obs} \cup v|\boldsymbol{\theta}) \propto \exp[\psi(\boldsymbol{\theta}|y_{obs}) - \psi(\boldsymbol{\theta})] \tag{3.11}$$

which can then be maximized with respect to $\boldsymbol{\theta}$ by two sets of MCMC samples: the first term by a chain conditional on y_{obs} and the second term by a chain on the complete data. That said, the estimation of ERGM parameters $\boldsymbol{\theta}$ is only slightly more difficult than those discussed in Section 2.2.

Both the `ergm` package (Hunter et al., 2008b) of the `statnet` (Handcock et al., 2008) software suite for **R** (R Core Team, 2018) and software MPNet (Wang et al., 2014) have implemented

simulation-based algorithms for approximating the MLE of $\boldsymbol{\theta}$ under (3.11). Additionally, both allow simulations from the estimated parameters with or without missing data (the former is used in goodness of fit assessments).

In this work, we implement the HOPE procedure in **R** taking advantage of the functionalities available in the `ergm` package. To run conditional simulations from ERGMs, the observed part of the graph can be fixed using the `constraint` argument in the `simulate.ergm` function. The procedure is repeated for each set of held-out edge variables in observed data, and the simulated networks are then evaluated based on several criteria from various aspects. This makes HOPE analogous to cross-validation in that we are evaluating how well we can predict data that are not used for model training.

3.3.2 HOPE for Model Selection of ERGMs

While first introduced in Wang et al. (2016) as a powerful multiple imputation technique for missing edge data, HOPE was extended for the purpose of model selection for ERGM in Yin et al. (2019). We denote the entire index set of edge variables, $\mathcal{D} = \big\{(i,j)|i,j \in \mathbf{V}, i < j\big\}$ ($\mathcal{D} = \big\{(i,j)|i,j \in \mathbf{V}, i \neq j\big\}$, if Y is directed), which can be partitioned into M subsets, $A_1, \cdots, A_M$, where $\bigcup_{m=1}^{M} A_k \subseteq \mathcal{D}$. We let A_m^c represent the relative complement of A_m in $\mathcal{D}$. Intuitively, the HOPE procedure operates by holding out the values (or equivalently, states) of edge variables in one subset to create artificial missingness while fitting to the resulting partially observed data, using the resulting estimate to predict the values of the held-out data. Compared to conventional cross-validation, in which the variables themselves - and not merely their values - are held out. While this distinction is immaterial for independence models, it is consequential for typical ERGMs. While other schemes are also feasible, two natural options for holding out network data are random sampling of edges and removal of all edge values associated with a randomly chosen vertex,

- Random edge held-out: $\mathcal{D}$ is randomly divided into M non-overlapping [2], equally-sized batches of edge variables, with each batch being held-out in turn. A similar strategy was used by Hoff (2008) for the purpose of selecting the optimal number of dimensions in latent space structure for latent space models. Based on our empirical findings, we recommend to hold out $n-1$ edge variables in each batch, where n is the network size.

- Node held-out: All edge variables involving a particular node are simultaneously held-out. Koskinen et al. (2018) used a similar strategy to identify model-specific influential nodes under the ERGM framework.

The random edge removal is analogous to M-fold cross validation, and leave-one-out cross validation in the extreme case where $M = |\mathcal{D}|$. The general procedure for HOPE then proceeds as follows. For $m = 1, \cdots, M$:

1. Fit an ERGM to an artificially created partially observed data $y_{A_m^c}$ based on (3.11), which in turn gives the corresponding held-out data MLE, $\hat{\theta}^{(m)}$.

2. Obtain N_{sim} draws, $v^{m,1}, \ldots, v^{m,N_{sim}}$, from the conditional distribution $P(\cdot|Y_{A_m^c} = y_{A_m^c}, \hat{\boldsymbol{\theta}}^{(m)})$ defined in (3.10), which yield $\hat{y}^{m,k} = v^{m,k} \cup y_{A_m^c}, k = 1, \ldots, K$.

3. Evaluate the ability of the model to accurately predict the held-out data under error metrics of interests. The choice of which metrics to use will depend on which structural features users deem substantively important for their model.

Yin et al. (2019) provided two general strategies to hold out edge variables (node held-out and random edge held-out) and proposed a comprehensive list of metrics that can be used to evaluate a model's performance at different granularities (dyad, node and graph level). For the purpose of this work, we shall focus on the random edge held-out and the prediction

[2] It is a common practice in cross-validation to split the observed dataset into disjoint subsets, in order to reduce the correlation between estimated models and hence to reduce the variance.

accuracy of held-out edge variables. The prediction accuracy of the held-out edge variables under model $\mathcal{M}$ are evaluated via the following cost function C

$$C_{\mathcal{M}}(\hat{y}, y) = \sum_{m=1}^{M} \sum_{(i,j)\in A_m} \ell(\hat{y}_{ij}, y_{ij})$$

where $\hat{y}_{ij} = \frac{\sum_{k=1}^{N_{sim}} \hat{y}_{ij}^{m,k}}{N_{sim}}$ is a Monte-Carlo estimate of the conditional expectation $\mathbb{E}[Y_{ij}|Y_{A_m^c} = y_{A_m^c}]$, and $\ell(a, b)$, for example, can take the following forms

- Squared loss : $\ell(a, b) = |a - b|^2$
- Absolute loss : $\ell(a, b) = |a - b|$

The goal for HOPE is to search for the model $\mathcal{M}$ that minimizes the cost $C_{\mathcal{M}}$. As y_{ij}'s are binary random variables, we also consider the area under the curve (AUC) for receiver operating curve (ROC) as another candidate metric. Therefore, under the HOPE framework, we have three candidate methods, "HOPE-Square loss", "HOPE-Absolute loss" and "HOPE-ROCAUC".

We present a simple example based on a six-node undirected binary network to better illustrate HOPE. The observed network data have the adjacency matrix representation as follows [3]

[3]The lower triangle part and the diagonal elements of the adjacency matrix are omitted because the network is undirected and the self-loop is prohibited, respectively.

$$y = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ & 0 & 0 & 0 & 1 \\ & & 1 & 1 & 1 \\ & & & 0 & 1 \\ & & & & 0 \\ & & & & \end{pmatrix}$$

For a randomly chosen subset of edge variables $A_m = \big\{(1,3),(1,5),(2,5),(3,4),(4,6)\big\}$, the states of these chosen edge variables are set as missing (marked by "?"),

$$y_{A_m^c} = \begin{pmatrix} 1 & ? & 1 & ? & 0 \\ & 0 & 0 & ? & 1 \\ & & ? & 1 & 1 \\ & & & 0 & ? \\ & & & & 0 \\ & & & & \end{pmatrix}$$

and competing ERGM specifications are fit to the resulting partially observed data $y_{A_m^c}$. Based on the fitted models, we carry out conditional simulations to "predict" the states and conditional probability of the edge variables in A_m.

3.4 Simulation Studies

We conduct simulation studies to compare the traditional model selection methods (AIC, BIC) and the novel cross-validation analogue method HOPE. In particular, we consider two scenarios, closed-$\mathcal{M}$ and open-$\mathcal{M}$, where the former corresponds to the scenario under

which the true model belongs to the set of candidate models, while the latter corresponds to the scenario under which the true model does not belong to the set of candidate models. Under each scenario, we consider networks of varying sizes, and evaluate the model selection methods on various aspects as follows

- Model selection accuracy.
- Predictive deviance on independent test data.
- Prediction accuracy of edge variables on independent test data.

The general structure of the simulation study for the closed-$\mathcal{M}$ setting is as follows:

- Begin with the "ground truth" model for a given network and specify several competing models.
- Obtain N_{train} i.i.d draws from the "ground truth" model as the "training data" and N_{test} i.i.d draws from the "ground truth" model as the "test data".
- Fit the competing ERGM specifications to each network in the "training data" using MCMC MLE described in Section 2.2.3, and calculate AIC, BIC according to 3.2.1. Identify the *best* model in the set of competing models according to AIC and BIC values.
- Carry out the HOPE procedure introduced in Section 3.3. Identify the *best* model in the set of competing models according to the values of the cost functions.
- Evaluate the performance of each model selection method in correctly selecting the true model, along with the predictive deviance and predictive accuracy of edge variables on the entire "test data". In particular, when evaluating the prediction accuracy of edge variables, we hold out the same number of edge variables as that in HOPE procedure, and conduct conditional simulation to make predictions about those held-out edges.

where we have $N_{train} = 50$ and $N_{test} = 50$ in the second step.

3.4.1 Closed-$\mathcal{M}$

"Ground truth"

The "ground truth" models are ERGM distributions defined on the three most commonly used network sufficient statistics but with distinct parameters,

- $g_1(y) = \sum_{i<j} y_{ij}$, total number of edges.
- $g_2(y) = e^{\phi} \sum_{k=1}^{n-2} \left\{1 - (1 - e^{-\phi})^k\right\} EP_k(\mathbf{y})$, 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.
- $g_3(y; \mathbf{X}) = \sum_{i<j} y_{ij} \mathbb{1}_{\{\mathbf{X}_i = \mathbf{X}_j\}}$, total number of edges with endpoints sharing same value on node level covariate $\mathbf{X}$, often known as nodematch term.

We fix node-level covariate $\mathbf{X}$ to be a binary variable, and let half of the nodes take value 0, while the other half takes value 1 on $\mathbf{X}$. The network size is 40.

We have two different parameter settings

$$\boldsymbol{\theta}^{40}_{true} = \begin{pmatrix} \text{edges} & \text{gwesp}, \phi = 0.25 & \text{nodematch}(\mathbf{X}) \\ -4.95 & 2.5 & 0.25 \\ -3.75 & 1.25 & 1.25 \end{pmatrix}$$

to ensure that the simulated networks have similar mean degree ($\sim$ 10) across different parameter settings but represent distinct formation mechanism.

- "Strong triadic closure" (first numeric row of $\boldsymbol{\theta}^{40}_{true}$): strong triadic closure effect but weak homophily effect.
- "Intermediate triadic closure, homophily" (second numeric row of $\boldsymbol{\theta}^{40}_{true}$): intermediate triadic closure and homophily effect.

Therefore, for each fixed network size, we have two "ground truth" model that share the same set of sufficient statistics but are equipped with very different parameter values.

Candidate Models

The candidate models $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_{10}\}$ are listed in Table 3.1, where $\mathcal{M}_6$ corresponds to the "ground truth" model specification. The "diff = T" in Table 3.1 indicates the differential homophily statistics, $\sum_{i,j} y_{ij} \mathbb{1}(x_i = x_j = a)$, which are different from the uniform homophily statistics, $\sum_{i,j} y_{ij} \mathbb{1}(x_i = x_j)$, as the former allows each group to have a unique propensity for within-group ties. The graphletCount(1) counts the number of G_1 (i.e. a type of three-node induced subgraphs, where exactly two edges are present), which helps capture the total number of "open two-path" structures and is implemented in package `ergm.graphlets` (Nebil et al., 2015). It is also worth noting that $\mathcal{M}_2$ is a natural parsimonious version of the true model when the homophily effect does not play a leading role in the network formation process, and $\mathcal{M}_3$ is a natural parsimonious version of the true model when the triadic closure effect does not play a leading role.

Table 3.1: List of candidate models for both closed-$\mathcal{M}$ and open-$\mathcal{M}$ scenarios. ✓ indicates the corresponding term is included in the respective model.

	edges	nodematch($\mathbf{X}$)	gwesp,$\phi = 0.25$	graphletCount(1)
$\mathcal{M}_1$	✓			
$\mathcal{M}_2$	✓		✓	
$\mathcal{M}_3$	✓	✓		
$\mathcal{M}_4$	✓	✓, diff=T		
$\mathcal{M}_5$	✓			✓
$\mathcal{M}_6$	✓	✓	✓	
$\mathcal{M}_7$	✓	✓, diff=T	✓	
$\mathcal{M}_8$	✓	✓		✓
$\mathcal{M}_9$	✓	✓	✓	✓
$\mathcal{M}_{10}$	✓	✓, diff=T	✓	✓

Results

We first discuss the performance on model selection accuracy. Table 3.2 shows the distribution of selected models under different model selection methods and different true model coefficients. The key observations are summarized as follows. First of all, none of these model selection methods dominate all other methods. Second, the information-criterion-based methods, especially BIC, prefer models with fewer terms in general. Third, "HOPE-Absolute loss" and AIC are more capable of selecting the true model when triadic closure effect plays nearly a dominant role in the true network formation process; BIC yields the best model selection accuracy when neither triadic closure nor homophily plays a minor role. Thirdly, the "HOPE-Square loss" and "HOPE-ROC AUC" appear to be consistently incompetent in identifying the true models. Overall, BIC appears to be very robust under the closed-$\mathcal{M}$ scenario as it either selects the true model or falls back to the most natural parsimonious models $\mathcal{M}_2$ and $\mathcal{M}_3$.

Figures 3.2 and 3.3 present the mean predictive deviance for models selected by different model selection methods when the true data generating process is "strong triadic closure" and "intermediate triadic closure, homophily", respectively. The BIC seems to be able to consistently select model specifications that yield smallest predictive deviance, regardless of

Table 3.2: Selected models under different true model coefficients and model selection methods. Network size = 40. $\mathcal{M}_6$ corresponds to the true model specification.

model specifications / model selection techniques	AIC	BIC	HOPE-Sq. loss	HOPE-Abs. loss	HOPE-ROC AUC
Parameter setting : strong triadic closure					
Rank	2	3	5	**1**	4
$\mathcal{M}_1$: edges	0	0	4	0	1
$\mathcal{M}_2$: edges + gwesp.fixed.0.25	19	40	8	8	8
$\mathcal{M}_3$: edges + nodematch.x	0	0	4	0	2
$\mathcal{M}_4$: edges + nodematch.x.0 + nodematch.x.1	0	0	2	0	0
$\mathcal{M}_5$: edges + graphlet.1.Count	0	0	10	1	11
$\mathcal{M}_6$: edges + nodematch.x + gwesp.fixed.0.25	14	10	4	**15**	7
$\mathcal{M}_7$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25	2	0	5	9	2
$\mathcal{M}_8$: edges + nodematch.x + graphlet.1.Count	0	0	9	3	13
$\mathcal{M}_9$: edges + nodematch.x + gwesp.fixed.0.25 + graphlet.1.Count	14	0	4	11	5
$\mathcal{M}_{10}$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25 + graphlet.1.Count	1	0	0	3	1
Parameter setting : intermediate triadic closure, homophily					
Rank	2	**1**	5	3	4
$\mathcal{M}_1$: edges	0	0	0	0	0
$\mathcal{M}_2$: edges + gwesp.fixed.0.25	0	0	0	0	0
$\mathcal{M}_3$: edges + nodematch.x	0	5	18	0	5
$\mathcal{M}_4$: edges + nodematch.x.0 + nodematch.x.1	0	0	11	0	4
$\mathcal{M}_5$: edges + graphlet.1.Count	0	0	1	2	1
$\mathcal{M}_6$: edges + nodematch.x + gwesp.fixed.0.25	34	**45**	6	20	12
$\mathcal{M}_7$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25	4	0	2	13	3
$\mathcal{M}_8$: edges + nodematch.x + graphlet.1.Count	0	0	10	1	20
$\mathcal{M}_9$: edges + nodematch.x + gwesp.fixed.0.25 + graphlet.1.Count	8	0	2	7	3
$\mathcal{M}_{10}$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25 + graphlet.1.Count	4	0	0	7	2

the parameter settings of the true model, though AIC by design provides an approximation to the predictive deviance and hence is expected to select model specifications that yield the smallest predictive deviance. We note that "HOPE-Absolute loss" is comparable to those information-criterion based methods, while "HOPE-Square loss" and "HOPE-ROCAUC" are inferior to all other methods at this task as well.

Figures 3.4, 3.5 and 3.6 show the mean squared prediction error and mean absolute prediction error and mean AUC of the ROC curves across 50 replicates. The models selected by BIC seem to yield the smallest prediction errors and largest AUC of the ROC curves in general, and we note that AIC and "HOPE-Absolute loss" also give comparable performance. The predictive performance of "HOPE-ROCAUC" and "HOPE-Square loss" is still inferior, which might be due to their incompetency in selecting the true model in such setting.

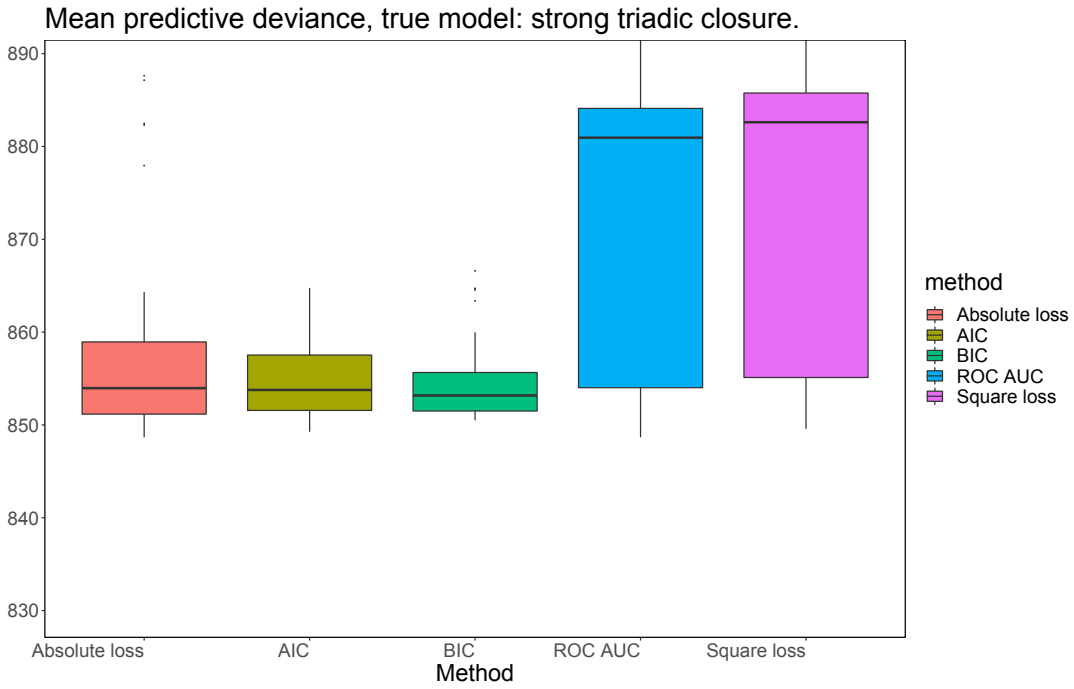

Figure 3.2: Boxplots for mean predictive deviance on the independent test data. True model: strong transitivity.

