METHOD OF MOMENTS FOR EXPONENTIAL RANDOM GRAPH MODEL SELECTION

Helal El-Zaatari

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2024

Approved by:
Michael R. Kosorok
Marcella Boynton
Didong Li
Fei Yu
John Preisser
ABSTRACT

Helal El-Zaatari: Method of Moments for Exponential Random Graph Model Selection
(Under the direction of Michael R. Kosorok)

Collaboration has become crucial in solving scientific problems in biomedical and health sciences. There is a growing interest in applying social network analysis to professional associations aiming to leverage expertise and resources for optimal synergy. As a set of computational and statistical methods for analyzing social networks, Exponential Random Graph Models (ERGMs) examine complex collaborative networks due to their uniqueness of allowing for non-independent variables in network modeling. This dissertation introduces a novel methodology for endogenous variable selection in Exponential Random Graph Models (ERGMs) to enhance the analysis of social networks across various scientific disciplines. Addressing critical challenges such as ERGM degeneracy and computational complexity, our method integrates a systematic stepwise feature selection process. This approach effectively manages the intractable normalizing constants characteristic of ERGMs, ensuring the generation of accurate and non-degenerate network models. An empirical application to ten real-life binary networks demonstrates the method’s effectiveness in accommodating network dependencies and providing meaningful insights into complex network interactions. Particularly notable is the adaptability of this methodology to both directed and undirected networks, overcoming the limitations of traditional ERGMs in capturing realistic network structures. The findings contribute significantly to network analysis, offering a robust framework for modeling and interpreting social networks and laying a foundation for future advancements in statistical network analysis techniques.
This dissertation is dedicated to my parents, Rima and Mohammad, and my brothers, Ziad and Bassil, whose words and actions inspired me to pursue a doctorate.
ACKNOWLEDGEMENTS

I would like to thank my family for their everlasting support. I am blessed and fortunate to have Dr. Micahel Kosorok as my dissertation adviser. Michael not only helped me develop professionally but also inspired me to become a better person and musician. During my graduate career, Michael’s calmness and levelheadedness made me believe that any problem I faced was manageable.

We gratefully acknowledge several collaborators on this research endeavor. Chapters 1 and 2 are a joint collaboration with Dr. Fei Yu. Her research project in 2021 introduced me to Exponential Random Graph Models and inspired me to learn more about these fascinating mathematical objects. I want to thank Fei for her guidance and advice. Chapter 3 is a joint collaboration with Dr. Didong Li and Yun Jin Park. I thank them for their wonderful ideas, time and enthusiasm.

I thank Dr. Marcy Boynton and Dr. John Preisser for their service and feedback on this work. I am inspired by their dedication to rigorous scientific output. I will always fondly remember the lunches with John as well as being a teaching assistant for Marcy.

It takes a village to raise a child, the same can be said about research endeavours. I want to thank Melissa Hobgood, Tiffanny Harris and Betsy Seagroves for helping me during my time as a UNC student. Thank you Tarek and Jesus for being wonderful classmates and friends.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1: LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Methods</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Literature Search</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Study Selection</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Results</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Study Design</td>
<td>5</td>
</tr>
<tr>
<td>1.3.2 Networks</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Discussion</td>
<td>8</td>
</tr>
<tr>
<td>1.4.1 Study Design</td>
<td>8</td>
</tr>
<tr>
<td>1.4.2 Networks</td>
<td>10</td>
</tr>
<tr>
<td>1.4.3 Limitations</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Conclusion</td>
<td>14</td>
</tr>
<tr>
<td>CHAPTER 2: ENDOGENOUS VARIABLE SELECTION</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Definition of ERGMs</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Endogenous Variable Selection for ERGMs</td>
<td>18</td>
</tr>
<tr>
<td>2.3.1 Obtaining an Initial Set of Endogenous Variable</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 Stochastic Forward Selection</td>
<td>20</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1.1 Data Extraction Schema ................................................................. 5
1.2 Summarized Report of the Literature Review ................................. 7
1.3 Network Characteristics ................................................................. 8
2.1 Average Network Motif Counts ...................................................... 27
2.2 Average Network Motif Counts for the ERGMs ................................. 28
3.1 Descriptive Statistics for 9 Networks.............................................. 40
3.2 Number of endogenous variables and ERGMs ................................. 42
LIST OF ABBREVIATIONS

AIC    Aikake Information Criterion
CD     Contrastive Divergence
CvM    Cramer von Mises
dsp    dyadwise shared partners
ERGM  Exponential Random Graph Model
esp    edgewise shared partners
endo   endogenous
exo    exogenous
FRC    Forman Ricci Curvature
gwesp  geometrically weighted edgewise shared partners
gwdsp  geometrically weighted dyad-wise shared partners
gwnsp  geometrically weighted non-edgewise shared partners
GOF    Goodness of fit
MCMC   Monte Carlo Markov Chain
nsp    non-edgewise shared partners
CHAPTER 1: LITERATURE REVIEW

Collaboration has become crucial in solving scientific problems in biomedical and health sciences. There is a growing interest in applying social network analysis to professional associations aiming to leverage expertise and resources for optimal synergy. As a set of computational and statistical methods for analyzing social networks, Exponential Random Graph Models (ERGMs) examine complex collaborative networks due to their uniqueness of allowing for non-independent variables in network modeling. This study took a review approach to collect and analyze ERGM applications in health sciences by following the protocol of a systematic review. We included a total of 30 studies. The bibliometric characteristics revealed significant authors, institutions, countries, funding agencies, and citation impact associated with the publications. In addition, we observed five types of ERGMs for network modeling (standard ERGM and its extensions: Bayesian ERGM, Temporal ERGM, Separable Temporal ERGM, and Multilevel ERGM). Most studies (80%) used the standard ERGM, which possesses only endogenous and exogenous variables examining either micro-level (individual based) or macro-level (organization based) collaborations without exploring how the links between individuals and organizations contribute to the overall network structure. Our findings help researchers (a) understand the extant research landscape of ERGM applications in health sciences, (b) learn to control and predict connection occurrence in a collaborative network, and (c) better design ERGM applied studies to examine complex relations and social system structure, which is native to professional collaborations.
1.1 Introduction

In biomedical and health sciences, a collaborative effort has become crucial in addressing scientific challenges (Ho et al. 2021), improving health service effectiveness (Harris et al. 2012), reducing health disparities (Okamoto 2015), and producing high impact scientific publications (Smith et al. 2021). The synergy usually leads to innovation and acceleration of sciences and services by leveraging the perspectives, expertise, and resources of professionals from different disciplines, institutions, and countries (Bennett et al. 2010). Therefore, researchers examine collaboration networks to disclose patterns and characteristics of extant collaborative endeavors, measure the associations between factors and variables that can influence the network structure, and predict the probability of future connections between nodes (e.g., professionals and organizations). Previous studies have widely used network analysis in analyzing and visualizing co-authorship (Yu et al. 2018, 2020) and health policies (Provan et al. 2005; Luke et al. 2013). Exponential Random Graph Models (ERGMs), also known as p* models, are a set of computational and statistical methods for analyzing social networks. They are predictive models of network structure, a new and rapidly evolving area of social network analysis. ERGMs assume that the presence or absence of other relationships or individual attributes influences the emergence of a connection/link. For example, A, B, and C are three nodes in a network. If A connects with B and B connects with C, the probability of the connection formation between A and C increases (i.e., transitivity). Compared with other standard regression methods and inferential tests that researchers have adopted to assess the relationship between measures of variables in a collaborative social network (e.g., linear regressions (Fujimoto et al. 2009; Kesternich and Rank 2022), weighted least squares regression (Provan et al. 2009; Retrum et al. 2013), and ordinary least squares regression (Puro and Kelly 2022)), ERGMs stand out because (1) practically, ERGMs can handle regression analyses and also simulate similar networks; and (2) theoretically, they are modified logistic regressions that allow for the probability of a connection/link to depend upon the presence or absence of other connections/links in the network (Stanford Hu-
man Evolutionary Ecology and Health; YousefiNooraie et al. 2014; van der Pol 2018). ERGMs can model a network relation with both endogenous (dependent) and exogenous (independent) variables. While exogenous variables are attributes of a node and are not related to the network structure, endogenous variables are strictly a function of the network modeling and the essence of a standard ERGM. Overall, ERGMs focus on the interaction between the network structures with endogenous variables (e.g., transitivity, reciprocity) and individual attributes of nodes with exogenous variables (e.g., age, gender, experience). As a relatively new approach, ERGMs are considered underutilized for investigating collaboration networks (Harris et al. 2012). Although there are literature surveys of ERGMs on concepts (Drobyshevskiy and Turdakov 2020), new versions (Goodreau 2007; Robins et al. 2007), and statistical schemes and applications (Ghafouri and Khasteh 2020), it is unknown the extent of EGRM application to specific fields (Ghafouri and Khasteh 2020). Mainly, there is a lack of examining applications of ERGMs on collaborations in biomedical and health sciences. Therefore, this study fills in this gap by systematically searching and screening literature, extracting data, and synthesizing research evidence pertinent to the applications of ERGMs to collaboration networks within biomedical and health sciences.

Random graph models are essential for building and analyzing complex networks. Thus, our findings help health sciences researchers understand ERGMs and learn to control and predict connection occurrence in a collaboration network (Drobyshevskiy and Turdakov 2020). The remainder of this paper is structured as follows. First, in Section 2, titled Method, we describe the rigorous review methodology we adopted, including the literature search, study selection, and data extraction processes. Our primary findings are presented in Section 3, Results, which includes the bibliometrics, study design, and network characteristics and features of the included studies. In Section 4, Discussion, we provide a detailed interpretation of the results, highlighting significant insights and addressing potential study limitations. Finally, Section 5, Conclusion, summarizes our main discoveries and their broader implications.
1.2 Methods

Since we aimed to produce research synthesis systematically, transparently, and timely, we chose a review approach. We followed the protocol of a systematic review (SR), a gold standard for knowledge synthesis. We reported results in descriptive summaries or data categorization instead of qualitative summaries or meta-analyses, typical in an SR. In addition, our synthesis process did not involve quality assessment.

1.2.1 Literature Search

We systematically searched for literature in PubMed and Scopus in June 2022. PubMed is the optimal citation database for publications in biomedical and health sciences (Falagas et al. 2008), while Scopus is the largest citation database in the world with a broad journal indexing scope (Schotten et al. 2017). We constructed the search query with identified ERGM name variations (e.g., exponential family random graph models, temporal exponential random graph model, p* model, etc.) and collaboration network terms (e.g., scientific network, team sciences, co-authorship, etc.). After testing and validating the query in PubMed, we translated it into a query for Scopus. The retrieved results were aggregated and deduplicated in Endnote. In addition, we conducted backward searching by manually reviewing relevant studies references and identified additional publications for screening (Fig. 1).

1.2.2 Study Selection

Two reviewers (FY & HE) went through both title, abstract and full text screening independently. We employed a set of inclusion and exclusion criteria for study selection. The exclusion criteria are (1) not published in the English language; (2) not in the domain of biomedical and health sciences; (3) not an original research article (e.g., commentary, editorial, erratum, abstract only, letter, note, review, etc.); (4) does not apply any ERGMs to the dataset; (5) does not study collaboration networks; (6) focuses on ERGM applications in epidemiology and disease spread-
ing (e.g., HIV infection, needle sharing, etc.); (7) focuses on a non-professional relationship like friendship, dating, or sexual relation, etc. The inter-rater reliability of the two reviewers was 74% (proportionate agreement) at the title/abstract level and 88% at the full text level. The two reviewers met and discussed all conflicts until they agreed on the final inclusion.

