The Pennsylvania State University
The Graduate School

THEORY AND APPLICATIONS OF ESTIMATION METHODS FOR EXPONENTIAL-FAMILY RANDOM GRAPH MODELS

A Dissertation in
Statistics
by
Christian Song-Hyo Schmid

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Christian Song-Hyo Schmid was reviewed and approved by the following:

David R. Hunter
Professor of Statistics
Dissertation Advisor
Chair of Committee

Bruce A. Desmarais
Professor of Political Science

Ethan Xingyuan Fang
Assistant Professor of Statistics

Murali Haran
Professor of Statistics
Department Head

Xiaoyue Maggie Niu
Associate Research Professor of Statistics

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Studies

Abstract

Over the past four decades, two distinct likelihood-based estimation methods for Exponential-family Random Graph Models (ERGMs) have emerged. One method tries, due to the complexity of the model, to approximate the maximum likelihood estimator (MLE) using Markov Chain Monte Carlo (MCMC) techniques, while the other method optimizes a simpler, but misspecified likelihood function that results in the so-called maximum pseudolikelihood estimator (MPLE). Interestingly, both estimators can be seen as opposite ends of a spectrum in multiple senses. In this dissertation, I examine the theory of these two estimation approaches and, in the process, develop an improved method for determining starting values for the MCMC algorithm that is based on the likelihood principle and uses simulated annealing. I also introduce two approaches for correctly approximating standard errors for the MPLE. These two approaches are based on parametric bootstrapping and the Godambe information matrix, respectively. This dissertation also provides empirical evidence that it is possible, by using an offset term that depends on the sample size, to embed some models in a sequence that renders the MPLE consistent and asymptotically normal.

I focus the applications of the proposed techniques on networks predominantly explored in the field of political science. In a first project, a citation ERGM is developed to analyze the citation network among US Supreme Court majority opinions. In addition to finding evidence of exogenous covariates, the model provides evidence for reciprocity, transitivity, and popularity in the network. The other project examines the network of bills introduced in the United State Senate, where ties between bills indicate the similarity of the set of legislators who co-sponsor them.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank all the people who contributed in some capacity to the work described in this thesis. First and foremost, I offer my sincerest gratitude to my advisor, Prof. David R. Hunter, for his guidance, understanding, patience, and most importantly, his friendship, during my graduate studies. Without his encouragement and effort, this thesis would not have been possible.

I would also like to sincerely thank Prof. Bruce A. Desmarais who has supported me throughout my research with his patience and knowledge while allowing me to think independently as well. I also want to thank him for giving me the opportunity to attend and present at two conferences.

Many thanks also to my remaining committee members, Prof. Murali Haran, Prof. Xiaoyue Niu, and Prof. Xingyuan Fang, for their valuable input and stimulating questions. Their insightful perspectives were critical for the success of this dissertation.

A very special thanks goes to the late Prof. Detlef Dürr, without whose motivation and encouragement I would not have considered a graduate career. It was under his tutelage that I became interested in mathematics and statistics at an early age, and over the course of my studies I came to consider him just as much a friend as a mentor. Sadly, he passed away a few months before my defense, but I will always remember and be grateful for his presence in my life.

I am deeply and forever indebted to my parents, Stefan and Jae-Sook Schmid, for their love, support and encouragement throughout my entire life. Their daily hard work made it possible for me to be as academically focused as I wanted to be. I am especially grateful for the countless hours they spent babysitting our son, Theodor, that allowed both my wife and I to concentrate on our studies during this busy time. I am also very thankful to my beautiful sisters, Jessica and Gloria, and my loony brothers Maximilian, Sebastian, Amadeus, and Godwin for all the great memories!

Also, I thank my wife's parents, Jin and Jenny Kim, as well as my wife's siblings, Catherine and Ben. I am grateful that I always felt welcome and loved in their presence and that we have their support at any given time.

Finally, and most importantly, I have saved the last word of acknowledgment for my wonderful wife, Jessica. Her support, encouragement, quiet patience, and unwavering love were undeniably the bedrock upon which the past years of my life have been built. I am also eternally grateful to her for giving birth to our wonderful son, Theodor, who brightens our lives every single day.

Chapter 1
Introduction

Networks are a form of relational data, which are data that not only capture attributes of individuals, like age, gender, and race, but also provide information about the relationships between these individuals. A network consists of nodes and edges, where the nodes represent the actors in a network - these could range from people, groups, and countries to papers in a citation network or proteins in a protein-to-protein interaction network - and the edges depict a relationship of interest between the nodes. These relationships, which are treated as the response in statistical models, often do not arise independently of each other. For example, if we consider a friendship network, a friend of a friend is likely to be a friend, which means that standard regression methods that assume the independence of observations, like logistic regression, are usually not appropriate. The dependence structure among the nodes is a systemic feature of the data, as opposed to an occasional coincidence, and is therefore a matter of particular interest.

Statistical network analysis tries to account for exactly these kinds of dependency structures, by not simply treating them as measuring errors, but rather including them as central components of the network model. This means that network models consider the patterns of ties as the response in a regression-like framework instead of looking at each edge individually and requiring independence among these edges.

This distinction, the lack of the independence assumption, calls for more complex models which consequently also results in a more complicated estimation of the model parameters. One of these more complex models is the exponential-family random graph model (ERGM), a probability model for networks that I will focus on in this dissertation. Over the past four decades, two distinct likelihood-based estimation methods have emerged. One method tries, due to the complexity of the model, to approximate the maximum likelihood estimator (MLE) using Markov Chain Monte Carlo (MCMC) techniques, while the other popular method treats the probabilities of ties, conditioned

on the rest of the network, as independent, which results in a misspecified, but easy-to -optimize likelihood function. The resulting estimator is called the maximum pseudolikelihood estimator (MPLE) and can be estimated in a logistic regression framework as I will discuss at a later point.

Between these two distinct estimation approaches there is general consensus that the approximate Markov Chain Monte Carlo MLE (MCMLE) is the superior of the two estimators. Instead of optimizing a potentially incorrect likelihood function, the MCMLE attempts to approximate and optimize the model's true likelihood function, resulting in an estimator that is expected to be close to the true MLE. The MLE has many desirable properties like satisfying the likelihood principle and defining a probability distribution that has the observed network's statistics as its expectation. (The likelihood principle states that all information in the data relevant to the model parameters is contained in the likelihood function.) However, MCML estimation is computationally expensive and not always successful (Hummel et al., 2012). These disadvantages lead scientists to consider the computationally simple MPLE as an alternative. Estimation is quick and can be done using any canned logistic regression software. However, the MPLE has acquired a bad reputation, mostly because standard errors obtained from logistic regression software were shown to be unreliable (van Duijn et al., 2009). In addition, unlike the MLE, the distribution defined by the MPLE does not necessarily have the observed network's statistics as its expectation and the MPLE does also not satisfy the likelihood principle.

1.1 ERGMs in Political Science

The applications of techniques that will be examined throughout this dissertation fall largely in the realm of political science, a field that has created a vast literature of statistical network applications in recent years. For example, Cranmer and Desmarais (2011) apply network theory to investigate militarized interstate disputes for the time period 1870 - 1996. The nodes in this undirected network represent countries while an edge indicates that there has been at least one militarized dispute between these two countries. The authors find that states are unlikely to start conflicts with countries with whom they share a common enemy.

A counterpart to conflict networks are alliance networks as they were studied by Cranmer et al. (2012). The formation and dissolution of alliances among nations is a complicated process that highly depends on the behavior of third states and is therefore

best investigated using a network approach. In contrast to the interstate conflict network, the authors find a significant tendency for states to cluster, meaning that states tend to form alliances with nations with whom they share a common ally.

Another popular type of networks is one that describes economic relationships between countries. Thurner et al. (2019) explore the international arms trade network between countries from 1950 - 2013 as a directed and binary network. A directed trade network means that the direction of an edge indicates which node acts as the seller and which as the buyer in any given transaction. By binarizing the monetary value of a transaction, the authors reduce the network to a form where a tie only indicates whether two nations trade weapons without accounting for the magnitudes of the deals. One result of the paper states that the arms trade network has a strong activity effect, meaning that the probability of having a tie from a node increases with the number of already existing ties from that node.

A different economic relationship network was discussed by Cranmer et al. (2014). The authors investigate the international economic sanction network, where a tie from nation A to B indicates that A imposed economic sanctions on B. A key finding is that economic sanctions between two countries are often mutual, meaning that sanctions from A to B increases the odds of B imposing sanctions on A as well.

Besides looking at nodes as nations or political territories, one can also consider the nodes of a network as representing smaller political entities, such as political parties or interest groups. With the existence of thousands of interest groups, it is difficult for a single interest group to have an impact on the formulation of public policies. Consequently, these groups form coalitions to increase their impact on political decisions. Heaney and Leifeld (2018) conceptualize the formation of coalitions between groups in the US health policy domain as a bipartite network, with interest groups and coalitions as two different classes of nodes. In a bipartite network, an edge can only appear between nodes of different classes. In this network, an edge between a group and a coalition indicates that this group has joined that coalition.

On a related topic, Heaney (2014) investigates the impact that US health policy interest groups' influence reputation has on policy making. The author examines a multiplex network, a network where actors are connected through more than just one type of edge. This means that the influence reputation of interest groups is studied by considering different types of networks, in particular, a communication network, a coalition overlap network, and an issue overlap network. The communication network is a directed and weighted network, where a tie from A to B can take the values $0, 1$

or 2 depending on how frequently members between the two groups communicate. The coalition network indicates whether two groups have been involved in a common health care coalition. The issue overlap network is defined as a weighted network where an edge between two interest groups takes the value of the number of health care issues both groups have defined as major priority on their agendas.

Hollway and Koskinen (2016) explore mulitlevel networks based on the global fisheries governance structure. Multilevel networks interlock unipartite, i.e., networks with ties on a single set of nodes, and bipartite networks. In this particular example, the network consists of two kinds of nodes, states and multilateral fisheries agreements, and different kinds of ties between the nodes. A tie between two states indicates a level of interaction in governance work between these two nations, while a tie between a state and fisheries agreement depicts the affiliation of a state to a fisheries agreement. Finally, a tie between two fisheries agreements displays related fisheries agreements.

1.2 A brief history of ERGMs

The beginning of statistical network analysis goes back to Erdös and Renyi (1959) and Gilbert (1959), who independently introduced a model where each tie/edge occurs independently with constant probability p, i.e., $P(Y_{ij} = 1) = p$, where $Y_{ij} = 1$ indicates that there is an edge from node i to j. I will focus on so-called unweighted networks, which means that a tie only has a binary outcome $Y_{ij} \in \{0, 1\}$.

A network on N nodes is represented by an adjacency matrix $A = (A_{ij}) \in \mathbb{R}^{N \times N}$ with $A_{ij} = 1$, if there is an edge or tie between i and j and $A_{ij} = 0$ otherwise, $i, j \in \{1, ..., N\}$. In this matrix every column and every row is assigned to a specific node and the ith column and the ith row are assigned to the same node.

Holland and Leinhardt (1981) proposed the $p1$-model, a network model for directed graphs that differs from the Erdös-Rényi-Gilbert model by including fixed effects that a) account for the reciprocity between two nodes and b) explain an individual node's tendency to attract and produce ties. Besides accounting for the model's general tendency for ties, which is measured by counting the number of ties in the observed network, the $p1$-model also considers the number of pairs of nodes with mutual ties, i.e., $Y_{ij} = Y_{ji} = 1$. Furthermore, a node i's tendency to attract ties is incorporated as the number of ties pointing to i, which we call the node's *outdegree*. Similarly, we include a node's tendency to produce ties as the number of ties pointing away from i, which is defined as the node's *indegree*.

The resulting $p1$-model is then defined as

$$P(Y = A) = \frac{\exp\Big(\theta \cdot \mathrm{e}(A) + \rho \cdot \mathrm{m}(A) + \sum_i \alpha_i \cdot \mathrm{out}_i(A) + \sum_j \beta_j \cdot \mathrm{in}_j(A)\Big)}{k(\theta, \rho, \{\alpha_i\}, \{\beta_j\})} \tag{1.1}$$

where $\theta, \rho, \{\alpha_i\}, \{\beta_j\}$ are the model's parameters and

$$k(\theta, \rho, \{\alpha_i\}, \{\beta_j\}) = \sum_{i<j}(1 + e^{\mu_{ij}} + e^{\mu_{ji}} + e^{\mu_{ij}+\mu_{ji}+\rho}) \quad , \quad \mu_{ij} = \theta + \alpha_i + \beta_j$$

is a normalizing constant that assures (1.1) to be a probability model. With $e(A)$, I denote the number of edges in A, with $m(A)$, the number of mutual ties, and with $\mathrm{out}_i(A)$ and $\mathrm{in}_i(A)$, I refer to node i's out- and indegree, respectively.

Define a dyad simply as a pair of nodes ij. Then, Holland and Leinhardt (1981) define the within-dyad relationship as

$$P(Y_{ij}, Y_{ji}|\theta, \rho, \alpha_i, \beta_j) = \frac{e^{\mu_{ij}y_{ij}+\mu_{ji}y_{ji}+\rho y_{ij}y_{ji}}}{1 + e^{\mu_{ij}} + e^{\mu_{ji}} + e^{\mu_{ij}+\mu_{ji}+\rho}}.$$

In addition, they assume the between-dyad relationship to be independent. This results in the following likelihood function:

$$\ell(\theta, \rho, \{\alpha_i\}, \{\beta_j\}) = \prod_{i<j} P(Y_{ij}, Y_{ji}|\theta, \rho, \alpha_i, \beta_j). \tag{1.2}$$

The authors propose estimating the parameters in (1.2) either by using Fisher scoring or by applying the generalized iterative scaling approach described by Darroch and Ratcliff (1972).

The $p1$-model only assumes dependence between edges within the same dyad and independence between edges of different dyads. Frank and Strauss (1986) consider undirected graphs and alleviate the independence assumption between different dyads by introducing Markovian dependence for networks. A graph has Markov dependence and is therefore called a Markov graph if nonincident dyads - dyads that do not share any nodes - are conditionally independent. This means that the probability of a tie between two nodes i and j only depends on dyads that involve nodes i and/or j. Frank and Strauss (1986) demonstrate that the probability of any undirected Markov graph can be defined as

$$P(Y = A|\{\sigma_k\}, \tau) = \frac{\exp(\tau \cdot t(A) + \sum_k \sigma_k \cdot s_k(A))}{k(\{\sigma_k\}, \tau)} \tag{1.3}$$

where $t(A)$ is defined as the graph's number of triangles, $s_k(A)$ as the graph's number of k-stars and τ and σ_k are the model's parameters. A triangle is defined as any set of edges (i, j), (j, l) and (l, i), $i \neq j \neq l$, while a k-star is defined as any set of k edges connected to the same node. Note that $s_1(A)$ equals $e(A)$, the number of edges in the network. Finally, $k(\{\sigma_k\}, \tau)$ is the model's normalizing constant, defined as the weighted sum of all possible networks on N nodes, which, unlike the $p1$-model, is intractable in most cases.

A special case of model (1.3) is the $\rho\sigma\tau$-model that will be discussed at a later point in this dissertation. It is defined as

$$P(Y = A|\rho, \sigma, \tau) = \frac{\exp(\rho \cdot \mathrm{e}(A) + \sigma \cdot s_2(A) + \tau \cdot t(A))}{k(\rho, \sigma, \tau)} \tag{1.4}$$

where ρ, σ, and τ refer to the model parameters.

Similar to the normalizing constant in (1.3) and unlike the $p1$-model, assuming between-dyad dependence comes at the price of an intractable normalizing constant, which makes maximum likelihood estimation extremely difficult in most cases. For this reason, Frank and Strauss (1986) only propose a trial-and-error method for the $\rho\sigma\tau$-model. This approach constrains the set of possible networks to those with a fixed number of edges and requires either σ or τ to be 0. This results in a univariate model where only either σ or τ have to be estimated. The suggested approach is based on the exponential family properties that the MLE is unique and that the expectation of the sufficient statistic of the distribution defined by the MLE equals the sufficient statistic of an observed network A^{obs}. For the case $\tau = 0$ this means $\mathbb{E}_{\hat{\sigma}}(s_2(Y)) = s_2(A^{obs})$. Based on this property, the authors propose to fix σ, simulate a large number of networks from the distribution defined by σ, and calculate the mean number of 2-stars. This process is then repeated for different values of σ until a value is found where the mean number of 2-stars is close enough to the observed number of 2-stars.

Frank and Strauss (1986) also briefly describe that estimation for dyad-dependence models could be done using logistic regression software, but it is Strauss and Ikeda (1990) who discuss the theory behind this approach in detail. Let θ be the set of parameters, e.g., for model (1.4) $\theta = (\rho, \sigma, \tau)$. Based on the fact that every entry of the adjacency matrix A can be taken as a outcome of a single Bernoulli variable Y_{ij}, Strauss and Ikeda (1990) show that

$$\mathrm{logit}(P_\theta(Y_{ij} = 1|Y_{ij}^c = A_{ij}^c)) = \theta^\top \cdot (g(A_{ij}^+) - g(A_{ij}^-)) \tag{1.5}$$

where A^c_{ij} is short for the network without dyad ij, i.e., A_{rs} for all $r, s \in \{1, ..., N\}$ with $rs \neq ij$, and A^+_{ij} and A^-_{ij} are defined as the adjacency matrix that emerges from A when $A_{ij} = 1$ and $A_{ij} = 0$, respectively. If one assumes the Y_{ij} are conditionally independent, equation (1.5) defines a logistic regression model. In general, the resulting likelihood function

$$L(\theta) = \prod_{ij \in \Omega(N)} P_\theta(Y_{ij} = 1 | Y^c_{ij} = A^c_{ij})^{A_{ij}} (1 - P_\theta(Y_{ij} = 1 | Y^c_{ij} = A^c_{ij})^{1-A_{ij}} \tag{1.6}$$

is not the model's true likelihood function and, as a consequence, the resulting maximum likelihood estimator is called the maximum pseudolikelihood estimator (MPLE). In equation (1.6), $\Omega(N) = \{ij \mid i, j \in \{1, \ldots, N\}, i \neq j\}$ is defined as the set of all dyads in a network of N nodes.

The ERGM or p^*-model, which generalizes the $p1$- model as well as Markov graph models, was first introduced by Wasserman and Pattison (1996). The model is defined as

$$P(Y = A|\theta) = \frac{exp(\theta^\top \cdot g(A))}{k(\theta)} \tag{1.7}$$

where $\theta \in \mathbb{R}^q$ is a vector of parameters, $k(\theta)$ is the normalization constant, and $g : A(N) \to \mathbb{R}^q\ ,\ A \mapsto (g_1(A), \ldots, g_q(A))$ is a vector of network statistics. The vector of statistics $g(A)$ enables the implementation of any network statistic used in models (1.1), (1.3), and (1.4) as well as any other potential statistic. In other words, the ERGM is not limited to within-dyad dependence or Markov dependence, and therefore, generalizes the models introduced by Holland and Leinhardt (1981) and Frank and Strauss (1986).

For the same reason as for the estimation of Markov graphs, straight forward maximum likelihood estimation is generally infeasible for ERGMs. Due to a lack of alternatives, Wasserman and Pattison (1996) propose the MPLE, as introduced by Strauss and Ikeda (1990), as the parameter estimation method. The MPLE remained the prevailing estimation approach until Snijders (2002) and Hunter and Handcock (2006) developed a competing alternative for approximating intractable likelihood functions that is based on an importance sampling technique introduced by Geyer and Thompson (1992). The idea of this approach is that for any $\theta_0 \in \mathbb{R}$ one can show that

$$\frac{k(\theta)}{k(\theta_0)} = \mathbb{E}_{\theta_0}\Big[\exp\Big((\theta - \theta_0)^\top \cdot g(Y)\Big)\Big]. \tag{1.8}$$

This equality states that the quotient of normalizing constants equals an expectation

with respect to θ_0. This is especially useful because it allows us to approximate the ratio of normalizing constants by simulating a large sample of networks $A_1, \ldots, A_S$ from the distribution defined by θ_0:

$$\frac{k(\theta)}{k(\theta_0)} \approx \frac{1}{L} \cdot \sum_{s=1}^{S} \exp\Big((\theta - \theta_0)^\top \cdot g(A_s)\Big). \tag{1.9}$$

This result can then be used to approximate the model's loglikelihood function

$$\ell(\theta) - \ell(\theta_0) \approx (\theta - \theta_0) \cdot g(A) - \log\left(\frac{1}{L} \sum_{s=1}^{L} \exp\left\{(\theta - \theta_0)^\top g(A_s)\right\}\right) \tag{1.10}$$

from which an approximate MLE, called MCMLE, is obtained.