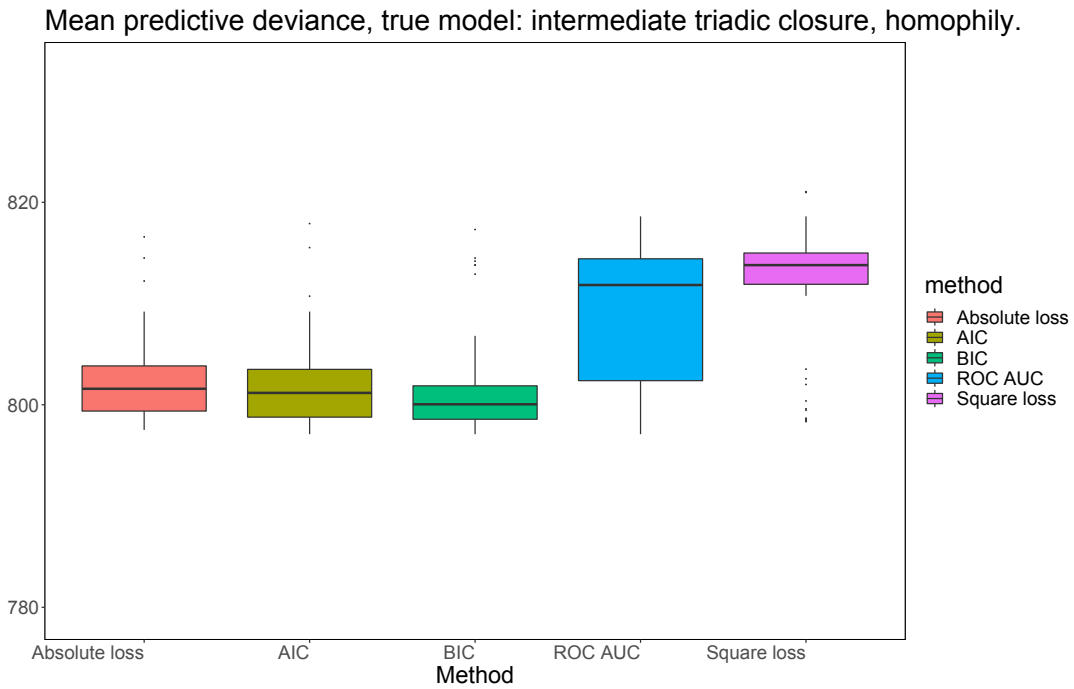

Figure 3.3: Boxplots for mean predictive deviance on the independent test data. True model: intermediate transitivity, homophily.

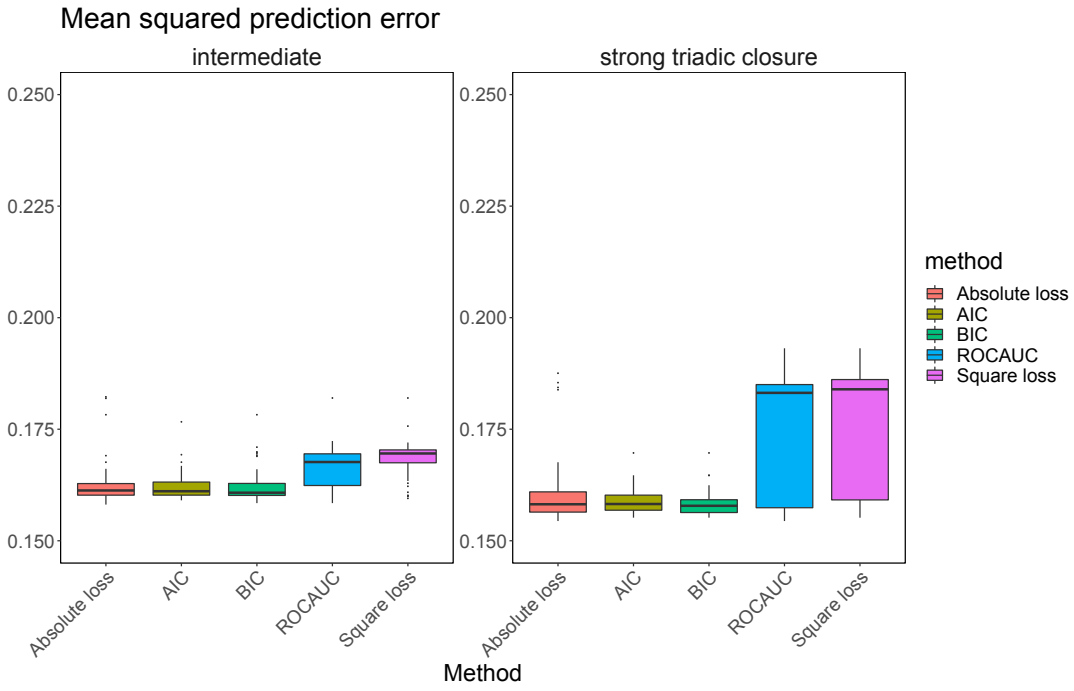

Figure 3.4: Boxplots for mean squared prediction errors on the independent test data, under different true model coefficients and model selection methods.

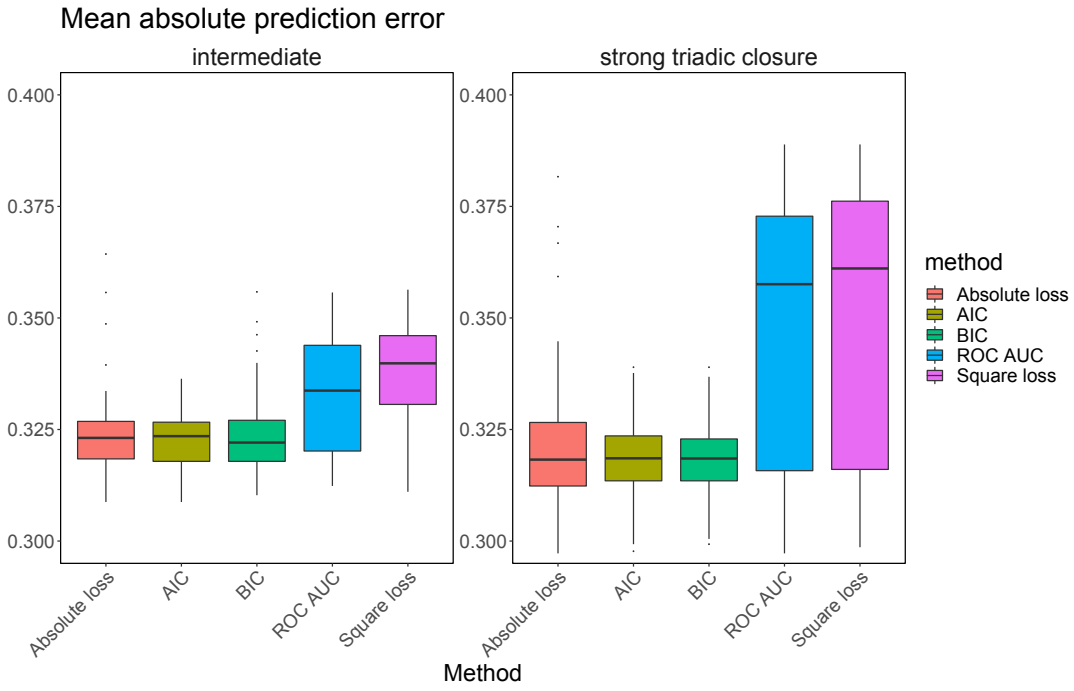

Figure 3.5: Boxplots for mean AUC of ROC curves on the independent test data, under different true model coefficients and model selection methods.

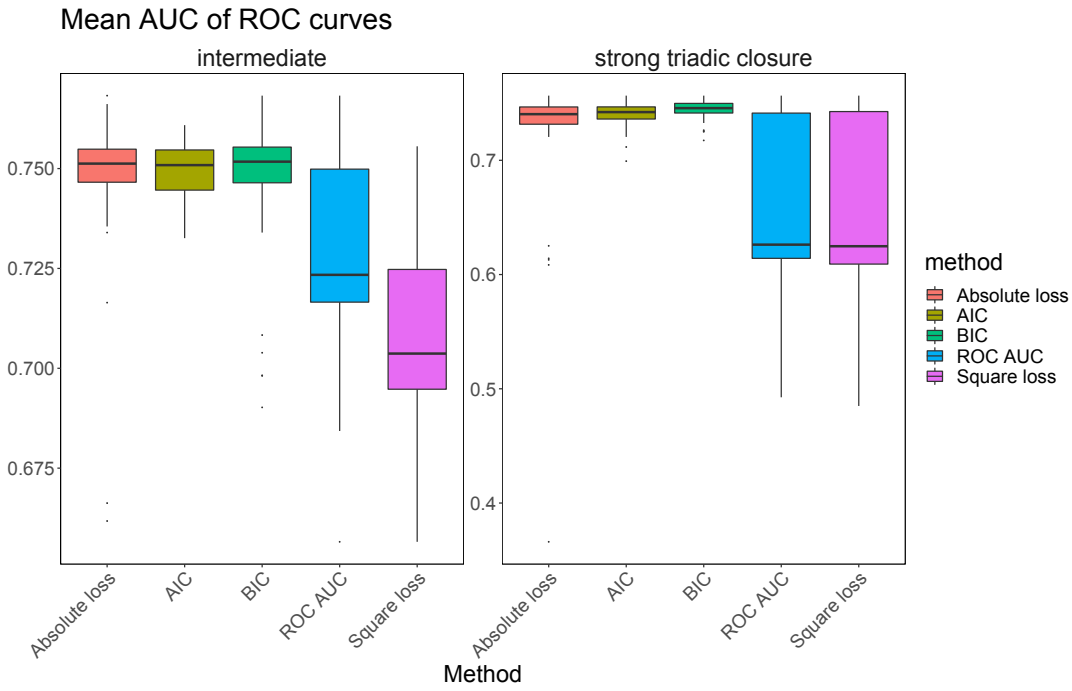

Figure 3.6: Boxplots for mean squared prediction errors on the independent test data, under different true model coefficients and model selection methods.

3.4.2 Open-$\mathcal{M}$

"Ground truth"

We consider a representative graph generated from the model "Intermediate triadic closure, homophily" in the Closed-$\mathcal{M}$ scenario and fit a latent space model (Hoff et al., 2002) with intercept α and 2-D Euclidean distance terms $|z_i - z_j|$

$$\begin{aligned} P(Y|Z,\alpha) &= \prod_{i<j} P(Y_{ij} = y_{ij}|z_i, z_j, \alpha) \\ &= \prod_{i<j} logit^{-1}(\alpha - |z_i - z_j|) \end{aligned}$$

to this representative graph using the `ergmm` function in R package `latentnet` (Krivitsky and

Handcock, 2008). The resulting latent space model serves as the "ground truth" model in the Open-$\mathcal{M}$ scenario. Although latent space models can be viewed as mixtures of Bernoulli random graph models (Schweinberger et al., 2019) (i.e., ERGMs with dyadic independent terms only), we note none of the candidate models in Table 3.1 is explicitly equivalent to the resulting latent space model, that is, the "ground truth" model. We simulate "training data" and "test data" from the "ground truth" model in a similar manner to that in Closed-$\mathcal{M}$ scenario.

Candidate Models

The "ground truth" model here indeed provides a latent-space-based compression of the unobservable original data generating process, that is, a combination of homophily effects and triadic closure effects. Therefore we still consider the set of candidate models shown in Table 3.1, but this time none of the candidate models is the exact true data generating model.

Results

Table 3.3 shows the distribution of selected models under different model selection methods (3 runs in which we encounter convergence issues fitting ERGMs are dropped). There is no true model in this open-$\mathcal{M}$ scenario. Instead of focusing on the model selection accuracy, we analyze the qualitative patterns of the selected models –

- First of all, we note that the information-criterion-based methods, AIC and BIC, prefer models with less terms in general, while the HOPE-based methods seem to be less confined to models with less terms.
- Second, different metrics can lead to considerable difference in model choices under

HOPE – Squared loss and AUC of the ROC curve seem to be more similar to AIC and BIC, while the absolute loss appears to be very different from other methods as it favors $\mathcal{M}_5$, "edges + graphletCount(1)".

Table 3.3: Selected models, open-$\mathcal{M}$. Network size = 40.

model specifications / model selection techniques	AIC	BIC	HOPE-Sq. loss	HOPE-Abs. loss	HOPE-ROCAUC
$\mathcal{M}_1$: edges	0	0	0	0	0
$\mathcal{M}_2$: edges + gwesp.fixed.0.25	0	0	0	0	0
$\mathcal{M}_3$: edges + nodematch.x	16	27	8	0	5
$\mathcal{M}_4$: edges + nodematch.x.0 + nodematch.x.1	10	2	10	0	6
$\mathcal{M}_5$: edges + graphlet.1.Count	0	1	5	26	7
$\mathcal{M}_6$: edges + nodematch.x + gwesp.fixed.0.25	1	0	1	1	3
$\mathcal{M}_7$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25	0	0	0	0	0
$\mathcal{M}_8$: edges + nodematch.x + graphlet.1.Count	14	13	18	8	17
$\mathcal{M}_9$: edges + nodematch.x + gwesp.fixed.0.25 + graphlet.1.Count	2	3	4	7	6
$\mathcal{M}_{10}$: edges + nodematch.x.0 + nodematch.x.1 + gwesp.fixed.0.25 + graphlet.1.Count	4	1	1	5	3
Total	47	47	47	47	47

Figures 3.7, 3.8, 3.9 and 3.10 present the mean predictive deviance, mean squared prediction errors, mean absolute prediction errors and mean predictive AUC of ROC curves under the open-$\mathcal{M}$ scenario. We have the following observations –

- All methods, except for "HOPE-Absolute loss", yield similar predictive performance with respect to deviance on the independent test data.
- Though "HOPE-Absolute loss" gives the worst predictive performance with respect to the deviance, it yields the lowest prediction errors and highest AUC values on the test data.
- Other HOPE-based methods ("HOPE-Square loss" and "HOPE-Absolute loss") also give better performance with respect to edge predictions compared to information-criterion-based methods.

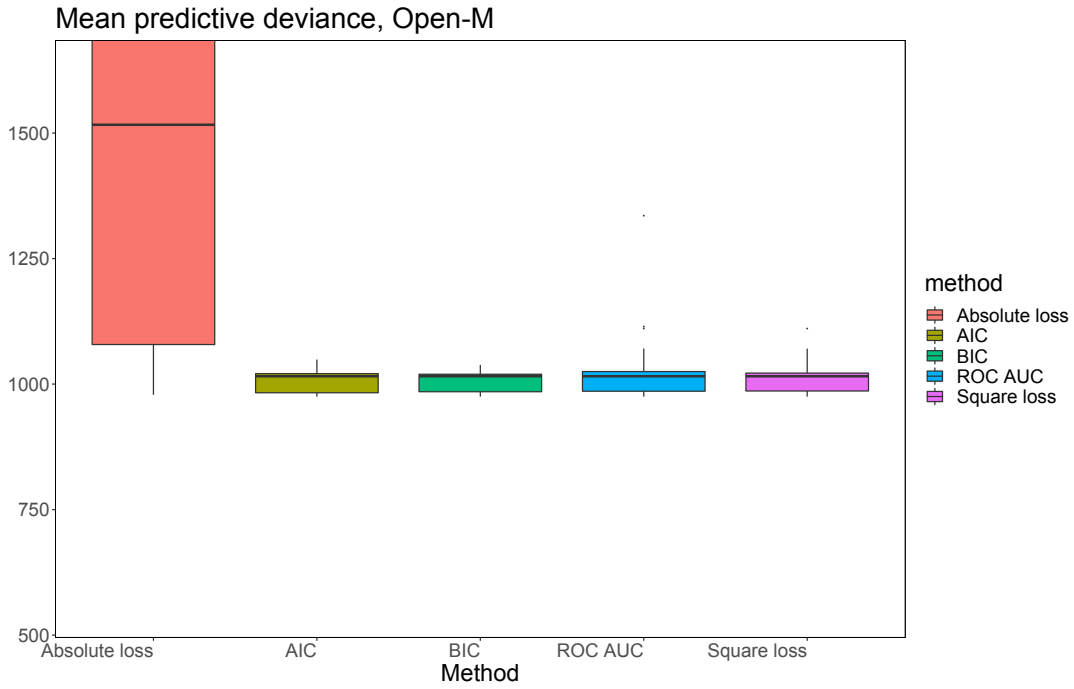

Figure 3.7: Boxplots for mean predictive deviance on the independent test data.

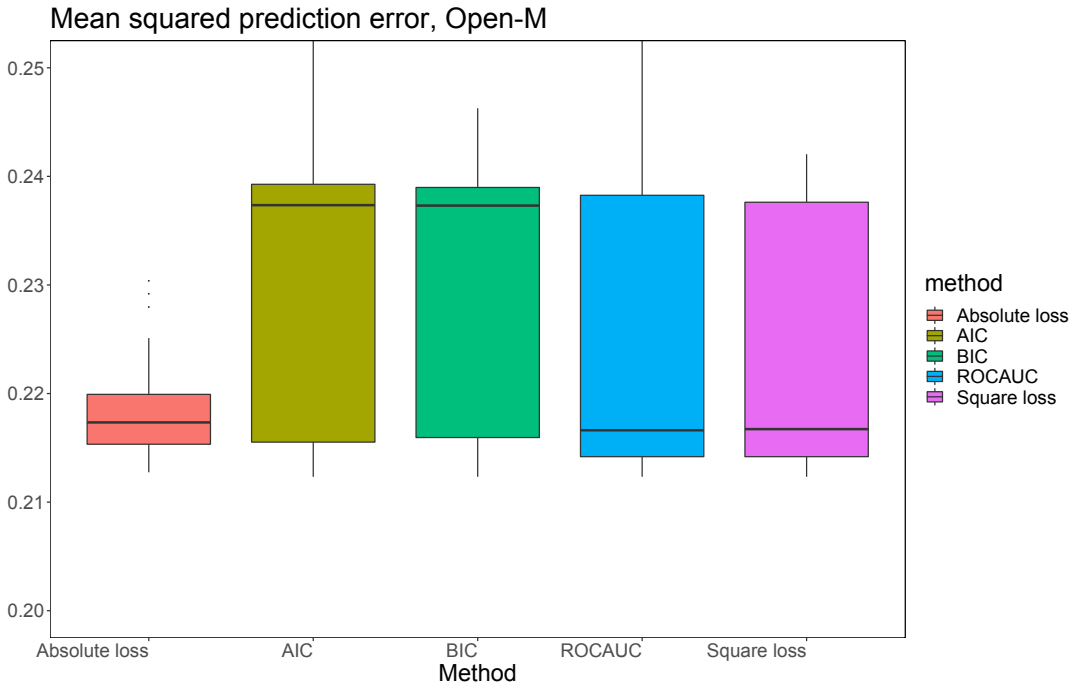

Figure 3.8: Boxplots for mean squared prediction errors on the independent test data.

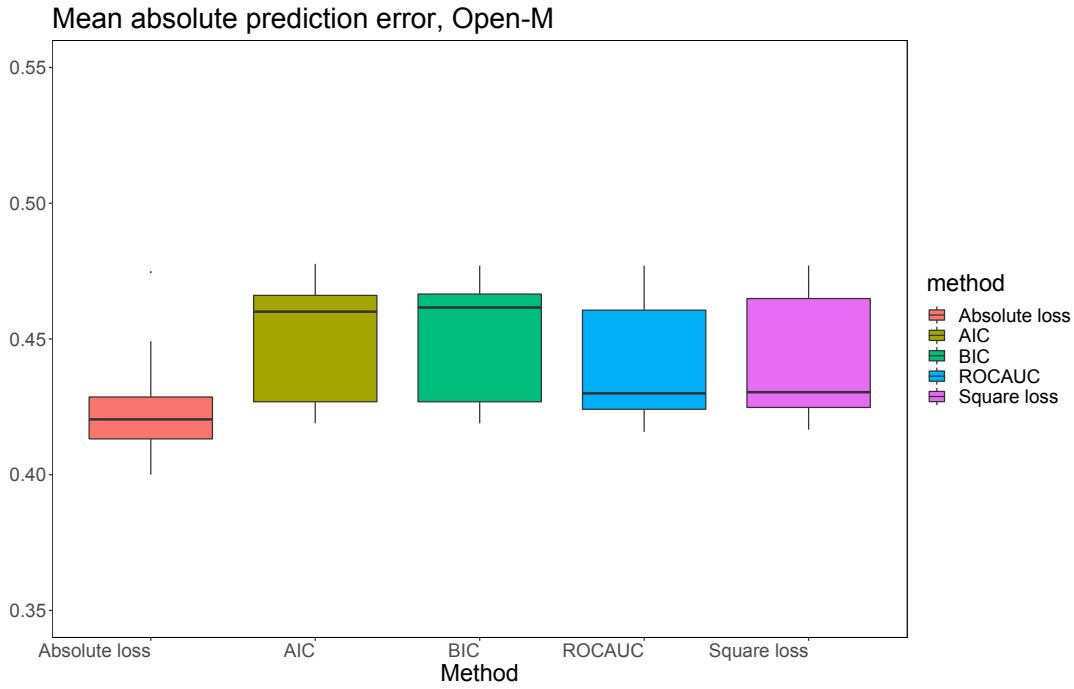

Figure 3.9: Boxplots for mean absolute prediction errors on the independent test data.