<table>
<thead>
<tr>
<th>Category:</th>
<th>Data Field:</th>
<th>Data Source:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Design</td>
<td>ERGM application, ERGM type, node type, data source, data analysis software</td>
<td>Full-text article</td>
</tr>
<tr>
<td>Networks</td>
<td># networks, average # nodes, network direction, # endogenous variables, # exogenous variables, density, centrality, Goodness of Fit</td>
<td>Full-text article</td>
</tr>
</tbody>
</table>

**Table 1.1: Data Extraction Schema**

### 1.3 Results

We include 30 studies that met our inclusion criteria (Fig. 1.1). The extracted data is presented in Table 2 and 3.

![Figure 1.1: Flow diagram of literature search and study selection (Page et al. 2021)](image)

#### 1.3.1 Study Design

We observed five types of ERGMs for network modeling: 1) Standard ERGM (STD) that consists of only endogenous and exogenous variables and assumes that the network structure is
fixed and does not change over time; 2) Bayesian ERGM (Bergm) that combines the principles of Bayesian statistics and ERGM for modelling complex networks that exhibit various patterns of connectivity such as clustering, reciprocity, and transitivity; 3) Temporal ERGM (Tergm) that accounts for the inter-temporal dependence in longitudinal collaborative network modeling; 4) Separable temporal ERGM (Stergm) that extends Tergm to account for the temporal dependencies among the network structures over time and can handle both within period component (i.e., within each time period) and between-period component (i.e., across time); 5) Multilevel ERGM (Mergm) that consists micro, meso, and macro-level network modelling accounting for the effects of both individual level and group level covariates on the network structure. Most studies (80%) used STD. As an extension to STD, Bergm, Tergm, Stergm, and Mergm were adopted to either address the limitations in STD or add additional parameters for network modeling. For example, only one study (Caimo et al. 2017) applied Bayesian computational methods to address the limitations of ERGM diffusion complexity. Tergm and Stergm predict the probability of maintaining a connection between two nodes concerning time. Incorporating this time dependent parameter, researchers gained insight into the processes that drive connection formation and change (Azondekon 2018a; Broekel and Bednarz 2019; Ho et al. 2021). Both Wang et al. (Wang et al. 2013) and McGlashan et al. (McGlashan et al. 2019) adopted a Mergm network to map and test complex interdependencies in a collaboration network. Wang was also a coauthor in McGlashan et al.’s study. While Wang et al. focused on the French cancer researcher collaboration network at laboratory and advice seeking levels, McGlashan et al. explored the complex collaborative system where steering committee members connected for community based public health interventions.

1.3.2 Networks

The networks reviewed in this study were modest in size, with only two studies (Harris 2013; Azondekon 2018a) that applied ERGMs to model a network with more than 1000 nodes. In addition, density and centrality (e.g., degree, betweenness, closeness) were typical network measures in 80% of the included studies. Further, researchers predominantly used simulation for
Goodness of Fit (GOF) evaluation. Notably, researchers used several endogenous variables to model network transitivity and centrality. For example, 22 studies modeled transitivity operating edgewise shared partners, dyadwise shared partners, their geometrically weighted counterparts, triangles and alternating triangles (Luke et al. 2010, 2013; Lomi and Pallotti 2012; Harris et al. 2012, 2015; Wang et al. 2013; Harris 2013; Uddin et al. 2013; Heaney 2014; Shearer et al. 2014; Okamoto 2015; Shoham et al. 2016; Caimo et al. 2017; Mascia et al. 2018; Bunger et al. 2018; Azondekon 2018b; Broekel and Bednarz 2019; McGlashan et al. 2019; Prochnow et al. 2020; Ho et al. 2021; Kesternich and Rank 2022; Bohnett et al. 2022). Similarly, 21 studies adopted k-stars and geometrically weighted degrees to model network centrality (Fattore and Salvatore 2010; Luke et al. 2010, 2013; Zappa and Mariani 2011; Zappa 2011; Lomi and Pallotti 2012; Harris et al. 2012, 2015; Wang et al. 2013; Harris 2013; Uddin et al. 2013; Heaney 2014; Shearer et al. 2014; Okamoto 2015; Shoham et al. 2016; Caimo et al. 2017; Mascia et al. 2018; Bunger et al. 2018; Azondekon 2018b; Broekel and Bednarz 2019; McGlashan et al. 2019; Prochnow et al. 2020; Ho et al. 2021; Kesternich and Rank 2022; Bohnett et al. 2022). Most studies (93 %) conducted GOF diagnostics for their ERGMs, in which two diagnostic measures were adopted. The first type is local by comparing the final ERGM with a null ERGM, while the second type is global by reaching the final ERGM with a network regression model constructed.
via a quadratic assignment procedure [33,32,3]. The majority of authors relied on simulation for the GOF diagnostic.

Table 1.3: Network Characteristics

<table>
<thead>
<tr>
<th>Study:</th>
<th># Networks</th>
<th>Avg. # Nodes</th>
<th>Direction</th>
<th># Endo Var</th>
<th># Exo Var</th>
<th>Density</th>
<th>Centrality</th>
<th>GOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fattore 2010</td>
<td>1</td>
<td>223</td>
<td>U</td>
<td>0</td>
<td>4</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Luke 2010</td>
<td>5</td>
<td>34.6</td>
<td>U</td>
<td>0</td>
<td>4</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Zappa 2011a</td>
<td>1</td>
<td>784</td>
<td>U</td>
<td>4</td>
<td>11</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Zappa 2011b</td>
<td>2</td>
<td>891</td>
<td>D</td>
<td>4</td>
<td>17</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Harris 2012</td>
<td>3</td>
<td>18</td>
<td>U</td>
<td>4</td>
<td>17</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Lomi 2012</td>
<td>1</td>
<td>91</td>
<td>D</td>
<td>10</td>
<td>15</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Harris 2013</td>
<td>1</td>
<td>2107</td>
<td>D</td>
<td>4</td>
<td>10</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Luke 2013</td>
<td>8</td>
<td>18.88</td>
<td>U</td>
<td>2</td>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Uddin 2013</td>
<td>85</td>
<td>26.22</td>
<td>U</td>
<td>5</td>
<td>0</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Wang 2013</td>
<td>1</td>
<td>97.82</td>
<td>D</td>
<td>8</td>
<td>0</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Heaney 2014</td>
<td>4</td>
<td>168</td>
<td>U(n=2), D(n=2)</td>
<td>5</td>
<td>9</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Yousef-Nooraie 2014</td>
<td>2</td>
<td>28</td>
<td>D</td>
<td>3</td>
<td>10</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Shearer 2014</td>
<td>6</td>
<td>69</td>
<td>D</td>
<td>4</td>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Beve 2015a</td>
<td>162</td>
<td>18</td>
<td>D</td>
<td>1</td>
<td>1</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Beve 2015b</td>
<td>162</td>
<td>27.77</td>
<td>D</td>
<td>1</td>
<td>1</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Harris 2015</td>
<td>4</td>
<td>49</td>
<td>U</td>
<td>3</td>
<td>8</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Okamoto 2015</td>
<td>2</td>
<td>167</td>
<td>D</td>
<td>4</td>
<td>15</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Shoham 2016</td>
<td>1</td>
<td>71</td>
<td>U</td>
<td>4</td>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Caimo 2017</td>
<td>1</td>
<td>110</td>
<td>D</td>
<td>9</td>
<td>10</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Azondekon 2018</td>
<td>1</td>
<td>1792</td>
<td>U</td>
<td>4</td>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Broekel 2018</td>
<td>1</td>
<td>910</td>
<td>U</td>
<td>3</td>
<td>12</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Bunger 2018</td>
<td>2</td>
<td>131</td>
<td>D</td>
<td>7</td>
<td>12</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Marchland 2018</td>
<td>4</td>
<td>168</td>
<td>U</td>
<td>1</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Mascia 2018</td>
<td>2</td>
<td>48.5</td>
<td>D</td>
<td>9</td>
<td>19</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>McGlashan 2019</td>
<td>2</td>
<td>19</td>
<td>D</td>
<td>13.21</td>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Prusaczzyk 2019</td>
<td>1</td>
<td>69</td>
<td>U</td>
<td>1</td>
<td>2</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Prochnow 2020</td>
<td>1</td>
<td>18</td>
<td>U</td>
<td>2</td>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Ho 2021</td>
<td>3</td>
<td>9.33</td>
<td>D</td>
<td>4</td>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bohnett 2022</td>
<td>42</td>
<td>19.4</td>
<td>U</td>
<td>2</td>
<td>10</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Kesternich 2022</td>
<td>2</td>
<td>113</td>
<td>U</td>
<td>9</td>
<td>5</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

1.4 Discussion

1.4.1 Study Design

There are two types of nodes in the reviewed ERGM networks: individuals and groups/organizations. Collaboration networks modeled upon individuals are at the micro-level. This review found that the micro-level network studies are primarily about physician collaborations (7 studies) for knowledge sharing or advice seeking ((Zappa and Mariani 2011; Zappa 2011; Mascia et al. 2018; Bunger et al. 2018), patient care (e.g., (Uddin et al. 2013), (Kesternich and Rank 2022)), or organization membership (Fattore and Salvatore 2010). The other micro-level col-
Collaborative network is researcher scientific collaborations (Wang et al. 2013; Okamoto 2015; Harris et al. 2015; Azondekon 2018a; Ho et al. 2021). Collaboration networks modeled upon organizations are at the macro-level, involving different types of institutes, such as public health organizations (Luke et al. 2010, 2013) (Bevc et al. 2015a, b), hospitals/healthcare organizations (Lomi and Pallotti 2012; Caimo et al. 2017; Prusaczyk et al. 2019; Bohnett et al. 2022), RD organizations in the industry (Broekel and Bednarz 2019), government public health departments (Harris 2013; YousefiNooraie et al. 2014), and health policy interest groups (Heaney 2014). In addition, we observed only two studies incorporating a multilevel network analysis approach, which characterized micro, macro, and meso-level networks and their interactions. The meso-level network models the network between individuals and organizations. Since networks rarely exist in isolation, the multilevel approach provides a complete picture of the mechanisms vital to the overall network structure, which shall be investigated in future research. One third of the included studies used secondary data from previous studies, projects, or existing datasets from databases, among which three studies supplemented the secondary data with additional data collected by interviews (Harris 2013; Luke et al. 2013) or interviews with a survey (Lomi and Pallotti 2012). Among all the data sources, The PARTNER database (Program to Analyze, Record, and Track Networks to Enhance Relationships) was adopted in three studies (Bevc et al. 2015a, b; Bohnett et al. 2022). As one of the most extensive network databases, PARTNER captures the connection of health organizations. It has been primarily used by practitioners engaged in inter-organizational collaborations in their communities for outcome evaluation. For cross-sectional studies, surveys were used in 10 studies to collect network related data. In addition, medical or health insurance claim data were used in two studies (Uddin et al. 2013; Kesternich and Rank 2022). Interestingly, two studies used citation data from literature database searching or researchers curriculum vitae (Marchand et al. 2018; Azondekon 2018a). The co-authorship from the citations was helpful for data modeling upon homophily, cluster assignment, and temporal dependencies. Twenty two of 30 studies disclosed the software they adopted for network analysis and modeling. The ERGM package in the R statnet is the most frequently used software
for ERGM fitting, especially for the studies which adopted the standard ERGM. Other R packages include R tergm, R btergm, R igraph, and R mixer. The R libraries for ERGMs like statnet are crucial for network modeling. However, we have not identified a similar library in Python, a widely used data analytics platform. Researchers are calling for an ERGM library in Python so that users who are more familiar with this environment can also conduct social network analysis (Ghafouri and Khasteh 2020).