One major disadvantage of the MPLE over the MCMLE is that model degeneracy is not automatically detected. This is because the estimation of the MPLE does not require the simulation of networks. Model degeneracy occurs when the model that is defined by specific values of θ puts most of its probability mass on just a few possible networks, usually either the empty or the full network (Handcock, 2003; Schweinberger, 2011). For this reason, the choice of network statistics that are included is of particular importance. Adding unsuitable statistics can lead to a degenerate model that returns an estimate of θ which is not generating networks that are similar to the observed graph. Handcock et al. (2008) demonstrate the problem of model degeneracy by discussing an ERGM that counts the number of edges and accounts for a cluster statistic that divides the number triangles by the number of two-stars in the network. The resulting probability distribution defined by the MLE turns out to be bimodal; one mode produces networks with a small number of edges and high clustering, while the other one defines networks with a large number of edges, but little clustering. Even though this model does, on average, produce networks with the same statistics as the observed network, due to the bimodality it almost never creates networks with statistics that are actually similar to the observed network. The authors conclude that a degenerate model indicates that either the MLE does not exist or that if it exists, it does not provide a good fit to the data.

Figure 1.1 visualizes a bimodal distribution that was obtained from a model of a network on 9 nodes that accounts for the number of edges and triangles. If an observed network has 20 edges and 21 triangles then the network was chosen small enough that the exact MLE can be calculated at $(-0.94269, 0.5069)$. Darker areas indicate higher probability. The resulting distribution favors networks with roughly 12 edges and 3 triangles, and networks with 35 edges and 77 triangles. On average, this yields networks

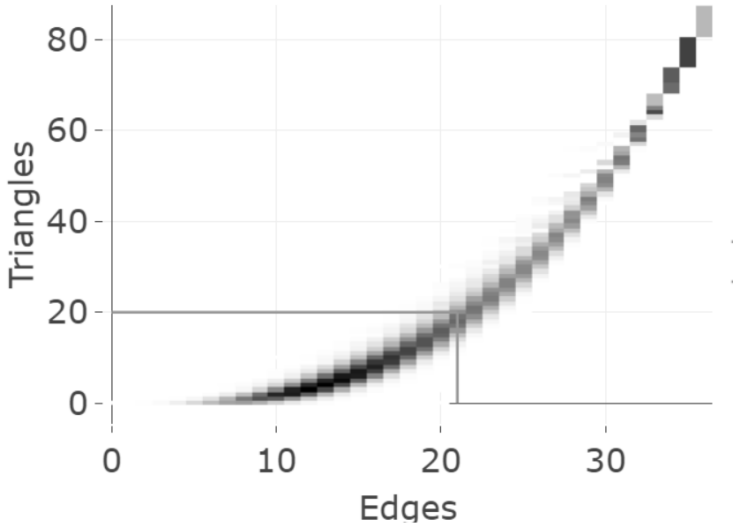

Figure 1.1. Distribution defined by the MLE of a 9-node network with 20 edges and 21 triangles. Darker gray indicates higher probabilities.

with 20 edges and 21 triangles, however, sampled networks do in general not resemble the original network.

The solution to model degeneracy is to specify a model that better fits the network's dependency structure. For this reason, Snijders et al. (2006) and Hunter and Handcock (2006) introduced curved ERGMs, a generalization of the ERGM that can account for a network statistic's entire distribution in a single statistic. Implementing these network statistics makes a model more robust against model degeneracy, yet still does not guarantee a non-degenerate model.

1.3 Contrastive Divergence

In the next two sections, I will demonstrate that the MLE and the MPLE are estimators of a larger class of estimators and that these two can in fact be seen as two endpoints of a continuum of potential estimators.

Contrastive divergence, first introduced by Hinton (2002), is an iterative likelihood estimation technique that has already produced a vast amount of literature in the machine learning community (Carreira-Perpiñán and Hinton, 2005; Tieleman and Hinton, 2009; Asuncion et al., 2010).

The ERGM's normalizing constant uses the sum of all possible networks on N nodes, a number so excessively large for even networks on a relatively small number of nodes

that direct maximum likelihood estimation is generally computationally infeasible. This problem results in parameter estimation for ERGMs that is usually accomplished by MPLE or MCMLE. MPLE is simple and fast, but has the reputation of performing badly in certain cases (Geyer and Thompson, 1992; van Duijn et al., 2009), while the MCMLE is approximately exact, meaning that for a larger MCMC sample size the MCMLE is expected to be closer to the MLE, but requires an increasing amount of computing power.

Contrastive divergence is an iterative optimization algorithm for a fixed step size n that I will denote as CD-n. For iteration t , $t \in \{1, \ldots, T\}$, the algorithm updates θ_t using a sample size of S networks that were all sampled by starting from the observed network A^{obs} and using n Gibbs updates. Hyvärinen (2006) shows that choosing a step size of $n = 1$ is a stochastic approximation of the MPLE, while Carreira-Perpiñán and Hinton (2005) prove that CD-∞, i.e., letting the Gibbs update theoretically run to infinity, results in the MLE. As a consequence, one can see the MPLE and MCMLE as two endpoints of a continuum within the contrastive divergence framework and we expect CD-n to result in an estimate somewhere between the MPLE and MCMLE (Asuncion et al., 2010).

The CD-n algorithm is defined the following way:

1. Choose θ_0; set the number of iterations T; set the number of sampled networks for each iteration S

2. Sample a new network A_s by starting with the observed network A^{obs}

   (a) Choose a dyad ij either randomly or systematically and update the dyad via Gibbs sampling using equation (1.5) and θ_{t-1}

   (b) Repeat (a) n times

3. Repeat the previous step S times

4. Approximate $\ell(\theta_t) - \ell(\theta_{t-1})$ with (1.10) using the S sampled networks and obtain θ_t via Fisher scoring

5. Repeat steps 2–4 T times.

As we can see by this algorithm, the main difference of CD-n compared to the MCMLE algorithm is that the MCMC chain for each sampled network A_s is started at the observed network A^{obs}. Furthermore, the first MCMC iterations are usually discarded due to burn-in, meaning that a large number of MCMC iterations have been calculated

before obtaining a sampled network. For this reason, usually only one MCMC chain is used for the MCMLE. CD-n, however, requires S separate MCMC chains, one for each newly sampled network, and each sample is started from the observed network A^{obs}. Consequently, step 3 is embarrassingly parallel and can therefore easily be parallelized to lessen the computational burden.

Additionally, the MCMLE's main idea is to reweight the sample of S networks from $P_{\theta_0}(Y = A)$ through importance sampling to approximate the likelihood function that leads to the estimation of $\hat{\theta}_{MLE}$. Theoretically, $\hat{\theta}_{MLE}$ is obtained after one iteration, but this algorithm can be repeated to obtain a more numerically stable estimate. CD-n, however, simulates S networks, each generated using n Gibbs updates, always starting from the observed network A^{obs}. In contrast to the MCMLE, the initially estimated $\hat{\theta}_1$ is not the final CD-n estimate, but is used to sample another S networks in the following iteration. This procedure is repeated T times until the final estimate $\hat{\theta}_{CD_n}$ is obtained.

For a step size greater than $n = 1$, $\hat{\theta}_{CD_n}$ is expected to be closer to the MLE than the MPLE and each CD-n iteration itself advances more towards $\hat{\theta}_{MLE}$. Therefore, Krivitsky (2017) proposes contrastive divergence for the MCMLE algorithm to obtain an improved starting value θ_0, compared to the conventionally used MPLE. Krivitsky shows that the additional computational burden for this new starting value is minimal, but that it has the potential to increase the success rate of the MCMLE algorithm.

I define CD-∞ as the above algorithm, when exact samples from the current model are obtained due to sufficiently many MCMC iterations. This would result in perfectly sampled networks and thus, the MLE. But how is CD-1 related to the MPLE? Consider the pseudologlikelihood

$$p\ell(\theta) = \sum_{ij \in \Omega(N)} \Big(A_{ij}\theta(\Delta A)_{ij} - \log(1 + \exp(\theta \cdot (\Delta A)_{ij})) \Big), \tag{1.11}$$

where $(\Delta A)_{ij} := g(A^+_{ij}) - g(A^-_{ij}))$ is the change statistic. Thus,

$$\begin{aligned} p\ell(\theta) - p\ell(\theta_0) \quad = \quad & \sum_{ij \in \Omega(N)} \Big(A_{ij}(\theta - \theta_0)(\Delta A)_{ij} \\ & - \log(1 + \exp(\theta(\Delta A)_{ij}) + \log(1 + \exp(\theta_0(\Delta A)_{ij})) \Big). \end{aligned} \tag{1.12}$$

The likelihood for CD-1, on the other hand, is given by

$$cd_1\ell(\theta) = \theta g(A) - \log(k_{cd1}(\theta)), \tag{1.13}$$

where the normalizing constant $k_{cd1}(\theta)$ is the sum over all networks that differ from the observed network by no more than one dyad. The CD-1 estimate $\hat{\theta}_{CD_1}$ is the vector that maximizes (1.13). Applying the same trick as in (1.9) and focusing on just one dyad ij for simplicity, one can show

$$\begin{aligned}
\frac{k_{cd1}(\theta)_{ij}}{k_{cd1}(\theta_0)_{ij}} &= \frac{\exp(\theta' g(A_{ij}^+)) + \exp(\theta' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \\
&= \frac{\exp(\theta_0' g(A_{ij}^+))}{\exp(\theta_0' g(A_{ij}^+))} \cdot \frac{\exp(\theta' g(A_{ij}^+))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \\
&+ \frac{\exp(\theta_0' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^-))} \cdot \frac{\exp(\theta' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \\
&= \sum_{A \in \{A_{ij}^+, A_{ij}^-\}} \exp((\theta - \theta_0)' g(A)) \cdot \left(\frac{\exp(\theta_0' g(A))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \right) \\
&= \sum_{A \in \{A_{ij}^+, A_{ij}^-\}} \exp((\theta - \theta_0)' g(A)) \cdot P_{\theta_0}(Y = A) \\
&= \mathbb{E}_{\theta_0 | Y_{ij}^c = A_{ij}^c} \Big[\exp((\theta - \theta_0)' g(Y)) \Big], \qquad (1.14)
\end{aligned}$$

where $\mathbb{E}_{\theta_0 | Y_{ij}^c = A_{ij}^c}$ defines the expectation with respect to the conditional distribution of Y_{ij}, given A_{ij}^c. This yields

$$cd_1\ell(\theta) - cd_1\ell(\theta_0) = \sum_{ij \in \Omega(N)} (\theta - \theta_0)' g(A_{ij}) - \log(\mathbb{E}_{\theta_0 | Y_{ij}^c = A_{ij}^c}[\exp((\theta - \theta_0) g(Y)]). \quad (1.15)$$

This means that showing (1.12) to be equal to (1.15) proves that CD-1 equals to a stochastic MPLE optimization.

Proof that (1.12) = (1.15)

Case 1: $A = A_{ij}^+ \;\Rightarrow\; A_{ij} = 1$

$$\begin{aligned}
& cd_1 k(\theta)_{ij} - cd_1 k(\theta_0)_{ij} \\
= \;& (\theta - \theta_0)' g(A_{ij}^+) - \log\left(\frac{\exp(\theta' g(A_{ij}^+)) + \exp(\theta' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \right) \\
= \;& (\theta - \theta_0)' g(A_{ij}^+) - (\theta - \theta_0)' g(A_{ij}^-) + (\theta - \theta_0)' g(A_{ij}^-) - \log\left(\frac{\exp(\theta' g(A_{ij}^+)) + \exp(\theta' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \right) \\
= \;& (\theta - \theta_0)' (\Delta A)_{ij} + \log\left(\frac{\exp(\theta g(A_{ij}^+))}{\exp(\theta_0 g(A_{ij}^+))} \right) - \log\left(\frac{\exp(\theta' g(A_{ij}^+)) + \exp(\theta' g(A_{ij}^-))}{\exp(\theta_0' g(A_{ij}^+)) + \exp(\theta_0' g(A_{ij}^-))} \right)
\end{aligned}$$

$$
\begin{aligned}
&= (\theta-\theta_0)'(\Delta A)_{ij} + \log\left(\frac{\exp(-\theta_0 g(A^+_{ij}))}{\exp(-\theta g(A^+_{ij}))}\cdot\frac{\exp(\theta_0' g(A^+_{ij}))+\exp(\theta_0' g(A^-_{ij}))}{\exp(\theta' g(A^+_{ij}))+\exp(\theta' g(A^-_{ij}))}\right)\\
&= (\theta-\theta_0)'(\Delta A)_{ij} + \log\left(\frac{\exp(1+\theta_0(\Delta A)_{ij})}{\exp(1+\theta(\Delta A)_{ij})}\right)\\
&= A_{ij}(\theta-\theta_0)'(\Delta A)_{ij} - \log\Big(1+\exp(\theta'(\Delta A)_{ij})\Big) + \log\Big(1+\exp(\theta_0'(\Delta A)_{ij})\Big)\\
&= p\ell(\theta)_{ij} - p\ell(\theta_0)_{ij}
\end{aligned}
$$

Case 2: $A = A^-_{ij} \;\Rightarrow\; A_{ij} = 0$

$$
\begin{aligned}
& cd_1 k(\theta)_{ij} - cd_1 k(\theta_0)_{ij}\\
&= (\theta-\theta_0)' g(A^-_{ij}) - \log\left(\frac{\exp(\theta' g(A^+_{ij}))+\exp(\theta' g(A^-_{ij}))}{\exp(\theta_0' g(A^+_{ij}))+\exp(\theta_0' g(A^-_{ij}))}\right)\\
&= (\theta-\theta_0)' g(A^-_{ij}) - (\theta-\theta_0)' g(A^-_{ij}) + (\theta-\theta_0)' g(A^-_{ij}) - \log\left(\frac{\exp(\theta' g(A^+_{ij}))+\exp(\theta' g(A^-_{ij}))}{\exp(\theta_0' g(A^+_{ij}))+\exp(\theta_0' g(A^-_{ij}))}\right)\\
&= 0 + \log\left(\frac{\exp(\theta g(A^+_{ij}))}{\exp(\theta_0 g(A^+_{ij}))}\right) - \log\left(\frac{\exp(\theta' g(A^+_{ij}))+\exp(\theta' g(A^-_{ij}))}{\exp(\theta_0' g(A^+_{ij}))+\exp(\theta_0' g(A^-_{ij}))}\right)\\
&= A_{ij}(\theta-\theta_0)'(\Delta A)_{ij} + \log\left(\frac{\exp(\theta g(A^+_{ij}))}{\exp(\theta_0 g(A^+_{ij}))}\right) - \log\left(\frac{\exp(\theta' g(A^+_{ij}))+\exp(\theta' g(A^-_{ij}))}{\exp(\theta_0' g(A^+_{ij}))+\exp(\theta_0' g(A^-_{ij}))}\right)\\
&\;\vdots\\
&\;\vdots \quad \text{just like Case 1}\\
&\;\vdots\\
&= A_{ij}(\theta-\theta_0)'(\Delta A)_{ij} - \log\Big(1+\exp(\theta'(\Delta A)_{ij})\Big) + \log\Big(1+\exp(\theta_0'(\Delta A)_{ij})\Big)\\
&= p\ell(\theta)_{ij} - p\ell(\theta_0)_{ij}
\end{aligned}
$$

1.4 Composite Likelihoods

We have seen that contrastive divergence is an algorithm with the full likelihood and the pseudolikelihood at two ends of a continuum. While CD-1 leads to MPLE, CD-∞ leads to MLE. *Composite likelihoods* are estimation techniques, first introduced by Lindsay (1988), that also put the MLE and the MPLE at two opposing ends of a spectrum. Lindsay defines a composite likelihood function as the product of conditional probabilities $P(Y_{U_c} = A_{U_c} | Y_{V_c} = A_{V_c})$ where A_{U_c} and A_{V_c} are two mutually exclusive subgraphs of network A, both with a non-zero probability.

Let $\Omega(N) := \{ij \mid i, j \in \{1, \ldots, N\},\ i \neq j\}$ be defined as the set of all possible dyads on N nodes, excluding pairs consisting of the same node. Then, any subset $U_c \subset \Omega(N)$, $U_c \neq \emptyset$, with $V_c = \Omega(N) \setminus U_c$ satisfies the definition of a conditional likelihood. Note that V_c is not required to be the complement of U_c, but can be defined as any subset of $\Omega(N)$, as long as $U_c \cap V_c = \emptyset$. However, I will focus on the special case where $V_c = \Omega(N) \setminus U_c$, which results in a likelihood also known as *conditional composite likelihood*.

As an example for a composite likelihood function in the ERGM framework, define $c = ij \in \Omega(N)$, $U_c = \{c\}$ and $V_c = \Omega(N) \setminus U_c$. This results in a composite likelihood defined as the product of probabilities of ties (i, j), conditioned on the rest of the network. This composite likelihood function turns out to be the pseudolikelihood function

$$c\ell(\theta) = \sum_{c \in \Omega(N)} \log P_\theta(Y_{U_c} = A_{U_c} | Y_{V_c} = A_{V_c}) = \sum_{ij \in \Omega(N)} \log P_\theta(Y_{ij} = A_{ij} | Y_{ij}^c = A_{ij}^c). \quad (1.16)$$

On the contrary, if we let $U_c = \Omega(N)$, and as a consequence $V_c = \emptyset$, then this results in a composite likelihood that is equal to the full likelihood

$$c\ell(\theta) = \log P_\theta(Y_{U_c} = A_{U_c} | Y_{V_c} = A_{V_c}) = \log P_\theta(Y = A | \emptyset). \quad (1.17)$$

We can see from both equations, (1.16) and (1.17), that the conditional composite likelihood has the pseudolikelihood as well as the full likelihood as two special cases. It is, however, even more interesting that the pseudo- and full likelihood functions are defined as two opposing ends of the conditional composite likelihood's spectrum. While the full likelihood accounts for dependencies in the entire network, the pseudolikelihood only accounts for dependencies within the block of a single dyad, which simply means assuming independence among dyads. However, more general forms of conditional composite likelihoods can account for dependencies within blocks of a network, resulting in an estimate somewhere between the MPLE and MLE. The more a network's dependency structure is accounted for by these composite likelihood blocks, the more similar we expect the maximum composite likelihood estimate (MCLE) to be to the MLE. On the other hand, if these blocks hardly capture any dependency structures in the network, then the MCLE is expected to be more similar to the MPLE. Liang and Jordan (2008) show that the MCLE's variance is smaller than the MPLE's, but larger than that of the MLE, which means that the MCLE can be seen as a trade off between the computationally fast, but inaccurate, MPLE and the computationally expensive, but efficient, MLE. Also

note that for dyad-independent networks, one has MPLE=MCLE=MLE.

The full advantage of the MCLE is revealed if a network's entire dependency structure is captured by the composite likelihood's blocks. In this case, MCLE=MLE, but for the MCLE we don't have to consider the full likelihood function.

1.5 Overview of Dissertation

In this dissertation, I examine the theory of MCMLE and MPLE estimation techniques for ERGMs by 1) improving the starting values of the MCMC algorithm to successfully find an approximate MLE and by 2) correctly approximating standard errors of the MPLE. In Chapter 2, the pseudolikelihood's seemingly unfavorable property of not satisfying the likelihood principle is turned into an advantage to improve starting values for MCML estimation. This technique, however, is not an universal remedy that always guarantees an improved starting value. Nevertheless, chapter 3 provides an example of an ERGM that could not be estimated using any other known MLE estimation technique except the approach described in Chapter 2. Chapters 4 and 5 investigate why MPLE standard errors are considered unreliable and provide two different approaches to fix it. Chapter 6 rounds up this dissertation by summarizing the conclusions of the previous chapters and discussing some remaining and newly emerged open questions. In summary, this dissertation develops methodology for improving estimation for both approximate MLE and exact MPLE, and applies these techniques to networks studied in the field of political science. In addition, it provides empirical evidence suggesting that consistent estimation for ERGMs, despite the work of Shalizi and Rinaldo (2013), which conclude that most ERGMs can not be embedded into a normal asymptotic framework, is possible. Even though many properties of dyad-dependent models are still unknown, proving this result would potentially be a very important result.

It should be noted that since Chapters 2-5 are four standalone papers, there are discrepancies in the notation. I decided to preserve the fidelity of the papers since they have already been published or submitted, rather than changing the notations for this dissertation.

Chapter 2
Improving ERGM Starting Values Using Simulated Annealing

This paper was written in collaboration with David R. Hunter and submitted for publication. The content and ideas in this paper were developed together. All simulations and visualizations were created by myself. I also did the majority of the writing.