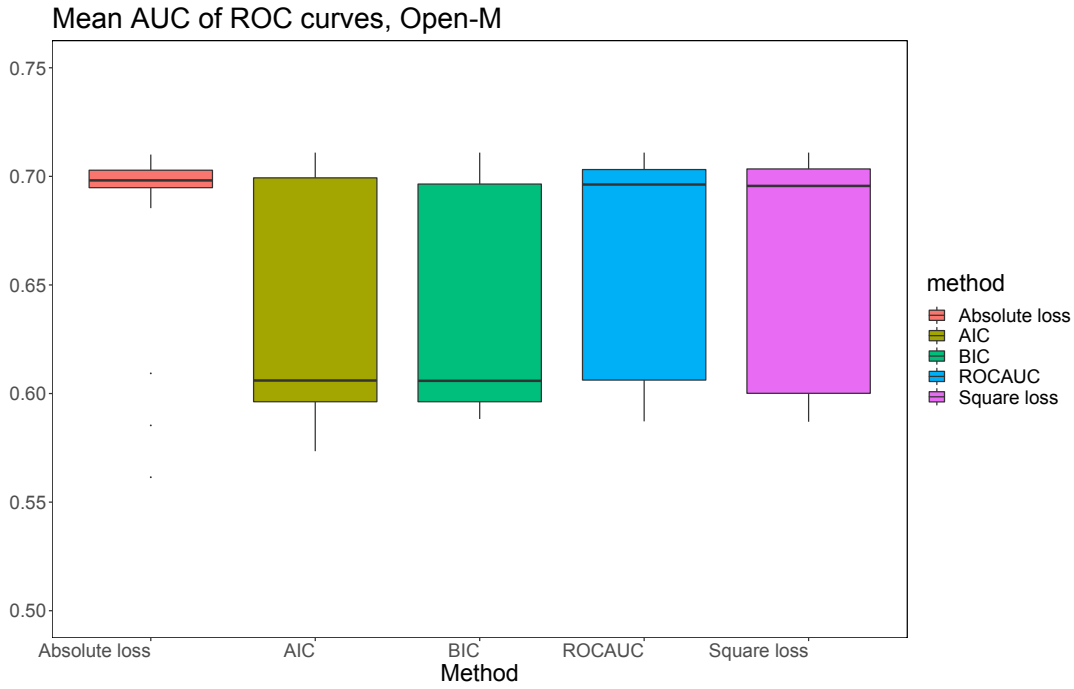

Figure 3.10: Boxplots for mean AUC of ROC curves on the independent test data.

3.5 Discussion and Conclusions

In this chapter, we first reviewed several conventional methods along with a recently proposed method, Held-Out Predictive Evaluation (HOPE), for selecting competing model specifications of ERGMs. We designed and conducted systematic simulation studies to compare their performance on two scenarios, closed-$\mathcal{M}$ and open-$\mathcal{M}$. The performance is evaluated with respect to several desiderata, including the model selection accuracy, predictive deviance, and prediction accuracy of edge variables.

The simulation studies offer important insights about the strengths and weaknesses of the model selection techniques under consideration. We first focus on the closed-$\mathcal{M}$ scenario. The information-criterion-based methods, especially BIC, appear to be superior in terms of selecting the true models under closed-$\mathcal{M}$ scenario. We note that AIC appears to be slightly better than the BIC provided that the network formation process is mainly governed by one dominant driving force, as BIC seems to penalize too much on the model complexity and hence is more prone to identify a simpler model. The pattern in predictive performance is similar to that of model selection accuracy, as the information-criterion-based methods are more capable of selecting the true model in the closed-$\mathcal{M}$ scenario, and this advantage naturally extends to the predictive performance. Overall, we recommend BIC under closed-$\mathcal{M}$ scenario because it is more robust than AIC in the sense that BIC either selects the true model or falls back to the most natural parsimonious models. However, AIC and BIC are both observed to work well overall, and HOPE with absolute loss delivers comparable performance with respect to predictive outcomes (albeit at greater computational cost); all three of these methods are hence reasonable options when predictive performance is of primary interest. By contrast, HOPE based on the squared error loss or AUC criteria performs quite poorly in our experiments, and we are unable to recommend either for closed-$\mathcal{M}$ model selection.

We now turn to the open-$\mathcal{M}$ scenario. As none of the candidate models is explicitly equivalent to the true model, we comment on qualitative patterns of the selected models. The information-criterion-based methods, AIC and BIC, prefer models with fewer terms in general, while the HOPE-based methods seem to be less confined to models with fewer terms. Different loss metrics for HOPE can lead to substantial difference in terms of selected models, with squared error loss and AUC of ROC curves behaving more similarly to the information criteria than to the absolute loss. With regards to the predictive performance, despite being the worst in terms of predictive deviance, the "HOPE-Absolute loss" clearly outperforms all other methods in terms of prediction accuracy of edge variables. We also note that the "HOPE-ROCAUC" and "HOPE-Square loss" are slightly better than information-criterion-based methods with respect to prediction accuracy of edge variables, and comparable to information-criterion-based methods with respect to the predictive deviance. In the open-$\mathcal{M}$ setting, then, HOPE appears to provide a strong alternative to more traditional methods.

Of course, it may be argued that, from a practical standpoint, the closed-$\mathcal{M}$ versus open-$\mathcal{M}$ distinction is moot: for real measurements of complex systems, *no* available model will plausibly be "true," and we are hence always in an open-$\mathcal{M}$ setting. Although we regard this as self-evidently true, we would suggest that there may be a useful heuristic distinction between settings where an available model may be very close to the data generating process (conditional on available covariates) and settings in which one is able to obtain at best very crude approximations thereto. Our open-$\mathcal{M}$ scenarios are of this latter type, while our closed-$\mathcal{M}$ scenarios may be viewed as more reflective of the former. From that standpoint, our findings suggest that information criteria will work quite well where an ERGM family is available that can relatively closely approximate the target graph distribution, while HOPE has strong advantages (with the exception of deviance prediction) when the ERGM families under consideration offer only rough approximations to the target. When it is unclear which regime one is in, the absolute loss HOPE procedure offers a sound compromise for most inferential goals (though, as noted, it can lead to poor deviance predictions when no close

approximation to the generating process is available).

While this work provides a first comparative examination of predictive and information-based methods for ERGM model selection, there is considerable room for further development in this area. Examples of fruitful directions include theoretical investigation of both the formal properties of the HOPE procedure and the effective degrees of freedom of ERGM distributions. Among the questions raised by the present simulation study include:

- *Why do information-criterion-based methods work well under typical ERGM use cases in which standard asymptotic theory does not clearly apply?* The answer suggested by recent work by Schweinberger and colleagues is that the conditions for concentration of the ERGM likelihood as a function of $N = |\mathbf{V}|$ may be fairly weak, in which case standard approximations to the behavior of the deviance may in fact hold despite dyadic dependence. While some sufficient conditions for such concentration properties to hold are known, finding a general characterization remains an open problem.

- *Why does BIC give good performance in terms of selecting true models under closed-$\mathcal{M}$ despite using a clearly inaccurate number for effective sample size?* As the effective data degrees of freedom for an ERGM with dependence terms is less than the nominal degrees of freedom, one would expect the (already conservative) BIC to be strongly biased towards low-dimensional models. The overall strong performance of the BIC in these tests suggests that this concern may be misplaced. One possible explanation is that, since the nominal degrees of freedom scale as $\mathcal{O}(N^2)$, any constant "deflation" of the degrees of freedom due to dependence will have a vanishing impact in relative terms. Specifically, if the effective degrees of freedom scale as αN^2 (for some $\alpha < 1$), then the ratio of the ideal versus nominal complexity penalty in the BIC will be $(\log \alpha + 2 \log N)/(2 \log N) = 1 + 2(\log \alpha)/(\log N)$, which goes to 1 as $N \to \infty$. While this is a plausible explanation so long as α is non-small, the conditions under which the latter holds have not been characterized.

- *Is it possible for us to have a better estimate of the effective sample size that can be used to calculate BIC (et al.)?* Expanding on the above, a formal characterization of the effective degrees of freedom of an ERGM distribution remains an open problem. Some bounds are available for special cases (e.g., models with block-local dependence), but a general characterization remains an open problem.

- *What theoretical guarantees can be provided for HOPE in terms of model selection and generalization errors?* Recent consistency and concentration results (as noted above) for some classes of ERGMs suggest a way forward here, although generalization to the missing data case is necessary. The impact of dependence on predictive accuracy is also subtle, among other things implying that one must carefully consider the predictive task (since e.g., the conditional distribution of a pair of edge variables can deviate substantially from the product of the full conditionals for each edge variable separately). Results for errors in conditional prediction of single edges would seem to be the most natural starting point.

As the above suggest, this is a rich problem space with many avenues for further exploration.

Chapter 4

Finite Mixtures of ERGMs for Modeling Ensembles of Networks

4.1 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., 2019b), 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 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 also 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).

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., Robins et al., 2007, for a review)), as base distributions. Faust and Skvoretz (2002) introduced 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 single populations 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). Nonparametric models (e.g., latent space or block models) have also been employed for studying sets of networks, e.g. hierarchical mixed membership stochastic blockmodels for multiple networks (Sweet et al., 2014). In general, those methods have either not posited a generative model for the parameters of the base distribution (as in descriptive meta-analytic approaches), have not attempted to jointly estimate population-level and network-level parameters (as in conventional meta-analysis), or have assumed a simple hierarchical form in which coefficients are taken to be drawn from a simple population distribution (often Gaussian) with common mean and variance. The latter work well for homogeneous (super)populations; but when the network ensemble reflects higher levels of heterogeneity, more structure is required. In contrast, work such as that of Durante and Dunson (2018); Lehmann (2019) explicitly considers heterogeneity within graph subpopulations, but assumes that the subpopulation labels are observed. 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 a mixture of ERGMs to model the generative process of ensembles of networks in which the group 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 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 the dyadic independence property (e.g., a priori stochastic blockmodel, 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 Metropolis-within-Gibbs algorithm to perform Bayesian inference for the proposed model, with both the subpopulation assignments and the ERGM parameters in the subpopulations being estimated simultaneously. Given that our primary focus is to develop a practical procedure that can obtain meaningful subpopulations, we employ a pseudo-likelihood approximation to the ERGM likelihood for efficient computation; while we show here that this approach can work well, more advanced MCMC techniques can also be deployed to obtain more accurate estimates when the interest lies mainly in the inference of subpopulations-specific parameters. (It is also possible to use the pseudo-likelihood when updating subpopulation assignment parameters and then use high-accuracy MCMC-based likelihood calculations to update subpopulation-specific parameters, offering additional options for speed/accuracy tradeoffs.) We approach the problem of choosing number of subpopulations from a model selection perspective, using a version of deviance information criterion.

The remainder of this chapter is structured as follows. In Section 4.2 we briefly introduce the exponential-family random graph models (ERGMs) and common estimation techniques. Section 4.3 describes the idea of mixtures of ERGMs, along with our estimation algorithms and our proposed method for selecting the number of subpopulations. Section 4.4 presents simulation studies showing that the proposed method can accurately recover the true subpopulation assignment and model parameters. Section 4.5 shows the results of our method applied to a political co-voting data analysis and Section 4.6 provides another case study

that aims at clustering advice-seeking networks among school teachers. Section 4.7 concludes with a discussion.

4.2 Exponential-family Random Graph Models (ERGMs)

4.2.1 Definition and Estimation

We consider the general formulation of ERGMs defined in (1.1),

$$\mathbb{P}_{\boldsymbol{\eta}}(\mathbf{Y}=\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\intercal}\mathbf{g}(\mathbf{y};\mathbf{X}) - \psi_{\mathbf{g},\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta})\right)h(\mathbf{y}), \quad \mathbf{y}\in\mathcal{Y}_n, \tag{4.1}$$

Exact evaluation of the normalizing factor, $\psi_{\mathbf{g},\boldsymbol{\eta},\mathbf{X},\mathcal{Y}_n}(\boldsymbol{\theta})$, involves integrating an extremely rough 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, 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 (see Hunter et al., 2012, for a review). To date, the most frequently used approaches include:

- Maximum pseudo-likelihood estimation (MPLE; Besag (1974)) adapted to ERGMs by Strauss and Ikeda (1990).
- Markov Chain Monte Carlo MLE (MCMC MLE; Geyer and Thompson (1992)) adapted to ERGMs by Handcock (2003); Hunter and Handcock (2006).
- Stochastic approximation (SA; Robbins and Monro (1951); Pflug (1996)) adapted to ERGMs by Snijders (2002).

- Fully Bayesian inference based on the approximate exchange algorithm (Caimo and Friel, 2011).

Recent developments on ERGM estimation have concentrated on: (1) finding better initial values for simulation-based MLE, including the *partial stepping* technique (Hummel et al., 2012) and *contrastive divergence* (CD,Hinton (2002))-based techniques adapted to ERGMs by Krivitsky (2017); and (2) more accurate tractable approximations to ERGM likelihood than pseudo-likelihood, such as the adjusted pseudo-likelihood (Bouranis et al., 2017, 2018) for fast Bayesian inference. Despite the computational challenges, these and related strategies have made ERGM inference practical for well-posed model families (e.g., see Schweinberger et al. (2019) for a recent review).

4.2.2 Size-adjusted Parameterizations

It is worth noting that the behavior of Equation (4.1) 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.,

$$\eta_1(\boldsymbol{\theta}) = \theta_1 - \log n, \tag{4.2}$$

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, although other size-adjusted parameterizations are also possible (e.g., Butts and Almquist, 2015; Kolaczyk and Krivitsky, 2015).

4.3 Finite Mixtures of ERGMs

We assume a population of networks $(\mathbf{Y}^{(1)}, \mathbf{V}^{(1)}, \mathbf{X}^{(1)}), \ldots, (\mathbf{Y}^{(m)}, \mathbf{V}^{(m)}, \mathbf{X}^{(m)})$, where $\mathbf{Y}^{(i)}$ is a graph structure on vertex set $\mathbf{V}^{(i)}$ with covariate set $\mathbf{X}^{(i)}$. Our interest is in modeling $\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(m)}$ given $(\mathbf{V}^{(1)}, \mathbf{X}^{(1)}), \cdots, (\mathbf{V}^{(m)}, \mathbf{X}^{(m)})$, where it will be assumed that the respective graph structures are conditionally independent given the generative process, vertex sets, and covariates.

4.3.1 Model Formulation

We model the generative process of the network ensemble as a finite mixture, with each mixture component (equivalently, subpopulation, or "cluster") being an ERGM distribution with cluster-specific parameters. (See Figure 4.1.) Given K clusters, the *a priori* probability for a network to belong to cluster k is τ_k for $k = 1, 2, \cdots, K$, and the probability law governing the formation of the network in group k is parameterized by Eq. (4.1) with cluster-specific parameter vector $\boldsymbol{\theta}_k \in \mathbb{R}^{q_k}$ and cluster-specific mapping to the natural parameters $\boldsymbol{\eta}_k(\boldsymbol{\theta}_k) = (\eta_{k,1}(\boldsymbol{\theta}_k), \cdots, \eta_{k,p_k}(\boldsymbol{\theta}_k)) \in \mathbb{R}^{p_k}$.

More specifically, the marginal likelihood for network $\mathbf{Y}^{(i)}$, with $|\mathbf{V}^{(i)}| \equiv n_i$, takes the fol-

lowing form

$$\mathbb{P}(\mathbf{Y}^{(i)} = \mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\tau}, \underline{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \tau_k \exp\left(\boldsymbol{\eta}_k(\boldsymbol{\theta}_k)^\intercal \mathbf{g}_k(\mathbf{y}^{(i)}; \mathbf{X}^{(i)}) - \psi_{\mathbf{g_k}, \boldsymbol{\eta}_k, \mathbf{X}^{(i)}, \mathcal{Y}_{n_i}}(\boldsymbol{\theta}_k)\right) h_i(\mathbf{y}^{(i)}), \mathbf{y}^{(i)} \in \mathcal{Y}_{n_i} \tag{4.3}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_K)$ and $\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ are the model parameters, and the former satisfies the constraint $\sum_{k=1}^{K} \tau_k = 1, \tau_k \geq 0$ for $k = 1, \ldots, K$.

The ensemble of networks consists of m independent observations $\underline{\mathbf{y}} = (\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(m)})$ with fixed covariate set $\underline{\mathbf{X}} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(m)})$ and fixed vertex set $\underline{\mathbf{V}} = (\mathbf{V}^{(1)}, \cdots, \mathbf{V}^{(m)})$, and hence the joint likelihood is

$$\mathbb{P}(\underline{\mathbf{Y}} = \underline{\mathbf{y}}|\underline{\mathbf{X}}; \boldsymbol{\tau}, \underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} \left[\sum_{k=1}^{K} \tau_k \exp\left(\boldsymbol{\eta}_k(\boldsymbol{\theta}_k)^\intercal \mathbf{g}_k(\mathbf{y}^{(i)}; \mathbf{X}^{(i)}) - \psi_{\mathbf{g_k}, \boldsymbol{\eta}_k, \mathbf{X}^{(i)}, \mathcal{Y}_{n_i}}(\boldsymbol{\theta}_k)\right) h_i(\mathbf{y}^{(i)})\right], \tag{4.4}$$

where we have absorbed the support constraint into the reference measure.