1.4.2 Networks

Networks modeled upon individuals (i.e., researchers or physicians) are always nested within a more extensive organizational network, such as academic institutions or hospitals. However, we found only two studies utilized this hierarchical approach by applying Mergm (Wang et al. 2013; McGlashan et al. 2019). All other studies modeled upon individuals at the micro-level or organizations at the macro-level without exploring how the links between individuals and organizations contribute to the overall network structure. Similarly, only three studies modeled a network with time as a parameter, while most studies investigated collaboration networks at a single or fixed time point. Since social networks are dynamic, constantly changing, and evolving over time, future studies are warranted in applying ERGMs to multilevel (micro-, macro-, and meso-levels) and longitudinal collaboration networks.

In an ERGM, inherent network dependencies are modeled via endogenous variables. Geometrically weighted edgewise shared partner (GWESP) triadic closures and alternating k-stars were used in multiple studies (Harris 2013; Uddin et al. 2013; Caimo et al. 2017). Despite leaving out the meso-level network, several researchers acknowledged this hierarchical structure by including organizational membership as an exogenous variable (Shearer et al. 2014; Okamoto 2015; Bevc et al. 2015b).

To select a high quality ERGM, all included studies commonly adopted the Akaike/Bayesian information criterion (AIC/BIC). After the ERGM was fitted, model diagnostics for that fitted ERGM were performed via simulation. Key network statistics summarizing the observed net-
works characteristics, such as the number of triangles and the number of edgewise shared partners, were recorded. The simulated network statistics values did not differ significantly from the observed network in all studies reviewed. However, only four studies evaluated whether ERGM was appropriate (Fattore and Salvatore 2010; Uddin et al. 2013; Heaney 2014; Okamoto 2015). They used a multiple regression model to achieve this goal with multiple regression quadratic assignment procedure as a popular choice.

The increased computational power of computers enables researchers to fit complex and computationally intensive models such as ERGMs. The leading utility of ERGMs lies in the specification of endogenous variables. Proper identification of network dependencies via the inclusion of endogenous variables can shed light on the characteristics and nature of a network, helping researchers answer complex questions regarding connection formation, change, and prediction. However, it is hard to discern a priori which endogenous variables to include as there are hundreds of possibilities used to model the same network dependence structure, such as transitivity and centrality. This challenge may explain the under-utilization of ERGMs in practice. Although we identified the most popular endogenous variables from the included studies were geometrically weighted edgewise shared partners and k-stars, its unknown why such endogenous variables were selected. Since the correct specification of endogenous variables is vital to GOF, for ERGMs to reach their full potential, a systematic and accessible guide to uncovering network dependencies via endogenous variables is needed. Further, for a network consisting of two types of nodes, organizational and individual level, researchers should consider using an Mergm. Additionally, a Tergm or Stergm is better for modeling dynamic networks that change significantly over time. Finally, researchers who wish to leverage domain specific expertise should consider a Bergm.

1.4.3 Limitations

Although we followed an SR protocol by systematically collecting, extracting, and summarizing research evidence pertinent to ERGM applications to collaborations in biomedical and
health sciences, we may miss studies that are not indexed in PubMed or Scopus. In addition, studies published in a language other than English were not included. Further, due to the heterogeneity of the studies, we did not conduct a quality assessment in this review. Last, we only focused on common network characteristics (nodes, density, centrality) for network data extraction and did not compare or discuss ERGM statistical formulas that researchers applied to their studies.

ERGMs are powerful tools in social network analysis, adept at capturing intricate network interactions. However, they are not without challenges. Their inherent complexity often makes parameter estimation and interpretation daunting, especially in expansive networks (Stivala et al. 2020; Albery et al. 2021). A significant concern of ERGMs is the proper selection of endogenous variables, which, if not accurately chosen, can lead to model misfits (An 2016). Moreover, the intractable normalizing constant in ERGMs adds to computational difficulties (Huang and Butts 2024). Another prevalent issue is model degeneracy, where ERGMs might yield unrepresentative or non-convergent results (Kei et al. 2023). While ERGMs are promising for network studies, researchers must tackle these challenges for their broader and more effective application.
Figure 1.2: Technical Drawing of ERGM Application Challenges in Collaboration Networks in Biomedical and Health Sciences
1.5 Conclusion

This review demonstrates the research landscape of how researchers have applied ERGMs to collaboration network analysis in biomedical and health sciences. The bibliometric, study design and network modeling characteristics of included studies provide insights and guidance for future research. Professional collaboration networks are essential to solving complex scientific problems and optimizing healthcare delivery and health policy. Our systematic synthesis contributes to the design, development, and utilization of ERGMs in studies examining complex relationships and system structures native to professional collaboration networks. Further, we advocate novel methodologies tailored for selecting endogenous variables and fitting ERGMs, addressing the pronounced challenges that currently impede the optimal use of ERGMs as shown in Figure 1.2.
CHAPTER 2: ENDOGENOUS VARIABLE SELECTION

This study introduces a novel methodology for endogenous variable selection in Exponential Random Graph Models (ERGMs) to enhance the analysis of social networks across various scientific disciplines. Addressing critical challenges such as ERGM degeneracy and computational complexity, our method integrates a systematic step-wise feature selection process. This approach effectively manages the intractable normalizing constants characteristic of ERGMs, ensuring the generation of accurate and non-degenerate network models. An empirical application to ten real life binary networks demonstrates the method’s effectiveness in accommodating network dependencies and providing meaningful insights into complex network interactions. Particularly notable is the adaptability of this methodology to both directed and undirected networks, overcoming the limitations of traditional ERGMs in capturing realistic network structures. The findings contribute significantly to network analysis, offering a robust framework for modeling and interpreting social networks and laying a foundation for future advancements in statistical network analysis techniques.

2.1 Introduction

Statistical analysis of social networks plays a pivotal role in various scientific disciplines, offering valuable insights into complex network interactions. Accurate modeling is particularly crucial when working with moderately sized networks, typically comprising a few thousand nodes, as it enables the explanation, analysis, replication, and prediction of network phenomena observed in nature. In the field of health sciences, social network analysis contributes to reducing health disparities Okamoto et al. (2015) and fostering collaboration and research efficiency.
leading to scientific innovations and discoveries Bennett et al. (2018). By uncovering patterns in collaboration networks, network analysis facilitates the prediction of future connections among individuals or organizations, which holds significant value for multiple stakeholders including health policy researchers, administrators, and research sponsors Yu et al. (2020) Provan et al. (2005) Luke et al. (2013).

The advancement in computational power of personal computers in the 21st century has empowered researchers to conduct sophisticated statistical modeling without relying on supercomputers Nordhaus (2001). One powerful technique widely used in social network research is Exponential Random Graph Models (ERGMs). ERGMs are particularly adept at capturing network dependencies by incorporating endogenous variables. However, a challenge arises when the chosen endogenous variables do not accurately capture the observed network structures, leading to what is known as ERGM degeneracy Li (2015); a state where networks become unrealistic and un-interpretable Krivitsky (2017) Li (2015) Bang-Jensen and Gutin (2018).

Addressing the weakness of ERGMs presents multiple challenges that require careful consideration. First, the dependency among network observations, similar to that in longitudinal studies, invalidates models that assume independence Kolaczyk and Csárdi (2014). Secondly, accurately modeling this dependency is crucial but challenging. While stochastic block models treat this dependency as a nuisance parameter, ERGMs aim to explicitly model and quantify it through endogenous variables. The complexity lies in selecting appropriate endogenous variables, given the vast array of choices and the risk of degeneracy through inappropriate selections. Researchers have reported at least five distinct types of ERGMs in their scholarly work, including the standard ERGM Uddin et al. (2013) Zappa and Mariani (2011), Bayesian ERGM Caimo et al. (2017), Temporal ERGM Azondekon (2018), Separable Temporal ERGM Ho et al. (2021) Broekel and Bednarz (2018), and Multi-level ERGM Wang et al. (2013) McGlashan et al. (2019). Researchers face the daunting task of choosing from thousands of potential endogenous variables Hunter et al. (2008), a process lacking systematic guidance or tools Yu et al. (2024).
Therefore, this study proposes and tests a novel methodology for endogenous variable selection in ERGMs, targeting the critical challenges in the field. Our approach encompasses key aspects of variable selection, degeneracy screening, and model fitting, providing a comprehensive solution to enhance the effectiveness and reliability of ERGM modeling, particularly in collaboration networks. We conduct empirical testing and rigorous analysis to contribute to advancing statistical techniques, aiming to facilitate more accurate and meaningful interpretations of network phenomena in various scientific disciplines Yu et al. (2024).

The paper is structured as follows: Section 2 offers a mathematical definition of an ERGM, setting the foundational understanding of this class of models. Section 3 delves into the details of the endogenous variable selection procedure, which includes establishing the initial set of variables, a novel step-wise variable selection process, and an innovative degeneracy screening method based on edge count. Section 4 applies our proposed algorithms to ten real-life binary networks, presenting numerical results that include the selection of endogenous variables, the count of potential pairwise ERGMs, average counts of edges, 2-stars, and triangles, and the efficacy of our degeneracy screening approach. Section 5 discusses the significance of our methodology and testing results, potential future research directions, and limitations, underscoring the implications of our findings for advancing the field of network analysis Yu et al. (2024).

2.2 Definition of ERGMs

A network or graph $G$ consists of nodes and edges denoted by $G = (V, E)$ respectively. The nodes are assumed to be finite with $V = \{1, \ldots, N\}$. The edges represent ties between two different nodes $i, j$. Modeling networks is centered around the edges $E$ of a graph. The outcome of interest $Y_{i,j}$ is defined for two separate nodes $i \in V$ and $j \in V$. Depending on the type of network, this outcome $Y_{i,j}$ can take on binary, discrete or real valued numbers. For example, a binary outcome where $Y_{i,j} = 1$ indicates an edge between nodes $i$ and $j$ while $Y_{i,j} = 0$ indicates no edge. Additionally, nodes $i \in V$ can possess a collection of attributes situated in Euclidean space.
Statistical modeling of a network involves defining a probability distribution over graph $G$. This model space comprises a set of these probability distributions, each indexed by a parameter space $\Theta$. The selected probability distribution will determine the complexity and details of the network model. Inspired by generalized linear models, exponential random graphs model the probability of a tie formation. Inspired by generalized linear models, exponential random graphs model the probability of a tie formation $Y_{i,j} = 1$ given the nodal attributes $X$. For a binary outcome, ERGMs are akin to logistic regression in network data analysis Handcock et al. (2015). Analogous relationships exist between ERGMs with discrete and continuous-valued ties and their generalized linear model counterparts, such as Poisson regression and Gamma regression, respectively. The formulation for an ERGM with a binary outcome is given below.

$$P_\theta(Y_{i,j} = y_{i,j}|\theta, x) = \psi(\theta_1, \theta_2) \exp\{\theta_1^T s(y) + \theta_2^T g(y, x)\}$$

(2.1)

In this model, the vector of regression coefficients is partitioned into two sub-vectors $\theta_1$ and $\theta_2$. The nodal attributes are represented by $x$. The computationally intensive normalizing constant is $\psi(\theta_1, \theta_2)$, and a vector of exogenous variables, $g(y, x)$, with their associated regression coefficients represented by $\theta_1$ and $\theta_2$ respectively.