Much of the theory of estimation for exponential family models, which include exponential-family random graph models (ERGMs) as a special case, is well-established and maximum likelihood estimates in particular enjoy many desirable properties. However, in the case of many ERGMs, direct calculation of MLEs is impossible and therefore methods for approximating MLEs and/or alternative estimation methods must be employed. Many MLE approximation methods require alternative estimates as starting points. We discuss one class of such alternatives here. The MLE satisfies the so-called "likelihood principle," unlike the MPLE. This means that different networks may have different MPLEs even if they have the same sufficient statistics. We exploit this fact here to search for improved starting values for approximation-based MLE methods. The method we propose has shown its merit in producing an MLE for a network data set and model that had defied estimation using all other known methods.

2.1 Maximum Likelihood Estimation for ERGMs

Given a network-valued random variable A and a set of sufficient statistics $T : \mathcal{A} \to \mathbb{R}^q$, $a \mapsto (T_1(a), \ldots, T_q(a))$, the exponential-family random graph model (ERGM) takes

the form

$$P_\theta(A = a) = \frac{\exp\left\{\theta^\top T(a)\right\}}{k(\theta)} \qquad \text{for } a \in \mathcal{A}, \tag{2.1}$$

where $\mathcal{A}$ is the sample space of allowable networks and $\theta \in \mathbb{R}^q$ is a vector of parameters. In many applications, $\mathcal{A}$ denotes the entire set $\{a \in \mathbb{R}^{N\times N}, a_{ij} = a_{ji} \in \{0, 1\}, a_{ii} = 0\}$ of possible undirected networks on N nodes, while in other applications $\mathcal{A}$ may be constrained to be a proper subset of this set. (If we drop the requirement that $a_{ij} = a_{ji}$, then the sample space consists of directed networks.) The sufficient statistics $T(a)$ play a central role in the model, since they enable the inclusion of traditional *exogenous* covariates like a node's age as well as *endogenous* statistics, i.e., statistics that allow for inference on the structure of the network. Popular endogenous statistics are a network's number of triangles or the number of pairs of ties that share one common node (two-stars). The normalizing constant

$$k(\theta) := \sum_{a^* \in \mathcal{A}} \exp\left\{\theta^\top T(a^*)\right\}, \tag{2.2}$$

a weighted sum over all possible networks in the sample space, assures that (2.1) defines a probability model.

Maximizing the log-likelihood function

$$\ell(\theta) = \theta^\top T(A^{\text{obs}}) - \log k(\theta) \tag{2.3}$$

of the ERGM results in the maximum likelihood estimator, or MLE. The MLE can be difficult to calculate directly due to the required calculation of $k(\theta)$, a sum which is only feasible for particular choices of $T(a)$ for which the sum may be simplified or sample spaces small enough to allow direct calculation. Even for a small number of nodes, the sample space can be prohibitively large; for instance, there are over 3.5×10^{13} networks on just $N = 10$ nodes, and an additional node inflates this number by a factor of over 1000. It is therefore often necessary to use an approximation method when an MLE is desired, as exact calculation is not possible.

An alternative method of estimation, known as maximum pseudolikelihood estimation (MPLE), has the desirable property that it is relatively easy to calculate even in cases where an exact MLE is elusive. This article first discusses approximate maximum likelihood estimation and then summarizes how MPLE works and explains why, even in cases where an MLE is desired, the approximation method often begins by calculating the MPLE. Then, in Section 2.4, we explain a novel method to augment the approximation

of the MLE by exploiting the failure of the MPLE to satisfy the so-called likelihood principle. We demonstrate the use of this idea in Section 2.5.

2.2 Approximate MLE and Exact MPLE

An expedient to the problem of maximizing the often-intractable likelihood function (2.3) is given by the Markov chain Monte Carlo maximum likelihood estimator (MCMLE). First proposed by Geyer and Thompson (1992) and then adapted to the ERGM framework by Snijders (2002) and Hunter and Handcock (2006). The MCMLE is based on the idea that for any $\theta_0 \in \mathbb{R}^q$,

$$\frac{k(\theta)}{k(\theta_0)} = \mathbb{E}_{\theta_0} \exp\left\{(\theta - \theta_0)^\top T(A)\right\}, \tag{2.4}$$

where the expectation is taken assuming that the random A has the distribution P_{θ_0}.

For a given θ_0, Equation (2.4) may be exploited by sampling a large number of networks $A_1, \ldots, A_L$ from the distribution P_{θ_0}. Then one may approximate and optimize the difference of log-likelihood functions

$$\ell(\theta) - \ell(\theta_0) \approx (\theta - \theta_0) \cdot T(A) - \log\left(\frac{1}{L}\sum_{i=1}^{L} \exp\left\{(\theta - \theta_0)^\top T(A_i)\right\}\right).$$

In theory, the MCMLE algorithm works for any starting value θ_0, however, Hummel et al. (2012) show that approximation (2.2) is best when θ is close to θ_0, and furthermore the approximation can degrade badly when these parameter values are not close. For this reason, in practice it is necessary to choose θ_0 close to a maximizer of $\ell(\theta)$ or else the MCMLE idea fails.

The most common choice of θ_0 is the maximizer of the so-called pseudolikellhood function. To construct the pseudolikelihood, we focus on the individual values of the tie indicators A_{ij}. A straightforward calculation shows that for a particular i, j, the conditional distribution of A_{ij} given the rest of the network A^c_{ij} is calculated from Equation (2.1) to be

$$P_\theta(A_{ij} = 1 | A^c_{ij} = a^c_{ij}) = \text{logit}^{-1}\left[\theta^\top (T(a^+_{ij}) - T(a^-_{ij})\right], \tag{2.5}$$

where the inverse logit function is defined by $\text{logit}^{-1}(x) = \exp\{x\}/(1 + \exp\{x\})$ and the networks a^+_{ij} and a^-_{ij} are formed by setting the value of a_{ij} to be one or zero, respectively, while fixing the rest of the network at a^c_{ij}. We define $(\Delta a)_{ij} := T(a^+_{ij}) - T(a^-_{ij})$ and refer

to $(\Delta a)_{ij}$ as the i, j vector of change statistics. We may therefore rewrite Equation (2.5) as

$$P_\theta(A_{ij} = 1 | A^c_{ij} = a^c_{ij}) = \text{logit}^{-1}\left[\theta^\top (\Delta a)_{ij}\right]$$

for all i, j. Under the additional assumption that the A_{ij} are all independent of one another, these equations together define a logistic regression model, and the maximum likelihood estimator for this logistic regression is known as the maximum pseudo-likelihood estimator (MPLE) because the assumption of independence is not justified in all cases and therefore the logistic regression likelihood function is sometimes misspecified. Despite this misspecification, the MPLE is frequently used as an estimator of θ because it is easily obtained using logistic regression software. Indeed, the MPLE has a lengthy history in the literature on ERGMs; see Schmid and Hunter (2021) for more details.

In particular, there is a substantial literature on the use of MPLE as an estimator in its own right. For example, Schmid and Hunter (2021) argue that when its covariance is properly estimated, the MPLE can allow for valid statistical inference just as the MLE can. On the other hand, for the most part it is assumed (see, e.g., van Duijn et al., 2009) that MLE is preferable to MPLE. Indeed, one might prefer MLE to MPLE simply based on the classical principle, as articulated by John Tukey for example, "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (Tukey, 1962). For the purpose of this article, the MPLE will serve merely as a value θ_0 used in Approximation (1.10) and we assume that the ultimate goal is to obtain an approximate MLE via MCMLE.

2.3 Comparison of MLE with MPLE

The theory of exponential family models is well-established in general; Barndorff-Nielsen (1978) gives an extensive book-length treatment of this theory. As a particularly useful example in the context of ERGMs, for exponential family distributions the expectation of the T vector under the MLE equals the sufficient statistic on the observed network, i.e., $\mathbb{E}_{\hat{\theta}}[T(A)] = T(A^{\text{obs}})$. In other words, the MLE equals the method of moments estimator. This aligns with one's general expectation that the distribution described by an estimate should on average describe the observed network. Furthermore, this fact provides a useful means for checking that a potential maximizer of the approximate log-likelihood function is in fact close to the MLE, as one may simulate networks from the distribution derived from the MLE and check that their sample mean is approximately equal to $T(A^{\text{obs}})$.

2.3.1 The Likelihood Principle

Another property of the MLE is that it satisfies the likelihood principle (Barnard et al., 1962; Birnbaum, 1962), which states that all information in the data relevant to the model parameters is contained in the likelihood function. In the ERGM context, this means that an estimator should depend on A^{obs} only through $T(A^{\text{obs}})$, as the likelihood itself depends on A^{obs} only through $T(A^{\text{obs}})$. This means that two networks with the same sufficient statistic will yield the same MLE. However, as Corander et al. (1998) point out, the MPLE does not satisfy the likelihood principle. This observation forms the basis of the remainder of this article.

The failure of the MPLE to satisfy the likelihood principle means that two networks A^1 and A^2 may have different MPLEs even if $T(A^1) = T(A^2)$. We see this fact illustrated in Figure 2.1, which depicts numerous possible MPLE values, each of which results from a network on 9 nodes with the same values of T. Also depicted in Figure 2.1 is the mean value parameter space, an alternative to the natural parameter space $\mathbb{R}^q$, which is defined as the interior of the sufficient statistic's sample space. Each parameter θ can be uniquely projected to a point g in mean value parameter space by a bijective function $\mu : \mathbb{R}^q \to \mathcal{C}$. For exponential family distributions, this function is defined as $\mu(\theta) = \nabla \log k(\theta) = \mathbb{E}_\theta[T(Y)]$. In other words, a parameter θ's corresponding point in mean value space is the expectation of the sufficient statistic with respect to the distribution defined by θ. This means that $T(A^{\text{obs}})$ is the MLE's projection into mean value parameter space, which is an interesting fact in its own right, since the MLE in natural parameter space is hard to find, but finding its counterpart in mean value space is trivially easy.

More specifically, Figure 2.1 depicts the MLE and multiple MPLEs in the natural as well as in mean value parameter space of networks on $N = 9$ nodes with exactly 18 edges and 13 triangles. The dashed lines in the lower plot visualize the boundary of the convex hull of the mean value parameter space. The networks were sampled from the (ρ, σ, τ)-model of Frank and Strauss (1986) with σ taken to be zero, which results in an ERGM with a 2-dimensional T vector. We fixed $\rho = -1$ and $\tau = 0.53$, which is equivalent to

$$P_\theta(A = a) \propto \exp\left\{-1 \cdot (\#\text{ edges}) + 0.53 \cdot (\#\text{ triangles})\right\}. \tag{2.6}$$

With only 9 nodes, it is computationally feasible to enumerate all possible $T(a)$ vectors for $a \in \mathcal{A}$, so it is possible to calculate the MLE exactly in this case. Schmid and Hunter

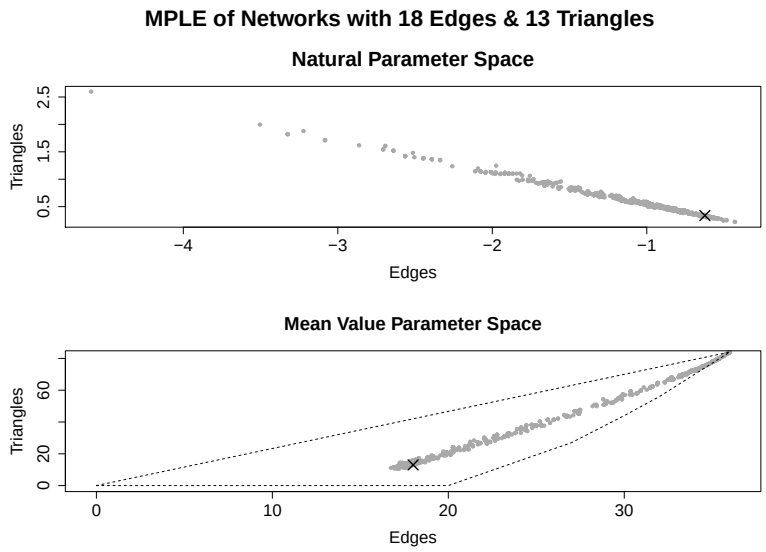

Figure 2.1. Every grey dot represents the MPLE of a network on $N = 9$ nodes with 18 edges and 13 triangles using Model 2.6. The 'X' visualizes the MLE.

(2021) demonstrate how to achieve this enumeration using the `ergm` package (Handcock et al., 2019) for the R computing environment (R Core Team, 2020). Although it is possible to enumerate all network statistics according to their multiplicity in this 9-node example, it is not computationally feasible to calculate every possible MPLE resulting from a network with 18 edges and 13 triangles. Thus, Figure 2.1 uses the simulated annealing method explained in Section 2.4 to generate multiple such networks randomly.

2.3.2 Rao-Blackwellization

Another potential way to exploit the failure of the MPLE to satisfy the likelihood principle is via the so-called Rao-Blackwellization of the MPLE. The Rao-Blackwell theorem (Lehmann and Casella, 1998) states that any estimator that is not a function of the sufficient statistic $T(a)$ is inadmissible, meaning that for any estimator $\tilde{\theta}$ that is not a function of $T(a)$, one can find an improved estimator θ^* by taking the expectation of $\tilde{\theta}$ conditional on $T(a)$. This new estimator has lower risk than $\tilde{\theta}$ by any convex lost function, e.g., mean squared error.

In principle, computing the Rao-Blackwellized MPLE would require the distribution of the MPLE conditional on $T(A^{\text{obs}})$. To put this idea into practice, a sample from this distribution could be obtained using, say, the simulated annealing method of Section 2.4. However, since the stochastic properties of the simulated annealing algorithm are not

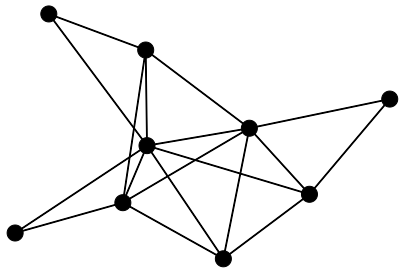

Figure 2.2. A network with 18 edges and 13 triangles.

well understood, we have not explored this particular idea in the current manuscript.

2.4 New Starting Values Via Simulated Annealing

As described earlier, maximum likelihood estimation by MCMC requires the selection of an auxiliary parameter θ_0. Even though in theory the algorithm operates with any choice of θ_0, in practice, a parameter close to the MLE is essential for the MCMLE algorithm to work successfully. Hummel et al. (2012) demonstrate on the basis of a simple Erdös-Rényi model that given an unfavorable starting value it can be very difficult to find the MLE. If the observed statistics do not fall inside the interior of the convex hull of the network statistics simulated from the model defined by θ_0, then the approximate likelihood function has no maximizer and the algorithm will not finish anywhere near the MLE. Theoretically, any θ_0 could produce a sample that includes the observed statistic inside the convex hull, but in practice getting such a sample can be extremely improbable if θ_0 and the MLE are far apart. The commonly used starting value for θ_0 is the MPLE, since it is simple to calculate via logistic regression; even though other methods have been proposed (e.g., Krivitsky, 2017), MPLE remains the predominant method. As shown in Figure 2.1, however, the MPLE can in some cases be very different from the MLE, which typically results in the failure of the MCMLE algorithm. In our 9-node example, the MCMLE algorithm failed for about a third of the MPLEs as starting values shown in Figure 2.1.

2.4.1 Failure of the MCMLE algorithm: An Illustrative Example

We demonstrate the problems of the MPLE as starting value for the MCMLE algorithm by taking a particular network on 9 nodes with 18 edges and 13 triangles as depicted in Figure 2.2. The MPLE of this particular network is $\tilde{\theta} = (-1.3, 0.702)$, while the MLE can be exactly calculated as $\hat{\theta} = (-0.623, 0.337)$. Figure 2.3 shows MCMLE results for the same network, where the algorithm was initialized by different $\tilde{\theta}$ values at each trial from among the values depicted in Figure 2.1. In four out of ten trials, the algorithm stopped due to model degeneracy. As first sketched by Handcock (2003) and later studied in detail by Schweinberger (2011), degeneracy occurs when the distribution defined by $P_{\hat{\theta}}$—ostensibly the best possible parameter value, in some sense—places most of its probability mass on just a few networks, usually the empty and the full network. In this scenario, the simulated networks are so different from the observed network that the MCMLE algorithm fails. In six cases shown in Table 2.3, the algorithm provided an MCMLE. Yet three trials yielded estimates further away from the MLE than the MPLE, leading to estimates in natural parameter space close to the boundary of the convex hull and therefore clearly different from the true MLE based on the observed network. The only glimmer of hope is trial 5, which leads to an estimate somewhat similar to the MLE. This example shows that it is necessary to check that an MLE is producing values of T, on average, close to $T(A^{\text{obs}})$. Once this takes place, additional techniques such as those detailed in Hummel et al. (2012) may be employed to fine-tune the MLE. However, even using such techniques, many of the trials from Table 2.3 did not end successfully.

2.4.2 Simulated Annealing and MPLE

As noted earlier, Hummel et al. (2012) show that a poorly chosen θ_0 can make successful MCML estimation almost impossible. Consequently, the more different the MPLE is from the MLE, the more difficult it is to find the MCMLE, since the MPLE is commonly taken as the algorithm's starting value θ_0 due to its simple and fast calculation. Inspired by Figure 2.1, we propose a novel approach for finding an improved starting value for the MCMLE algorithm, namely, to search for networks that yield the same—or at least nearly the same— statistics as the observed network, then consider these networks' MPLEs as potential starting values.

We propose searching for networks with the same statistics as the observed network using the *simulated annealing* algorithm (Kirkpatrick et al., 1983). This algorithm was initially inspired from the practice of annealing metal, a procedure in which the material

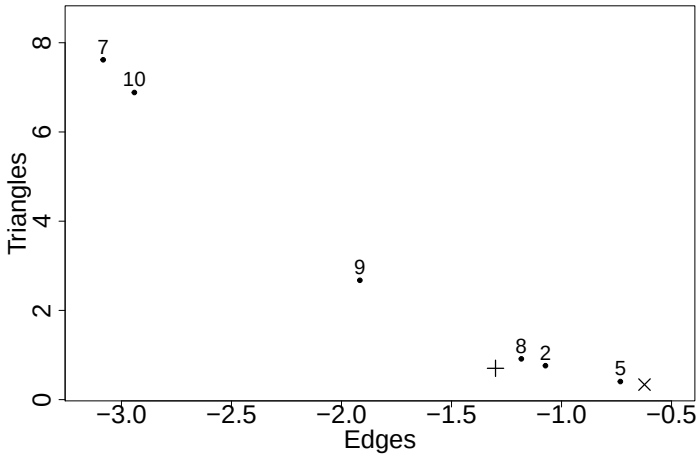

Figure 2.3. MCMLEs of the network in Figure 2.2 using Model (2.6) for ten independent trials. '+' indicates the MPLE, 'x' the MLE.

is heated and then cooled very slowly, allowing it to be gradually shaped into the desired form. The process of heating up and cooling down is translated into the simulated annealing approach by allowing interim results initially (during the "heated" phase) to be worse than previous results; that is, the algorithm is more stochastic than deterministic in its early phase. Gradually, the algorithm "cools," becoming more deterministic and less prone to random jumps. The hope is that the algorithm does not get stuck at local maxima that it otherwise might not be able to leave if a purely deterministic optimization algorithm were used. By starting in an initial random phase followed by increasingly deterministic behavior, the algorithm can find and then focus on areas of the search space that include globally optimal solutions.

The simulated annealing algorithm for networks is implemented through the *san* function in the `ergm` package. The goal is to minimize the energy function, which is defined as

$$E_W(a) = (T(a) - T)^\top W(T(a) - T)$$

where T is the target statistic and W is a symmetric, positive definite matrix of weights. For a given temperature t the algorithm is then defined as

1. If $E_W(a) = 0$ exit
2. Generate random network a^*

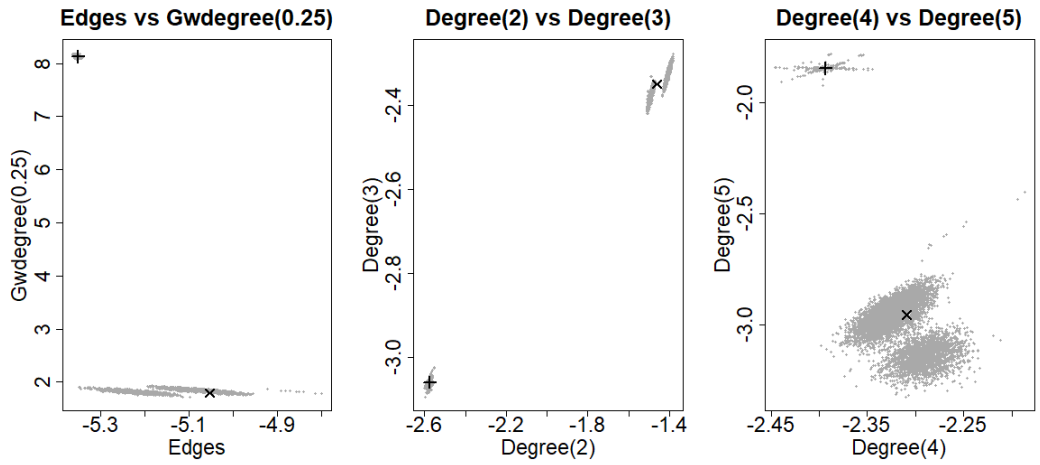

Figure 2.4. MPLEs of networks with the same sufficient statistics as the observed E. coli network. The MPLE of the observed network is marked with an '+', the MCMLE with an 'x'.