To facilitate statistical inference, we consider the representation of (4.4) from a latent variable perspective. Let $Z_i, i = 1, \cdots, m$ be latent variables following a categorical distribution with K values and probability parameter $\boldsymbol{\tau}$, such that $Z_i = k$ if $\mathbf{Y}^{(i)}$ belongs to cluster k. We may then treat $\mathbf{Y}^{(i)}$ as arising from a process in which Z_i is first drawn from Categorial($\boldsymbol{\tau}$), and $\mathbf{Y}^{(i)}$ is then drawn from the ERGM distribution corresponding to cluster Z_i. While one could allow the reference measure to also vary by cluster, we rely on the case of ERGMs specified relative to the Krivitsky reference measure if the sizes of the networks vary. In addition, we focus on canonical ERGM (i.e., $\eta(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}$) as it comprises a wide range of models that have been found successful for modeling many real world networks. We note that the generalization to curved ERGM is straightforward conceptually, but the development of

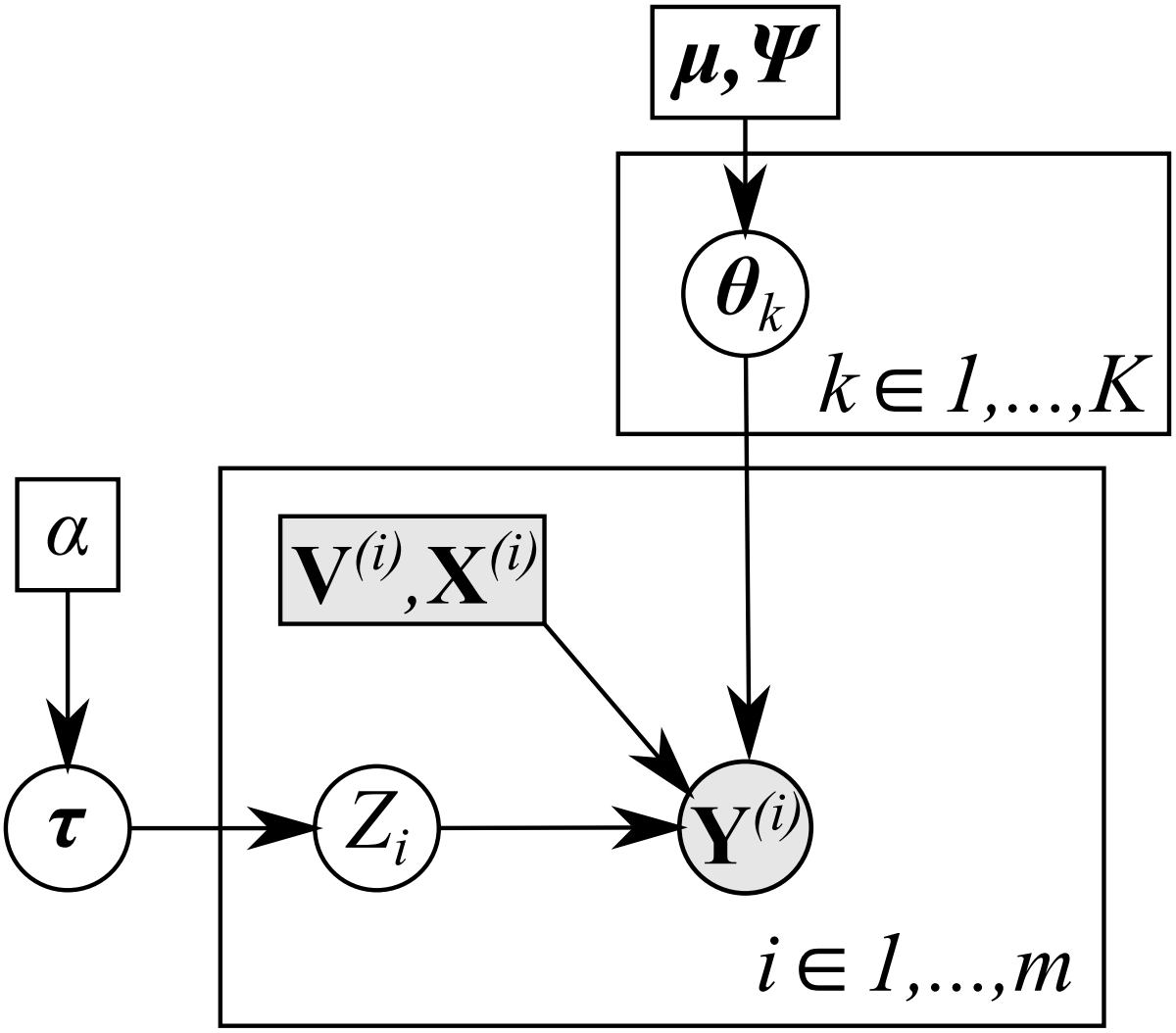

Figure 4.1: Structure of the graph mixture model. Random quantities are depicted within circles, fixed quantities within rectangles; observables are shaded.

reliable estimation algorithm is a very challenging problem by itself which we leave for future research.

4.3.2 Bayesian Estimation

Bayesian estimation is a natural choice for parameter inference 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. Our strategy is to employ Metropolis-within-Gibbs sampling to obtain MCMC samples from the joint posterior distribution of $\underline{\boldsymbol{\theta}}$ and $\boldsymbol{\tau}$.

We specify prior distributions for the parameters as follows,

$$\boldsymbol{\tau} \sim \text{Dirichlet}(\boldsymbol{\alpha}),$$

$$\boldsymbol{\theta}_k \overset{i.i.d.}{\sim} \text{MVN}_p(\boldsymbol{\mu}, \Psi), \quad k = 1, \cdots, K,$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$, $\boldsymbol{\mu}$ and Ψ are hyper-parameters to be specified by the user. For typical use cases, a reasonable choice of hyperparameters are $\alpha_1 = \ldots = \alpha_K = 3$, which puts low probability on any group being extremely small, $\boldsymbol{\mu} = \mathbf{0}$, and $\Psi = 25 I_p$, which is fairly flat over the typical range of variation for common parameterizations.

As noted, we perform posterior inference via MCMC. Our algorithm iterates over the model parameters $(\underline{\boldsymbol{\theta}}, \boldsymbol{\tau})$ with the priors given above, and the latent variables $\mathbf{Z} = (Z_1, \cdots, Z_m)$. 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. On the basis of some experimentation, we use the symmetric proposal $\mathcal{N}(\boldsymbol{\theta}, \sigma^2 I_q)$, where $\sigma = 0.05$. At each MCMC iteration, we permute the labels to impose ordering constraints on the first common element of the parameter vectors (e.g., total number of edges), $\theta_{11} < \theta_{21} < \cdots < \theta_{K1}$ for model identifiability purposes. Simulation studies and case studies show that the ordering constraints can work well, though other

Algorithm 4 Metropolis-within-Gibbs sampler for the ERGM mixture model

Initialization: Set $\boldsymbol{\tau}^0$, $\underline{\boldsymbol{\theta}}^0$ and $\boldsymbol{Z}^0$ to initial values (e.g., prior means).
for $t = 1, 2, \cdots, T$ **do**
  Generate Z_i^t $(i = 1, \cdots, m,\ k = 1, \cdots, K)$ from
    $\mathbb{P}(Z_i^t = k | \eta_k^{t-1}, \boldsymbol{\theta}_k^{t-1}, \mathbf{y}^{(i)}) \propto \eta_k^{t-1} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{t-1})$
  Compute $\nu_k^t = \sum_{i=1}^m \mathbb{1}_{Z_i^t = k};\ k = 1, \cdots, K$
  Generate $\boldsymbol{\tau}^t$ from Dirichlet$(\alpha_1 + \nu_1^t, \cdots, \alpha_K + \nu_K^t)$
  for $k = 1, \cdots, K$ **do**
    Propose $\boldsymbol{\theta}_k^{'} \sim q(\cdot | \boldsymbol{\theta}_k^{t-1})$
    Accept $\boldsymbol{\theta}_k^{'}$ with probability equal to
    $$\frac{\pi(\boldsymbol{\theta}_k^{'}) \prod_{Z_i^t = k} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{'}) q(\boldsymbol{\theta}_k^{t-1} | \boldsymbol{\theta}_k^{'})}{\pi(\boldsymbol{\theta}_k^{t-1}) \prod_{Z_i^t = k} \mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{t-1}) q(\boldsymbol{\theta}_k^{'} | \boldsymbol{\theta}_k^{t-1})}$$
  end for
end for

post-processing techniques (e.g., Kullback-Leibler relabeling algorithm (Stephens, 2000) and Pivotal Reordering algorithm (Marin et al., 2005), etc.), can be used depending on practitioners' preference.

To deal with the intractability of $\mathbb{P}(\mathbf{y}^{(i)} | \mathbf{X}^{(i)}; \boldsymbol{\theta})$, there are at least three possible solutions in the ERGM literature:

- Work with a tractable approximation in place of the ERGM likelihood, e.g., pseudo-likelihood (Strauss and Ikeda, 1990), fully adjusted pseudo-likelihood (Bouranis et al., 2018), or other composite likelihoods (Austad and Friel, 2010; Asuncion et al., 2010);
- Use importance sampling to approximate the ERGM likelihood (Koskinen, 2004, 2008);
- Use auxiliary-variable based MCMC algorithms to eliminate the intractable normalizing factor in ERGM likelihood (Caimo and Friel, 2011).

In fact, updating $\boldsymbol{\theta}_k$'s using the Metropolis-Hastings ratio in (8) is a *doubly-intractable* problem, which can be approached using various advanced MCMC techniques (see Park and Haran, 2018, for a review). However, these advanced techniques all require simulating networks from ERGMs at each MCMC iteration to approximate the true likelihood (4.1), which

can be expensive for large networks. When the major goal is clustering instead of estimation of cluster-specific parameters, we propose to work with the most common form of tractable approximation, the pseudo-likelihood, in which the full likelihood of each network is approximated by a product of full conditional distributions of edge variables y_{ij} in $\mathbf{y}$,

$$f_{PL}(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}) = \prod_{(i,j)\in\mathcal{D}} \mathbb{P}(y_{ij}|y_{-ij};\mathbf{X};\boldsymbol{\theta}) = \prod_{(i,j)\in\mathcal{D}} \frac{1}{1+\exp\left\{-\boldsymbol{\eta}(\boldsymbol{\theta})^{\intercal}\Delta_{i,j}\mathbf{g}(\mathbf{y};\mathbf{X})\right\}}, \tag{4.5}$$

where $\Delta_{i,j}\mathbf{g}(\mathbf{y};\mathbf{X}) = \mathbf{g}(y_{ij}^{+};\mathbf{X}) - \mathbf{g}(y_{ij}^{-};\mathbf{X})$ 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, $\mathcal{D} = \{(i,j)|i,j \in \mathbf{V}, i \neq j\}$, while for undirected networks, $\mathcal{D} = \{(i,j)|i,j \in \mathbf{V}, i < j\}$. In the frequentist paradigm, maximizing (4.5) 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) (especially 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 MLE in such cases (Desmarais and Cranmer, 2010). Similar logic has motivated the use of Bayesian bootstrap estimation based on "pseudo-MAP" estimates using the PL approximation to the likelihood (Grazioli et al., 2019b).

4.3.3 Choosing the Number of Clusters

We recast the problem of choosing the number of clusters as a model selection problem, as different numbers of clusters result in distinct statistical models. Therefore, we use 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 models with latent variables. Given posterior draws $\boldsymbol{\tau}^l, \underline{\boldsymbol{\theta}}^l = (\boldsymbol{\theta}_1^l, \cdots, \boldsymbol{\theta}_K^l)$ and observed ensemble of networks $\underline{\mathbf{y}} = (\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(m)})$, the observed DIC is defined by

$$DIC_K = -4\mathbb{E}_{\underline{\boldsymbol{\theta}}}[\log \mathbb{P}(\underline{\mathbf{y}}|\underline{\mathbf{X}}; \underline{\boldsymbol{\theta}})|\underline{\mathbf{y}}] + 2\log \hat{\mathbb{P}}(\underline{\mathbf{y}}|\underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}), \tag{4.6}$$

where

$$\hat{\mathbb{P}}(\underline{\mathbf{y}}|\underline{\mathbf{X}}; \underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} \hat{\mathbb{P}}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} \left(\frac{1}{m} \sum_{l=1}^{L} \sum_{k=1}^{K} \tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{(l)}) \right),$$

and

$$\mathbb{E}_{\underline{\boldsymbol{\theta}}}[\log \mathbb{P}(\underline{\mathbf{y}}|\underline{\mathbf{X}}; \underline{\boldsymbol{\theta}})|\underline{\mathbf{y}}] = \frac{1}{m} \sum_{l=1}^{L} \sum_{i=1}^{m} \log \left\{ \sum_{k=1}^{K} \tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k^{(l)}) \right\}.$$

As practitioners often seek for parsimonious models to represent the population, we present a rule-of-thumb to identify the point where there is diminishing return by further increasing the number of clusters, and hence to avoid potential over-fitting. Define the relative difference (RD) in DIC as

$$RD(k) = \frac{DIC_k - DIC_{k-1}}{DIC_{k-1}}, k = 2, 3, \cdots .$$

We define the optimal number of clusters given by a pre-specified cut-off value ϵ as $k_{opt}(\epsilon) = \min_k \left\{k | RD(k) \geqslant \epsilon\right\}$, based on the reasoning that the optimal number of clusters should be the first k resulting in limited relative improvement in terms of DIC. Simulation studies in Section 4.4 show empirical evidence supporting that $\epsilon = -0.005$ can be a reasonable rule-of-thumb for selecting the number of clusters.

We note that having an ensemble of networks makes it possible to assess the out-of-sample performance of mixture of ERGMs using the traditional statistical principle of cross-validation (CV), and there is work on using CV to estimate the number of clusters for observations with continuous values (Fu and Perry, 2019). In particular, to reduce the possibility of accidentally dropping all graphs in a single cluster by holding out too many graphs simultaneously, leave-one-out CV should be favored. The loss function for the cross-validation procedure can be negative log-likelihood evaluated on the held-out data as well as prediction error with respect to any structural properties of interest (obtained by simulating from estimated model using training data). Though the CV is not Bayesian and violates the likelihood principle, it is easy to implement and obviates the need to choose a threshold for when to stop adding clusters based on the predictive power of the model.

4.3.4 Posterior Probability of Cluster Membership

An appealing aspect of mixture modeling is that the posterior probability of individuals belonging to each cluster (alternately: graphs having been generated by a particular process)

can be conveniently obtained as

$$\mathbb{P}(Z_i = k|\mathbf{y}^{(i)}) = \int \frac{\tau_k \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \tau_k \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k)} \pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\tau}|\underline{\mathbf{y}}) d\underline{\boldsymbol{\theta}} d\boldsymbol{\tau}, \tag{4.7}$$

where $\pi(\underline{\boldsymbol{\theta}}, \boldsymbol{\tau}|\underline{\mathbf{y}})$ is the posterior distribution of $\underline{\boldsymbol{\theta}}, \boldsymbol{\tau}$. The integral in (4.7) 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,

$$\hat{\mathbb{P}}(Z_i = k|\mathbf{y}^{(i)}) = \frac{1}{L} \sum_{l=1}^{L} \frac{\tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^{K} \tau_k^l \mathbb{P}(\mathbf{y}^{(i)}|\mathbf{X}^{(i)}; \boldsymbol{\theta}_k^l)}. \tag{4.8}$$

The posterior mode, i.e., $\hat{Z}_i = \arg\max_k \hat{\mathbb{P}}(Z_i = k|\mathbf{y}^{(i)})$ can be used as the output for cluster analysis, provided that the goal is to obtain a deterministic cluster assignment.

4.4 Simulation Studies

We conduct extensive simulation studies to show that the proposed approach is capable of selecting the true number of clusters, recovering the true cluster memberships and true model parameters.

4.4.1 Experiment Settings

The ground truth is available for the synthetic data, as we simulate networks from mixtures of ERGM distributions defined on the three most commonly used network sufficient statistics

but with distinct parameters,

- $g_1(\mathbf{y}) = \sum_{i<j} y_{ij}$, 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.
- $g_3(\mathbf{y}; \mathbf{X}) = \sum_{i<j} y_{ij} \mathbb{1}_{\{\mathbf{x}_i = \mathbf{x}_j\}}$, total number of edges with endpoints sharing the same value on node-level covariate $\mathbf{X}$, often known as 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 three treatments

- Network size: 40, 100, 250.
- Number of clusters: 2, 3.
- Cluster size: 10, 20, 50.

We thus have a total of 18 experimental conditions, each of which is run for 50 replicates. The true cluster-specific parameters are specified as

$$\boldsymbol{\theta}^{40}_{true} = \begin{pmatrix} -1.15 & 0 & 0 \\ -2.85 & 0.25 & 2.25 \\ -4.95 & 2.5 & 0.25 \end{pmatrix}, \ \boldsymbol{\theta}^{100}_{true} = \begin{pmatrix} -2.20 & 0 & 0 \\ -4.15 & 0.25 & 2.25 \\ -5.85 & 2.5 & 0.25 \end{pmatrix}, \ \boldsymbol{\theta}^{250}_{true} = \begin{pmatrix} -3.20 & 0 & 0 \\ -4.95 & 0.25 & 2.25 \\ -6.42 & 2.5 & 0.25 \end{pmatrix}$$

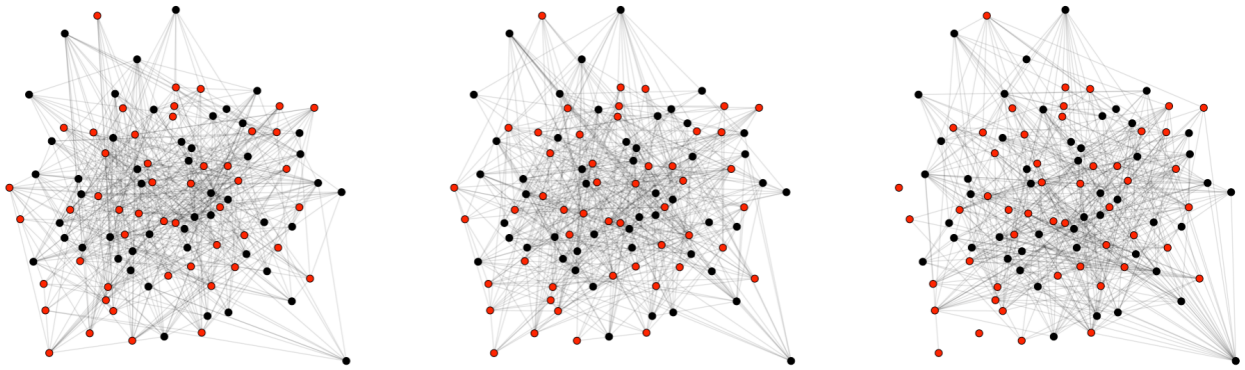

Figure 4.2: Representative networks from clusters 1 (left), 2 (middle), and 3 (right). Network size: 100. Color indicates nodal covariate value: 0 (black), 1 (red). Despite the apparent similarity of the networks produced by the three generative processes, we are able to infer the latter from the observed ensemble.

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 clusters and network sizes; and (ii) represent three most common-yet-intuitive patterns in real-world networks (parameter settings in the first row corresponds to the cases in which ties are independent Bernoulli draws, and parameter settings in the second row corresponds to the cases in which there is a strong homophily effect but a weak triadic closure effect, while the parameter settings in the third row correspond to the case in which there is a strong triadic closure effect but weak 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 simulate networks using first two rows of the parameter matrices when the number of clusters is 2. Identifying subpopulations from ensembles of networks produced by this model is by no means a trivial task, especially as the cluster-specific parameters are chosen to produce networks of similar mean degrees (≈ 0.10) as shown in Figure 4.2. While these networks appear superficially similar, we can recover the distinct processes that generated them.

We apply the proposed Algorithm 4 to analyze the synthetic data sets, allowing the candidate values for the number of clusters to range from 1 to one greater than the true number of

clusters (i.e., to 4, if the true number of clusters is 3; and 3, if the true number of clusters is 2). We assign random initial values to the latent indicator membership $\boldsymbol{Z}_i^0$, weight parameters $\boldsymbol{\tau}^0$ according to the prior, and set the parameters associated with the edge term as -2 (i.e., $\theta_{11} = \cdots = \theta_{K1} = -2$), while all other elements in $\underline{\boldsymbol{\theta}}$ are drawn independently from a uniform distribution $\mathcal{U}(-0.1, 0.1)$. It is worth noting that our experiments suggest that better initial values can result in faster convergence and more stable performance for large networks. One effective way to initialize the proposed Algorithm 4 is to first find the MPLE for each network in the ensemble separately, then cluster these MPLE estimates with K-means algorithm to initialize $\boldsymbol{Z}_i^0$ and calculate the intra-cluster mean MPLE estimates to determine the starting value of cluster-specific model parameters for each cluster. Table 4.1 presents the MCMC settings, prior and proposal distribution for the experiments. The thinning interval is chosen as 50 for all MCMC chains to obtain high-quality, weakly correlated draws from the posterior. All computations in this paper are implemented in **R** (R Core Team, 2018), and we use software suite `statnet` (Handcock et al., 2008) to generate networks from ERGMs.