2.3 Endogenous Variable Selection for ERGMs

In this section, we introduce a novel step-wise feature selection methodology for ERGMs, which adapts the classical statistical method of forward selection technique Effroymson (1960) to address the challenges posed by the complexity of selecting endogenous variables for ERGMs. Our approach starts by focusing on ERGMs with two predictors and employs the AIC for model assessment. This methodological framework guides the initial selection of endogenous variables, their subsequent evaluation, and categorization based on AIC impacts, and concludes with advanced degeneracy screening techniques. Sections 3.1-3.4 detail each step of this comprehensive process, ensuring a robust and accurate selection of endogenous variables for ERGM analysis.
2.3.1 Obtaining an Initial Set of Endogenous Variable

The selection of endogenous variables for an ERGM presents a significant challenge due to the vast array of available options, including hundreds of pre-defined or user-customized variables. These variables are integral for modeling different network structures Hunter et al. (2013). In this study, we start with thirteen commonly used endogenous variables, as identified from the ERGM package in R Handcock et al. (2015). These variables, selected for their relevance to binary un-directed networks include kstar, degree-wise shared partners (dsp), non-edgewise shared partners (nsp), edgewise shared partners (esp), triangle, isolates, sociality, degree cross product, degree popularity, geometrically weighted edgewise shared partners (gwesp), geometrically weighted non-edgewise shared partners (gwnsp), geometrically weighted dyad-wise shared partners (gwdsp) and geometrically weighted degree.

We employ a systematic method to select endogenous variables by establishing an informed upper bound, thus providing a structured way to refine choices, particularly for variables requiring a natural number input. For example, the dsp variable is a network statistic equal to the number of dyads with k shared partners Hunter et al. (2008), and demands a selection of an input \( k \in \mathbb{N} \). Given the vast range of possibilities, it is necessary to set an upper limit for k. Similarly, variables such as kstar, esp, and nsp also require a natural number to be well-defined Handcock et al. (2015). Different natural numbers k lead to distinct network structures, as illustrated by the difference between kstar(2) kstar(3) (Figure 1).

Addressing the impracticality of considering an infinite yet countable number of endogenous variables, our approach involves sequentially fitting uni-variate ERGMs. Starting from \( k = 1 \) and progressing until we reach a specific cutoff point, \( k = N_k \). This cutoff, \( N_k \) is determined when we achieve three or more consecutive parameter estimates reaching infinity. Beyond \( N_k \), further consideration of endogenous variables becomes impractical, as they lead to coefficient estimates of infinity. The identification of \( N_k \) plays a pivotal role in dictating the size of the initial set of endogenous variables denoted by \( S_N \). Applying this method to an observed network results in a
list of N candidate endogenous variables, $S_N$, with the sets size determined by the upper bounds of variables like dsp, esp, nsp and kstar.

Figure 2.1: Illustration of the dyadwise shared partner endogenous variable with $k = 1$, $k = 2$ and $k = 3$ respectively. Source: van der Pol (2019).

### 2.3.2 Stochastic Forward Selection

This section delves into the Stochastic Forward Selection process, the core of our proposed methodology. We start with a basic ERGM featuring only an edge term, akin to the intercept in a linear model. This baseline model, also known as a Bernoulli Random Graph assumes that the probability of a tie formation between two nodes follows a Bernoulli distribution independent of other ties within the network Kolaczyk and Csárdi (2014). This assumption does not account for observation dependence, an important factor in network data.

The process involves the following steps, detailed in Algorithm 1 (Bounding the input Parameter k) and Algorithm 2 (Stochastic Forward Selection for Endogenous Variables).

Building upon the Algorithm 1, we categorize the endogenous variables variable $s_i(y) \in S_N$ based on their observed relative AIC changes during the stochastic forward selection process. This categorization is essential in discerning variables that significantly enhance the model and those that may lead to degenerate models or yield ambiguous results.

- **Category D1**: Endogenous variables that consistently lower the AIC compared to the null ERGM, indicating a positive contribution and suggesting their inclusion in the model.
- **Category D2**: Variables that lead to degenerate ERGMs, marked by a very negative relative AIC change or a consistently negative mean relative AIC change, suggesting their exclusion.
Algorithm 1 Bounding the Input Parameter $k$

1. **Endogenous Variable Requirement**: Start with an endogenous variable $s_i(y, j)$ with an input parameter $j \in \mathbb{N}$

2. **Sequential Fitting**: Fit uni-variate ERGMs with an edge term and endogenous variable $s_i(y, j)$ beginning with $j = 2$.

3. **Upper Bound Determination**: Identify the upper bound number $N_k$ when the parameter estimates for $s_i(y, N_{k+2})$, $s_i(y, N_{k+1})$, $s_i(y, N_k)$ all yield negative infinity.

4. **Variable Iteration**: Repeat 1-3 for the following endogenous variables: dsp, esp, nsp and kstar, marking their respective upper bounds by $N_1$, $N_2$, $N_3$ and $N_4$.

5. **Final Set Formation**: Obtain a final set of endogenous variables $S_N$.

- Category D3: Variables with ambiguous impact on the AIC, possibly due to poor initial parameter estimates or lack of predictive power. Here, the 10th percentile of the relative AIC change, denoted as $\hat{b}_i$, is crucial. A positive $\hat{b}_i$ indicates potential predictive power, while a negative value suggests exclusion.

Therefore, the null model serves as a baseline for comparing candidate ERGMs. We systematically fit uni-variate ERGMs for each element in $s_i(y) \in S_N$ for $i \in \{1, \ldots, N\}$ as shown below.

$$P_\theta(Y_{ij} = y|\theta) = \psi(\theta_0, \theta_i)\exp\{\theta_0 s_0(y) + \theta_i s_i(y)\} \quad i \in \{1, \ldots, N\} \quad (2.2)$$

In this formula, $Y_{ij}$ denotes the binary outcome of tie formation between two nodes, and $y$ is the observed network at a given state. The models normalizing constant is denoted by $\psi$. The edge term and its associated regression coefficient are represented by $s_0(y)$ and $\theta_0$, respectively. The endogenous variable under consideration is $s_i(y)$, with its regression coefficient $\theta_i \in \mathbb{R}$. The null ERGMs AIC is denoted by $AIC_0$, and the AIC for each uni-variate ERGM is by $AIC_i$ for $i \in \{1, \ldots, N\}$. The selection of the endogenous variable, $s_i(y)$ is contingent upon the calculated relative AIC change as below.
\[ b_i = \frac{AIC_0 - AIC_i}{AIC_0} \quad i \in \{1, \ldots, N\} \] (2.3)

**Algorithm 2** Stochastic Forward Selection for Endogenous Variables

1. **Initial Set Formation:** Start with a set \( S_N = \{s_1(y), \ldots, s_N(y)\} \) consisting of \( N \) candidate endogenous variables.

2. **Null Model Fitting:** Fit a null model ERGM with only an edge term, denoting the AIC value by \( AIC_0 \).

3. **Variable Assessment:** For \( i \in \{1, \ldots, N\} \):
   - Sequentially fit uni-variate ERGMs with one endogenous variable at a time; \( P_\theta(Y_{i,j} = 1) = \psi(\theta_0, \theta_i)exp(\theta_0s_{0}(y) + \theta_is_{i}(y)) \).
   - Record the estimate of the AIC for each uni-variate ERGM by \( AIC_i \) and compute the relative AIC change \( b_i \).
   - Refit the uni-variate ERGM \( M \) times and compute the 10th percentile of the relative AIC change denoted by \( \hat{b}_i \).

4. **Variable Exclusion:** If \( \hat{b}_i \leq 0 \) then remove \( s_i(y) \) from the set \( S_N \).

This forward selection strategy effectively categorizes endogenous variables based on their influence on the AIC. Variables that consistently elevate the AIC are considered less informative and excluded. The 10th percentile of the relative AIC change, \( \hat{b}_i \), plays a crucial role in this process. If \( \hat{b}_i \) is positive, it suggests that this endogenous variable can predict network structure; otherwise, it should not be included.

**2.3.3 Degeneracy Screening**

After categorizing and selecting endogenous variables for inclusion, we now address a critical aspect of model refinement in ERGMs: degenerate screening. This step is vital to ensure that our model remains robust and accurate, free from the distortions of multi-collinearity often observed in ERGMs with multiple endogenous variables Li (2015).

Degenerate networks, characterized by unrealistic network structures, emerge as a significant challenge in ERGMs when multi-collinearity occurs due to the inclusion of numerous endoge-
nous variables. To counteract this, our approach analyzes network motifs — small, statistically significant graph patterns typically comprising up to 6 nodes Masoudi-Nejad et al. (2012). These motifs serve as a barometer for assessing the realism and practicality of the networks generated by our ERGM.

To discard degenerate ERGMs, model selection will be based on network motif counts. Network motifs which are small graphs, typically not exceeding 6 nodes, that are statistically significant patterns found in a larger network Masoudi-Nejad et al. (2012).

To recommend non-degenerate ERGMs we count the number of edges for an observed network and compare that count to the average number of edges a candidate ERGM produces. We then compute the relative error of these counts in the last step.

Algorithm 3 Degeneracy Screening Algorithm

1. **Edge Count Assessment**: For an observed network, compute the observed edge counts denoted by $H_O$.

2. **Model Comparison**: For a proposed ERGM $M \in \mathcal{M}$, compute the average edge count denoted by $H_M$.

3. **Discrepancy Calculation**: Compute the difference $|H_M - H_O|$.

4. **Model Exclusion**: If $\frac{|H_M - H_O|}{H_O} \geq 4$ then discard $M$ from the set of possible models.

The last step of this algorithm is critical: if a candidate ERGMs average edge count significantly deviates from the observed count (either overestimates or underestimates), it is deemed degenerate and thus excluded from our set of potential models. This step ensures the models we select not only statistically represent the observed network structure but also adhere to realistic network formations.

By integrating degeneracy screening into our methodology, we enhance the model’s reliability, ensuring that it reflects the true nature of the network data and remains free from the distortions of multicollinearity. This process, combined with our earlier steps of variable selection
and categorization, culminates in an ERGM that is both robust and representative of the complex
dynamics inherent in network structures.

2.3.3.1 Network Motifs used for Model Selection

The use of network motifs for model selection is an extension of the degeneracy screening,
which enables us to delve deeper into the structural analysis of the networks.

While the edge network motif played a central role in screening for degenerate models, our
model selection process employs additional network motifs to refine the selection criteria further.
Specifically, we focus on the counts of 2-stars and triangles alongside the edge counts. The ratio-
nale behind incorporating these motifs is to compare the distribution of these specific patterns in
the candidate ERGMs against their occurrence in the observed network.

The selection of models is based on the alignment of the mean number of edges, 2-stars, and
triangles in the ERGMs with their corresponding counts in the observed network. The closer
these averages are to the observed counts, the more representative and accurate the model is
considered. This approach ensures that the selected ERGM not only avoids degeneracy but also
closely mirrors the actual network structure in terms of these key motifs.

2.4 Numerical Studies

This section presents the application of our algorithms to ten real-life networks Handcock
et al. (2015) Caimo and Friel (2012), varying in complexity and size (i.e., The nodes vary from
16 to 418). These networks, all un-directed with binary outcome values, are categorized into three
main categories based on their structure and size, allowing us to comprehensively evaluate the
potential and limitations of our method across varied network types. (1) Small Networks: This
category includes networks with up to 20 nodes and a maximum of 30 edges. Representative
networks in this group are the Florentine marriage, Florentine business and Molecule networks
Padgett (2011). (2) Moderately Complex Networks: Networks in this category are slightly larger
and more complex, with node counts ranging from 20 to 80 and edge counts between 50 and 200.
Examples include the Lazega lawyer network, Kapferer tailor shop networks 1 and 2, and the Zach karate networks Kapferer (1972) Lazega (2001). (3) Highly Complex Networks: the final category encompasses the largest and most complicated network with nodes ranging from 80 to 418 and edges counts from 200 to 556 Resnick et al. (1997). A notable network in this category is the challenging Ecoli network Shen-Orr et al. (2002), known for its complexity.