3. Calculate acceptance probability

$$\alpha = \exp[-(E_W(a^*) - E_W(a))/t]$$

4. Update t and W, and repeat.

For more details, we refer to Krivitsky et al. (2021).

2.5 Applications

We demonstrate the simulated annealing-based approach to MCMLE using the E. coli transcriptional regulation network of Shen-Orr et al. (2002), which is based on the RegulonDB data of Salgado et al. (2001). The nodes in this network represent operons, while an edge from operon i to j indicates that i encodes a transcription factor that regulates j. Even though this is originally a directed network that contains self-edges, i.e., operons that regulate themselves, we follow Saul and Filkov (2007) and treat this network as undirected and without self-edges. This results in a network with 519 edges and 418 nodes. We study the same ERGM on these data as Hummel et al. (2012), a model which yields MCMLEs considerably different from the MPLE, making estimation difficult. The model's statistics consist of the number of edges; the numbers of nodes with degrees two, three, four, and five; and the geometrically weighted degree distribution with the decay parameter fixed at 0.25. As demonstrated by Hummel et al. (2012), initializing the MCMLE algorithm with the MPLE does not produce successful results.

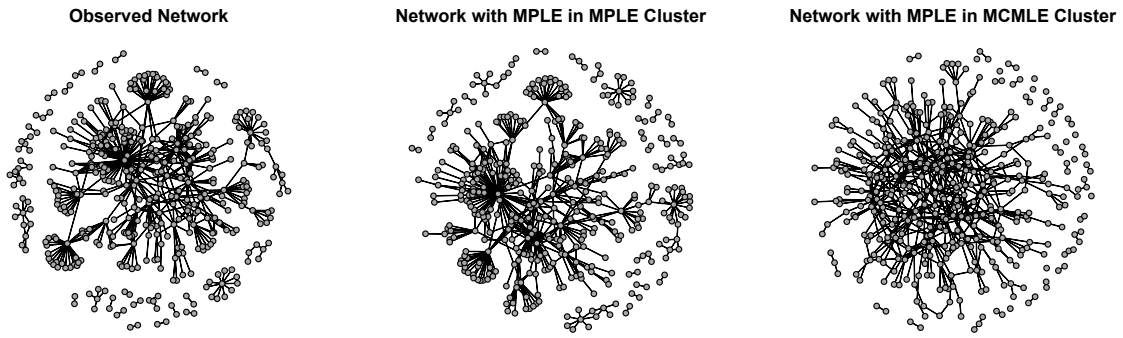

Figure 2.5. The left visualizes the E.coli transcriptional regulation network of Shen-Orr et al. (2002). The center depicts a network whose MPLE falls into the MPLE cluster of Figure 2.4. The right depicts a network whose MPLE falls into the MPLE cluster of Figure 2.4. All three networks have the same sufficient statistics vector.

Figure 2.4 depicts the MPLEs of 10,000 networks (in grey) that have the same sufficient statistics as the original network, including the MPLE of the original network as well as the MCMLE. All $10,000$ networks were obtained using the simulated annealing algorithm, always starting with the observed network. It is remarkable that the MPLEs of these networks essentially form two clusters, one cluster that includes the MPLE and another one that includes the MCMLE of the observed network. Among other things, this figure illustrates why finding the MCMLE by beginning the algorithm at the observed network's MPLE is a difficult task. The MPLE is simply not close enough to the MLE for the approximate log-likelihood using an MPLE-based sample of networks to be effective. On the other hand, setting the starting value to one of the MPLEs that form the cluster around the MCMLE evidently makes the algorithm more likely to succeed.

A natural question that arises when considering Figure 2.4 is how the networks in the two clusters differ. Figure 2.5 visualizes networks of each cluster as well as the original E. coli network. Even though all three networks have the same sufficient statistics, it is possible to discern that the network in the MPLE cluster maintains some of the distinctive structures of the original network, while the network in the MCMLE cluster is more reminiscent of an Erdös-Rényi network, i.e., a network where each tie occurs independently with same probability p. The fact that the occurrence of a tie in an Erdös-Rényi network does not depend on the occurrence or absence of any other tie results in a model in which the MPLE and the MLE are the same. Consequently, we conjecture that it is advantageous to find a network that resembles an Erdös-Rényi network and that, in addition, yields the same sufficient statistics as the observed network.

Simulating networks from the probability distribution defined by the obtained MCMLE

results in networks that resemble the rightmost network in Figure 2.5, rather than the observed network. This consequently casts doubt on whether the statistics of this ERGM appropriately capture the unique structure of the observed network. Stated differently, if the MLE, ostensibly the gold standard among estimators, yields unsatisfying results, the model itself should be reconsidered. It is important to remember that in such cases, as indeed in this case, generally the MPLE also fails to result in simulated networks resembling the original network; thus, a poor model cannot generally be mended by using a different estimation technique.

We repeat the simulation study that resulted in Figure 2.4, with the exception that we start the simulated annealing algorithm with an Erdös-Rényi-generated network with similar density to the original network, i.e., where p is defined as the ratio of the number of ties to the number of possible ties. With this modification, the simulated annealing algorithm only finds networks where the MPLE is in the proximity of the MCMLE, meaning that the MPLE of any of these networks is an improved starting value for the MCMLE algorithm compared to the original MPLE.

In summary, we suggest finding an improved starting value θ_0 the following way:

1. For a given network A^{obs}, find the sufficient statistics $T(A^{\text{obs}})$
2. Simulate a network G using an Erdös-Rényi model on the same nodes as A^{obs} and with p given by the edge density of A^{obs}
3. Simulate a new network A^* with $T(A^{obs}) = T(A^*)$ with simulated annealing and start with the algorithm at G. If $T(A)$ includes continuous statistics, find a network A^* with $T(A) \approx T(A^*)$
4. Take the MPLE of A^* as the improved starting value θ_0

This approach was successfully applied by Schmid et al. (2021a) to a citation network based on all US Supreme Court majority opinions written between 1937 and 2015. The majority opinion of each court case is considered a node, and a citation from case i to case j is defined as a directed tie. The network includes 10,020 nodes and the ERGM used includes 14 endogenous and exogenous statistics—complicated enough to be computationally challenging. Ultimately, Schmid et al. (2021a) find that multiple different approaches fail to yield a successful MCMLE, among them the stepping method of Hummel et al. (2012) and the standard MCMLE algorithm using the MPLE as starting value. Yet the approach described here—which reproduces the observed network statistics

only approximately due to the inclusion of several continuous variables among these statistics—does produce a successful MCMLE; it is the only known method that does so in this example. This, however, does not mean that the simulated annealing method assures an improved starting value for all model. In fact, there is yet no method that works for all cases. It is possible to construct examples where the proposed method is not useful, but at this point it is not clear what distinguishes cases that are amenable to the simulated annealing method to those that are not.

2.6 Discussion

The basic idea of this article—searching for networks that have the same, or approximately the same, ERGM statistics as the observed network and then using the MPLEs from these networks as starting values for a traditional MCMLE algorithm—is relatively simplistic. Yet it has already demonstrated its value in solving previously unsolvable ERGM-based estimation problems.

It seems there is still much to learn about this methodology. For instance, is there a way to implement Rao-Blackwellization, and does this approach lead to estimates that are closer in their behavior to the MLE? Also, how important is it to search the sample space of all networks thoroughly, and is an alternative to simulated annealing possible for this purpose? Figure 2.4 suggests that initializing the simulated annealing algorithm at the observed network and the Erdös-Rényi-generated network generate wholly distinct clouds of MPLE values, which leads to the question of whether *all* possible MPLE values for a particular set of statistics is actually bimodal, or whether in fact there is a vast as-yet-unexplored set of MPLE values that could potentially be of use both as starting MCMLE values and as sample points in a Rao-Blackwellization scheme.

That such a simple idea can prove so effective relative to all other known methods suggests that there exists immense untapped potential for improving upon approximate likelihood-based inference using MPLEs.

Chapter 3
Generative Dynamics of Supreme Court Citations: Analysis with a New Statistical Network Model

This paper was written with Ted H. Chen and Bruce A. Desmarais and has been accepted for publication in Political Analysis. All authors contributed equally to the development of the model and the write-up of the paper. My main contributions were the merging of the various data sets, the coding and estimation of the models, the visualization of the results, and the creation of the resulting R-package ***cERGM****.*

The significance and influence of US Supreme Court majority opinions derive in large part from opinions' roles as precedents for future opinions. A growing body of literature seeks to understand what drives the use of opinions as precedents through the study of Supreme Court case citation patterns. We raise two limitations of existing work on Supreme Court citations. First, dyadic citations are typically aggregated to the case level before they are analyzed. Second, citations are treated as if they arise independently. We present a methodology for studying citations between Supreme Court opinions at the dyadic level, as a network, that overcomes these limitations. This methodology—the citation exponential random graph model, for which we provide user-friendly software—enables researchers to account for the effects of case characteristics and complex forms of network dependence in citation formation. We then analyze a network that includes all Supreme Court cases decided between 1950 and 2015. We find evidence for dependence processes, including reciprocity, transitivity, and popularity. The dependence effects are

as substantively and statistically significant as the effects of exogenous covariates, indicating that models of Supreme Court citation should incorporate both the effects of case characteristics and the structure of past citations.

3.1 Introduction

United States Supreme Court opinions exercise authority and influence, in part, through their roles as precedents affecting future jurisprudence in the US. The findings regarding the nature of the influences of precedent on the Supreme Court have been mixed, but the balance of the literature finds that past decisions exert some form of influence on the justices' decision making (Knight and Epstein, 1996; Gillman, 2001; Richards and Kritzer, 2002; Hansford and Spriggs, 2006; Bailey and Maltzman, 2008, 2011; Pang et al., 2012). Despite a considerable body of research that focuses on how precedents shape decision making on the Court, relatively little work has focused on understanding which past opinions are cited by an opinion. Our focus in this paper is to provide what is, to our knowledge, the first comprehensive analysis of exactly which cases are cited in an opinion. We follow an emerging body of work on legal citations, and treat the system of citations as a network (e.g., Caldeira, 1988; Fowler et al., 2007; Fowler and Jeon, 2008; Bommarito II et al., 2009; Lupu and Voeten, 2012; Pelc, 2014; Ethayarajh et al., 2018).

We are not the first to ask what predicts the citations in US Supreme Court Opinions. Indeed, a voluminous body of work has sought to explain how many times an opinion is cited (e.g., Cross, 2010; Benjamin and Desmarais, 2012; Fix and Fairbanks, 2019), when an opinion is cited (e.g., Black and Spriggs, 2013; Spriggs and Hansford, 2001), and how many cases are cited by an opinion (e.g., Lupu and Fowler, 2013)—all focused on the US Supreme Court. One common feature of the research design in all of these studies is that the observations are at the case or case-year level. The outcome variables in these analyses are defined as measures of the number of citations to a case over a period of time, the number of citations to a case at a particular time, or a measurement on the cases cited by a case. These are case-level studies in that, based on the unit of analysis, it is impossible to determine both the origin and target case of a citation that contributes to the dependent variable.

An alternative approach to case-level analysis of citations would be to model them in the directed dyadic form through which they arise. A case decided at time t can cite (or not cite) each case decided previously, and in the US Supreme Court, each other case decided at time t. We are aware of one prior study, Clark and Lauderdale (2010), in which

a statistical model is used to analyze directed dyadic citations between cases. However, Clark and Lauderdale (2010) use a dyadic latent variable model in order to estimate ideal points for Supreme Court Opinions, but do not use this model to understand the relationships between explanatory variables and the formation of citation ties between opinions. We build upon the literature on citation analysis both methodologically and substantively. Methodologically, we develop a novel extension of a statistical model for networks, which we adapt to the network structural constraints of court citations. Second, we apply this methodology to a half-century of directed dyadic citations between U.S. Supreme Court opinions.

There are two benefits of analyzing citations at the directed dyad level. The first is that directed dyadic analyses can test both dyadic and case-level hypotheses. For example, case-level analyses can model whether opinions supported by a liberal majority coalition are more likely than those supported by a conservative majority coalition to be cited heavily in the future, but they cannot precisely model whether liberal cases will be cited more by liberal cases than by conservative cases. Thus, the first reason for analyzing citations at the dyadic level is to expand the set of hypotheses that can be tested. The second reason for studying citations at the directed dyadic level is that, as articulated in the growing literature on legal citation networks, citations form complex networks in which a citation at one point in time may influence future citations. This phenomenon of complex dependence is very common in networks of many types, but processes specific to Supreme Court citations create interdependence in citations. For example, if opinion i relies heavily on opinion j as precedent, opinion i is likely to discuss the legal basis for opinion j, and as a consequence, cite some of the opinions cited by opinion j. Suppose opinion k is cited by opinion j. Opinion k is more likely to be cited by opinion i because opinion i relies heavily on j, and opinion j cites k. This is a special case of a very common process on networks referred to as "triad closure". Complex dependence is theoretically interesting on its own merits, but the effects of covariates cannot be reliably identified—either in terms of coefficient values or standard errors—without accounting for the interdependence inherent in networks (Cranmer and Desmarais, 2016).

In this paper we develop a theoretical case that citations on the US Supreme Court are characterized by forms of complex dependence that are common in networks. We then develop an extension of a model—the exponential random graph model (ERGM)—that can incorporate both exogenous covariates and complex forms of interdependence into a directed dyadic analysis of citations. Finally, we develop and estimate a specification of

this model in an analysis of US Supreme Court citations between 1950 and 2015. We find robust support for the inherent complexity underpinning the formation of citation ties, and show that incorporating complex dependence into the model of citation formation significantly improves the model's fit.

We offer three contributions. First, we advance our understanding of the factors that drive citations between U.S. Supreme Court opinions. Second, for those who study judicial citations in general (e.g., U.S. state supreme courts (Hinkle and Nelson, 2016), international courts (Lupu and Voeten, 2012), German lower courts (Berlemann and Christmann, 2020)), we illustrate network-theoretic considerations that are likely to apply beyond the context of Supreme Court opinions. Third, we offer a novel extension of a statistical model for networks, and disseminate this model as a package for the R statistical software (Schmid et al., 2020), which can be used for any form of citations. For example, patent citations are used to measure both causes and consequences of innovation in the field of political economy (Akcigit and Kerr, 2018; Dincer, 2019), citations in academic journal articles and syllabi have been used to study gender bias in political science (Dion et al., 2018; Maliniak et al., 2013; Hardt et al., 2019; Atchison, 2017), and citations in documents produced in the policymaking process have been used to study links between public policy and scientific expertise (Costa et al., 2016; Koontz and Thomas, 2018; Pattyn et al., 2020). Our contributions are both substantive and methodological. We focus most heavily on the first—the study of the U.S. Supreme Court citation network, since we see it as the most substantial innovation with respect to the existing literature that we offer in the current paper.

3.2 Network Processes in Supreme Court Citations

When it comes to the development and testing of theory, the defining feature of networks is that the micro-level unit of analysis—the relationship between two entities (i.e., the citation from one opinion to another) is a component of a complex system of relations. The formation (or lack thereof) of that relationship cannot be fully understood without considering how the relationship fits into the system. Analytical designs that account only for covariates in explaining tie formation are incomplete theoretically, and, as a consequence, are subject to a form of omitted variable bias (Cranmer and Desmarais, 2016). Citations in legal opinions are unique in terms of the windows into network dependencies offered by the texts of the opinions. A number of common structural dependencies that are found in networks are likely to apply to citations in Supreme

Court opinions. In this section we present these dependence forms, and document the mechanisms by which they arise through archetypal passages in example opinions.

We should note that we do not distinguish between positive and negative citations in this theoretical framework. The dynamics we outline are not specific to a particular type of citation. The only distinction we draw (in the empirical analysis) is the instance in which a case has been overruled. In the extreme instance of overruled precedents, we assume that (and test whether) a case is much less likely to be cited after it has been overruled.

The first network property that we theorize in the context of Supreme Court citations is transitivity. In a network of directed relations (e.g., i cites j, but j doesn't cite i) transitivity refers to the tendency for i to send a tie to k if i sends a tie to j and j sends a tie to k (Holland and Leinhardt, 1971; Hallinan and Kubitschek, 1990). In undirected networks, transitivity is simply the process by which friends of friends become friends (i.e., a friend of a friend is a friend). The term, "transitive closure" refers to a tie forming from i to k in response to extant ties from i to j and j to k. When writing opinions, Supreme Court justices present the legal bases for their rulings, which often involves discussing the most primary/relevant precedents underpinning these legal bases, but also the precedents and legal rules on which the primary precedents were based. This process of presenting several layers/levels of precedent in an opinion follows the structure of transitive closure exactly—opinion i cites opinion j as a primary precedent, and then cites opinion k because opinion j cites opinion k. The two examples presented below illustrate this process.

In the first example, a passage from Kansas v. Marsh (548 U.S. 163, 2006)—a case considering the constitutionality of a death sentence statute in Kansas. In this example, the case Stringer v. Black is cited by Kansas v. Marsh as a case that is quoted by Sochor v. Florida. The primary precedent under discussion in this passage of the opinion is Sochor v. Florida, but Stringer v. Black is cited as a result of its role in the Sochor v. Florida opinion.

> The statute thus addresses the risk of a morally unjustifiable death sentence, not by minimizing it as precedent unmistakably requires, but by guaranteeing that in equipoise cases the risk will be realized, by "placing a 'thumb [on] death's side of the scale,' " Sochor v. Florida, 504 U. S. 527, 532 (1992) (quoting Stringer v. Black, 503 U. S. 222, 232 (1992); alteration in original).

The second example, which we illustrate visually in Figure 3.1 is a passage from Seminole

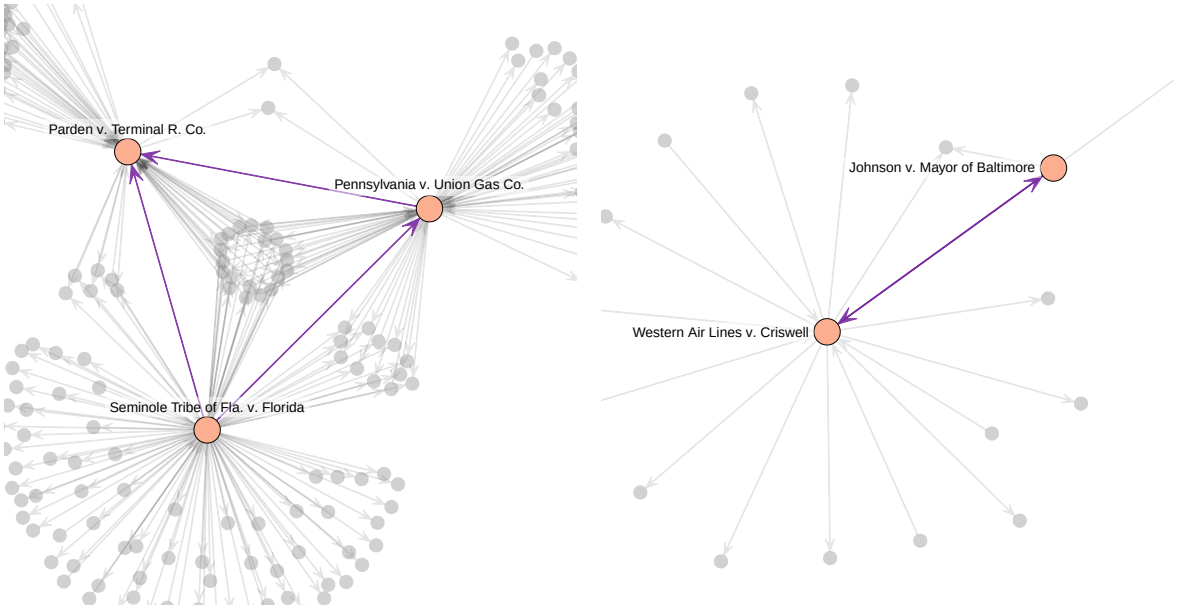

Figure 3.1. Illustrations of transitive triangle connecting US Supreme Court opinions through citations (left) and a reciprocal tie between two US Supreme Court opinions (right).

Tribe of Fla. v. Florida, (517 U.S. 44 1996)—a case addressing the rights of groups and citizens to sue states in federal court. In this example, Pennsylvania v. Union Gas Co (491 U.S. 1, 1989) is the primary precedent being critiqued, and several cases are cited and discussed in terms of their roles as precedents in the Union Gas opinion. We highlight one—Parden v. Terminal R. Co., which is cited and discussed in both the Seminole and Union Gas opinions.