Table 4.1: Total number of iterations, burn-in size, initialization method, prior hyper-parameters and covariance matrix for random-walk Metropolis-Hastings update of $\underline{\boldsymbol{\theta}}$ in simulation studies

	Total iterations	Burn-in	Initialization	$\boldsymbol{\mu}$	Ψ	Prop. Cov
40, 2	17500	7500	Random	(-1,0,0)	$25I_3$	$0.0025I_3$
40, 3	20000	10000	Random	(-1,0,0)	$25I_3$	$0.0025I_3$
100, 2	17500	7500	Random	(-1,0,0)	$25I_3$	$0.0025I_3$
100, 3	20000	10000	MPLE, K-means	(-1,0,0)	$25I_3$	$0.0025I_3$
250, 2	22500	12500	MPLE, K-means	(-1,0,0)	$25I_3$	$0.0016I_3$
250, 3	25000	15000	MPLE, K-means	(-1,0,0)	$25I_3$	$0.0016I_3$

4.4.2 Recovery of True Number of Clusters and Cluster Membership

We analyze the performance of proposed method in terms of its ability to identify the true number of clusters and cluster memberships. Figures 4.3 and 4.4 show that selecting the

number of clusters according to the point beyond which there is diminishing return ($\epsilon = -0.005$) is unanimously superior to the minimum DIC criterion ($\epsilon = 0$), as the latter tends to be in favor of more complex models (i.e., with more clusters) than is optimal. Under DIC criterion with $\epsilon = -0.005$, we note that one has an 90% or higher chance of identifying the true number of clusters when the true number is 2, and such chance is about 80% when the true number of clusters is 3. Compared to identifying true number of clusters, recovering cluster memberships can be a more meaningful task in real-world applications, which we evaluate using adjusted rand index (ARI) (Hubert and Arabie, 1985), a corrected-for-chance measure of the similarity between two clustering assignments, which yields a value of 1 for perfect cluster assignments and has an expected value of 0 for completely random cluster assignments. ARI is employed as an accuracy measure for cluster assignments here because the ground truth is available in the simulation study. Table 4.2 gives the mean ARI calculated across 50 replicates within each experiment setting, it shows that the proposed method can work well on the task of cluster assignments as all the mean ARI values are higher than 0.90 when the true number of clusters is 2 and 0.85 when the true number of clusters is 3 (a rule-of-thumb threshold value for "good clustering" is 0.80). We note that the mean ARI scores in Table 4.2 includes those calculated on the runs in which the true number of clusters is falsely identified, indicating that the proposed method is robust. In other words, the method fails gracefully, as it tends to completely combine two clusters or split one entire cluster into two when it errs, rather than mixing two clusters.

Table 4.2: Mean ARI calculated across 50 replicates within each experiment setting. The true number of clusters is denoted as K.

	K=2			K=3		
	10	20	50	10	20	50
40	0.940	0.980	0.900	0.902	0.942	0.924
100	0.980	0.996	0.980	0.902	0.869	0.905
250	1.000	1.000	1.000	0.884	0.939	0.905

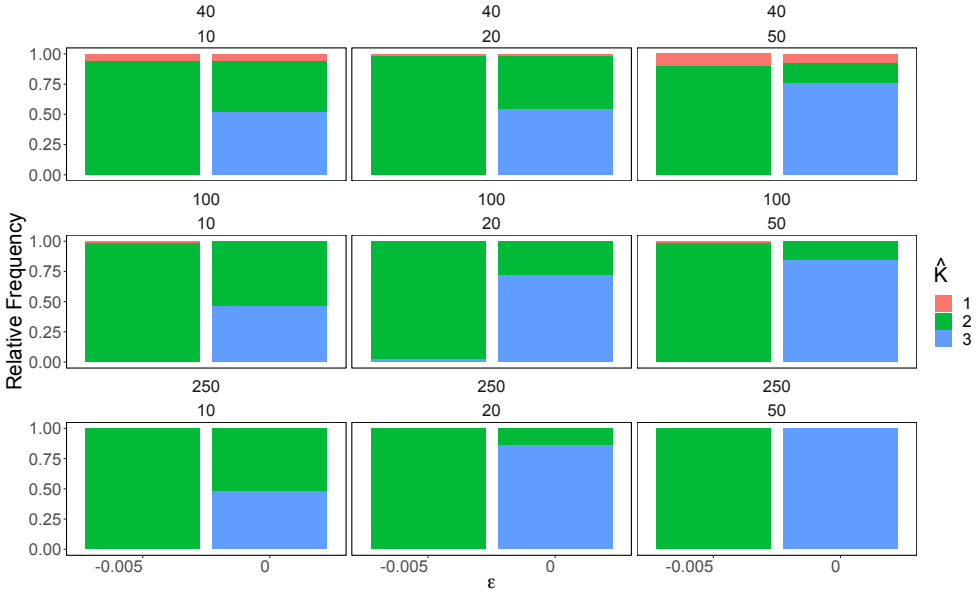

Figure 4.3: Relative frequency of $\hat{K}$ selected by DIC criterion with $\epsilon = 0$ and $\epsilon = -0.005$. True number of clusters $(K) = 2$

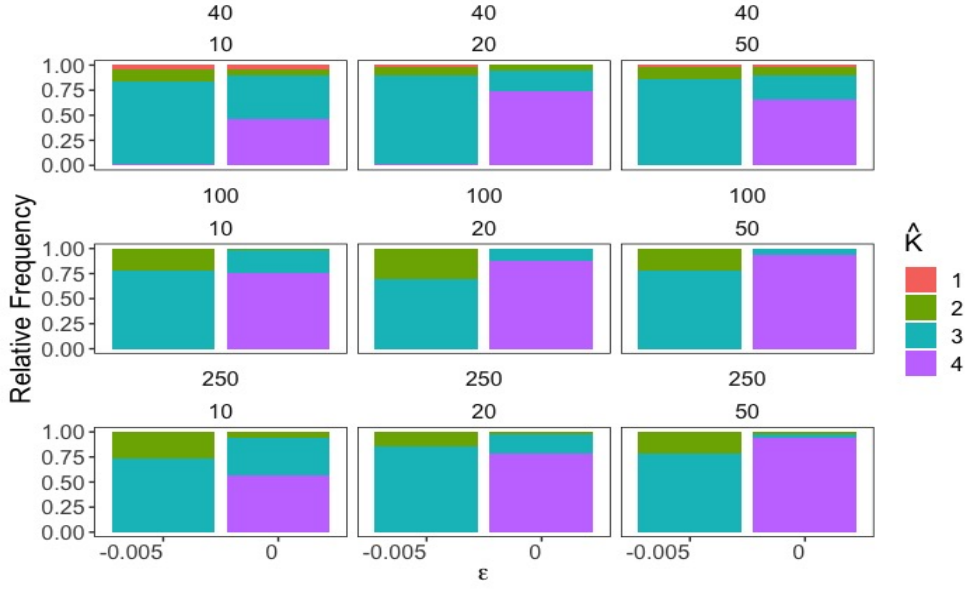

Figure 4.4: Relative frequency of $\hat{K}$ selected by DIC criterion with $\epsilon = 0$ and $\epsilon = -0.005$. True number of clusters $(K) = 3$.

4.4.3 Estimation Accuracy

Given a correctly identified number of clusters, one natural question to ask is whether the proposed algorithm can accurately estimate the cluster-specific parameters. Specifically, we evaluate the estimation accuracy by examining the bias of posterior means.

Table 4.3 summarizes the bias for cluster-specific model parameters under all experimental settings. We notice that the bias is in general small, especially for large networks, though there is slightly higher bias when the true number of clusters is 3. Large bias is mostly seen in the clusters in which there is strong dyadic dependence among edge variables (i.e., large coefficients associated with gwesp term), as expected. However, such bias becomes smaller and also less variable as sample size increases, indicating that larger sample size can mitigate the bias induced by the adoption of pseudo-likelihood. These findings offer implications to practitioners as estimated parameters are more reliable when large sample size is available or when the size of networks of interests is large.

4.4.4 Posterior Predictive Assessments

One of the most appealing aspects of mixture modeling framework is that one can use simple probability distributions as building blocks to approximate complex probability distributions (e.g., mixtures of Gaussians are often used to approximate multimodal distributions). It is of substantial interest to see whether mixtures of ERGMs can provide an adequate fit to complex graph distributions. Although the selection of metrics should be guided by the particular properties of interests in practice, we consider four widely used metrics that characterize different aspects of graph structure as follows

- 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

Table 4.3: Mean (standard deviation) of bias across replicates in which the true number of clusters is correctly identified by DIC criterion ($\epsilon = -0.005$) within each experimental setting.

	K=2						K=3								
size	edges	gwesp	nodematch	edges	gwesp	nodematch	edges	gwesp	nodematch	edges	gwesp	nodematch	edges	gwesp	nodematch
40	-2.85	0.25	2.25	-1.15	0	0	-4.95	2.5	0.25	-2.85	0.25	2.25	-1.15	0	0
10	-0.041 (0.132)	0.025 (0.087)	0.01 (0.073)	0.002 (0.095)	0 (0.052)	-0.004 (0.047)	-0.06 (0.358)	0.051 (0.25)	-0.014 (0.049)	0.009 (0.153)	0.007 (0.107)	-0.017 (0.07)	0.02 (0.114)	-0.009 (0.058)	-0.005 (0.058)
20	-0.014 (0.128)	0.002 (0.078)	0.011 (0.05)	-0.001 (0.069)	0.002 (0.034)	-0.001 (0.034)	-0.021 (0.256)	0.02 (0.174)	-0.014 (0.034)	0.016 (0.121)	-0.004 (0.078)	-0.009 (0.051)	0.002 (0.077)	0.005 (0.042)	-0.008 (0.042)
50	-0.003 (0.077)	-0.001 (0.049)	0.006 (0.039)	0.005 (0.046)	-0.003 (0.025)	0.002 (0.02)	-0.046 (0.154)	0.032 (0.109)	-0.006 (0.022)	0.003 (0.069)	0 (0.043)	-0.003 (0.033)	0.006 (0.053)	-0.003 (0.029)	0.001 (0.028)
100	-4.15	0.25	2.25	-2.20	0	0	-5.85	2.5	0.25	-4.15	0.25	2.25	-2.20	0	0
10	-0.003 (0.048)	0.003 (0.027)	-0.001 (0.051)	0.01 (0.033)	-0.003 (0.018)	0 (0.03)	0.001 (0.107)	0.002 (0.075)	0.002 (0.035)	-0.007 (0.049)	0 (0.029)	0.006 (0.057)	0.002 (0.044)	-0.001 (0.021)	0.003 (0.026)
20	0.004 (0.034)	-0.002 (0.021)	-0.002 (0.036)	0.008 (0.026)	-0.003 (0.013)	-0.001 (0.021)	-0.005 (0.081)	0.006 (0.056)	-0.004 (0.029)	-0.005 (0.042)	0.002 (0.023)	0 (0.037)	-0.006 (0.03)	0.003 (0.015)	0.001 (0.022)
50	0.005 (0.02)	-0.004 (0.015)	0.001 (0.027)	0.002 (0.017)	0 (0.007)	-0.001 (0.015)	-0.019 (0.055)	0.013 (0.037)	0 (0.017)	0.003 (0.022)	-0.001 (0.014)	-0.002 (0.02)	-0.004 (0.015)	0.001 (0.007)	0 (0.016)
250	-4.95	0.25	2.25	-3.20	0	0	-6.42	0.25	2.25	-4.95	0.25	2.25	-6.42	2.5	0.25
10	0.004 (0.027)	-0.002 (0.011)	-0.001 (0.027)	-0.002 (0.015)	0 (0.012)	0 (0.019)	-0.012 (0.056)	0.009 (0.042)	-0.001 (0.026)	0.004 (0.025)	-0.001 (0.012)	-0.001 (0.028)	0.001 (0.013)	0.001 (0.012)	-0.006 (0.022)
20	-0.003 (0.019)	0 (0.008)	0.003 (0.021)	0.001 (0.01)	-0.001 (0.009)	0 (0.011)	-0.009 (0.046)	0.006 (0.032)	0.001 (0.017)	-0.001 (0.018)	0.001 (0.009)	-0.001 (0.021)	-0.001 (0.012)	0.001 (0.007)	-0.002 (0.013)
50	0.001 (0.012)	0 (0.006)	0 (0.013)	0 (0.007)	0 (0.006)	-0.001 (0.007)	-0.003 (0.028)	0.003 (0.019)	-0.002 (0.009)	0 (0.014)	0 (0.005)	0.001 (0.015)	-0.001 (0.007)	0 (0.005)	-0.002 (0.008)

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 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 geodesic distances: a measure of the overall closeness between nodes in a graph.

We focus on the experimental settings in which we have the most observations (3 clusters, 50 networks in each cluster) in this section. As each ensemble of networks in the synthetic data sets contains a total of 150 graphs, we also generate 150 networks using posterior samples with the data generating mechanism described in Figure 4.1. The simulated networks based on posterior samples and those synthetic networks are summarized by the four graph-level metrics, and their discrepancies are quantified in terms of the Hellinger distance, a commonly used metric for quantifying the distance between two probability distributions. We use function `CalcHellingerDist` in package `textmineR` (Jones, 2019) to calculate the empirical Hellinger distance between two sample vectors. Table 4.4 summarizes the mean and standard deviation of Hellinger distance evaluated across all replicates, regardless of whether the number of clusters selected by DIC criterion ($\epsilon = -0.005$) under the experimental settings of interests (i.e., true number of clusters is 3) is correct. The discrepancy

[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).

between posterior predictive samples and synthetic data sets increases as the model selection accuracy decreases, from network size 40 to 100 and then to 250. To better understand the connections between Hellinger distance values and underlying visual difference in distributions in terms of histograms, we consider two representative replicates when the network size is 250. Figure 4.5 corresponds to a case in which the true number of clusters is selected and with Hellinger distance close to the average – it is clear that the posterior predictive distribution of metrics of interest is very close to that of synthetic data. Figure 4.6 corresponds to a representative case in which the number of clusters is underestimated to be 2 – the key observation is the resulting mixture model successfully captures the bimodal feature of mean eigenvector centrality and the left-skewed feature of mean of inverse geodesic distribution, and also identifies two of the three modes for standard deviation of degree distribution and transitivity. Although the result does not seem to be ideal, one key observation is that the resulting mixture model converges to the "middle ground" between two clusters, indicating that the possible reason for the model to choose two clusters over three is that the algorithm gets stuck at a local optimum, which might be mitigated by running MCMC chains longer or a more efficient proposal distribution for the Metropolis-Hastings step of Algorithm 4. At a higher level, these results suggest the potential of mixtures of ERGMs as a tool to approximate complex graph distributions as one can view ERGMs as an analogue to "kernel" in density estimation.

Table 4.4: Mean (standard deviation) of Hellinger distance

	Mean EC	Transitivity	SD of deg. dist.	Mean of inverse geodesic distance
40, 3	0.045 (0.002)	0.086 (0.007)	0.083 (0.006)	0.011 (0.001)
100, 3	0.076 (0.005)	0.154 (0.009)	0.123 (0.006)	0.025 (0.003)
250, 3	0.145 (0.015)	0.271 (0.016)	0.137 (0.010)	0.074 (0.016)

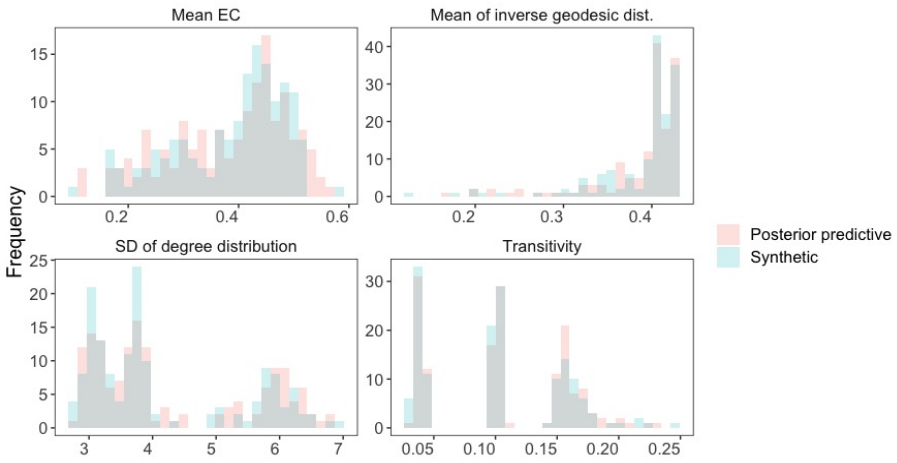

Figure 4.5: Distribution of metrics of interests for posterior predictive samples and synthetic data, with corresponding Hellinger distance values : 0.150 (upper left), 0.283 (upper right), 0.141 (lower left), 0.076 (lower right).

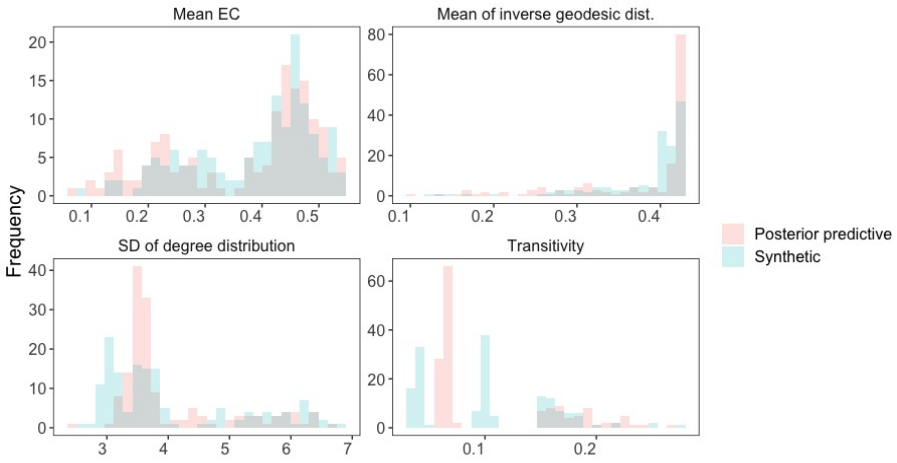

Figure 4.6: Distribution of metrics of interests for posterior predictive samples and synthetic data, with corresponding Hellinger distance values: 0.173 (upper left), 0.270 (upper right), 0.125 (lower left), 0.105 (lower right).

4.5 Case Study: Political Co-voting Networks among U.S. Senators

In this section, we apply the proposed method to cluster the co-voting patterns among U.S. Senators from 1867 (start year of Congress 40) to 2014 (end year of Congress 113), which was a subset of the data first analyzed by Moody and Mucha (2013) using modularity and role-based blockmodels. The co-voting tendencies are represented by networks 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]. The nodes in the co-voting network represent Senators and an edge is placed between two nodes if the corresponding Senators vote concurrently (both yay of 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.