2.4.1 Stochastic Forward Selection:

In applying Algorithm 1, we obtained an initial set of endogenous variables, establishing a model space for each of the 10 networks (Table 2.1). The model space consists of ERGMs with either one or two distinct endogenous variables alongside the edge term. Aiming to select endogenous variables that accurately reflect the observed network structure, we applied Algorithm 2, leading to a significant reduction in the model space for all networks. Notably, for networks with fewer than 30 nodes, such as the Florentine marriage network, the algorithm consistently proposed fitting ERGMs. An example of such an ERGM, utilizing the kstar(2) and kstar(3) terms is visualized below. This visualization represents three networks simulated from this ERGM, showcasing Algorithm 2s effectiveness in capturing network dynamics.

Transitioning from the broader application of our proposed stochastic forward selection approach above, the following is a specific example highlighting the importance of the relative AIC percentile in our algorithm. We applied algorithm 1 to a transcription regulation network for Ecoli Hummel et al. (2010) Salgado et al. (2001) Shen-Orr et al. (2002), a case that exemplifies the nuances of variable selection in Category D3. In this instance, the relative AIC change between AIC change between the null ERGM and the uni-variate ERGM was recorded M = 90 times. The focal endogenous variable was the degree cross product. The results, depicted in Figure 2.3, show that in 5 out of 90 instances, there was a substantial increase in AIC compared to the null model. Crucially, the 10th percentile of the relative AIC for the degree cross product is positive with a median value of 0.0265. This observation underscores the potential risk of incor-
Figure 2.2: Upper left network denotes the observed Florentine marriage Network. The remaining three networks are draws from the ERGM in equation $P_{\theta}(Y_{i,j} = 1 | X) = \psi(\theta) \exp\{\theta_0 \times \text{edges} + \theta_1 \times \text{kstar}(2) + \theta_2 \times \text{kstar}(3) + \theta_3 g_1(x) + \theta_4 g_2(x) + \theta_5 g_3(x)\}$. $g_1(x)$ denotes the exogenous variable of familial wealth, $g_2(x)$ denotes the number of seats on the civic council and $g_3(x)$ denotes the total number of business and marriage ties.

rectly excluding significant variables based on a single AIC calculation, thus validating the need for a multi-faceted evaluation approach as embodied in our methodology.

2.4.2 Degeneracy Screening and Model Selection Results

Our focus of model selection was primarily on the counts of edges, 2-stars and triangles as determined by Algorithm 3. This approach allowed us to systematically identify and exclude degenerate ERGMs, which are characterized by edge counts that significantly diverge from those observed in the actual networks.

For this reduced model space, the average number of edges, 2-stars and triangles generally aligned with the observed counts. However, the number of edges tended to be overestimated across all networks, a bias often inherent in exponential family models Efron (1978). To address
this discrepancy, it is advisable to allow ERGMs to include more than 3 endogenous variables, thereby enhancing model accuracy. The variations in the counts of 2-stars and triangles compared to the observed networks further illustrate the complex dynamics within these networks and the necessity of a flexible and robust modeling approach.

Figure 2.3: Frequency of the relative AIC change for the uni-variate ERGM with degree cross product as the main predictor. The histogram on the left is an example of relative AIC fluctuation due to poor initial starting points. The histogram on the right exhibits relative AIC fluctuation that mimics random noise.

<table>
<thead>
<tr>
<th>Network</th>
<th># of Nodes</th>
<th># of Observed Edges</th>
<th># of 2-stars</th>
<th># of Triangles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazega</td>
<td>36</td>
<td>115</td>
<td>1852</td>
<td>240</td>
</tr>
<tr>
<td>Kapferer</td>
<td>39</td>
<td>158</td>
<td>3132</td>
<td>402</td>
</tr>
<tr>
<td>Kapferer2</td>
<td>43</td>
<td>190</td>
<td>4074</td>
<td>504</td>
</tr>
<tr>
<td>Zach</td>
<td>34</td>
<td>78</td>
<td>1056</td>
<td>90</td>
</tr>
<tr>
<td>Wind Surfers*</td>
<td>95</td>
<td>556</td>
<td>21966</td>
<td>2890</td>
</tr>
<tr>
<td>Molecule</td>
<td>20</td>
<td>28</td>
<td>120</td>
<td>12</td>
</tr>
<tr>
<td>Faux Mesa High School</td>
<td>205</td>
<td>203</td>
<td>1318</td>
<td>124</td>
</tr>
<tr>
<td>Ecoli*</td>
<td>418</td>
<td>519</td>
<td>10580</td>
<td>84</td>
</tr>
<tr>
<td>Florentine Marriage</td>
<td>16</td>
<td>20</td>
<td>94</td>
<td>6</td>
</tr>
<tr>
<td>Florentine Business</td>
<td>16</td>
<td>15</td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.1: Average Network Motif Counts

For this reduced model space, the average number of edges, 2-stars and triangles tend towards the observed count. The number of edges were overestimated for all networks. This bias is inherent to exponential family models Efron (1978) and can be corrected if more endogenous variables are used. On the other hand, the average number of 2-stars and triangles were sometimes above the observed count and below the observed count for different networks. A remedy
<table>
<thead>
<tr>
<th>Network:</th>
<th># of Models</th>
<th>Average # of Edges:</th>
<th>Average # of 2-stars</th>
<th>Average # of Triangles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazega</td>
<td>85</td>
<td>224.03</td>
<td>1924</td>
<td>174.2</td>
</tr>
<tr>
<td>Kapferer</td>
<td>128</td>
<td>288.8</td>
<td>2387</td>
<td>207.14</td>
</tr>
<tr>
<td>Kapferer2</td>
<td>150</td>
<td>293.2</td>
<td>2792</td>
<td>222</td>
</tr>
<tr>
<td>Zach</td>
<td>16</td>
<td>117.6</td>
<td>1061</td>
<td>81.3</td>
</tr>
<tr>
<td>Wind Surfers</td>
<td>324</td>
<td>1043</td>
<td>12741</td>
<td>770</td>
</tr>
<tr>
<td>Molecule</td>
<td>69</td>
<td>29.68</td>
<td>123.4</td>
<td>3.98</td>
</tr>
<tr>
<td>Faux Mesa High School</td>
<td>39</td>
<td>344.4</td>
<td>754.3</td>
<td>96.90</td>
</tr>
<tr>
<td>Ecoli</td>
<td>104</td>
<td>869</td>
<td>2057</td>
<td>9.54</td>
</tr>
<tr>
<td>Florentine Marriage</td>
<td>1</td>
<td>24.97</td>
<td>147.5</td>
<td>12.59</td>
</tr>
<tr>
<td>Florentine Business</td>
<td>2</td>
<td>22.80</td>
<td>59.13</td>
<td>23.93</td>
</tr>
</tbody>
</table>

**Table 2.2:** Average Network Motif Counts for the ERGMs

to this discrepancy is allowing the candidate ERGM to possess more than 3 endogenous variables.
2.5 Discussion

The results of the numerical studies have affirmed the effectiveness of our proposed methodology in generating well-fitting and non-degenerate ERGMs. Our approach has been successfully applied to a diverse range of networks, from small-scale networks with fewer than 20 nodes to complex networks like the Ecoli network, encompassing both directed and undirected structures. This versatility in application demonstrates the robustness of this approach.

The novelty of our method lies in its adaptive use of a step-wise feature selection process tailored specifically for ERGMs. This approach, which meticulously evaluates the impact of each endogenous variable using the AIC, marks a significant departure from traditional methods. It addressed the inherent complexity of ERGMs, especially the challenges posed by their intractable normalizing constants, and offers a comprehensive framework that enhances the models accuracy and interpretability.

Our methodological framework opens new avenues in network analysis, particularly in handling directed networks, which traditionally pose a challenge due to their complex structures. The ability to incorporate directed endogenous variables, though resulting in an expanded model space, paves the way for more nuanced analyses of directed networks. This capability is crucial for future studies to understand the directional dynamics within networks, offering potential for groundbreaking discoveries in fields such as social network analysis, epidemiology, and beyond.

One notable limitation of our approach is the inflated model space that emerges when incorporating directed endogenous variables. This expansion complicates the model selection process, particularly when models include more than three variables. Addressing this limitation necessitates the development of sophisticated model selection procedures capable of navigating this increased complexity.
2.6 Conclusion

In conclusion, the method we introduced in this study represents a significant advancement in network analysis. By providing a robust and adaptable framework for ERGMs, our methodology not only ensures the generation of accurate and non-degenerate models but also enhances the potential for their application in more complex network types. While the challenge of an expanded model space presents an opportunity for further research, it also highlights the fertile ground for further advancements in the analytical capabilities and applications of ERGMs. As we build on this work, the potential for new insights and understandings of complex network structures becomes increasingly attainable.
CHAPTER 3: MODEL SELECTION

3.1 Introduction

Correct modelling of networks is important for various scientific fields providing insights into complex interactions and phenomena. Multiple stakeholders such as administrators, research sponsors and health policy researchers may use a social network model to predict future ties among individuals or organizations (Yu et al., 2020). The accuracy of such insights and predictions is contingent on the choice of model for a given observed network. A statistical model for a network begins with defining a probability distribution over graphs which consist of edges and nodes. For example a social network can be regarded as a graph since the nodes can represent individual actors or organizations and the edges can represent collaboration.

The main difficulty in modelling networks is the dependency of the edges to one another. This dependency is vital for a given network’s structure and properly accounting for it is imperative. There are two groups of statistical models for networks. The first group considers this dependence as a nuisance parameter and attempts to project it away. A well known example of statistical models for this group are Stochastic Block Models (SBM) (Abbe, 2017). The second group considers this dependence to be a phenomena that should be explicitly modeled and quantified. Exponential Random Graph Models (ERGMs) model the dependence of observations via endogenous variables (Hunter et al., 2008).

The inclusion of endogenous variables in the ERGM family of statistical network is both a major strength and weakness. Incorporating endogenous variables that do not accurately capture the dependency structure present in an observed network leads to a phenomena known as degeneracy (Krivitsky and Handcock, 2014). This phenomenon is characterized by the proposal of very
dense or sparse networks in which every node is connected to another node or nearly all nodes are isolated. Given a list of endogenous variables that are reflective of an observed network’s structure, it is possible to induce degeneracy by incorporating different combinations of endogenous variables. The degeneracy in this case is due to col-linearity among the variables. Additionally, directed networks require unique endogenous variables to accommodate for the added structure due to directional ties. This results in an inflated model space for candidate ERGMs.

The main contribution of the paper is the use of edge Forman Ricci Curvature (FRC) as a model screening tool as opposed community detection. Network analogues of Ricci Curvature such as edge Forman Ricci Curvature have been used extensively for community detection in the literature (Sia et al., 2019) (Ni et al., 2019). The computational efficiency of edge Forman Ricci Curvature allows us to consider tens of thousands of candidate ERGMs for a given observed network. By computing the edge FRC for an observed network, one obtains a projection of the observed network topology. This characterization manifests itself in the form of a cumulative density function of edge FRC values. Then for a given ERGM, we draw networks and compute the cumulative density function edge FRC for each draw. The Cumulative Distribution Functions (CDFs) are then compared to the edge FRC CDF of the observed network via the Cramer von Mises (CvM) distance. If the average CVM distance obtained from the candidate ERGM is large then we discard the candidate ERGM from our model space.