> Never before the decision in Union Gas had we suggested that the bounds of Article III could be expanded by Congress operating pursuant to any constitutional provision other than the Fourteenth Amendment. Indeed, it had seemed fundamental that Congress could not expand the jurisdiction of the federal courts beyond the bounds of Article III. Marbury v. Madison, 1 Cranch 137 (1803). The plurality's citation of prior decisions for support was based upon what we believe to be a misreading of precedent. See Union Gas, 491 U. S., at 40-41 (SCALIA, J., dissenting). The plurality claimed support for its decision from a case holding the unremarkable, and completely unrelated, proposition that the States may waive their sovereign immunity, see id., at 14-15 (citing Parden v. Terminal Railway of Ala. Docks Dept., 377 U. S. 184 (1964)), and cited as precedent propositions that had been merely assumed for the sake of argument in earlier cases, see 491 U. S., at 15 (citing

> Welch v. Texas Dept. of Highways and Public Transp., 483 U. S., at 475-476, and n. 5, and County of Oneida v. Oneida Indian Nation of N. Y., 470 U. S., at 252).'

The second network property that we hypothesize is reciprocity (Erikson, 2013) among cases that are decided in the same term. We expect that opinions that are written within the same term, and cover highly similar cases will cite each other. The first case in our example reciprocal dyad is a passage from Western Air Lines v. Criswell (472 U.S. 400, 1985)—a case considering mandatory retirement in the context of age discrimination laws. The second case in the dyad, Johnson v. Mayor of Baltimore (472 U.S. 353, 1985) is another case considering whether mandatory retirement violates the Age Discrimination in Employment Act. The mutual edge connecting these two cases is visualized in Figure 3.1. These cases addressed very similar legal questions, which increased the likelihood that they would inform each other, and the opinions were written within the same term, which made it possible for them to cite each other.

> *From Western Air Lines:* On a more specific level, Western argues that flight engineers must meet the same stringent qualifications as pilots, and that it was therefore quite logical to extend to flight engineers the FAA's age 60 retirement rule for pilots. Although the FAA's rule for pilots, adopted for safety reasons, is relevant evidence in the airline's BFOQ defense, it is not to be accorded conclusive weight. Johnson v. Mayor and City Council of Baltimore, ante at 472 U. S. 370-371. The extent to which the rule is probative varies with the weight of the evidence supporting its safety rationale and "the congruity between the ... occupations at issue." Ante at 472 U. S. 371. In this case, the evidence clearly established that the FAA, Western, and other airlines all recognized that the qualifications for a flight engineer were less rigorous than those required for a pilot.

> *From Johnson:* The city, supported by several amici, argues for affirmance nonetheless. It asserts first that the federal civil service statute is not just a federal retirement provision unrelated to the ADEA, but in fact establishes age as a BFOQ for federal firefighters based on factors that properly go into that determination under the ADEA, see Western Air Lines, Inc. v. Criswell, post p. 472 U. S. 400. Second, the city asserts, a congressional finding that age is a BFOQ for a certain occupation is dispositive of that determination with respect to nonfederal employees in that occupation.

The third, and final, network property we consider in the context of Supreme Court citations is popularity. Popularity, also termed "preferential attachment" is the tendency for ties to be sent to nodes to which many ties have already been sent (Barabási and Albert, 1999; Chayes, 2013). Citations to an opinion signal both the Court's awareness of the legal reasoning of the case and the Court's evaluation that the opinion is an authoritative precedent. The more citations, the stronger this signal. Landmark cases, or those that establish new legal rules, are particularly authoritative and accrue citations from most future opinions that follow the respective line of reasoning. The passage below, from Oregon v. Mitchell (400 U.S. 112, 1970)—a case on the legality of state age restrictions on voting in federal elections—illustrates this popularity dynamic. In this opinion passage Baker v. Carr is cited in reference to its role as a landmark precedent, and noted for the number of other cases by which it has been followed. and for which an authoritative opinion is referenced, and even discussed in terms of the number of other cases by which it was followed. The language in this passage suggests that the attention to Baker v. Carr in previous Court opinions is in part responsible for its authority in Oregon v. Mitchell. The citations to Baker v. Carr are visualized in Figrue 3.2.

> The first case in which this Court struck down a statute under the Equal Protection Clause of the Fourteenth Amendment was Strauder v. West Virginia, 100 U. S. 303, decided in the 1879 Term. [Footnote 2/1] In the 1961 Term, we squarely held that the manner of apportionment of members of a state legislature raised a justiciable question under the Equal Protection Clause, Baker v. Carr, 369 U. S. 186. That case was followed by numerous others, e.g.: that one person could not be given twice or 10 time the voting power of another person in a state-wide election merely because he lived in a rural area..."

These three properties—transitivity, reciprocity, and popularity—form the core of our network theory of U.S. Supreme Court citations. We also seek to model the effects of covariates (i.e., case features) on citation formation. Next we introduce a modeling framework that can incorporate all of these effects into a single model, and adapt to the structural constraints of citation networks.

3.3 The Citation Exponential Random Graph Model

We develop a methodology that can be used to jointly test for the effects of covariates on citations—as have been studied in prior research, and test for dependence effects, as we

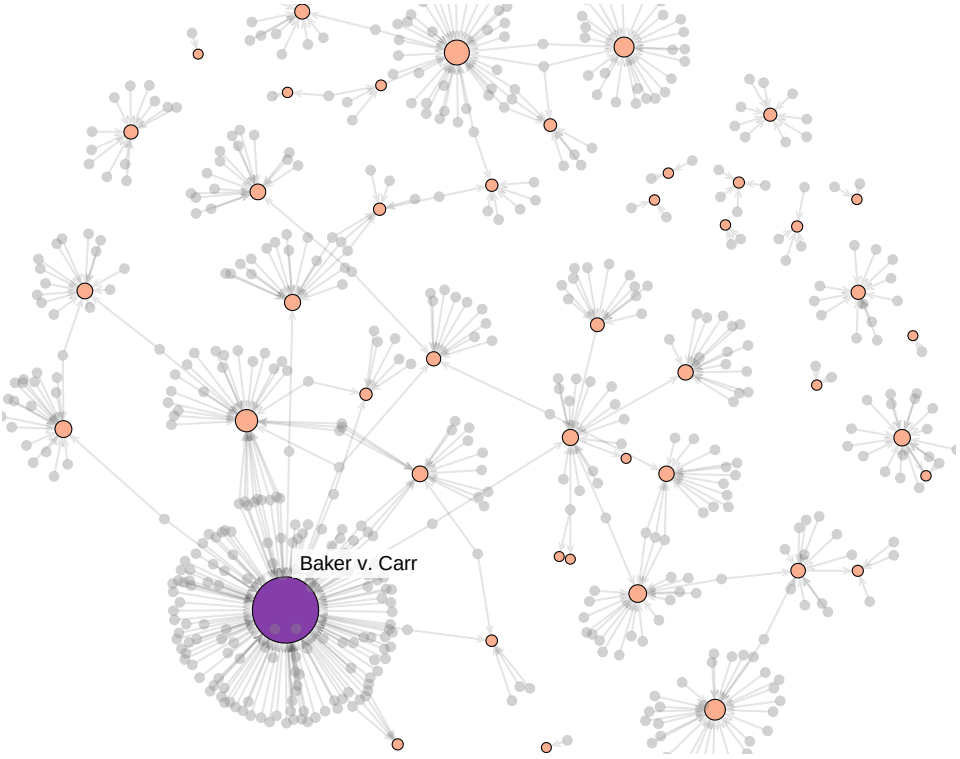

Figure 3.2. Illustration of ties sent to a landmark Supreme Court opinion via citations.

hypothesize above. To accomplish this, we extend a model that has been developed to jointly represent covariate and dependence effects in network data, and has seen extensive application in recent political networks research—the exponential random graph model (ERGM) (e.g., Bratton and Rouse, 2011; Box-Steffensmeier and Christenson, 2014; Duque, 2018; Osei, 2018). The network structures for which ERGMs are currently designed are insufficient to account for the structure of citation networks.

We develop the citation ERGM (cERGM), to account for the structural constraints that apply to the network of Supreme Court citations. These structural constraints amount to three departures from the structure of networks for which ERGMs are currently designed. First, the citation network is partially acyclic. If two cases are decided during the same term, they can cite each other, forming a mutual edge (or two-cycle). However, if case i is decided before case j, case j can cite case i, but case i cannot cite case j. Second, new edges can be created over time, but cannot be eliminated. Unlike in, e.g., an alliance network, in which two countries can dissolve an alliance, once a citation exists in a citation network it cannot be dissolved. Third, the set of nodes in the network must increase for new edges to be created. In conventional ERGMs, the number of nodes in the network can increase or decrease in each time period, and is typically stable over

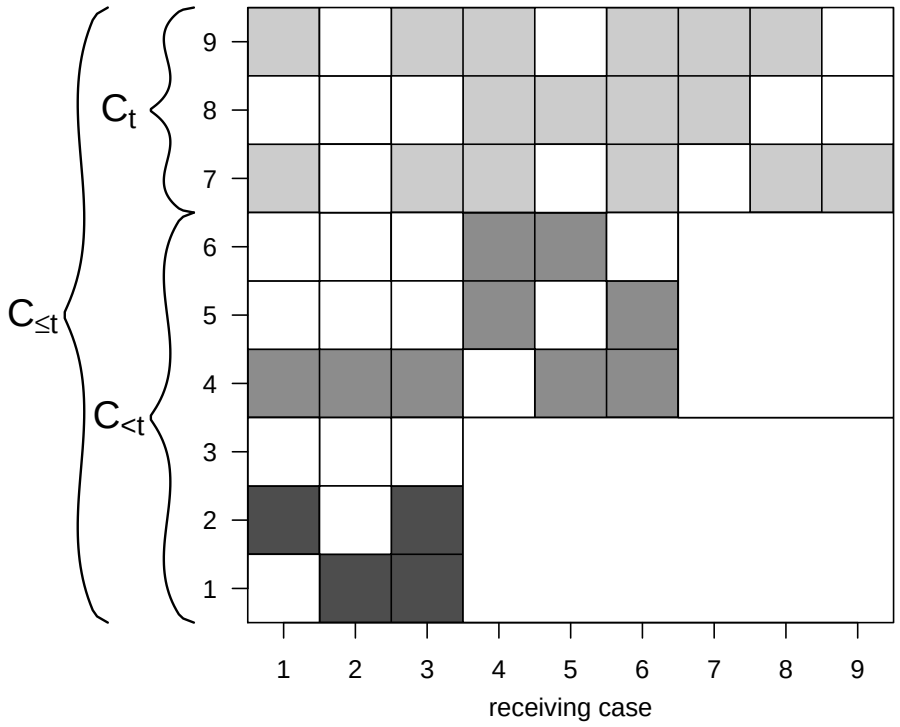

Figure 3.3. Illustration of temporal structure of the Supreme Court Citation Network. $C_{\leq t}$ is the entire set of citations (and non-citations) on which citations and non-citations at time t (i.e., C_t) depend. C_t are conditioned on the citations and non-citations established before time t (i.e., $C_{<t}$). The shaded small squares are hypothetical observed citations, and the white small squares are citations that could have been observed but were not. The regions of the matrix that are represented by large white rectangles are citations that could not have been observed since the citing case would have been decided in a term that preceded the term of the cited case. The citing case ID is given in the row and the prospective cited case is given in the column.

time. In a citation network, new edges (and non-edges) are introduced over time via the introduction of new nodes. The structure of the citation network is depicted in Figure 3.3—a hypothetical citation network established over three time periods, with three cases decided in each time period. We denote C_t to be the set of citation and non-citations added to the network at time t (i.e., via the addition of three cases), $C_{<t}$ to be the citations among cases decided before time t, and $C_{\leq t}$ to denote the entire set of citations and non-citations on which C_t can depend through the cERGM specification.

The likelihood function of the cERGM is given by

$$l(\boldsymbol{\theta}, C_{\leq t}) = \frac{\exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t, C_{<t})\right]}{\sum_{C_t^* \in \mathcal{C}_t} \exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t^*, C_{<t})\right]}, \tag{3.1}$$

where $\boldsymbol{\theta}$ is a vector of real-valued parameters, and $\boldsymbol{h}(C_t, C_{<t})$ is a vector of scalar-valued functions that each quantify a feature of the citation network (e.g., the relationship between citation ties and a case attribute, the number of mutual edges in the citation network). $\exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t, C_{<t})\right]$ is a positive weight that is proportional to the probability of observing any particular form of the citations and non-citations added to the network at time t. The denominator of Equation 1 represents a normalizing constant, in which the positive weight is summed over all possible configurations of C_t, from the network in which the cases added to the network at time t send no citations at all, to the network in which the cases added at time t cite every possible case, and everything in between.

Though the likelihood function of the cERGM is quite different from those of conventional regression models, analysis of the conditional probability of a single citation from case i to case j reveals that we can interpret the parameters similar to logistic regression coefficients.

$$P(C_{ij,t} = 1 | C_{-ij,t}, C_{<t}) = \frac{\exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 1)\right]}{\exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 1)\right] + \exp\left[\boldsymbol{\theta}'\boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 0)\right]},$$

$$= \frac{1}{1 + \exp\left[-\boldsymbol{\theta}'\left(\boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 1) - \boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 0)\right)\right]},$$

where $C_{ij,t} = 1$ indicates that case i cites case j, $C_{ij,t} = 0$ indicates that case i does not cite case j, $C_{-ij,t}$ is the observed elements of C_t except $C_{ij,t}$, and

$$\left(\boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 1) - \boldsymbol{h}(C_t, C_{<t} | C_{ij,t} = 0)\right)$$

is the change in $\boldsymbol{h}(C_t, C_{<t})$ that results from toggling $C_{ij,t} = 0$ to $C_{ij,t} = 1$. This rearrangement illustrates that the parameters can be interpreted in terms of the change in the log odds of a citation from i to j given a one-unit increase in the corresponding element of $\boldsymbol{h}$, conditional on the other citations observed in the network. For example, if the value of θ corresponding to an element of $\boldsymbol{h}$ that counts the number of mutual edges in the network is 0.5, then the log odds of observing $C_{ij,t} = 1$ increases by 0.5 if case i is cited by case j (as compared to the configuration in which case j does not cite case i). The logit form conditional probability is well known for the ERGM family (Goodreau et al., 2009).

3.4 Empirical Analysis

Our three data sources for this study include the Supreme Court Database (SCDB) (Spaeth et al., 2014), Martin-Quinn scores (Martin and Quinn, 2002), and Supreme Court citation data provided by the CourtListener Free Law Project (Lissner and Carver, 2010). In the next section we explain the variables we construct using these data sources[1]. We limit the Supreme Court terms included in our analysis to those that are covered by all three of these data source (1937–2015).[2] The Supreme Court citation network from $1937 - 2015$ consists of $10,020$ cases. The breakdown of the data by the Court's Chief Justice is presented in Table 3.1. The network has a total of $112,939$ citation ties.[3] The in- and outdegree distributions (i.e., the distributions of the number of citations sent and received by cases, respectively), are visualized in Figure 3.4. The maximum indegree (i.e., number of cases citing to a case) is 230 and the maximum outdegree (i.e., number of cases cited by a case) is 162. The majority of cases cite to and/or are cited by twenty or fewer other cases.

The degree distributions indicate that there is a long tail to both the number of citations sent and received. These long-tailed (i.e., high kurtosis) distributions provide preliminary evidence of substantial heterogeneity in the features that drive citations to and from cases (Strogatz, 2001). Figure 3.4 displays the citation network network from 1937–2015. We see here that the densest rates of tie formation tend to be between consecutive courts (e.g., the Stone Court is much more tightly tied to the Hughes Court than the Rehnquist Court is to the Hughes Court). This pattern lends preliminary support to the hypothesis that the rate of citations to a case decreases over time.

We fit the cERGM separately for every Supreme Court term between 1950 and 2015, meaning that we have 66 models where the outcome variables are the citations sent during the given term. Our approach of fitting separate term-by-term models is motivated by the temporally-dynamic data generating process. First, as our results will

[1]Data and replication code are provided by Schmid et al. (2021b)

[2]There were 145 cases that were listed in the SCDB but could not be matched to a case in the CourtListener data. We decided to exclude these 145 cases from our analysis. Additionally, since the most commonly used data on Supreme Court citations in political science come from Fowler et al. (2007), we compared the Fowler data to the CourtListener data in the time interval that they overlap (1937–2001). We found considerable agreement—over 95% of the citations recorded in the CourtListener dataset were also in the Fowler dataset, and over 96% of the citations recorded in the Fowler data were also in the CourtListener dataset.

[3]In order to focus on citation actions that were not intended to totally invalidate an opinion, we exclude 315 citations that caused the cited case to be overruled. The data on overruling citation came from US. Congress. Senate (2016). Our results are virtually unchanged if we include the overruling citations.

Table 3.1. For the time range of interest (1937 - 2015) this table displays the chief justices, the time range they served as chief justice, the number of cases in their time range as well as the average number of cases per year.
* CE Hughes served as chief justice from 1930 - 1941.
** J Roberts still serves as chief justice (retrieved 5/2020).

Chief Justice	Terms	Total Number Cases	Cases/Term
CE Hughes*	1937 - 1941	628	125.6
HF Stone	1942 - 1946	756	151.2
FM Vinson	1946 - 1953	789	98.63
E Warren	1954 - 1969	2149	126.41
WE Burger	1970 - 1986	2805	155.83
W Rehnquist	1987 - 2005	2022	106.42
J Roberts **	2006 - 2015	871	87.1

show, the temporal context and justice composition of a court will lead to differences in how opinions are written and therefore how citations are used. Second, based on the findings of Shalizi and Rinaldo (2013), we know that the set of parameter values that best fits a network should change as more nodes are added to the network. This means it would be inappropriate to assume a constant set of parameter values as more cases are added each term. Practically, our dataset is large enough to support this approach, with each term introducing the potential for hundreds-of-thousands of new citations. This gives us the statistical power necessary to estimate a separate set of parameter values for each term. Further, while the entire set of citations can in theory be fit in a single network with varying parameters, it is technically difficult. While the network for the 1950 term consists of a manageable $1,962$ nodes, the size of the network increases to $10,020$ nodes in the 2015 term, making estimation more challenging.

Estimation of the cERGM parameters, which is done with a computationally-intensive, simulation-based approach, is presented in detail in the appendix. The basis of our computation is the **ergm**-package (Hunter , David R. et al., 2008) in **R** (R Core Team, 2020). In addition, we developed a wrapper **R**-package called **cERGM** (Schmid et al., 2020) to conveniently fit the cERGM.

3.4.1 cERGM specification

The cERGM we estimate includes two classes of terms—one that captures the effects of covariates on tie formation, and another that captures the complex dependence processes that we expect to observe in the Supreme Court Citation Network. We describe our model specification by defining the terms within these two classes.

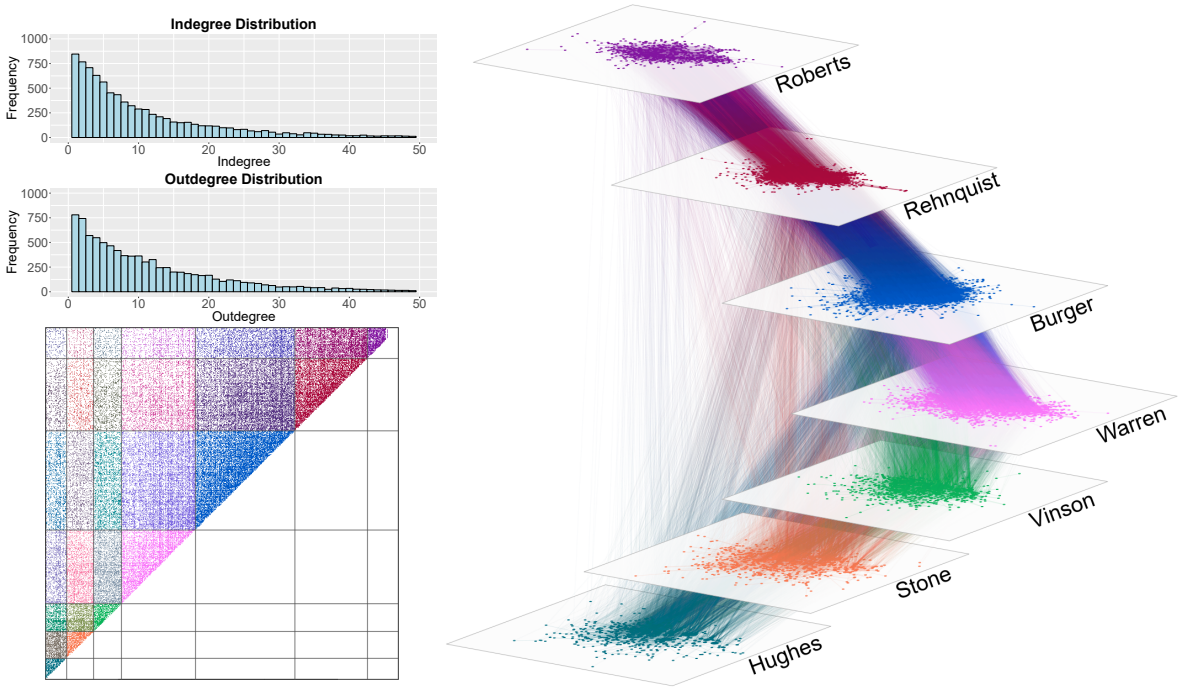

Figure 3.4. Supreme Court Citation Network, 1937-2015. Network visualization on the right. Nodes are Supreme Court cases, color-coded based on the chief justice presiding over the court. On the top left is the in- and outdegree distribution of the network. There are cases with an in- or outdegree >50, but they are not captured in this figure. The bottom left shows the citation data in adjacency matrix format following 3.3.