4.5.1 Model Specification and Estimation

Figure 4.7 shows that the co-voting networks vary in structure on different years, and that party-affiliation appears to be a key factor affecting the co-voting patterns among Senators. Therefore we consider an ERGM model with following sufficient statistics

$$g_1(\mathbf{y}) = \sum_{i<j} y_{ij}, \text{ total number of edges;}$$

$$g_2(\mathbf{y}; \mathbf{X}) = \sum_{i<j} y_{ij}, \mathbb{1}_{\{\mathbf{X}_i = \mathbf{X}_j = D\}}, \quad \text{total number of edges between Democrats;}$$

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

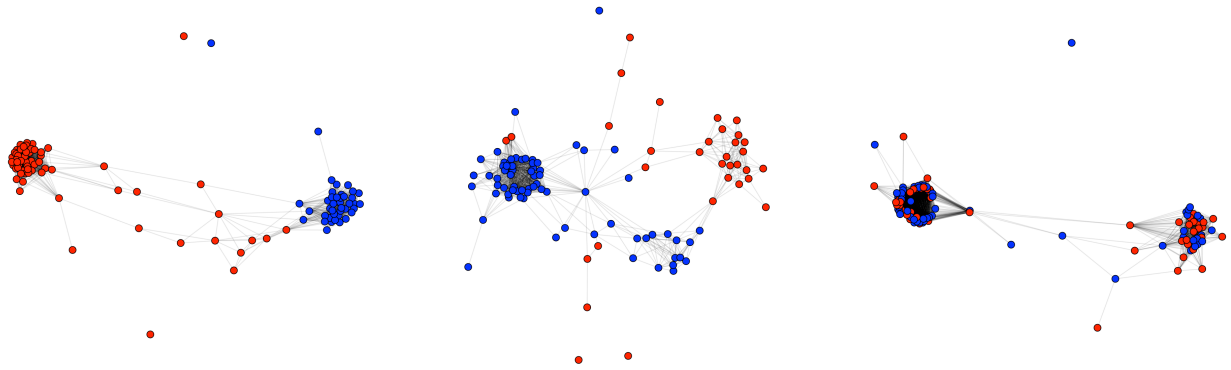

Figure 4.7: Co-voting networks of 61st, 89th and 111th Congress, which were formed in the year of 1909, 1965 and 2009, respectively. Colors indicate Senators' party affiliations, blue = Democrats(D), red = Republican(R).

$$g_3(\mathbf{y};\mathbf{X}) = \sum_{i<j} y_{ij}\mathbb{1}_{\{\mathbf{x}_i=R,\mathbf{x}_j=D\}}, \text{ total number of edges between Democrats and Republicans;}$$

$$g_4(\mathbf{y}) = e^{\phi}\sum_{k=1}^{n-2}\left\{1-(1-e^{-\phi})^k\right\}EP_k(\mathbf{y}), \text{ GWESP statistic}$$

The decay parameter of GWESP term is fixed as $\phi = 0.25$ as often used in ERGM literature. We note that these networks vary in size (range: $69-112$) and thus include an offset term (4.2) to adjust for network size. (This is equivalent to using the Krivitsky reference measure, which provides a parameterization with constant baseline expected degree.) We use the prior specification in Section 4.3, and run long MCMC chains (total iterations = 80000, burn-in = 30000, thinning interval = 50) with random initial values.

4.5.2 Results

Figure 4.8 indicates that the DIC reaches its minimum at $K = 3$, and hence $K = 3$ appears to be a plausible choice for the number of clusters. Under $K = 3$, visual inspections on the traceplots suggest that the chains converge very fast and mix well (see Figure 4.9 for traceplots of edges parameter; other traceplots also show similar pattern, but are omitted in the interest of space). The posterior mean estimates of cluster-specific parameters are

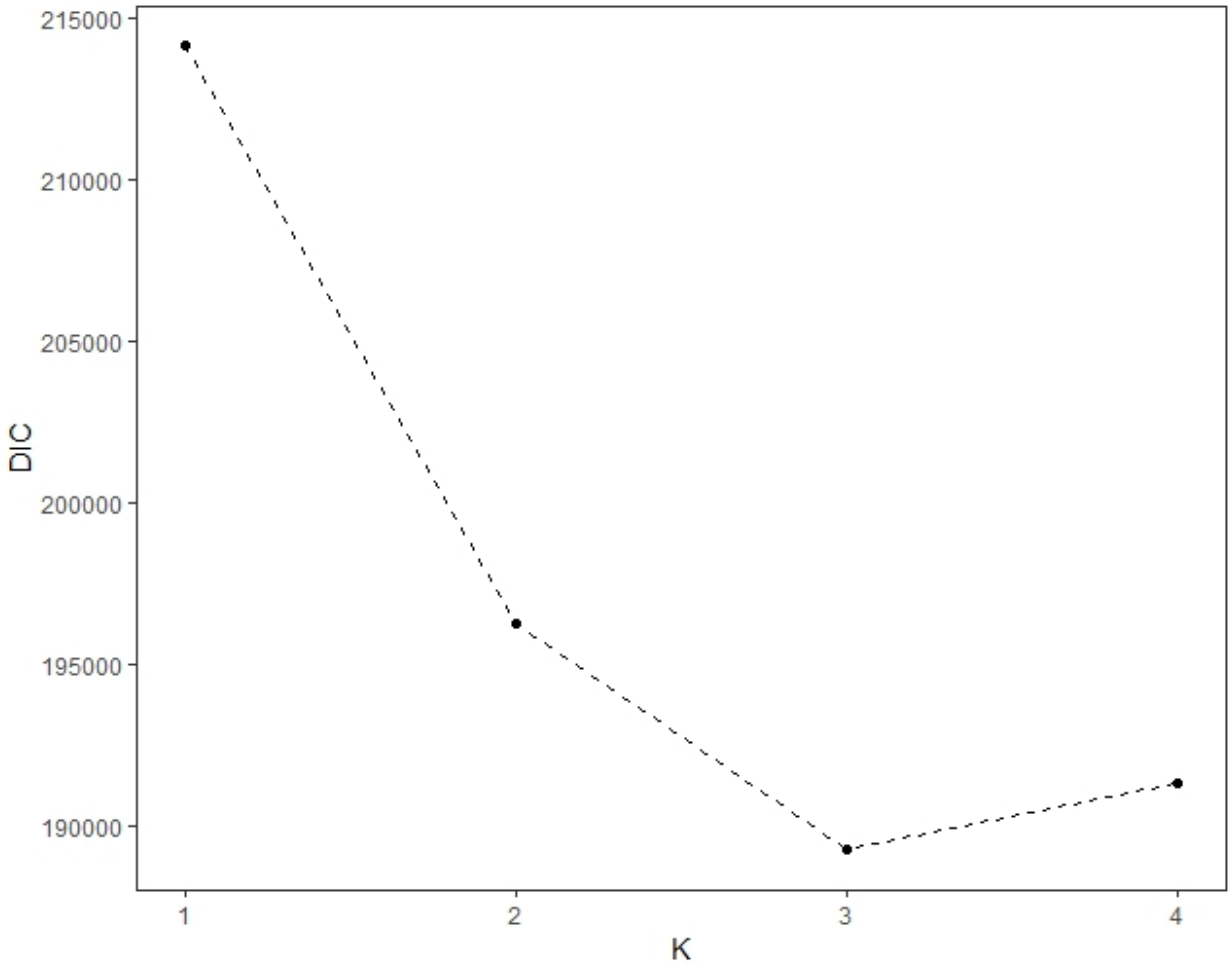

Figure 4.8: DIC vs Number of clusters, U.S. Congress co-voting networks

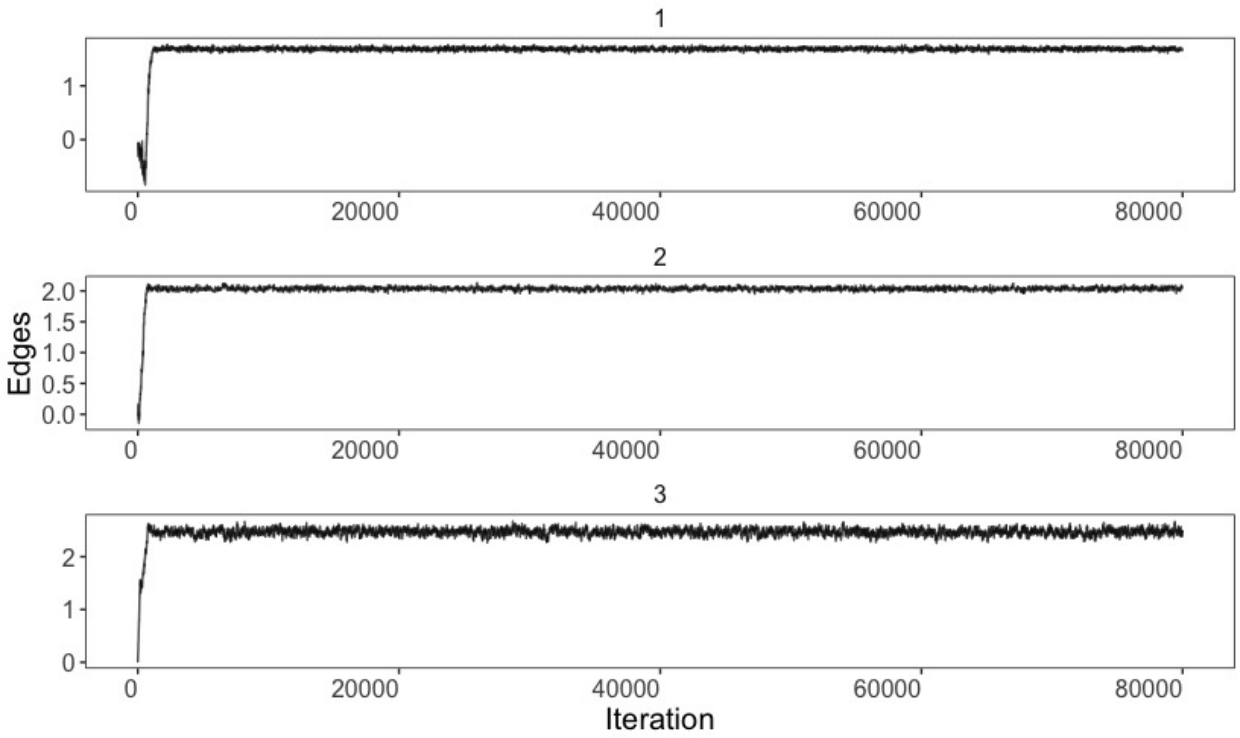

Figure 4.9: Traceplots for parameters associated with edges term for 3 clusters.

$$\hat{\boldsymbol{\tau}} = \begin{pmatrix} 0.36 \\ 0.47 \\ 0.17 \end{pmatrix} \underline{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 1.69 & 0.01 & -2.49 & 1.42 \\ 2.04 & -0.12 & -3.09 & 2.14 \\ 2.47 & 0.92 & -4.47 & 2.63 \end{pmatrix}$$

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 5.5 to 12 across clusters prior to consideration of other effects.

Based on these estimates, we have the following observations regarding the co-voting patterns. Across all clusters, we see both inhibition of cross-party ties (third column) and strong triadic closure (fourth column). Clusters do differ, however. Cluster 1 shows essentially symmetric behavior by party (column two), with lower levels of cross-group inhibition

and triadic closure bias than in the other clusters; overall, cluster 1 suggests a relatively low level of polarization by party, with voting only loosely restricted by party lines. By contrast, cluster 2 reflects a much more polarized regime, with more activity overall and co-voting being more concentrated within party. Like cluster 1, however, cluster 2 shows little party asymmetry (apart from a fairly weak tendency towards lower levels of co-voting among Democrats). Such asymmetry is much more strongly pronounced within cluster 3, with intraparty Democratic ties being approximately 2.5 times as likely (ceteris paribus) as ties within the GOP. This cluster also reflects extremely high levels of polarization, with cross-party co-voting being strongly inhibited and high levels of triadic closure. Over the period studied here, the most common pattern (probability 0.47) is the symmetric polarization of cluster 2, with the loose, low polarization pattern of cluster 1 also being fairly common (probability 0.36). The asymmetric, highly polarized regime of cluster 3 is less common, but is still estimated to account for approximately 17% of the observed cases. Interestingly, we do not see a corresponding asymmetric pattern in which the GOP shows high intraparty vote density, as might be anticipated; thus, there appear to be latent differences in how the two parties behave during the period that, while not manifest in every congress, always have the potential to arise.

One advantage of working with a fully generative model is the ability to perform "what-if" analyses that separate effects due to observed covariates from differences in structure arising from differences in generative processes. To probe the impact of the three behavioral regimes inferred from the co-voting data, 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. To perform such an analysis, we first 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 100 "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, 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. Figure 4.10 shows the realized proportion of intra-party triangles in simulated networks, and Figure 4.11 shows the realized proportion of 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 formation of very few voting coalitions, either

within party or between party). As suggested by the parameter values, we see little difference in coalition formation between the two parties, both having little cohesion.

- By contrast, the regime of cluster 2 shows a much higher incidence of intra-party coalition formation, with roughly 10-20% of the potential intra-party coalitions being present. Coalition incidence differs little by party, with at best a small average increment in the rate of coalition incidence for Republicans versus Democrats. Interestingly, this regime also shows the highest rate of cross-party coalition formation; while the rate is very low overall, it is considerably higher than that observed under cluster 1.

- Finally, the regime of cluster 3 favors extremely high levels of intra-party cohesion, with rates approaching 50% of the maximum possible for Republicans and 75% 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 GOP. 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 2. That said, the much higher incidence of intra-party coalition formation under cluster 3 leads inter-party coalitions to be a smaller fraction of the total coalition set than under cluster 2, potentially making them less critical to the legislative process.

Taken together, these observations suggest that the cluster 1 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 2 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

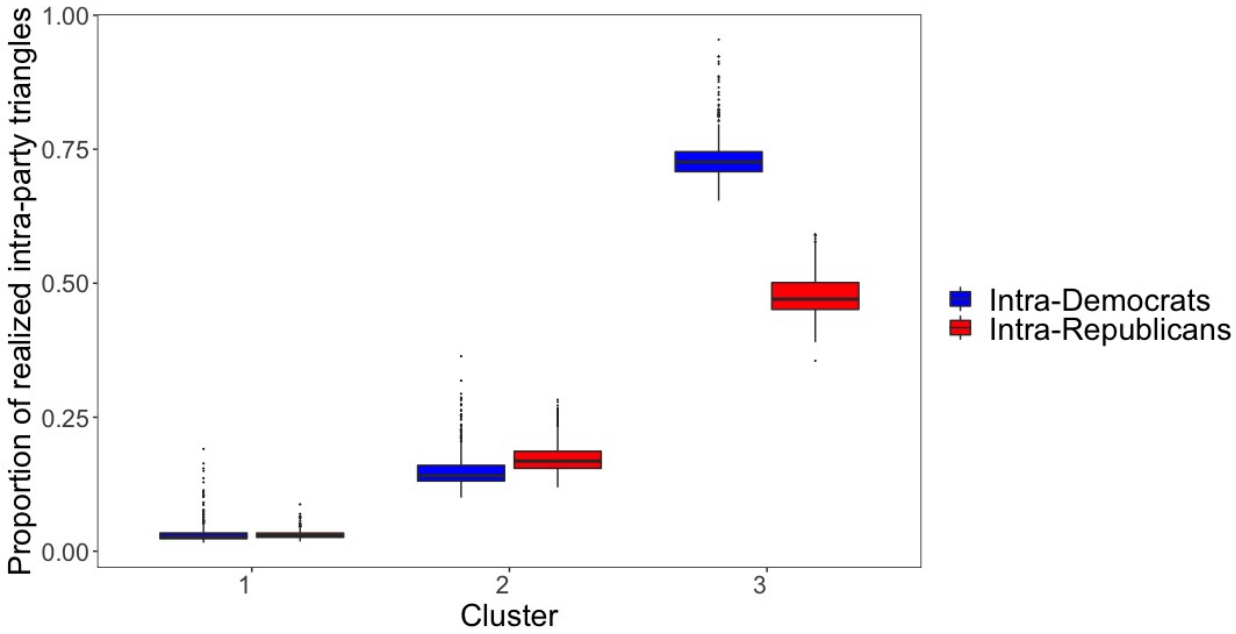

Figure 4.10: Proportion of realized intra-party triangles in simulated networks. Colors indicate the party affiliation (blue = Democratic (D), red = Republican (R)).

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 bullwark against within-party influence. Finally, the regime of cluster 3 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.

In addition to examining the potential impact of different behavioral regimes on voting networks, our model also provides insight into the incidence of these regimes over time. For instance, Figure 4.12 shows maximum probability cluster assignments over the study period. We see that the relatively symmetric cultures represented by cluster 1 and cluster 2 alternate in the nineteenth and twentieth centuries, while the culture of asymmetric polarization represented by cluster 3 becomes dominant after late 1990's. This finding is in line with

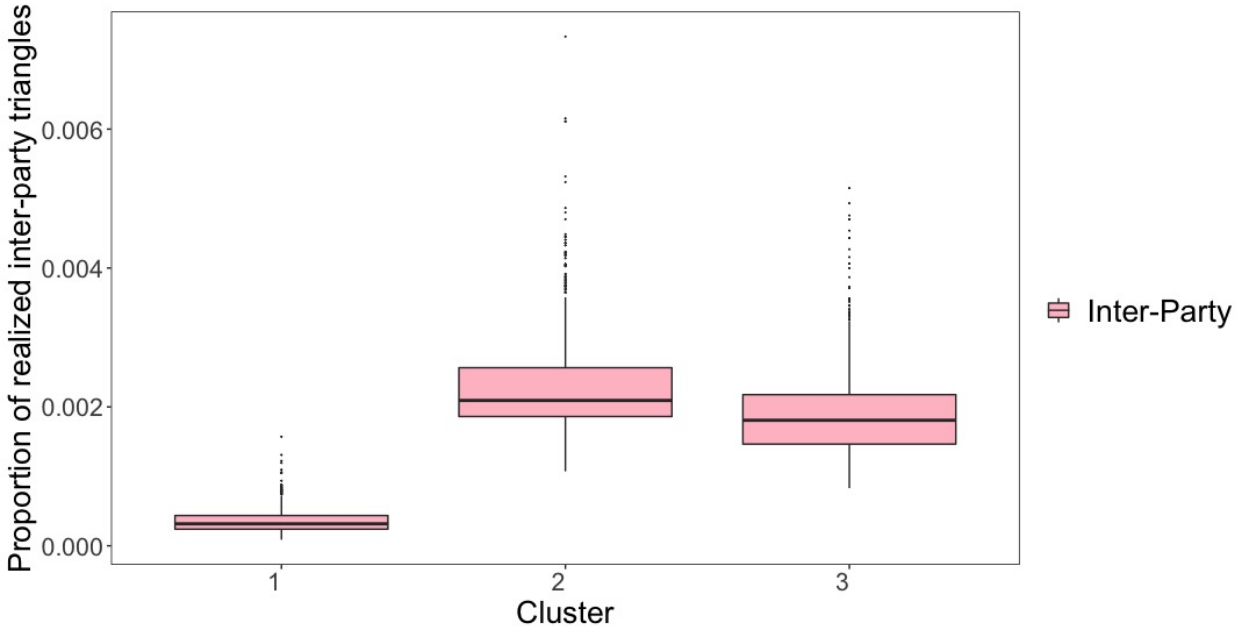

Figure 4.11: Proportion of realized inter-party triangles in simulated networks.