The model selection procedure for ERGMs proposed in this paper is as follows. First, we use Forman-Ricci Curvature to obtain a quantification of complex network structures. Second, two node and three node network motifs counts such as the number of edges, triangles and isolated nodes are used to further characterize network structures not captured via FRC. After these two procedures, a candidate ERGM’s average network motif counts are compared with the observed network’s. Finally, a scoring system based on the network motif counts is implemented to rank different candidate ERGMs.
3.2 Mathematical Background

A network can be represented as a graph \( (V, E) \). The nodes represent the actors in a given network, which can be individuals or organizations. Denote the set of nodes by \( V \). An edge is a tie between two distinct nodes \( i \) and \( j \) and the set of all edges is denoted by \( E \). For a graph of order \( n \in \mathbb{N} \), all the information of which two nodes are connected or not is can all be found in an adjacency matrix. As such, modeling networks is centered around the elements of the adjacency matrix which are the edges \( E \) of a graph. The richness and complexity of the statistical model is determined by the choice of probability distribution over the graph with order \( n \).

3.2.1 ERGM

Recall, that ties in a network are dependent on one another leaving any statistical model that assumes independent observations invalid. For example consider, A, B, and C are three nodes in a network. If A connects with B and B connects with C, the probability that a tie is formed between A and C increases. An ERGM possesses the following probability distribution defined over a graph \( g \in G \) of order \( n \).

\[
P_{\theta}(G) = \frac{1}{\psi(\theta)} \exp \{\theta^T T(G)\} \tag{3.1}
\]

The notorious normalizing constant is denoted by \( \psi(\theta) \). The vector of real regression coefficients is given by \( \theta \). The vector of features is given by \( T(G) \). The vector of features consists of both endogenous variables and exogenous variables. Endogenous variables are attributes of the network. Exogenous variables are attributes of the node. Popular endogenous variables include network statistics such as edges, triangles and k-stars (Hunter et al., 2008).

3.2.2 Forman Ricci Curvature

Ricci curvature is a tool developed in differential geometry to quantify how a local surface differs geometrically from the flat geometric properties of \( \mathbb{R}^n \). To study the local geometric prop-
erties of a given graph, multiple discretizations of Ricci curvature have been proposed (Fesser et al., 2023). We selected the edge based Forman Ricci Curvature (FRC) as it is well defined for both directed and un-directed networks. Indeed, the Forman-Ricci Curvature (FRC) can be regarded as the network analogue of the Bochner-Weitzenbock formula (Samal et al., 2018). The extension is defined for an edge $e_{i,j}$ that connects two nodes $i$ and $j$ together Topping et al. (2021). The formula is given below.

$$F(e_{i,j}) = 4 - d_i - d_j + 3|\#\Delta|$$  \hspace{1cm} (3.2)

Here, $d_i$ and $d_j$ denote the degrees of nodes $i$ and $j$ respectively. The degree of a node is the total number of edges it possesses. Additionally, $|\#\Delta|$ is the number of triangles with base $e_{i,j}$.

For more details on this augmentation please see Topping et al (Topping et al., 2021). For the remainder of the paper, we will refer to Augmented Forman-Ricci Curvature as edge FRC.

### 3.2.3 Illustrative Example

Consider a social network of a karate club consisting of 34 members (Zachary, 1977). The nodes in this case would represent each unique member and the edges represent social interactions that occur outside of the karate club. This social network possesses 78 edges. For each edge there is a respective edge FRC value. These edge FRC values then form a distribution.

### 3.2.4 Two Sample Cramer von Mises Distance

This distance is used to judge the goodness of fit of two cumulative density functions (Darling, 1957). It is regarded as a weighted $L^2$ norm between two distribution functions (Baringhaus and Henze, 2017). Denote the first cumulative density function by $F_1$ and the second cumulative density function by $F_2$. Consider $X_1, \ldots, X_{n_1}$ independent observations used to obtain an empirical cumulative density function of $F_1$. Similarly, consider $Y_1, \ldots, Y_{n_2}$ independent observations used to obtain an empirical cumulative density function of $F_2$. The test the two sample Cramer-von Mises distance is defined below.
\[ \Delta (\hat{F}_1, \hat{F}_2) = \int_{-\infty}^{\infty} (\hat{F}_1(x) - \hat{F}_2(x))^2 d\hat{F}_1(x) \]

\( \hat{F}_1 \) and \( \hat{F}_2 \) denote the empirical cumulative density function of \( F_1 \) and \( F_2 \) respectively. The distance \( \Delta (\hat{F}_1, \hat{F}_2) \) is used to for testing the equivalence of the two cumulative density functions. The hypothesis test is \( H_0 : F_1(x) = F_2(x) \) vs. \( H_A : F_1(x) \neq F_2(x) \). The decision on whether to reject or fail to reject the hypothesis hinges on the value Cramer-von Mises statistic given by

\[ \omega^2_{n_1+n_2} = (n_1 + n_2) \Delta (\hat{F}_1, \hat{F}_2) \]  
(Baringhaus and Henze, 2017). Large values of \( \omega^2_{n_1+n_2} \) indicate that we should reject the null hypothesis and smaller values indicate a failure to reject the null.

### 3.3 Exponential Random Graph Model Selection

Our model space consists of ERGMs that possess one to six endogenous variables. For a given network, apply the stochastic step-wise feature selection method to obtain an initial set of endogenous variables \( S \) (El-Zaatari et al., 2023). The explosive combinatorial complexity requires us to limit the size of this set for ERGMs that possess 3 or more endogenous variables. For illustration, if the size of \( S \) is 70, then the number of possible bi-variate ERGMs is 2415.
For ERGMs that possess 6 endogenous variables, the number of possible models is 131,115,985. This motivates the use of Forman-Ricci curvature in order further refine the initial set of endogenous variables. Then a model selection procedure is implemented for the few thousand remaining candidate ERGMs. Model selection consists of an initial screening that uses Forman-Ricci curvature and a scoring procedure. The scoring procedure is based on the average count of 2 and 3 node network motifs produced by the ERGMs. The average counts are then compared to the observed number from a given network. The remaining ERGMs are ranked based on their respective scores. For the top performing ERGMs, their respective edge variances are compared and then the ERGM with the least variance is chosen.

3.3.1 Initial Screening

The computation of the edge FRC CvM distance is feasible for tens of thousands of candidate ERGMs. However, when the number of candidate ERGMs is in the hundreds of thousands then a reduction in the model space is required. This reduction is obtained by removing endogenous variables from $S$. Begin with a set of endogenous variables $S$ and consider only uni-variate and bi-variate ERGMs. For a given ERGM, draw $T$ networks and compute the FRC values for each network. The computation will generate $T$ cumulative density functions for each network respectively. The $T$ edge FRC cumulative density functions are compared to the observed edge FRC cumulative density function obtained from the observed network. We compute the two sample CvM distance for each of the $T$ cumulative density functions with respect to the observed cumulative density function and take their average. This average CvM distance is associated with a candidate ERGM and is considered a goodness of fit measure. A set of unique endogenous variables are extracted from the ERGMs that possess the lowest CvM distance. By restricting the maximum size of the refined set to be at most 20. We guarantee that the model space will consist of a maximum of 60249 candidate ERGMs possessing 3 to 6 endogenous variables.

The model space $M$ consists of tens of thousands ERGMs. To help narrow down our search for the optimal ERGM, we compute the edge FRC for each model and compare their cumulative
density function to the observed edge FRC via CvM distance. We then take the average CvM
distance for a given ERGM. Candidate ERGMs are then clustered based on their average CvM
distance via K-means. The number of clusters $K$ needed will depend on the ratio of the between
clusters sum of squares over the total sum of squares. The value of $K$ is the minimum number
needed to achieve a ratio of 80%. Below is a detailed description of the initial screening proce-
dure.

**Algorithm 4** Edge FRC for ERGM Initial Screening

0: Input: $\mathcal{M}$, $T$
0: Output: $\mathcal{M}_1 \subset \mathcal{M}$
0: for $M_i \in \mathcal{M}$ do
  0: for $t = 1, \ldots T$ do
    0: Draw $A_t \sim M_i \in \mathcal{M}$
    0: Compute $\text{FRC}(A_t)$
    0: Denote the CDF of $\text{FRC}(A_t)$ by $F_t$
    0: Compute the CvM distance $\Delta(F_t, F_0) \equiv d_t$
  0: end for
0: Compute $\frac{1}{T} \sum_{t=1}^{T} d_t = \bar{d}_{M_i}$.
0: end for
0: Number of clusters $K = \arg\min_L \{L(\text{Kmeans}(\bar{d}_M))|L(\cdot) > 0.8\}$.
0: $\mathcal{M}_1 = \arg\min_{C_k \in C_k} \{\bar{d}_{M_i}|M_i \in C_k\}$

3.3.2 Model Scoring via Network Motif Counts

One drawback of the use of curvature such as Forman Ricci Curvature is that isolated com-
munity structures are not captured. Graph curvature is only useful for capturing the local struc-
ture of well connected communities. Isolated nodes do not possess any local structure and thus
their presence can not be determined via edge or node curvatures (Coupette et al., 2022). Another
drawback of relying solely on FRC values is that they are not sensitive to the overall number of
edges. It is possible for an ERGM to produce an edge FRC cumulative distribution function that
closely resembles the observed network’s edge FRC cumulative distribution function and yet
vastly overestimate the number of edges. This motivates the use of network motif counts as a
model selection tool by scoring different ERGMs based on the following difference of the aver-
age network motif counts generated by a candidate ERGM and the observed network motif count of the network. Un-directed networks will possess 4 total network motifs namely edges, 2-stars, triangles and number of isolated nodes. Directed networks will possess a total of 16 network motifs. The two node network motifs are one directional and bi-directional edges. There are 13 different directional three node network motifs and their configurations are found below. We will also count the number of isolated nodes for directional networks. In order to count the number of isolated nodes for a given network, we used the Leiden algorithm for community detection (Traag et al., 2019). The resolution parameter $\gamma$ was set to be a large negative number.

![Network motifs](image)

**Figure 3.2:** All 13 possible network motifs with 3 nodes and directed ties. Source: Bentley (2017)

### 3.3.2.1 Score Formula

Consider the difference between the average network motif count generated by an ERGM and the observed network motif count. For un-directed networks, denote the edge difference by $\text{edge}_e$, the 2-star difference by $\text{star}_e$, triangle difference by $\text{tri}_e$ and isolated node difference by $\text{iso}_e$. Denote the observed network motifs by $\text{edge}_o$, $\text{star}_o$, $\text{tri}_o$ and $\text{iso}_o$. The formula used to to score candidate ERGMs is found below. To penalize generation of excess edges we multiplied the ratio $\frac{\text{edge}_e}{\text{edge}_o}$ by 4. Note if observed count for a network motif is less than 10, then the metric is not included in the score.

$$\text{SCORE} = (4 \times \frac{\text{edge}_e}{\text{edge}_o}) + \frac{\text{star}_e}{\text{star}_o} + \frac{\text{tri}_e}{\text{tri}_o} + \frac{\text{iso}_e}{\text{iso}_o}$$ (3.4)
For directed networks, denote the bi-directional edge difference by $edge_{1e}$, the one directional edge difference by $edge_{2e}$. For 3 node network motifs, denote the first configuration difference by $config_{1e}$, the second 3 node network motif configuration by difference by $config_{2e}$, ... until reaching $config_{13e}$. Their observed counterparts are denoted by $edge_{1o}$, $edge_{2o}$, $config_{1o}$, $config_{2o}$, ..., $config_{13o}$. Denote the the isolated node difference by $iso_e$ and observed count of isolated nodes by $iso_o$. The formula to rank candidate directed ERGMs is found below. To penalize generation of excess one-directional edges we multiplied the ratio $\frac{edge_{2e}}{edge_{2o}}$ by 5. Note if an observed network motif count is less than 10 then is not included in the score.

\[
SCORE_d = \frac{edge_{1e}}{edge_{1o}} + (5 \times \frac{edge_{2e}}{edge_{2o}}) + \frac{config_{1e}}{config_{1o}} + \cdots + \frac{config_{13e}}{config_{13o}} + \frac{iso_e}{iso_o}
\]  