3.4.1.1 Covariate terms

Covariate effects are accounted for in the cERGM via the term that is used to specify the effect of covariates in other ERGM family models, as

$$h_{covariate}(C_t, C_{<t}, X) = \sum_{ij} C_{ij,t} X_{ij}.$$

Since $C_{ij,t}$ is a binary indicator of whether case i cites case j, $h_{covariate}$ amounts to the sum of covariate values among directed dyads for which we observed a citation. The dyadic interpretation of the coefficient attributed to this term is the change in the log odds of a tie from i to j given a one-unit increase in X_{ij} (i.e., exactly the interpretation of the effect of a covariate in logistic regression). We include several exogenous covariates based on this standard formulation.

The first covariate we incorporate into the model accounts for the degree to which cases cite those that are similar in terms of the ideological positions of the justices who supported the decision. We account for this effect following Spriggs and Hansford (2001),

Table 3.2. Assigned numbers for the variable *Issue Area*. This information originates from the Supreme Court Database code book.

1	Criminal Procedure	8	Economic Activity
2	Civil Rights	9	Judicial Power
3	First Amendment	10	Federalism
4	Due Process	11	Interstate Relations
5	Privacy	12	Federal Taxation
6	Attorneys	13	Miscellaneous
7	Unions	14	Private Action

who find that cases are more likely to be overruled when the Court is ideologically distant from the median justice in the majority coalition that decided the case. Clark and Lauderdale (2010) find that the majority opinion falls at the ideal point of the median member of the majority coalition in the case. We include a covariate term in which X_{ij} is the absolute difference between the Martin-Quinn scores (Martin and Quinn, 2002) of the median justices in the majority coalitions for cases i and j. We expect this variable to have a negative effect, which would mean that cases cite those that are ideologically similar.

We include one covariate that accounts for the rate at which cases that fall under the same issue area cite each other. The variable X_{ij} indicates whether cases i and j share the same issue area, e.g., $X_{ij} = 1$ if cases i and j are classified into the same issue area, and 0 otherwise. The Issue area data comes from the Supreme Court Database (SCDB) (Spaeth et al., 2014). We include these variables because Cross (2010) finds that the number of citations to Supreme Court opinions depends heavily on the issue area of the case. Table 3.2 reports the issue areas covered in our data.

We also include an indicator variable that models the rate at which justices cite themselves. Similar to the same issue area variable, this effect is modeled with a single indicator variable in which $X_{ij} = 1$ if the majority opinions in cases i and j were written by the same justice. The data is taken from the SCDB (Spaeth et al., 2014).

We model the way in which citations to a case depend upon the age of a case. For this we use a second-order polynomial in which one covariate X_{ij} is defined as the age of case j at the time that case i is decided, and another term in which X_{ij} is the squared age of case j at the time that case i is decided. We include these covariates because Black and Spriggs (2013) find that the number of citations to a Supreme Court case over time depends significantly on the age of the case, characterized by a sharp drop off and leveling out with age.

Benjamin and Desmarais (2012) study the propensity for cases to be overruled and

cited in other negative ways. They find that cases with majority coalitions that are large and ideologically broad are less likely to be cited negatively. In our data we do not differentiate between negative and positive citations, but since the overwhelming majority of citations are positive, we hypothesize that the effects they found will be reversed in our analysis. We include one covariate in which X_{ij} is the number of justices in the majority coalition for case j. We also include another covariate in which X_{ij} is the absolute difference between the maximum and minimum ideal points of the justices in the majority coalition for case j. We expect these covariates to have positive effects.

The final control variable we include in the model is one in which $X_{ij} = 1$ if case j was overruled prior to the term in which case i was decided. This variable, quite simply, models the effect of being overruled on the rate of citation to a case after the overruling citation. Fowler and Jeon (2008) find that citations to a case drop off quickly after the case has been overruled.

3.4.1.2 Dependence terms

We include dependence terms to account for each of the dynamics that we hypothesize will characterize the Supreme Court Citation Network—transitivity, reciprocity, and popularity. The common dependence effects for these three dynamics are the number of triangles, the number of mutual dyads, and the number of in-2-stars. A triangle is a configuration in which there are citations connecting all three cases in a triad. From the perspective of a single tie closing a triad, a positive triangle coefficient indicates that a tie is more likely to form between nodes i and j if nodes i and j have citation ties with one or more cases k. A mutual dyad is a configuration in which two nodes within a dyad exchange directed ties. From the perspective of a node closing a mutual dyad with a tie, a positive coefficient indicates that case i is more likely to cite case j if case j has cited case i (mutual dyads are only possible within the same term). The in-2-star configuration is one in which two nodes in a triad send ties to the same third node. From the perspective of a case sending a tie to close an in-2-star, a positive coefficient indicates that case i is more likely to send a tie to case j if one or more other cases also cite case j. Unfortunately, adding the triangle or in-2-star statistic causes model degeneracy for the Supreme Court Citation Network—a common issue with these terms (Handcock, 2003). Model degeneracy is a common obstacle when fitting ERGMs, and occurs when the model places most of its probability mass on just a few networks, usually the empty and the full network. In this scenario, the simulated networks are too different from the observed network, making the underlying distribution defined by the model extremely

unrealistic.

The statistic included in the model for reciprocity counts the number of mutual dyads (i.e., dyads in which cases i cites case j, and case j cites case i) in $C_{\leq t}$. In practice, this is the number of mutual dyads in $C_{\leq t}$, since mutual dyads must emerge among opinions that were written in the same term. The log odds interpretation of the reciprocity statistic is that if the opinion in case i cites case j, the log odds that case j also cites i change by the estimated coefficient relative to the configuration in which case j does not cite case i. We expect the reciprocity effect to be positive.

Accounting for the popularity effect through the in-2-star statistic is prone to cause model degeneracy. In an attempt to stabilize the model against model degeneracy as well as to still capture the popularity effect, Snijders et al. (2006) introduces the *alternating-in-k-star* statistic, which was shown to be equivalent to the *geometrically weighted indegree distribution* (gwidegree) statistic introduced by Hunter and Handcock (2006). The indegree distribution is the distribution of the number of ties sent to nodes, across all nodes in the network. Define $D_r(C_t, C_{<t})$ as the number of nodes with indegree $r, r \in \{0, \ldots, N_t - 1\}$, where N_t is the number of cases in the network at time t, then gwidegree represents a network's indegree distribution in a single statistic by geometrically weighting the degree distribution

$$h_{gwidegree}(C_t, C_{<t}, \lambda) = \lambda \sum_{r=1}^{N_t-1} \left(1 - \left(1 - \frac{1}{\lambda}\right)^r\right) D_r(C_t, C_{<t}). \tag{3.2}$$

The decay parameter $\lambda \in [0, \infty]$ controls the decline of the weight put on each node in the network based on their indegree ($D_r(C_t, C_{<t})$) as the indegree value (r) increases, which means that the higher λ the more the statistic weighs cases with a high indegree. We fixed $\lambda = 1$. We chose this value for two reason. First, it results in a fairly straightforward interpretation of the coefficient for gwidegree, as Equation 3.2 reduces to a function that counts the number of nodes that receive at least one tie. Our second reason for selecting $\lambda = 1$ is that it results in an accurate fit in the indegree distribution of the observed network (as demonstrated in the appendix). At $\lambda = 1$, the weight attributed to each case is equivalent regardless of the indegree value, as long as the indegree is greater than zero. This means that the statistic does not grow with the addition of high indegree cases to the network. It only grows through the addition of cases with at least one in-citation. With $\lambda = 1$, the coefficient associated with the gwidegree statistic governs the degree to which the edges are sent to a small number of popular nodes, or distributed fairly evenly across the nodes. A positive coefficient encourages edges to be distributed evenly across

a large number of nodes—assuring that as many nodes as possible receive at least one tie. A negative gwidegree coefficient discourages network configurations in which many nodes receive ties, placing higher probability instead on configurations in which many of the ties are sent to a relatively small number of popular nodes. As stated by Levy (2016), this means for the interpretation that a negative parameter value indicates the centralization of edges (i.e., popularity) while a positive parameter indicates the dispersion of edges (i.e., new ties going to less popular nodes). Since we expect highly cited cases to be more likely to be cited again, we expect the gwidegree parameter to be negative.

We include two statistics to account for different types of transitivity. The first transitivity statistic calculates the number of transitive ties, $i \rightarrow j$, where there is a directed path of length two from i to j through at least one case k, and j and k were written during a different term than i. We will refer to this statistic as the *different term transitivity* statistic. This statistic captures the central form of transitivity that we hypothesize above—when a justice writes the opinion for case i, cites a past opinion (k) as precedent, and then cites one or more of the opinions (j) that were cited in opinion k. The two-path from opinion i to j is created when opinion i cites an opinion (k) that cites an earlier opinion (j). The closed transitive triangle is created when opinion i cites directly to j, one of the precedents used in opinion k. The edge $i \rightarrow j$ counts as a "transitive tie" since it closes at least one transitive triangle. The *different term transitivity* statistic is the count of the number of transitive ties in the network that connect opinions written in different terms. As illustrated in Section 2, we theorize Supreme Court citations to exhibit a high degree of transitivity. As a result, we expect positive parameter estimates for the *different term transitivity* statistic.

Whereas the first transitivity statistic is focused on modeling triangles that form through the citation of past cases, the second transitivity statistic we include captures clustering that includes ties formed between cases decided in the same term. We both expect and observe a higher level of transitivity among same-term cases since justices have the opportunity to confer about cases and coordinate citations among related opinions. For modeling within-term transitivity, we use the *geometrically weighted edgewise shared partners* (gwesp) statistic introduced by Hunter and Handcock (2006). We use the form of the gwesp statistic that captures configurations in which cases i and j are connected by a citation (i cites j, j cites i, or both), and also cite r common other cases (i.e., shared partners). Butts (2008) call this conceptualization of r the number of *outgoing shared partners*. This form of gwesp captures the degree to which two related cases in the same term are likely to be based on similar bodies of precedent. If the gwesp coefficient is

positive, we interpret the effect as saying that the more shared partners r to which cases i and j send ties, the greater the likelihood that there is a citation connecting cases i and j. As with the gwdegree term, there is a decay parameter. The gwesp statistic exhibits decreasing marginal returns with respect to the number of shared partners (e.g., the fourth shared partner between i and j does not contribute as much to gwesp as does the first shared partner between i and j). The decay parameter governs the degree of marginal return decrease, with a value of 0 indicating that additional shared partners do not contribute anything, and a value of ∞ indicating a linear (non-decreasing) function. We fix the decay parameter at $\lambda = 0.25$, as this accurately captures the edgewise shared partner distribution, and avoids degeneracy (which would be a problem at higher values of λ). In general, researchers can adjust the gwesp parameter down to avoid degeneracy, and increase it to more accurately capture the shared partner distribution in the network. As with cross-term transitivity, we expect the gwesp statistic to also carry a positive coefficient value.

We include one more structural network effect to adjust for inherent differences across cases in terms of the overall scope of the legal issues addressed in a case. Fowler and Jeon (2008) finds that cases vary in the degree to which they serve as "hubs"—citing to many other cases. We expect that new cases will be more likely to cite cases that themselves cited many cases, because cases that themselves cite a large number of cases address a broader array of legal issues. We include the outdegree (i.e., the number of citations sent) of every node that entered the network prior time t as a receiver effect. We expect the coefficient on this term to be positive, as a positive coefficient would mean that cases that send more citations are likely to also receive more citations. This is not technically a dependence term, as it is formulated as a covariate, but we discuss it in the current section because it is purely a network structure effect.

3.4.2 Model Fit

Our case for studying legal citations at the directed dyadic level hinges upon the contribution to modeling offered by incorporating network dependence. To quantify this contribution, we fit one model that incorporates covariate effects only, which we term the *Independent Model*—excluding the reciprocity, transitivity, and popularity terms. The full model is the one we present below. We use both AIC and BIC to compare the fit of the two models, with lower values indicating better fit.

Figure 3.5 depicts the AIC and BIC comparison for the two different models for each term from 1950 to 2015. We see that the AIC and BIC is considerably smaller for the

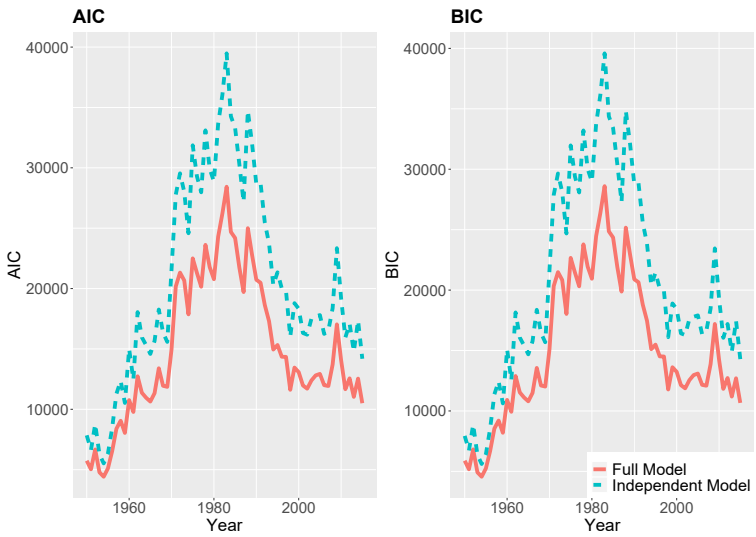

Figure 3.5. AIC and BIC for the full and the independent model for the time frame 1950-2015

full model throughout the period of consideration. On average the independent model yields a 5,525 point higher AIC and an average 5,582 point higher BIC. The maximum difference is reached in 1983 with a difference of 10,990 points for the AIC and 10,930 points for the BIC. This fit comparison provides robust evidence that the dependence terms contribute significantly to the explanatory power of the model. In the appendix we present conventional goodness-of-fit plots that are used to assess the structural fit for ERGM (Hunter et al., 2008), as well as degeneracy diagnostics (Mukherjee et al., 2020). We find that the full model fits the degree and edgewise shared partner distributions well, and is not degenerate.

3.4.3 cERGM Coefficient Estimates

The estimates from the cERGM are presented in Figures 3.6 and 3.7. In these figures we depict coefficient estimates and 95% confidence intervals for all of the effects for all the variables we discussed in section 3.4.1.1. The coefficients in an ERGM family model can be interpreted in the same way as logistic regression coefficients—they give the change in the log odds of a tie from i to j given a one-unit increase in the respective variable. In our figures, a circle indicates that the estimate was significant at an $\alpha = 0.05$ level, while an square translates to a p-value between 0.05 and 0.1. A triangle signals that the estimate for a given year was not significant at an $\alpha = 0.1$ level. In each panel, the background shading indicates different chief justice courts, with the two darker shades representing, respectively, the Warren (1954-1969) and Rehnquist (1987-2005) courts.

Finally, note that the panels showing results of the reciprocity statistic and the *Overruled Cases* variable do not provide estimates for every term. In these missing terms there were no two cases that cited each other or no overruled cases being cited, making the estimate of these terms equivalent to $-\infty$.

3.4.3.1 Dependence Terms

We first discuss the dependence effect estimates in Figure 3.6. In most of the years that it is estimable, the reciprocity effect is statistically significantly positive, and substantial in magnitude. In a typical year, the presence of a citation from case i to case j increases the log odds of a citation from case j to case i by approximately 0.50. Both transitivity effects—different term transitivity and gwesp—are estimated to be positive and statistically significant in every year, and are even greater in magnitude than the reciprocity effect. The log odds of a citation from case i to case j increases by more than 2 if there is at least one third case, k that is cited by i, and cites to j. Results regarding the popularity effect (i.e., gwidegree) are generally supportive of the hypothesis of a popularity effect in Supreme Court citations, as the gwidegree estimate is negative and statistically significant at conventional levels in nearly every term.

Contrary to expectations, we find that the hub effect is negative. That is, each additional citation sent by a case reduces the likelihood that the case is cited. For every citation sent, the log odds of any citation to a case decreases by an average of approximately 0.025—a result that is statistically significant in each term. We initially expected receiver outdegree to have a positive effect because opinions with broad scopes can be useful citation hubs. Instead, our finding indicate that high outdegree more importantly reflects case characteristics that negatively impact incoming citations. One likely characteristic is case complexity. Prior studies show that justices tend to take a more balanced approach when faced with complex cases (Collins Jr, 2008; Lindquist et al., 2007), which can yield opinions that cite widely (Wilkinson III, 2005). At the same time, it has been shown that complex cases are more likely to be overruled (Spriggs and Hansford, 2001). To the extent, then, that outdegree count captures case complexity, our finding suggests that justices cite more direct alternatives in anticipation of complex opinions being overruled.

The negative effect of receiver outdegree, which seems to result from strategic anticipation by the justices, ostensibly decreases with time. Specifically, term to term variation aside, there is an upward trend toward zero in the parameter estimates. We believe this finding is particularly illustrative of why our approach of fitting separate

models for each term is appropriate. As time passes, the number of nodes in the network (i.e., cases) grows. With the increasing number of citable cases and a relatively stable baseline rate of citation (i.e., the edges effect), there is an increase in outdegree mean and variation. This increasing difference between outdegree count by case makes it unlikely for the per unit change effect in outdegree to remain at a constant level. We have then an instance of network growth leading to changes in the model parameters even if the underlying network process remains unchanged. Approaches that cannot encompass these complexities are likely to produce poor fitting models.

3.4.3.2 Covariate Terms

Effects of the exogenous covariates included in the cERGM are presented in Figure 3.7. Most of the exogenous covariate effects align with expectations and have relatively stable parameter estimates over time. We find that cases are more likely to cite those that (1) have been decided most recently, (2) were authored by the same justice, (3) are in the same issue area, and (4) have not been overruled. Surprisingly, we do not find consistent effects for majority coalition size or for the two covariates that involve justice ideology. The effect of majority coalition size was generally not statistically significant until the later portion of the period examined; ideological distance between two cases had signs and significance levels that are highly inconsistent; and the effect of the ideological breadth of the potentially cited case was effectively zero for the entire period.

Our results leave open the question of whether and how citations are shaped by ideological factors. A purely measurement-based explanation of our findings is that we need a more precise measure of the ideological position of an opinion, as far as citation behavior goes, than the median ideal point of the justice in the majority coalition. Clustering that results from ideological homophily that is not effectively accounted for by our measure would be picked up, in part, by our measures of transitivity, which are statistically and substantively significant throughout the period studied.

Beyond potential measurement issues, we do observe connections between our findings and the literature on decision-making on the Supreme Court. These connections are evident from a closer look at temporal dynamics in the effects. In this discussion, we focus on the shift in the effect of ideological distance beginning with the Rehnquist court and extending into the Roberts court.

The existing body of evidence indicates that chief justices impact the courts they preside over beyond casting one of the nine votes (Cross and Lindquist, 2005; Danelski and Ward, 2016). It is therefore within expectations that there are discernible differences in

citation patterns across different courts. With regard to the effect of ideological distance, we see that the Vinson (1950-1953) and Warren (1954-1969) courts were highly volatile, the Burger court (1970-1986) had a generally higher likelihood of citing opinions that were ideologically closer to itself, and the more recent Rehnquist (1987-2005) and Roberts (2006-2015) courts showed some tendency toward citing cases that are ideologically distant. The latter finding might be initially puzzling, but can be explained in light of observations from legal scholars that the Rehnquist court was characterized by a 'split-the-difference' jurisprudence toward its later half (Wilkinson III, 2005; Basiak Jr, 2006). This means that the court strove to take moderate positions that compromised on both sides of the debated issues, which in large differed from the jurisprudential approach of prior courts. Less has been written about the Roberts court on this topic, but Roberts shares many similarities with Rehnquist especially in comparison to previous chief justices, and has himself noted the desirability of widely accepted rulings even at the expense of their scope (Pomerance, 2018; Sunstein, 2008).