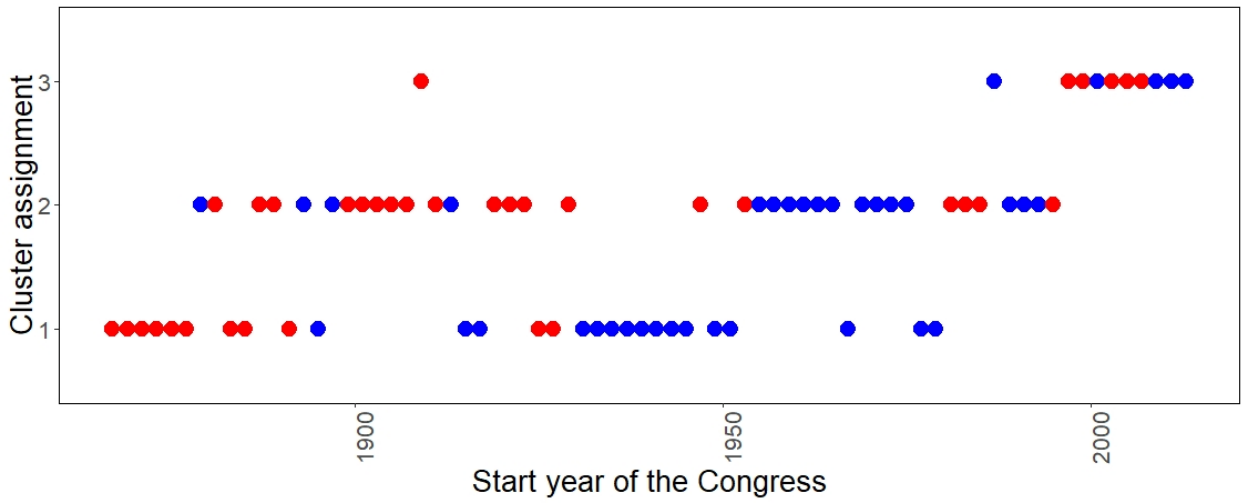

Figure 4.12: Maximum probability cluster assignments over study period. Colors indicate the majority party in the corresponding Congress (blue = Democratic (D), red = Republican (R)). Regimes of voting behavior are visibly correlated over time.

the current trend of political party polarization (Moody and Mucha, 2013). Table 4.5 shows the breakdown of congresses into 3×2 sub-categories according to the estimated co-voting pattern and the observed majority party. We examine the independence of co-voting pattern assignment and the majority party using Pearson's χ^2 test, and we fail to reject the null hypothesis that the majority party is independent of the co-voting patterns ($\chi^2_2 = 1.07$, p-value = 0.58). Thus, while the regimes of party behavior are quite visibly autocorrelated, this pattern does not seem to be related to which party has control of congress at any given time.

Table 4.5: Tabulation of co-voting pattern by majority party (from Congress 40 to Congress 113). Majority party is not significantly related to voting regime.

Co-voting Pattern	Democratic	Republican
1	16	11
2	17	19
3	5	6

4.5.3 Model Assessment

To assess the adequacy of the resulting model, we consider a simulation-based method suggested by Hunter et al. (2008a), with the basic insight that a fitted ERGM model should be able to reproduce in simulation structural properties similar to those of the observed networks. Instead of simulating from a single point estimate, we propose to simulate networks from estimated posterior distribution, following practices of posterior predictive assessment in the Bayesian literature (Gelman et al., 1996). The structural property of interest here is the modularity score (Newman, 2006) (assessed by party), which can be interpreted as a measure of the polarization of networks with respect to party structure. By definition, the modularity score ranges from -1 to 1, with larger values indicating higher levels of polarization. We replicate the following evaluation procedure 100 times:

1. For each vertex set, we first randomly draw a latent membership indicator using pos-

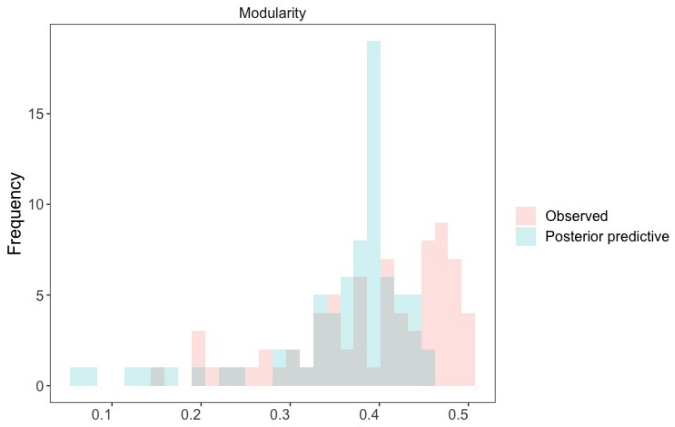

Figure 4.13: Modularity scores of simulated and observed ensemble of networks. Hellinger distance: 0.096.

terior samples of $\boldsymbol{\tau}$, then simulate a network from the corresponding component using posterior samples of $\boldsymbol{\theta}$.

2. Compute the modularity score of the observed ensemble of networks and the simulated networks.

We compare the distribution of modularity scores of simulated networks to that of observed networks using Hellinger distance. We obtain the mean of Hellinger distance values as 0.095 and standard deviation of Hellinger distance values as 0.002.

Figure 4.13 shows the distribution of modularity scores for a replicate that has average-case performance (Hellinger distance value : 0.096). We see that the resulting mixture model can capture not only the left-skewed feature of the modularity scores in the observed data but also the variation of the observed modularity scores to a large extent. The remaining discrepancy between observed modularity scores and those of simulated networks might be mitigated by more accurate but expensive estimation algorithms for cluster-specific parameters (e.g., using importance sampling to approximate ERGM likelihood rather than the pseudo-likelihood) and more sophisticated but potentially less interpretable model specifications.

4.6 Case Study: Advice-seeking Networks among School Teachers

We apply our method to cluster an ensemble of school-level teacher advice-seeking networks [3] (Pitts and Spillane, 2009). This dataset consists of 15 directed networks of sizes ranging from 12 to 76, with each directed tie indicates the seeking of professional advice, from the teacher who seeks the advice to the teacher who offers the advice. These 15 schools are operated independently and hence these advice-seeking networks are independent of each other. Multiple node-level and edge-level covariates are available. As demonstrated in Sweet et al. (2013), the binary edge-level covariate indicating whether the two incident nodes (i.e., teachers) of each edge teach the same grade, named "same grade", plays an important role in explaining the advice-seeking patterns among teachers. Figure 4.14 shows that these networks are different in their structures, and we shall identify clusters in these networks using our method.

4.6.1 Model Specification and Estimation

We choose the following sufficient statistics for analyzing the advice-seeking networks

$$g_1(\mathbf{y}) = \sum_{i<j} y_{ij}, \text{ total number of edges;}$$

$$g_2(\mathbf{y}; \mathbf{X}) = \sum_{i<j} y_{ij}, \mathbb{1}_{\{\mathbf{X}_i = \mathbf{X}_j\}}, \quad \text{total number of within-grade advice-seeking;}$$

$$g_3(\mathbf{y}) = e^{\phi} \sum_{k=1}^{n-2} \left\{1 - (1 - e^{-\phi})^k\right\} EP_k(\mathbf{y}), \text{ GWESP statistic}$$

[3] This dataset is publicly available in R package `HLSM` (Adhikari et al., 2020).

The decay parameter of the GWESP term is fixed as $\phi = 0.25$ as often used in the ERGM literature, and we include the offset term (4.2) to adjust for network size. The underlying idea of this model specification is that we use the edges term to control for the overall propensity of forming edges in these networks, and rely on the number of edges from a teacher to another teacher teaching the same grade to account for the additional propensity of forming edges between teachers from the same grade, and capture the triadic closure effect via the GWESP term. We use the prior specification in Section 4.3, and run the proposed MCMC algorithm for a total of 40000 iterations where the first 20000 draws are discarded as burn-in, and every 100-th iteration after the burn-in is stored as posterior samples.

4.6.2 Results

DIC selects the two cluster model (Figure 4.16), and Figure 4.15 shows that the chain converges well under model $K = 2$. We also compute the Gelman-Rubin statistic (Gelman and Rubin, 1992) using multiple chains with random initial values, the result of which indicates the chosen setting is sufficient for chains to converge. The posterior mean and standard deviation of model parameters are summarized in Table 4.6, and it appears that the key distinction between cluster 1 and cluster 2 lies on the propensity of seeking advice from teachers from the same grade (latter is higher).

Figure 4.17 displays the estimated posterior probability of cluster memberships, which highlights the benefits of model-based clustering, as we are able to know the uncertainty in the cluster assignments. We notice that School 2, 7, 8 and 15 are more likely to belong to cluster 2, and School 12 and 14 are equally likely to belong to the two identified clusters. This finding suggests that within-grade advice-seeking is more prevalent in School 2, 7, 8 and 15, and is lacking to some extent in School 1, 3, 4, 5, 6, 9, 10, 11, and 13. In addition, we note that the coefficients associated with the gwesp terms are quite similar in these two

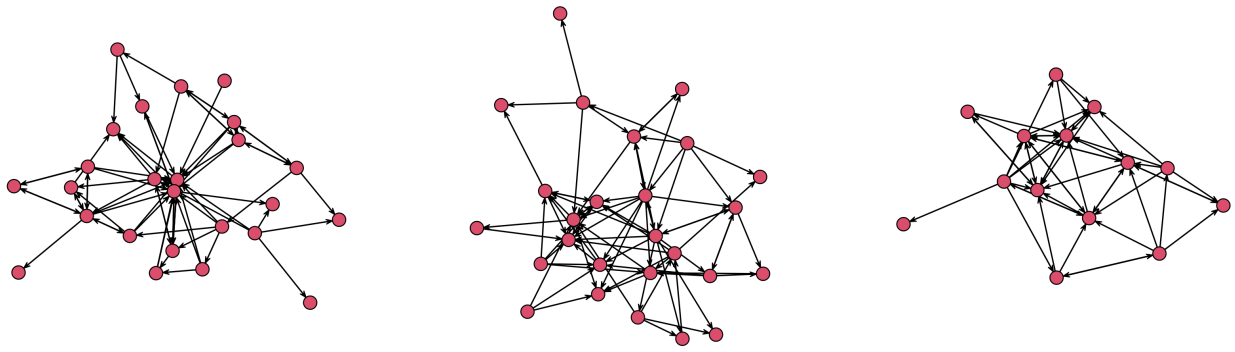

Figure 4.14: Advice-seeking networks among school teachers. School 3 (left), School 5 (middle) and School 14 (right).

Table 4.6: Posterior mean (standard deviation) of model parameters.

	edges + offset	edgecov	gwesp
Cluster 1	0.17 (0.07)	0.81 (0.26)	0.99 (0.10)
Cluster 2	0.28 (0.07)	1.02 (0.24)	1.04 (0.08)

clusters and well above 0, indicating that the triadic closure effect is a critical factor that drives the advice-seeking among school teachers. Although the paucity of publicly available background information limits our ability to gain more insights from this dataset, it is clear that the proposed method enables policy makers to identify schools that are comparatively lacking in the within-grade advice-seeking and hence take further actions accordingly to improve communication between teachers.

4.7 Discussion and Conclusions

In this chapter, we have proposed a mixture of ERGMs approach for modeling the generative process leading to heterogeneous network ensembles. We developed a Metropolis-within-Gibbs algorithm to fit ERGM mixtures and obtained Bayesian estimates of clustering assignment probabilities and the cluster-specific ERGM parameters. To account for the difference in the size of the observed networks, we used a size-adjusted parameterization for ERGMs. We also tailored a version of observed DIC and defined an empirical rule to select the num-

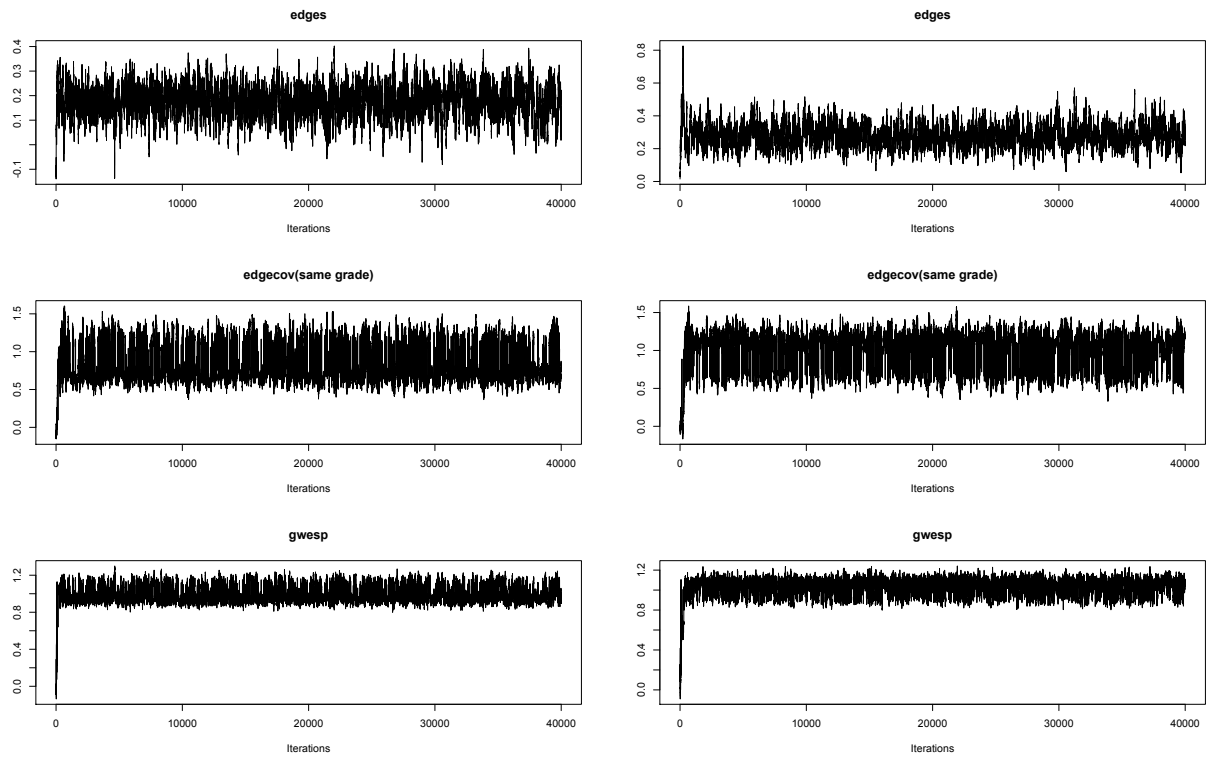

Figure 4.15: Posterior probability of cluster memberships, advice-seeking networks

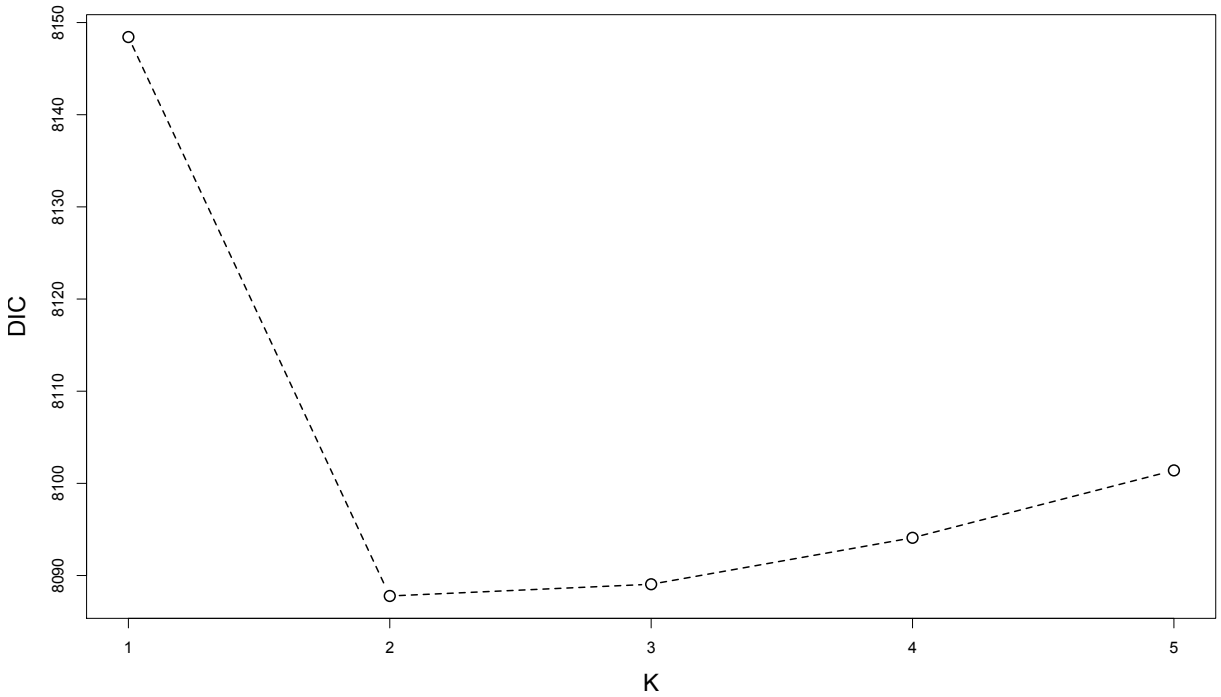

Figure 4.16: DIC vs Number of clusters, advice-seeking networks

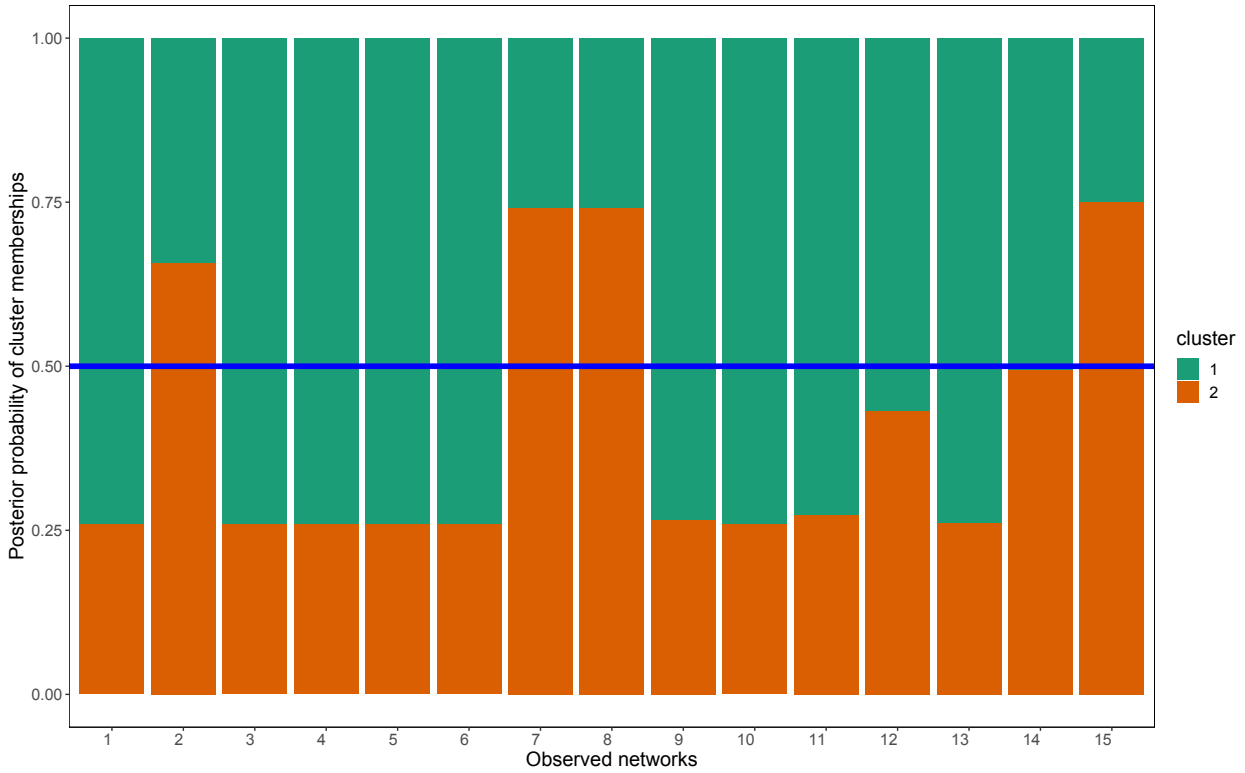

Figure 4.17: Posterior probability of cluster memberships, advice-seeking networks

ber of clusters, which is proved to be effective in a simulation study. The simulation studies also showed that the proposed approach can accurately recover the cluster membership and cluster-specific parameters, without requiring much effort on initialization.