(3.5)

3.3.2.2 Ranking Algorithm

Based off these scores, we can rank different candidate ERGMs with lower scores indicating better performance. Inevitably, a handful of ERGMs will possess very similar low scores indicating that they all perform reasonably well. In this scenario, the variance for the number of edges will act as a tie breaker and dictate which ERGM gets chosen. The ERGM that possesses the lowest edge variance relative to the other candidate ERGMs will be selected. For directed networks, the variance of the one direction edge term is considered only.
Algorithm 5 Optimal ERGM selection

0: Input: $M_1, T$
0: Output: $\text{ERGM}_{opt}$
0: for $M_i \in \mathcal{M}$ do
0: for $t = 1, \ldots, T$ do
0: Draw $A_t \sim M_i$
0: Compute $\text{SCORE}(M_i)$
0: Compute $\bar{S}(M_i) = \frac{1}{T} \sum_{t=1}^{T} \text{SCORE}(M_i)$
0: Compute $\text{Var}_c(M_i)$
0: end for
0: end for
0: Select number of clusters via $K = \arg \min_L \{(\text{Kmeans}(\bar{S}(M_i))|L(\cdot) > 0.98\}$
0: Perform Kmeans($\bar{S}(M_i)$)
0: $C_1 = \arg \min_{C_i \in C_K} \{\bar{S}(M_i)|M_i \in C_K\}$
0: $\text{ERGM}_{opt} = \arg \min_{M_i \in C_1} \text{Var}_c(M_i) = 0$

3.4 Numerical Studies

We explore the limitations and potential of the above methods by applying them to 9 networks. Five of the networks are un-directed and the remaining four networks are directed. The number of nodes for the networks varies from 20 to 423. The number of edges for the networks varies from 28 to 1197.

<table>
<thead>
<tr>
<th>Network</th>
<th>Type</th>
<th>Number of Nodes</th>
<th>Number of Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule</td>
<td>Un-directed</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>Zach</td>
<td>Un-directed</td>
<td>34</td>
<td>78</td>
</tr>
<tr>
<td>Lazega</td>
<td>Un-directed</td>
<td>36</td>
<td>115</td>
</tr>
<tr>
<td>Kapferer</td>
<td>Un-directed</td>
<td>39</td>
<td>158</td>
</tr>
<tr>
<td>Mesa</td>
<td>Un-directed</td>
<td>205</td>
<td>203</td>
</tr>
<tr>
<td>Bunt</td>
<td>Directed</td>
<td>32</td>
<td>56</td>
</tr>
<tr>
<td>Desert</td>
<td>Directed</td>
<td>107</td>
<td>439</td>
</tr>
<tr>
<td>Dixon</td>
<td>Directed</td>
<td>248</td>
<td>1197</td>
</tr>
<tr>
<td>Ecoli</td>
<td>Directed</td>
<td>423</td>
<td>519</td>
</tr>
</tbody>
</table>

Table 3.1: Descriptive Statistics for 9 Networks.

The observed edge FRC distribution and it’s associated cumulative density function depend on the number of edges the network possess. Smaller networks such as the molecule network will not yield a well defined edge FRC distribution. As such, we recommend skipping the initial screening algorithm if the observed network possesses less than 15 edges and directly apply
the model scoring and ranking algorithm. The edge FRC CvM distances obtained from various ERGMs are all dependent on the observed empirical cumulative density function (ecdf) of the observed network.

![Figure 3.3: The observed distribution of the edge FRC values and their associated empirical cumulative density function for the Zach network, Kapferer network, Bunt network and Ecoli Network.](image)

The Methods of this paper were applied to four different representative networks. The first is the Zach network. It is a social network consisting of 34 karate club members that regularly interacted outside of the club (Zachary, 1977). The network is small and un-directed with only 78 edges in total. The second network is the Kapferer network consisting of frequent interactions between tailor shop workers collected by Bruce Kapferer in Zambia in the summer of 1965 (Kapferer, 1972). This network is larger than the first with 158 edges and represents larger un-directed networks. The third network is a directed friendship network of college freshman collected from 1965 to 1966 (Van de Bunt et al., 1999). This network is a representative for small directed networks. Finally, the fourth network is an Ecoli network with nodes as operons and directed edges indicate that the transcription factor has been encoded from one operon to the other (Salgado et al., 2001). The ecoli network represents a large and directed network.
<table>
<thead>
<tr>
<th>Network</th>
<th>Number of Endogenous Variables</th>
<th>Number of Models</th>
<th>Number of Bi-variate ERGMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule</td>
<td>9</td>
<td>465</td>
<td>36</td>
</tr>
<tr>
<td>Zach</td>
<td>6</td>
<td>63</td>
<td>15</td>
</tr>
<tr>
<td>Lazega</td>
<td>10</td>
<td>847</td>
<td>45</td>
</tr>
<tr>
<td>Kapferer</td>
<td>13</td>
<td>4095</td>
<td>78</td>
</tr>
<tr>
<td>Mesa</td>
<td>7</td>
<td>126</td>
<td>21</td>
</tr>
<tr>
<td>Bunt*</td>
<td>12</td>
<td>2509</td>
<td>66</td>
</tr>
<tr>
<td>Desert</td>
<td>61</td>
<td>62,034,255</td>
<td>1830</td>
</tr>
<tr>
<td>Dixon</td>
<td>57</td>
<td>40,901,281</td>
<td>1596</td>
</tr>
<tr>
<td>Ecoli</td>
<td>30</td>
<td>768,211</td>
<td>435</td>
</tr>
</tbody>
</table>

**Table 3.2:** Number of endogenous variables and ERGMs

Based off the total number of candidate ERGMs in column 3 of the above table. The Desert, Dixon and Ecoli network require a smaller set of endogenous variables if we are to consider ERGMs with more than two endogenous variables. Therefore, we applied algorithm 1 to these three networks but only for the family of bi-variate ERGMs in order to obtain a reduced set of endogenous variables. The variables that produce the smallest edge FRC CvM distance are selected for further screening. Initially, these three networks possessed candidate ERGMs in the hundreds of thousands. After applying the size of the model space was reduced to a manageable size of a few thousand candidate ERGMs. More information can be found in the Numerical Studies Additional Results section in the appendix.

### 3.4.1 Initial Screening

For each candidate ERGM in our model space $\mathcal{M}$, we sampled 25 adjacency matrices and calculated the edge FRC values for each matrix. The edge FRC values result in an empirical cumulative density function which was then compared to the observed network cumulative density function via the two sample Cramer Von Mises distance. This results in 25 edge FRC CvM distances. The average edge FRC CvM distance was calculated for each ERGM and results in a distribution of these average distances across the model space. These values were then clustered via K-means and the model space was significantly reduced by considering the cluster that possessed the lowest average edge FRC CvM distance.

After the initial screening, the model space can potentially contain hundreds of ERGMs. The initial screening algorithm quantifies the difference between the observed cumulative density
function of edge FRC values and the sampled sumulative density function of edge FRC values obtained from a candidate ERGM in our model space. However, the edge FRC values do not capture network structures such as number of isolated nodes or number of edges. In order to select the optimal ERGM, a model selection and scoring procedure is implemented that is based upon network motif counts.

3.4.2 Model Selection and Scoring via Network Motif Counts

We begin with a detailed discussion for results of the model scoring procedure for the Bunt network. The representative of small directed networks. The procedure described below was the same for the other networks. However, the scoring formula is different for un-directed and directed networks as illustrated by the equations for $SCORE_u$ and $SCORE_d$ in the previous section. First the distribution of scores for each ERGM in $\mathcal{M}$ is clustered until the top handful are identified. Denote the set of top performing ERGMs by $C_1$. The edge variance is then calculated for each ERGM in $C_1$ and the ERGM with the lowest edge variance is selected.

The Bunt network possesses over a hundred candidate models. After scoring each one, four best performing ERGMs were identified as indicated by the vertical red line in figure 4. The next
3.4.3 Optimal ERGM Performance

The score provides an aggregate metric on the general performance of a candidate ERGM. However, it does not provide a complete picture of performance. For example, the score does not provide details into which motifs were over or under estimated. In this section, we provide detailed analysis of the optimal ERGM performance for the Bunt network. The directed model
score comprises of 16 different counts of network motifs. The model score for the Bunt network was comprised of 15 network motifs. Since the Bunt network did not contain more than 10 isolated nodes, the count of isolated nodes did not factor into the equation of the model score.

The Bunt network had an optimal ERGM with equation

\[ P_\theta(Y_{i,j} = 1) = \psi(\theta) \exp\{\theta_0 \times \text{edges} + \theta_1 \text{desp}(2, \text{type} = \text{OSP}) + \theta_2 \text{dgwnsp} + \theta_3 \text{istar}(3) + \theta_4 \text{m2star} + \theta_5 \text{simmelian}\}. \]

The average count of 15 network motifs and their associated standard deviations obtained from this ERGM is found in figure 6. The points on the figure represent the observed network motif counts. For 14 network motifs the points lie within the interquartile range of the box plots with bi-edges being the only network motif that did not satisfy this property.

**Figure 3.6:** Upper left network is the observed Bunt network. The remaining three networks are draws from the ERGM model with equation

\[ P_\theta(Y_{i,j} = 1) = \psi(\theta) \exp\{\theta_0 \times \text{edges} + \theta_1 \text{desp}(2, \text{type} = \text{OSP}) + \theta_2 \text{dgwnsp} + \theta_3 \text{istar}(3) + \theta_4 \text{m2star} + \theta_5 \text{simmelian}\} \]
Figure 3.7: A box plot of the mean network motif count generated by the optimal ERGM for the Bunt network. "C" is used to denote configuration and the number associated with it refers to the specific network motif found in Figure 1. The horizontal line represents the mean count generated by the ERGM and the points represent the observed network motif counts.

3.5 Discussion and Conclusion

The model selection methods presented in this paper can be extended, with no substantial changes, to other families of statistical network models and random graphs such as stochastic block models. Moreover, one can apply these methods to networks possessing count or valued ties by using a compatible graph curvature (Weber et al., 2017). Model selection for bi-partite, multi-level, temporal and other types of ERGMs can be performed by modifying the set of endogenous variables before the first algorithm.

The methods of this paper relied heavily on the Forman-Ricci notion of curvature which is defined via the edges of a graph. However, many other edge-based notions of curvature exist such as Olliver-Ricci and Bakery-Emery (Lin and Yau, 2010). Additionally, one can use the above notions of curvatures and define them via the nodes of a graph instead of edges (Pouryahya et al., 2017). Various modifications and adaptations of a single graph curvature exist, recall that the methods of this paper used an augmented version of the edge Forman-Ricci notion. Each of these curvatures emphasize different graph structures and these differences are represented
Figure 3.8: Upper left network is the observed Ecoli network. The remaining three networks are draws from the ERGM model with equation

\[ P_\theta(Y_{i,j} = 1) = \psi(\theta) \exp(\theta_0 \times \text{edges} + \theta_1 ddsp(2, \text{type} = 'ISP') + \theta_2 ddsp(3, \text{type} = 'OSP') + \theta_3 dmsp(2, \text{type} = 'OSP') + \theta_4 \text{idegree}(16) + \theta_5 \text{idegree}(2) + \theta_6 \text{idegree}(24)) \]

in the distribution of the curvature values. A future avenue of research is the comparison of the similarities and differences of using these curvatures in the initial network model selection phase.