One way to achieve the kind of balance required for support is to cite broadly from both sides of the argument (Wilkinson III, 2005), which increases the likelihood of citing ideologically distant cases. Consider for example Hamdi v. Rumsfeld (542 U.S. 1, 2004) which ruled on the question of whether a U.S. citizens can be detained as an "enemy combatant". The court, in a four-justice plurality opinion supported by two additional concurring justices, decided at the same time that "a state of war is not a blank check for the President" and the due process assessment must "pay keen attention to the particular burdens faced by the Executive in the context of military action". In total, the plurality opinion in Hamdi v. Rumsfeld cited 47 opinions, with cites as distant as 2.42 and 2.13 on either side of its median position.

In the above discussion, we identified an explanation for the observed temporal dynamics in the effect of ideological distance. Specifically, ideologically distant cases become more important as courts adopt the relatively new 'split-the-difference' jurisprudential approach which requires broad citations. We recognize that there are alternative explanations—perhaps one that is related to the coterminous change in the effect of majority coalition size—but a deeper look into the relevant literature is beyond the scope of this paper. Our discussion serves to illustrate that term-by-term cERGMs can uncover important temporal dynamics that courts researchers can use to study a much greater range of citation-related phenomenon beyond what we present here. To the extent that the US Supreme Court is a political entity that interacts with the broader society, the body of court opinions and citations between them is corpus that continues to grow

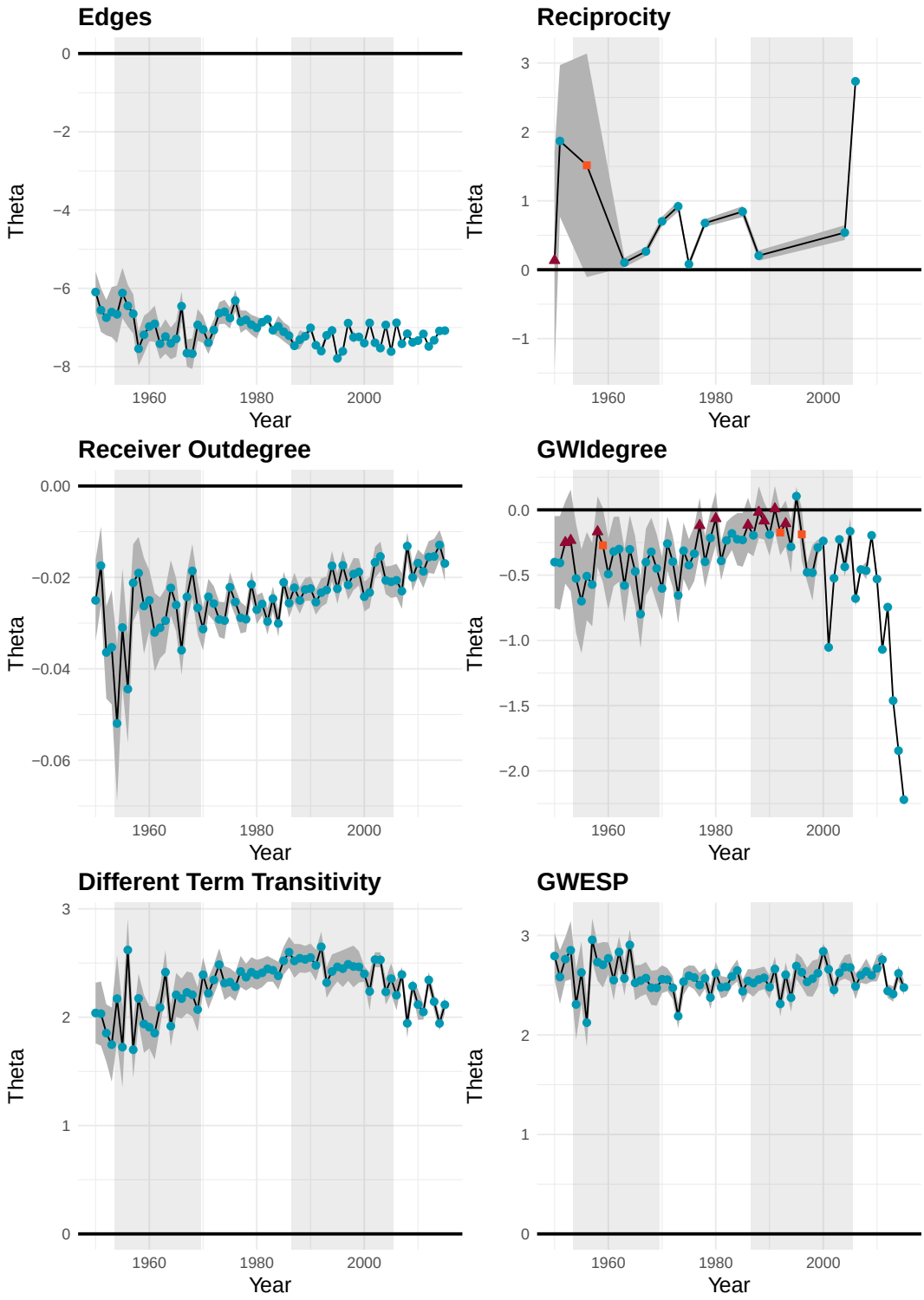

Figure 3.6. ERGM results for the dependence terms. Circles indicate a p-value smaller than 0.05, squares a p-value between 0.05 and 0.1 and triangles a p-value greater than 0.1. Different chief justice terms are indicated by shading in the background; the two grey areas indicate the Warren and Rehnquist courts.

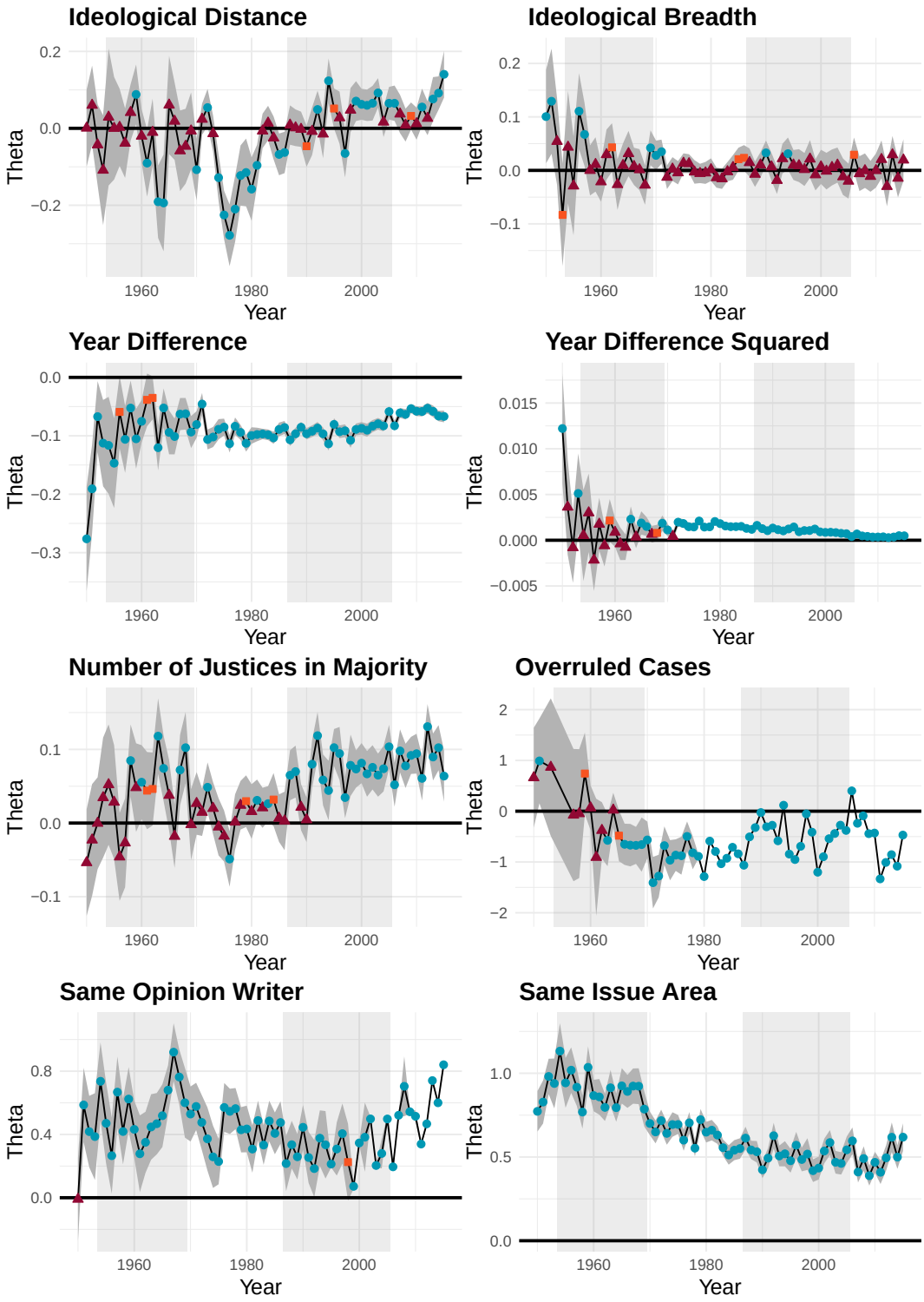

Figure 3.7. ERGM results for the covariate terms. Circles indicate a p-value smaller than 0.05, squares a p-value between 0.05 and 0.1 and triangles a p-value greater than 0.1. Different chief justice terms are indicated by shading in the background; the two grey areas indicate the Warren and Rehnquist courts.

based on both its history and its contemporary context. The methods we present here afford researchers the tools to better understand these complexities.

3.5 Conclusion

We present a methodology for studying citations between US Supreme Court opinions at the dyadic level, as a network. This methodology—the citation ERGM—enables researchers to include both exogenous covariates such as the ideological predisposition and age of a case, and dependence terms, such as transitivity and reciprocity, as explanations for citation formation. We apply this methodology to a network that includes all Supreme Court cases decided between 1950 and 2015. We find, somewhat counterintuitively, that Supreme Court citations are highly reciprocal. We also find that citations are driven by dependencies such as triad closure and popularity. The dependence effects that we identify are as substantively and statistically significant as the effects of the exogenous covariates we include in the model. The summary result from this analysis is that theoretical models of Supreme Court citation formation should consider both the effects of case characteristics and the structure of past citations.

Though we see the advancements in modeling the Supreme Court citations as our central contribution, we make two contributions to broader literature on judicial citations and citation analysis more generally. First, our arguments regarding the dependencies that shape Supreme Court citations are likely to apply to citation networks formed among other court opinions, and we have provided a road map and tools for modeling such dependencies. Second, we have provided an extended version of ERGM that is appropriate for all forms of citation network analysis, and is available in a convenient statistical software package.

Funding

This work was supported in part by NSF grants SES-1558661, SES-1637089, SES-1619644, and CISE-1320219.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

Acknowledgements

We thank Rachael Hinkle, Lauren Santoro, and Michael Nelson for helpful feedback on this project.

Data Availability Statement

The replication code and data for this paper can be found at Schmid et al. (2021b).

3.6 Appendix

3.6.1 cERGM Estimation

The normalizing constant in Equation 1 of the main text is intractable. For example, in the simple case of adding three cases to a network in which six cases already exists—like that depicted in Figure 4 of the main text—there are 16,777,216 unique configurations of C_t that could be observed. The typical Supreme Court term involves adding hundreds of cases to a network that already includes thousands of previous cases. This means that straightforward methods of maximum likelihood estimation (MLE) are infeasible with the cERGM.

The common alternative relies on Monte Carlo methods to approximate the normalizing constant by simulating a large set of networks (Hunter and Handcock, 2006; Hummel et al., 2012). The resulting estimator, the Monte Carlo MLE (MCMLE), is approximately consistent, meaning that it converges to the MLE as the sample size, i.e., the number of simulated networks, increases. However, one drawback is that with the number of nodes in the Supreme Court citation network being in the order of $10,000$, obtaining the MCMLE is computationally expensive (Schmid and Desmarais, 2017) and the success of the algorithm heavily relies on the starting parameter vector θ_0, which is ideally chosen in the proximity of the unknown MLE (Hummel et al., 2012). The prevailing choice for θ_0 is the maximum pseudo-likelihood estimation (MPLE) (Strauss and Ikeda, 1990), a fast estimation method that is defined as maximizing the log product of the conditional probability of each citation (and non-citation), conditional on the other elements of the observed citation network. The joint probability of all citations is replaced by the product over conditional probabilities, which, as we demonstrated in the main text, assume a logit form. The MPLE is simple to obtain, but does not guarantee

a starting value close to the MLE (Schmid and Hunter, 2021).

The MCMLE of networks up until the 90s was obtainable in a reasonable time frame starting at the MPLE and sampling $10,000$ networks to approximate the normalizing constant. However, the estimation of most networks in the 90s with the MPLE as starting values was not feasible in an reasonable time frame anymore. Instead, we improved the choice of starting value θ_0 by fixing it at the MCMLE of the previous term $t-1$ and successfully obtained the MCMLE of the network at term t. But even this approach started to fail for networks around the turn of the millennium. Neither the MPLE nor the MCMLE of previous terms as starting values led to successful estimation, and neither did the Stepping algorithm (Hummel et al., 2012). The MCMLE for these large citation networks was obtained by setting the starting value according to an novel approach introduced by Schmid and Hunter (2021). This method is based on the fact that the MLE of exponential family distributions is solely a function of the vector of sufficient statistics $h(C_t, C_{<t})$, meaning that the MLE of two networks A and B is equal if $h(A) = h(B)$. However, the MPLE of networks with the same sufficient statistics is not necessarily the same. Instead of starting the MCMLE algorithm at the network's MPLE, Schmid and Hunter (2021) propose searching for a new network C_t^* on the same nodes as the observed network that satisfies $h(C_t, C_{<t}) = h(C_t^*, C_{<t})$, and has a weak dependence structure among unfixed ties. Such a network can be found using simulated annealing algorithms (Kirkpatrick et al., 1983). For networks with a weak dependence structure among unfixed ties, the MPLE is similar to the MLE, in addition, the same sufficient statistic between C_t^* and the observed network C_t guarantees the same MLE between these two networks. This makes the MPLE of C_t^* an effective starting value for the MCMLE algorithm. Since for some networks the MCMLE was only obtainable using simulated annealing method to find a starting value, the final results in the paper have all been estimated using simulated annealing. The simulated annealing algorithm for finding an improved starting value for cERGMs was implemented in the **cERGM**-package for **R** (R Core Team, 2020) and can be found at `http://github.com/schmid86/cERGM`.

3.6.2 Goodness-of-Fit

We evaluate the goodness-of-fit of the model following Hunter et al. (2008) by examining the distribution of four hyper statistics, e.g., the out- and indegree distribution and the distribution for two different edgewise shared partners statistics. OTP stands for *outgoing two-paths* and refers to the number of cases r that are cited by case i and that cite case j, while j is also directly cited by i. The second ESP statistic is the OSP

specification that has been introduced in section 4.1.2 in the paper. Figure 3.8 visualizes the goodness-of-fit results for the citation network for the 1950 (top) and 2015 (bottom) term. The solid black line indicates the statistic's distribution in the Supreme Court citation network of that given term and the boxplots depict the statistic's distribution of 1000 networks that have been simulated from the ERGM defined by the MCMLE. This means that in the ideal case the solid black line passes through ever single boxplot.

We see that our models do a good job capturing the out and indegree distribution of the citation network, since the black line falls almost exclusively within the ranges spanned by the boxplots. For the ESP distributions we can observe that the number of ties with $r = 0$ shared partners is captured well for both the OTP as well as for the OSP statistic. However, the model overestimates the number of $r = 1$ shared partners and then, especially in the 2015 term network, underestimates the number of ties with more than $r = 1$ shared partners.

3.6.3 Checking for Model Degeneracy

A common challenge when fitting ERGMs is model degeneracy. Model degeneracy occurs when the probability distribution defined by the parameter vector does not predominantly yield networks with similar statistics as the observed network. Generally, model degeneracy results in simulated networks with no ties or all possible ties. In a non-degenerate model the statistics of the networks that were simulated from the probability distribution defined by the MCMLE fall in the proximity of the observed network's statistics. Figures 3.9 and 3.10 depict trace and density plots for the dependence terms in the 1950 and 2015 term citation network. The histograms on the left visualize a statistic's density from 1000 simulated networks, while the right side shows the statistic's trace plot of the same 1000 networks. The solid black line indicates the statistic's value in the actual citation network. Both figures indicate that this model is non-degenerate and that the simulated network's statistics fall almost evenly around the observed statistic. The density and trace plots for the ERGM of the terms not depicted provide similar results.

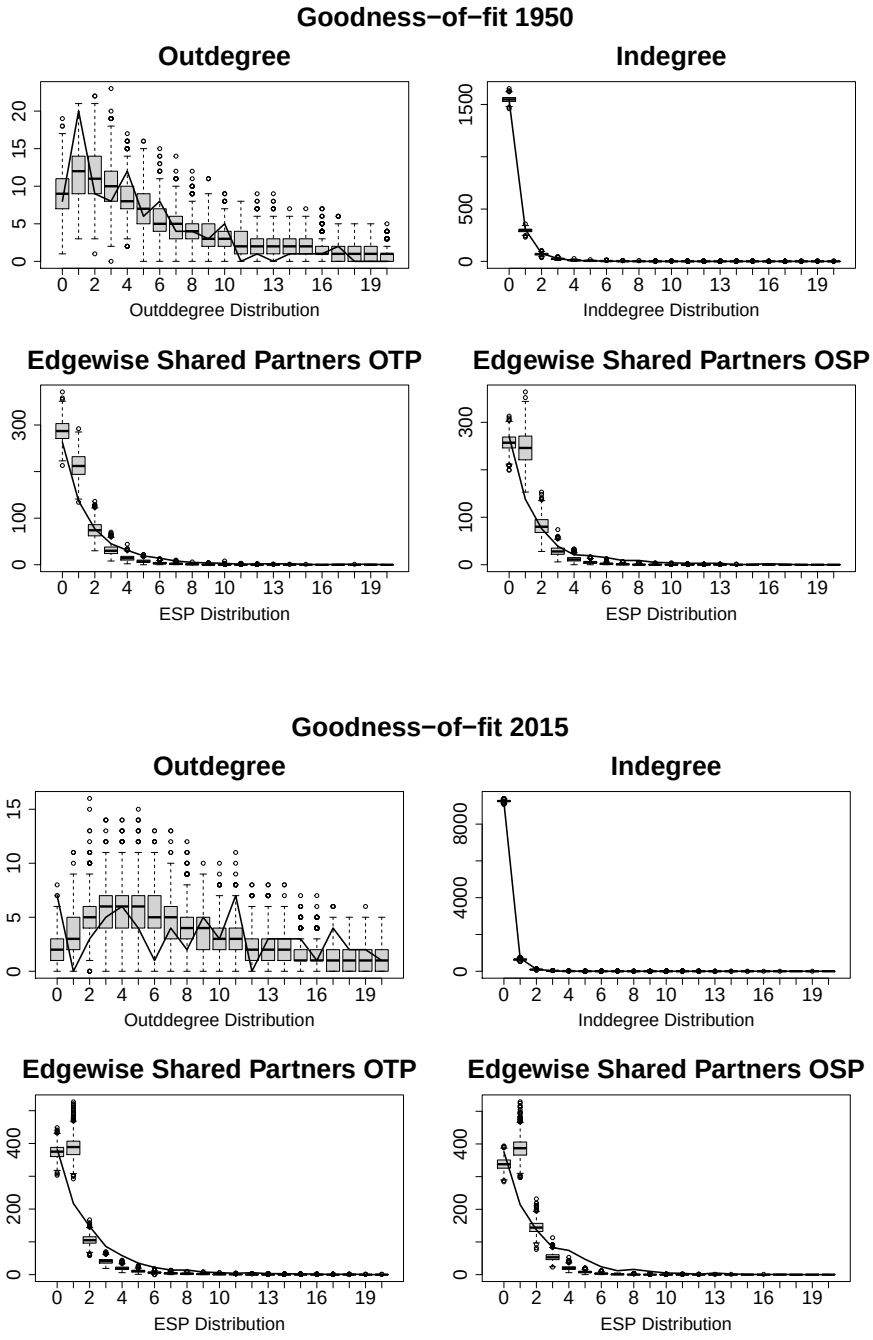

Figure 3.8. Goodness-of-fit diagnostic for the 1950 network (top) and the 2015 network (bottom).