We applied the proposed approach to study the political co-voting networks among U.S. Senators, and identified three clusters that represent vastly different co-voting patterns. After matching the clusters with temporal information, we observed that one symmetric co-voting pattern and another mildly asymmetric co-voting patterns alternate in nineteenth and twentieth century, and there appeared to be an abrupt shift in the co-voting pattern towards the direction of political party polarization in last two decades. We also applied the proposed approach to study the advice-seeking networks among school teachers, and identified several schools that are a bit lacking in the within-grade advice seeking. This application provides more insights about how teachers collaborate in schools, and further justifies the utility of the proposed approach.

Compared to other methods in the literature, our proposed method allows straightforward statistical inference for the generative processes of heterogeneous ensembles of networks with edgewise dependence, and is conveniently interpretable. We believe that the proposed method can prove to be a highly effective tool for both exploratory and inferential analysis of ensembles of networks.

In closing, we comment on three important directions of future research that could prove beneficial to the modeling of ensembles of networks: the development of more sophisticated size-adjusted parameterizations, more accurate tractable approximations of the ERGM likelihood and Dirichlet Process mixtures of ERGMs. It is worth mentioning that the sizes of the US congresses between 1867 and 2014 range from 69 to 112, and the sizes of advice-seeking networks range from 12 to 76, 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. Where 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 our simulation study using the easily computed pseudo-likelihood approximation. In particular, the main deficiency of the pseudo-likelihood is excessive sharpness near the mode, which could in principle encourage the over-production of mixture components. While we do not see this effect here, more accurate likelihood approximations that are inexpensive enough to perform at each MCMC step for large models would be desirable. As such improved approximations become available, they can be easily integrated into the posterior simulation framework described here. Last but not least, a natural further extension of the finite mixture modeling framework could be Dirichlet Process mixtures 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.

Chapter 5

Conclusion and Future work

In this thesis, I have made several key contributions to advance the field of exponential-family random graph models (ERGMs): computation, model selection, methodology. Chapter 2 proposes a highly parallel algorithm for fast Bayesian inference of ERGMs based on approximate Bayesian computation (ABC). With the growing popularity of multi-core processors, we envision the proposed algorithm as a practical alternative to the state-of-the-art MCMC-based algorithm. Chapter 3 offers a systematic investigation on the performance of various existing model selection methods for ERGMs based on extensive simulation studies. Chapter 4 extends the ERGM framework from modeling the generative process of single network observation to that of multiple networks with heterogeneity by proposing a mixture of ERGMs. The proposed novel methodology is also useful as a tool for exploratory analysis and cluster analysis of graph data.

Detailed summary and discussion of the aforementioned contributions are provided in the respective chapters. Here I describe four primary findings that advance the research of ERGMs.

- Chapter 2 shows that fast Bayesian inference for ERGMs is possible by parallel com-

putation under the framework of approximate Bayesian computation. The proposed novel algorithm can yield comparable posterior estimates and inferences to state-of-the-art MCMC-based algorithms, while cutting the wallclock runtime by half with 5 cores, and by 80% with 30 cores.

- Chapter 3 provides results from simulation studies showing that none of the model selection techniques (AIC, BIC and HoPE) dominate the others, they all have their own strengths and weaknesses. Information-criterion-based methods are better at identifying the true model under the scenarios in which the true model is among the candidate models, while HoPE with absolute loss seems to be more capable of selecting a model that yields better predictive performance when the true model is not among the candidate models.

- Chapter 4 proposes a novel methodology for modeling the generative process of ensembles of networks. This novel methodology fills an important gap in the current literature as a highly interpretable framework for characterizing the complex generative process of ensembles of networks as well as model-based clustering, and has been successfully applied to the modeling of generative process of an ensemble of co-voting networks among U.S. Senators and model-based clustering of an ensemble of advice-seeking networks among school teachers.

- The simulation studies in 4 show that the bias in estimating ERGM parameters based on the pseudo-likelihood can be mitigated by multiple network observations. The results also suggest that the biases are smaller for large and sparse networks, *ceteris paribus*.

As a final note, the results presented herein suggest a series of potential future directions. We believe these directions are important for researchers studying ERGMs, and serve the purpose of advancing the field as a whole.

- Although parsimoniously modeling dependencies in networks is the primary objective for ERGMs, we look forward to the development of efficient Bayesian inference on higher-dimensional ERGMs. With the recent development of high-dimensional ABC algorithms ((Nott et al., 2014; Li et al., 2017)), we envision ABC as a promising framework.

- The proposed mixture of ERGMs relies on the Krivitsky reference measure to adjust for the difference in network sizes, which is particularly suitable for social systems with fairly similar mean degrees. To facilitate mixture modeling in more general scenarios, more sophisticated size-adjusted parameterizations and reference measures are a potential avenue for future research.

- A natural further extension of the finite mixture modeling framework could be the development of Dirichlet Process mixtures 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.

- More accurate tractable approximations of the ERGM likelihood are also potential avenues for future research. In mixture modeling, likelihood evaluation is required at each step, a tractable likelihood can be of great value for the development of any practical algorithms for statistical inference.

- Theoretical investigations on HOPE and the effective sample size of general ERGMs are also critical. The former can establish theoretical guarantees on the model selection consistency and bounds on generalization errors. The latter can offer insights about why the current version of BIC for ERGMs can perform well with respect to selecting the true model the under closed-$\mathcal{M}$ scenario and shed light on how the BIC could be further improved for selecting competing ERGM specifications.

Bibliography

S. Adhikari, B. Junker, T. Sweet, and A. C. Thomas. *HLSM: Hierarchical Latent Space Network Model*, 2020. URL `https://CRAN.R-project.org/package=HLSM`. R package version 0.8.2.

H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

Z. W. Almquist and C. T. Butts. Bayesian analysis of dynamic network regression with joint edge/vertex dynamics. *Bayesian inference in the social and natural sciences. New York City, NY: John Wiley & Sons*, 2014.

V. Amati, A. Mol, T. Shafie, C. Hofman, and U. Brandes. A framework for reconstructing archaeological networks using exponential random graph models. *Journal of Archaeological Method and Theory*, pages 1–28, 2019.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

A. Asuncion, Q. Liu, A. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 33–40, 2010.

H. Austad and N. Friel. Deterministic Bayesian inference for the p* model. In *13th International conference on Artificial Intelligence and Statistics (AISTATS, 2010), Chia Laguna Resort, Sardinia, Italy, May 13-15 2010*. Journal of Machine Learning Research (JMLR), 2010.

D. Banks and K. M. Carley. Metric inference for social networks. *Journal of Classification*, 11(1):121–149, 1994.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley, New York, 1978.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, New York, 1994.

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

M. G. Blum. Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, 2010.

M. G. Blum and O. François. Non-linear regression models for approximate Bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

M. G. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

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

L. Bouranis, N. Friel, and F. Maire. Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *Journal of Computational and Graphical Statistics*, 27(3):516–528, 2018.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*, volume 50. Cambridge University Press, 2019.

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

C. T. Butts. A novel simulation method for binary discrete exponential families, with application to social networks. *The Journal of Mathematical Sociology*, 39(3):174–202, 2015.

C. T. Butts. Baseline mixture models for social networks. *arXiv preprint arXiv:1710.02773*, 2017.

C. T. Butts. A perfect sampling method for exponential family random graph models. *The Journal of Mathematical Sociology*, 42(1):17–36, 2018.

C. T. Butts and Z. W. Almquist. A flexible parameterization for baseline mean degree in multiple-network ERGMs. *The Journal of Mathematical Sociology*, 39(3):163–167, 2015.

C. T. Butts and K. M. Carley. Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory*, 11(4):291–305, 2005.

M. Byshkin, A. Stivala, A. Mira, R. Krause, G. Robins, and A. Lomi. Auxiliary parameter MCMC for exponential random graph models. *Journal of Statistical Physics*, 165(4): 740–754, 2016.

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

A. Caimo and N. Friel. Bayesian model selection for exponential random graph models. *Social Networks*, 35(1):11–24, 2013.

A. Caimo and N. Friel. Bergm: Bayesian exponential random graphs in R. *Journal of Statistical Software*, 61(2):1–25, 2014. URL `http://www.jstatsoft.org/v61/i02/`.

A. Caimo, F. Pallotti, and A. Lomi. Bayesian exponential random graph modelling of interhospital patient referral networks. *Statistics in Medicine*, 36(18):2902–2920, 2017.

G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673, 2006.

K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

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

B. Dabbs and B. Junker. Comparison of cross-validation methods for stochastic block models. *arXiv preprint arXiv:1605.03000*, 2016.

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

K. A. Doksum and A. Y. Lo. Consistent and robust bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1):443–453, 1990.

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

T. J. Fararo and M. Sunshine. *A Study of a Biased Friendship Network*. Syracuse University Press, Syracuse, NY, 1964.

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

I. Fellows and M. S. Handcock. Exponential-family random network models. *arXiv preprint arXiv:1208.0121*, 2012.

S. M. Fitzhugh, J. E. Pixley, and C. T. Butts. A life history graph approach to the analysis and comparison of life histories. *Advances in Life Course Research*, 25:16–34, 2015.

J. H. Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487, 2006.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

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

N. Friel. Evidence and bayes factor estimation for gibbs random fields. *Journal of Computational and Graphical Statistics*, 22(3):518–532, 2013.

W. Fu and P. O. Perry. Estimating the number of clusters using cross-validation. *Journal of Computational and Graphical Statistics*, pages 1–12, 2019.

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule. In *Advances in Neural Information Processing Systems*, pages 1737–1745, 2011.

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185, 1998.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 1995.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760, 1996.

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

W. R. Gilks, G. O. Roberts, and E. I. George. Adaptive direction sampling. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(1):179–189, 1994.

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

G. Grazioli, R. W. Martin, and C. T. Butts. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Frontiers in Molecular Biosciences, Biological Modeling and Simulation*, 6(42), 2019a. doi: 10.3389/fmolb.2019.00042.

G. Grazioli, R. W. Martin, and C. T. Butts. Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Frontiers in Molecular Biosciences*, 6:42, 2019b.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

C. Groendyke, D. Welch, and D. R. Hunter. A network-based analysis of the 1861 Hagelloch measles data. *Biometrics*, 68(3):755–765, 2012.

M. S. Handcock. Assessing degeneracy in statistical models of social networks. Technical report, Center for Statistics and Social Sciences, University of Washington, 2003. `https://www.csss.washington.edu/Papers/wp39.pdf`.

M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, 2010.

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

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

R. He and T. Zheng. GLMLE: graph-limit enabled fast computation for fitting exponential random graph models to large social networks. *Social Network Analysis and Mining*, 5(1): 8, 2015.

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

N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664, 2008.

P. D. Hoff and M. D. Ward. Modeling dependencies in international relations networks. *Political Analysis*, 12(2):160–175, 2004.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

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

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

R. M. Hummel, D. R. Hunter, and M. S. Handcock. Improving simulation-based algorithms for fitting ERGMs. *Journal of Computational and Graphical Statistics*, 21(4):920–939, 2012.

D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008a.

D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 2008b.

D. R. Hunter, P. N. Krivitsky, and M. Schweinberger. Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882, 2012.

T. Jones. *textmineR: Functions for Text Mining and Topic Modeling*, 2019. URL `https://CRAN.R-project.org/package=textmineR`. R package version 3.0.4.

G. Karabatsos and F. Leisen. An approximate likelihood perspective on ABC methods. *Statistics Surveys*, 12:66–104, 2018.

E. D. Kolaczyk and P. N. Krivitsky. On the question of effective sample size in network modeling: an asymptotic inquiry. *Statistical Science*, 30(2):184, 2015.

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

J. Koskinen, P. Wang, G. Robins, and P. Pattison. Outliers and influential observations in exponential random graph models. *Psychometrika*, 83(4):809–830, 2018.

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

J. H. Koskinen, G. L. Robins, and P. E. Pattison. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, 7(3):366–384, 2010.

P. N. Krivitsky. Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Computational Statistics & Data Analysis*, 107:149–161, 2017.

P. N. Krivitsky and M. S. Handcock. Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24, 2008.

P. N. Krivitsky, M. S. Handcock, and M. Morris. Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4):319–339, 2011.

B. C. L. Lehmann. *Inferring differences between networks using Bayesian exponential random graph models*. PhD thesis, University of Cambridge, 2019.

J. Li, D. J. Nott, Y. Fan, and S. A. Sisson. Extending approximate Bayesian computation methods to high dimensions via a gaussian copula model. *Computational Statistics & Data Analysis*, 106:77–89, 2017.

T. Li, E. Levina, and J. Zhu. Network cross-validation by edge sampling. *Biometrika*, 107 (2):257–276, 2020.

B. G. Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008. ISBN 0387763694, 9780387763699.

A. Løland, R. B. Huseby, N. L. Hjort, and A. Frigessi. Statistical corrections of invalid correlation matrices. *Scandinavian Journal of Statistics*, 40(4):807–824, 2013.

D. Lusher, J. Koskinen, and G. Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, 2013.

Z. Maoz, R. D. Kuperman, L. Terris, and I. Talmud. Structural equivalence and international conflict: A social networks analysis. *Journal of Conflict Resolution*, 50(5):664–689, 2006.

J.-M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25:459–507, 2005.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

D. A. McFarland, J. Moody, D. Diehl, J. A. Smith, and R. J. Thomas. Network ecology and adolescent social structure. *American Sociological Review*, 79(6):1088–1121, 2014. doi: 10.1177/0003122414554001.

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

K. L. Mengersen, P. Pudlo, and C. P. Robert. Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4):1321–1326, 2013.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087–1092, 1953.

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

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

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

I. Murray, Z. Ghahramani, and D. J. C. MacKay. Mcmc for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pages 359–366, Arlington, Virginia, United States, 2006. AUAI Press. ISBN 0-9749039-2-2. URL `http://dl.acm.org/citation.cfm?id=3020419.3020463`.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Y. Nebil, S. M. Fitzhugh, M. Kurant, A. Markopoulou, C. T. Butts, and N. Prulj. ergm. graphlets: A package for erg modeling based on graphlet statistics. *Journal of Statistical Software*, 65(i12), 2015.

M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

D. J. Nott, Y. Fan, L. Marshall, and S. Sisson. Approximate Bayesian computation and bayes' linear analysis: toward high-dimensional abc. *Journal of Computational and Graphical Statistics*, 23(1):65–86, 2014.

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

L. E. Ortiz and L. P. Kaelbling. Adaptive importance sampling for estimation in structured domains. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 446–454. Morgan Kaufmann Publishers Inc., 2000.

J. Park and M. Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018.

T. Pennanen and M. Koivu. An adaptive importance sampling technique. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 443–455. Springer, 2006.

G. C. Pflug. *Optimization of Stochastic Models. The Interface Between Simulation and Optimization*. Boston: Kluwer Academic., 1996.

V. M. Pitts and J. P. Spillane. Using social network methods to study school leadership. *International Journal of Research & Method in Education*, 32(2):185–207, 2009.

D. Prangle. Adapting the abc distance function. *Bayesian Analysis*, 12(1):289–309, 2017.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 01 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl301. URL `https://doi.org/10.1093/bioinformatics/btl301`.

W. Pu, J. Choi, Y. Hwang, and E. Amir. A deterministic partition function approximation for exponential random graph models. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25: 111–163, 1995.

A. Rapoport. Contribution to the theory of random and biased nets. *Bulletin of Mathematical Biology*, 19(4):257–277, 1957.

A. Rapoport. A probabilistic approach to networks. *Social Networks*, 2(1):1–18, 1979.

A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3: 446–484, 2009.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2): 192–215, 2007.

J. S. Rosenthal. Optimal proposal distributions and adaptive mcmc. *Handbook of Markov Chain Monte Carlo*, 4(10.1201), 2011.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

D. B. Rubin. Comment on 'the calculation of posterior distributions by data augmentation'by ma tanner and wh wong,". *Journal of the American Statistical Association*, 82(398):543–46, 1987.

D. B. Rubin. Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics*, 3:395–402, 1988.

R. Y. Rubinstein and D. P. Kroese. *The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-carlo Simulation (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2004. ISBN 038721240X.

M. Salter-Townshend and T. B. Murphy. Role analysis in networks using mixtures of exponential random graph models. *Journal of Computational and Graphical Statistics*, 24(2): 520–538, 2015. doi: 10.1080/10618600.2014.923777.

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

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

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

M. Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

M. Schweinberger and M. S. Handcock. 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, 2015.

M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. Stewart. Exponential-family models of random graphs: Inference in finite-, super-, and infinite-population scenarios. *Statistical Science*, page Forthcoming, 2019.

C. R. Shalizi and A. Rinaldo. Consistency under sampling of exponential random graph models. *Annals of Statistics*, 41(2):508, 2013.

J. Shore and B. Lubin. Spectral goodness of fit for network models. *Social Networks*, 43: 16–27, 2015.

M. Signorelli and E. C. Wit. Model-based clustering for populations of networks. *Statistical Modelling*, 20(1):9–29, 2020.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

S. L. Simpson, S. Hayasaka, and P. J. Laurienti. Exponential random graph modeling for complex brain networks. *PLoS One*, 6(5):1–11, 2011.

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

S. Sisson, Y. Fan, and M. Beaumont. Overview of abc. *Handbook of Approximate Bayesian Computation*, pages 3–54, 2018.

S. A. Sisson and Y. Fan. *Likelihood-free MCMC*. Chapman & Hall/CRC, New York.[839], 2011.

Ø. Skare, E. Bølviken, and L. Holden. Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737, 2003.

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

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

T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

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

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 64(4):583–639, 2002.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

J. Stewart, M. Schweinberger, M. Bojanowski, and M. Morris. Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Social Networks*, to appear, 2019.

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

T. M. Sweet, A. C. Thomas, and B. W. Junker. Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38(3):295–318, 2013.

T. M. Sweet, A. C. Thomas, and B. W. Junker. Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. *Handbook on mixed membership models and their applications*, pages 463–488, 2014.

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

L. S. Tan and N. Friel. Bayesian variational inference for exponential random graph models. *Journal of Computational and Graphical Statistics*, forthcoming, 2020.

C. J. ter Braak and J. A. Vrugt. Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446, 2008.

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

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

M. A. Van Duijn, K. J. Gile, and M. S. Handcock. 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, 2009.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

G. G. V. Von, A. Slaughter, and K. de la Haye. Exponential random graph models for little networks. *Social Networks*, forthcoming, 2020.

C. Wang, C. T. Butts, J. R. Hipp, R. Jose, and C. M. Lakon. Multiple imputation for missing edge data: a predictive evaluation method with application to Add Health. *Social Networks*, 45:89–98, 2016.

J. Wang and Y. F. Atchadé. Approximate Bayesian computation for exponential random graph models for large social networks. *Communications in Statistics-Simulation and Computation*, 43(2):359–377, 2014.

P. Wang, G. Robins, and P. Pattison. PNet. *Program for the Simulation and Estimation of Exponential Random Graph (p*) Models*, 2009.

P. Wang, P. Pattison, and G. Robins. Exponential random graph model specifications for bipartite networks - a dependence hierarchy. *Social Networks*, 35(2):211–222, 2013a.

P. Wang, G. Robins, P. Pattison, and E. Lazega. Exponential random graph models for multilevel networks. *Social Networks*, 35(1):96–115, 2013b.

P. Wang, G. Robins, P. Pattison, and J. Koskinen. *MPNet, Program for the simulation and estimation of (p*) exponential random graph models for Multilevel networks: USER MANUAL.* Melbourne School of Psychological Sciences The University of Melbourne Australia, 2014.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, Cambridge, U.K, 1994.

S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61(3):401–425, 1996.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

F. Yin, N. E. Phillips, and C. T. Butts. Selection of exponential-family random graph models via Held-Out Predictive Evaluation (HOPE). *arXiv preprint arXiv:1908.05873*, 2019.

W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

W. Zhu, J. M. Marin, and F. Leisen. A Bootstrap likelihood approach to Bayesian computation. *Australian & New Zealand Journal of Statistics*, 58(2):227–244, 2016.

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