The performance of a candidate ERGM was dictated exclusively by its score formula given in section 3. The score attempts to quantify the overall performance of a candidate ERGM with respect to the various network motifs. By changing the weights associated with each network motif ratio one can penalize different network structures. However, it does not provide network motif specific performance. Furthermore, graph curvature was only used in the initial model selection phase and not a part of the model scoring methods. With this in mind, the deviations observed in section 4.4 might be mitigated if such modifications are implemented to the ERGM scoring system. However, such deviations might be shortcomings of the ERGM itself.
CHAPTER 4: FUTURE RESEARCH

In this chapter, we discuss the mathematical tools that will enable us to understand the theoretical properties of the ERGMs proposed by this work. Then, we discuss potential applications.

4.1 Ensuring a Stable ERGM proposal

Rigorously proving that the ERGMs proposed by the methods of our work reliably produce non-degenerate ERGMs, as well as establishing other performance guarantees, is left as an avenue for future research. In this section, we discuss key mathematical tools that will be useful for this endeavour.

4.1.1 Hamiltonian Operator

The Hamiltonian operator $H(Y)$ plays a central role in limiting the degeneracy phenomena exhibited by ERGMs. For a tie, $e$, one can decompose the operator into $H(Y) = A_e(Y) + B_e(Y)$ where $Y$ is our network, $A_e(Y)$ consists network structures dependent on the tie $e$ and $B_e(Y)$ consists of network structures independent of the tie $e$. During an ERGM draw, one starts with an initial state $Y_0$ for the network and updates each potential tie in an iterative manner. One can characterize the future state $Y_{t+1}$ by either $Y_{e+}$ or $Y_{e-}$. Here $Y_{e+}$ is identical to the network $Y_t$ except that it possess an extra tie $e$. Similarly, $Y_{e-}$ is identical to $Y_t$. The transition probability from $Y_t \longrightarrow Y_{t+1}$ is completely determined by the derivative of the Hamiltonian $H(Y)$ Bhamidi et al. (2008). Explicitly we have the probability of obtaining state $Y_{e+}$ is given by $\frac{\exp\{\partial H(Y)\}}{1+\exp\{\partial H(Y)\}}$. Thus, the probability of generating degenerate networks from a proposed ERGM is intimately related to the smoothness properties of the Hamiltonian operator $H(Y)$ and the behavior of it’s
derivatives. In essence, the procedures of this paper choose the Hamiltonian in an ”informed” way thus significantly reducing the probability of model degeneracy.

4.1.2 Phase Shifts

After obtaining a candidate ERGM, one can investigate the the parameter coefficient estimates of the endogenous variables to obtain conclusions about the stability of the proposed ERGM. In other words this is a sensitivity analysis for ERGMs. Specifically, one investigates the following variational problem numerically. Let $k$ be the total number of endogenous variables for the proposed ERGM and $i = 1, \ldots, k$. Let $\theta_i$ be the coefficient estimate for the endogenous variable $H_i$. Finally, we define $e(H_i)$ be the number of edges that the endogenous variable possesses. For example, if our endogenous variable is a triangle then the number of edges would be 3.

$$
\sup_{0 \leq u \leq 1} \sum_{i=1}^{k} \theta_i u^{e(H_i)} - \frac{1}{2} u \log(u) - \frac{1}{2} (1 - u) \log(1 - u)
$$

When we have a unique solution, $u^*$, to the above variational problem then our proposed ERGM is stable. In the sense that the proposed ERGM can be used to model larger observed networks that possess the same dependency structure as our original observed network. However when the solution is not unique this indicates that our proposed ERGM may not be readily applied to similar networks as it is prone to degeneracy. In such a case we obtain a phase shift. Consider a simple ERGM with an edge term and triangle given below.

$$
P_\beta(Y_{i,j} = 1) = \psi(\theta) exp\{\beta_1 \times \text{edges} + \beta_2 \text{triangle}\}
$$

When the coefficient for the edge term $\beta_1$ is fixed at $-0.45$, a phase shift occurs when the coefficient for the triangle term approaches 0.65. If the proposed ERGM has such a combination of coefficients, for example $\beta_1 = -0.45$ and $\beta_2 = 0.62$, this would indicate that the ERGM is prone to degeneracy if we attempt to use it for another observed network with the same dependance structure as the original network. A similar scenario if found for $\beta_1 = -0.8$, a $\beta_2$ value near 0.9
Figure 4.1: The x-axis represents $\beta_2$ values and the y-axis represents the solution $u^*$. Source: Chatterjee and Diaconis (2013)

indicates troublesome stability properties. On the other hand, for $\beta_1 = 0.2$ and $\beta_1 = -0.35$ all of the $\beta_2$ coefficient values from 0 to 1.5 indicate a stable ERGM.

4.2 Asymptotic Properties

In this section, we discuss the mathematical tools that will be needed to establish asymptotic performance guarantees for the methods of this work.

4.2.1 Graph Homomorphism Densities

Network motifs are small graphs, typically not exceeding 6 nodes, that are statistically significant patterns found in a larger network Masoudi-Nejad et al. (2012). From these network motif counts one can construct homomorphism densities for an observed network and candi-
date ERGM. Homomorphisms are maps that preserve the algebraic structure found in a group 
Dummit and Foote (2004). Formally, for some group operation $\odot$, the map $f : A \rightarrow B$ is a 
homomorphism when it satisfies the relation below.

$$f(x \odot_A y) = f(x) \odot_B f(y) \quad x, y \in A$$  \hspace{1cm} (4.3)

For finite un-directed graphs, $G = (V(G), E(G))$ and $H = (V(H), E(H))$, a homomor-
phism $f : G \rightarrow H$ between the two graphs implies that for any two vertices $v_1$ and $v_2$ present in
$E(G)$ we automatically have that $f(v_1)$ and $f(v_2)$ belong in $E(H)$ Hahn and Tardif (1997). One
can thus count the number of homomorphisms for a given network motif $H$ into an observed net-
work $G$. Practically this counts the number of edge-preserving maps between vertex sets $V(H)$
and $V(G)$. Then after counting the number of maps, we can define the homomorphism density.

$$t(H, G) = \frac{|\text{hom}(H, G)|}{|V(G)||V(H)|}$$  \hspace{1cm} (4.4)

This density represents the probability that any arbitrary mapping $t : V(H) \rightarrow V(G)$ is

**4.2.2 Graph Limits**

Graph limits extend the idea of convergence to graphs by using homomorphism densities.
A graphon, or graph function, is a symmetric measurable function from $[0, 1]^2 \rightarrow [0, 1]$. Any
random graph can be converted into a set of graphons by using their associated homomorphism
densities. This set of homomorphism densities then characterizes a random graph. With graphons,
one can begin to envision the notion of convergence of graphs. Indeed, a sequence of graphs $G_n$ is converges if there exists a graphon $w$ such that for any network motif $H$ the following holds.

$$\lim_{n \rightarrow \infty} t(G_n, H) = t(W, H)$$  \hspace{1cm} (4.5)
One can see that convergence for graphs is defined using homomorphism densities. To obtain asymptotic performance guarantees, one must show that the sequence of graphs generated by the algorithms of this work eventually converges to the observed network (graph) after sufficient iteration. In this case the candidate graphon, \( W \), is the observed network. The idea is that the proportion of simple network motifs such as edges, 2-stars and triangles approaches the proportion found in the observed network. However, the difficulty in obtaining asymptotic performance guarantees lies in defining \( n \) in a meaningful way.

4.3 Potential Applications for Organizational Managers

Uncovering the dependence structure of an observed network has a lot of value for organizational managers. One can use the methods of this work to obtain concrete quantification of the organization’s network structure. Managers can then use the proposed ERGMs to investigate how a previous policy affected their organization. Similarly, they can use the proposed ERGM to help inform new decisions, policies and information flow by understanding their organization’s network structure. Additionally, this allows managers to investigate the impact of theoretical policies and decisions without actually having to implement them. The proposed ERGM aids managers in translating the given organizational network structure to concrete business decisions.
APPENDIX A: APPENDIX FOR CHAPTER 2

A.1 Results of Bounding the input Parameter

<table>
<thead>
<tr>
<th>Network</th>
<th>Upper Bound of kstar</th>
<th>Upper Bound of esp</th>
<th>Upper Bound of nsp</th>
<th>Upper Bound of dsp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazega</td>
<td>14</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Kapferer</td>
<td>23</td>
<td>11</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Kapferer2</td>
<td>24</td>
<td>12</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Zach</td>
<td>16</td>
<td>9</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Wind Surfers</td>
<td>57</td>
<td>35</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td>Molecule</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Faux Mesa High school</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Ecoli</td>
<td>71</td>
<td>6</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Florentine Marriage</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Florentine Business</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
APPENDIX B: APPENDIX FOR CHAPTER 3

B.1 Results Visualization

Figure B.1: Upper left network is the observed Molecule network. The remaining three networks are draws from the ERGM model with equation $P_\theta(Y_{i,j} = 1) = \psi(\theta) exp\{\theta_0 \times edges + \theta_1 degree 1.5 + \theta_2 gwdsp + \theta_3 kstar(2) + \theta_4 kstar(3) + \theta_5 kstar(5)\}$

Figure B.2: Upper left network is the observed Zach network. The remaining three networks are draws from the ERGM model with equation $P_\theta(Y_{i,j} = 1|X) = \psi(\theta) exp\{\theta_0 \times edges + \theta_1 nsp(2) + \theta_2 esp(3) + \theta_3 SOCIALITY + \theta_4 CLUB + \theta_5 FACTION + \theta_6 ROLE\}$
Figure B.3: Upper left network is the observed Lazega network. The remaining three networks are draws from the ERGM model with equation $P_{\theta}(Y_{i,j} = 1|X) = \psi(\theta) exp(\theta_0 \times \text{edges} + \theta_1 \text{esp}(4) + \theta_2 \text{gwnsp} + \theta_3 \text{isolates} + \theta_4 \text{AGE} + \theta_5 OFFICE + \theta_6 \text{PRACTICE})$.

Figure B.4: Upper left network is the observed Kapferer network. The remaining three networks are draws from the ERGM model with equation $P_{\theta}(Y_{i,j} = 1) = \psi(\theta) exp(\theta_0 \times \text{edges} + \theta_1 \text{dsp}(6) + \theta_2 \text{esp}(4) + \theta_3 \text{sociality})$. 

55
Figure B.5: Upper left network is the observed Mesa network. The remaining three networks are draws from the ERGM model with equation
\[ P(\theta Y_{ij} = 1|X) = \psi(\theta) \exp(\theta_0 \times \text{edges} + \theta_1 \text{esp}(3) + \theta_2 \text{nsrp}(2) + \theta_3 \text{nsrp}(4) + \theta_4 \text{GRADE} + \theta_5 \text{RACE} + \theta_6 \text{GENDER}) \]

Figure B.6: Upper left network is the observed Desert network. The remaining three networks are draws from the ERGM model with equation
\[ P(\theta Y_{ij} = 1|X) = \psi(\theta) \exp(\theta_0 \times \text{edges} + \theta_1 \text{esp}(3) + \theta_2 \text{nsrp}(2) + \theta_3 \text{nsrp}(4) + \theta_4 \text{GRADE} + \theta_5 \text{RACE} + \theta_6 \text{GENDER}) \]
Figure B.7: Upper left network is the observed Dixon network. The remaining three networks are draws from the ERGM model with equation

\[ P_\theta(Y_{i,j} = 1) = \psi(\theta) exp\{\theta_0 \times \text{edges} + \theta_1 ddsp(2, \text{type} = 'ISP') + \theta_2 ddsp(4, \text{type} = 'ISP') + \theta_3 desp(2, \text{type} = 'ISP') + \theta_4 desp(2, \text{type} = 'OTP') + \theta_5 desp(3, \text{type} = 'ISP') + \theta_6 \text{simmelian}\} \]
REFERENCES