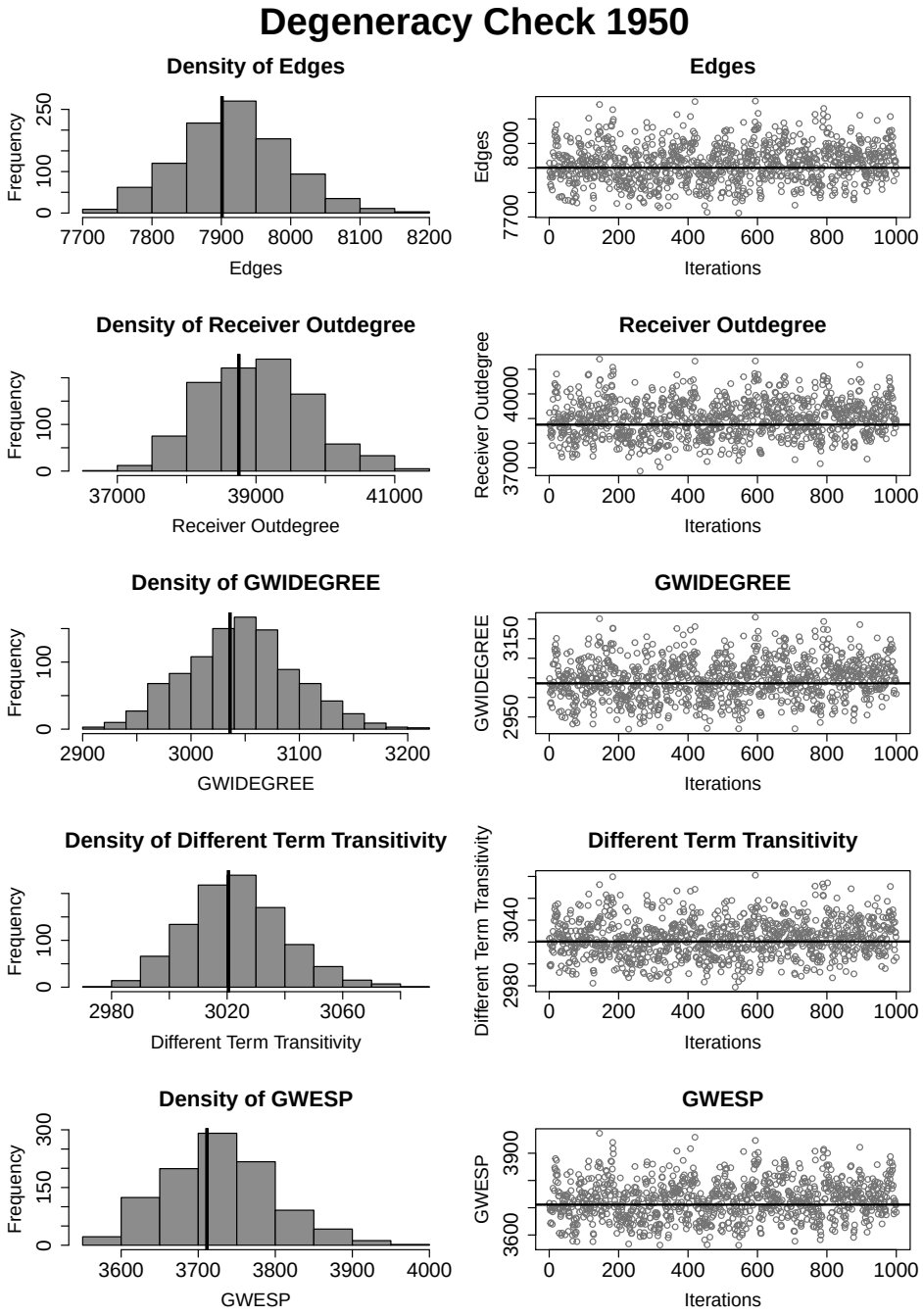

Figure 3.9. Density and trace plots for the dependency terms of the 1950 term citation network.

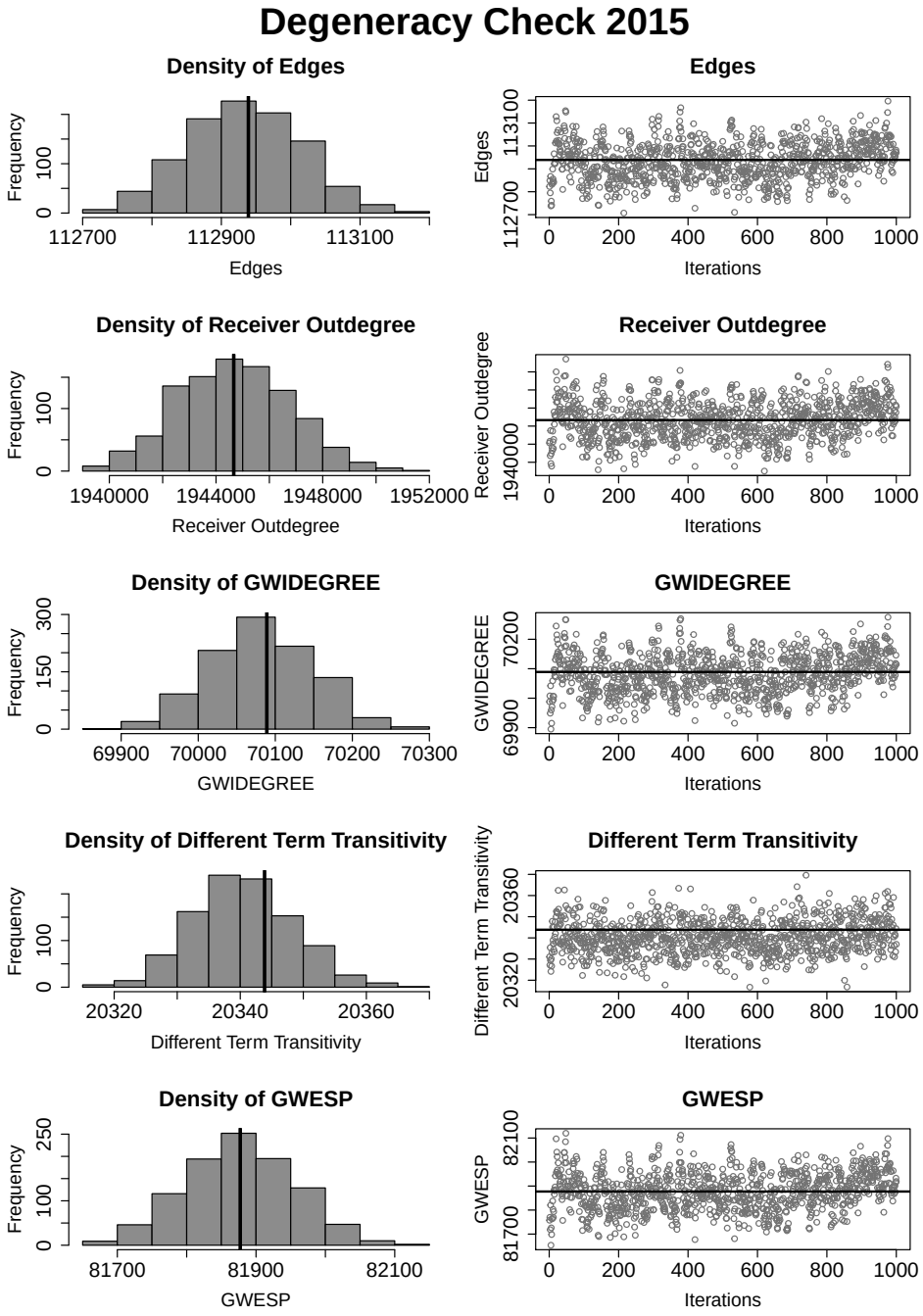

Figure 3.10. Density and trace plots for the dependency terms of the 2015 term citation network.

Chapter 4
Exponential Random Graph Models with Big Networks: Maximum Pseudolikelihood Estimation and the Parametric Bootstrap

This paper was written together with Bruce A. Desmarais and has been published at the IEEE International Conference on Big Data (Schmid and Desmarais, 2017). The idea to use parametric bootstrap to improve confidence intervals for MPL estimation and hence, for the paper, came from Bruce A. Desmarais. My part in this project was the coding and the simulation and visualization of the results. We both contributed in the write up equally. I have implemented the proposed parametric bootstrapping method into the ***ergm*** *R-package.*

With the growth of interest in network data across fields, the Exponential Random Graph Model (ERGM) has emerged as the leading approach to the statistical analysis of network data. ERGM parameter estimation requires the approximation of an intractable normalizing constant. Simulation methods are used in the state-of-the-art approach to approximating the normalizing constant, which is the foundation of estimation by Monte Carlo maximum likelihood (MCMLE). MCMLE is accurate when a large sample of networks is used to approximate the normalizing constant. However, MCMLE is computationally expensive, and may be prohibitively so if the size of the network is on the order of 1,000 nodes (i.e., one million potential ties) or greater. When the network is large, one option is maximum pseudolikelihood

estimation (MPLE). The standard MPLE is simple and fast, but generally underestimates standard errors. We show that a resampling method—the parametric bootstrap—results in accurate coverage probabilities for confidence intervals. We find that bootstrapped MPLE can be run in 1/5th the time of MCMLE. We study the relative performance of MCMLE and MPLE with simulation studies, and illustrate the two different approaches by applying them to a network of bills introduced in the United State Senate.

4.1 Introduction

The field of network science faces a double-edge sword when it comes to computationally intensive research. First, the availability of digital source data has led the growth in network science to be synonymous with the growth in research on big data. Second, analytical methods are growing more sophisticated, increasingly involving iterative and/or simulation-based optimization, rather than simple descriptive calculations (Snijders, 2011). Closing the gap in terms of the size of the networks to which it is feasible to apply the most sophisticated methods of network modeling requires research into scalable methods of inference. We propose a method of statistical inference for one of the most popular models for networks—the exponential random graph model (ERGM), in which both parameter estimates and confidence intervals are derived, that can require less than half the compute time of currently used methods.

The ERGM is a probabilistic model for networks (Wasserman and Pattison, 1996; Robins et al., 2007). They can be used for link prediction (Lu et al., 2010), simulating network adjacency matrices (Hackney and Axhausen, 2006), and testing theories regarding the processes underlying tie formation (Goodreau et al., 2009). The ERGM was first introduced by Holland and Leinhardt (1981). However, due to the intractable normalizing constant in the likelihood function of the ERGM, it did not see widespread and complete use until the 2000s, following the development of algorithms and software for efficient simulation-based methods for working with ERGM (Snijders, 2002). Training ERGM using simulation-based methods is computationally expensive, and can still be prohibitively burdensome with data on big networks. Approximate methods of estimation, which are much more feasible with large networks, have existed for some time, but these methods perform poorly when it comes to characterizing the uncertainty in parameter estimates, which is necessary when assessing risk in predictions or simulation, or in hypothesis testing.

The ERGM takes the adjacency matrix of an observed network G^{obs}, which is a matrix-valued random variable. This means that a network of N nodes can be defined as a adjacency matrix $G = (g_{ij}) \in \{0,1\}^{(N\times N)}$, where $g_{ij} \in \{0,1\}$ for all $i,j \in \{1,\dots,N\}$. $g_{ij} = 1$ means that there is an edge between actors i and j, while $g_{ij} = 0$ indicates that these actors are not directly connected. Since the model does not consider loops, one has $g_{ii} = 0$ for all $i \in \{1,\dots,N\}$. Furthermore, define $\mathcal{G}(N)$ as the set of all possible networks on N nodes without loops. Note that the cardinality of set $\mathcal{G}(N)$ is increasing exponentially for every newly included actor, which results in $2^{N(N-1)/2}$ total elements. For this reason calculating the likelihood function of the ERGM, which requires evaluating a normalizing constant on $\mathcal{G}(N)$ is either extremely time-consuming or with today's technology not achievable. As a consequence, many approximation methods have been provided by the literature, with the most popular method making use of Markov Chain Monte Carlo (MCMC) methods (Hunter et al., 2012), as we will introduce in the next section.

The probability function for the ERGM is defined as

$$\mathbb{P}_\theta(G) = \frac{\exp(\theta^T \cdot \Gamma(G))}{\sum_{G^* \in \mathcal{G}(N)} \exp(\theta^T \cdot \Gamma(G^*))} \tag{4.1}$$

where $\theta \in \mathbb{R}^q$ is a $q-$dimensional vector of parameters, $\Gamma : \mathcal{G}(N) \to \mathbb{R}^q$, $G \mapsto (\Gamma_1(G),\dots,\Gamma_q(G))^T$ is a q-dimensional function of different network statistics and $c(\theta) := \sum_{G^* \in \mathcal{G}(N)} \exp(\theta^T \cdot \Gamma(G^*))$ is a normalization constant which ensures that (4.1) defines a probability function on $\mathcal{G}(N)$. The generative processes captured by a model (e.g., density, reciprocity, popularity, clustering) are informed by the decision regarding which network statistics (i.e., $\Gamma(\cdot)$) are incorporated. The flexibility of the ERGM in capturing virtually any network generative process has led to it being applied broadly across several fields, including sociology (Smith et al., 2016), economics (Lomi and Fonti, 2012), political science (Cranmer and Desmarais, 2011), ecology (Dey and Quinn, 2014), and neuroscience (Simpson et al., 2011).

4.2 Estimation

The first method proposed in the literature for estimating ERGM parameters was maximum pseudolikelihood estimation (Strauss and Ikeda, 1990). Under maximum pseudolikelihood estimation (MPLE), the joint distribution is replaced by the product over conditional distributions (Besag, 1986). The conditional probability of a tie in

ERGM reduces, conveniently, to a logistic regression form given by

$$\mathbb{P}_\theta\left(g_{ij}=1|G_{-ij}\right)=\text{logit}^{-1}\left(\theta^T\cdot\delta(\Gamma(G))\right),$$

where G_{-ij} is the adjacency matrix, excluding element ij, $\delta(\Gamma(G))$ is the "change statistic" given by the difference in the network statistics when the ij element is toggled from 0 to 1 (i.e., $\Gamma(G|g_{ij}=1))-\Gamma(G|g_{ij}=0)$), and $\text{logit}^{-1}(x)=1/(1+\exp(-x))$ (Goodreau et al., 2009). For the ERGM, the pseudolikelihood function can be maximized using logistic regression software, in which the dependent variable is given by the elements of the adjacency matrix, and the covariates are given by the values of the change statistics corresponding to each element of the adjacency matrix.

Despite the computational efficiency underlying the implementation of MPLE, existing methods for assessing uncertainty with respect to the MPLE perform poorly (van Duijn et al., 2009). Estimating the uncertainty in parameter estimates (e.g., standard errors, confidence intervals), is a critical step in using the results from a statistical model. Estimates of uncertainty are used to test hypotheses about parameters, estimate variance (i.e., risk) in model predictions, and estimate effect sizes. The current conventional approach to estimating θ, introduced by (Snijders, 2002), is based on a Markov Chain Monte Carlo (MCMC) approximation of the MLE. This Monte Carlo maximum likelihood method (MCMLE) is based on a more direct attempt to approximate the log-likelihood function derived from (4.1). The log-likelihood function is not evaluated directly, rather, the log ratio of the likelihood under a proposed value of the parameters θ, and an initial value of the parameters θ_0, is approximated using L networks simulated from the ERGM with parameter values θ_0. The approximation, detailed in (Snijders, 2002) is given by

$$\text{loglik}(\theta)-\text{loglik}(\theta_0)\approx-\log\left(\frac{1}{L}\sum_{i=1}^{L}\exp\left((\theta-\theta_0)^T\Gamma(G_i)\right)\right)$$

As we demonstrate below, the MCMLE grows more accurate as L increases. Indeed, MCMLE approaches the MLE as the number of networks simulated goes to infinity.

4.3 Efficiency of MPLE and MCMLE

As mentioned in the previous section the MPLE approaches the MLE as the size of the networks increase and as a consequence, is a consistent estimator (see Lindsay (1988), Strauss and Ikeda (1990), Hyvärinen (2006), Desmarais and Cranmer (2012, 2010)). This

implies that for an increasing number of nodes, the MPLE converges in probability to the MLE, meaning that for large enough networks the MPLE performs as well as or better than MCMLE, and requires less compute time. To illustrate the relative efficiency of MPLE and MCMLE we run a simulation study. Desmarais and Cranmer (2012) show the MPLE outperforms the MCMLE if the number of simulated networks used to approximate the likelihood in MCMLE is not large enough. It is even more remarkable that the number of simulated networks needed for the MCMLE, in order to surpass the MPLE increases as the size of the network (i.e., the number of nodes in the network) increases. This means that, for very large networks, it becomes difficult to determine the number of simulated networks required for the MCMLE to outperform the MPLE. In other words, the larger the network, the more computationally intensive it becomes to use MCMLE in a way that out-performs MPLE.

To demonstrate this disadvantage of the MCMLE we conduct a simulation study using Goodreau's Faux Mesa High School data (Hunter , David R. et al., 2008), which represents a simulation of an in-school friendship network among 203 students as well as the Faux Magnolia High School data, representing an in-school friendship network among 1451 students. The data for both networks originates from Resnick et al. (1997). For both networks, we first calculate the MCMLE and treat the estimated coefficients as the network's true values θ. Then, we take the same parametrization, using the number of edges, the nodal attribute for gender, and the geometrically weighted edgewise shared partners (gwesp) distribution (Hunter and Handcock, 2006) where we fix the decay parameter λ at 0.25. The gwesp statistic is used to model the tendency towards triangles and clustering in a network.

We simulate $m = 500$ new networks using the 'true' coefficients and estimate the MPLE as well as the MCMLE of these simulated networks. For every single simulated network the MCMLE calculation is being repeated several times for 25 to $10,000$ simulated networks used in the likelihood approximation. Based on these results, we compute the root mean square error, which is a measure of the accuracy of an estimator, combining both the bias and the variance. Mathematically written, the RMSE for an estimator $\hat{\theta}$ is defined as $RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\theta - \hat{\theta}_i)^2}$ implying that the smaller the RMSE, the more accurate is the estimator. Since the MCMLE has higher efficiency and converges to the MLE, the RMSE decreases as the number of simulated networks used for the likelihood approximation increases. On the other hand, the RMSE of the MPLE is a constant value since no network simulations are required. In order to compare the RMSE of the two estimation techniques, we take the log of the ratio of the MCMLE to the

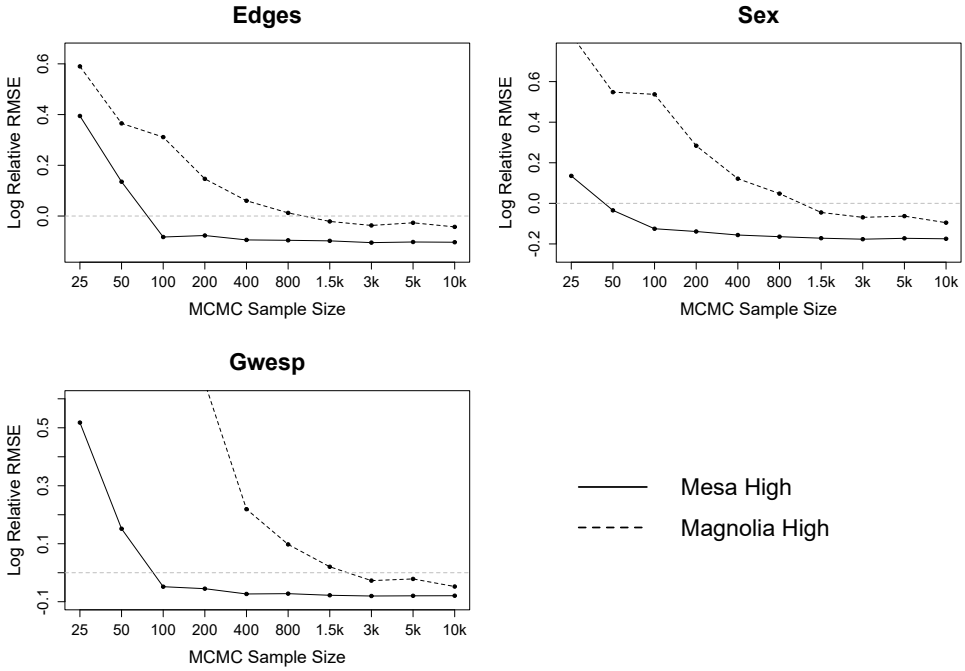

Figure 4.1. The log of the ratio of the RMSE for the MCMLE to the MPLE for different sample sizes and two different networks, Faux Mesa High and Faux Magnolia High

MPLE. As a result, a negative value indicates a better MCMLE performance, while a positive value indicates a better MPLE performance. Figure 4.1 visualizes the results of the simulation study. The solid line illustrates the results of the log relative RMSE of the Faux Mesa High network, while the dashed line illustrates the corresponding results of the Faux Magnolia High network. The plots support the fact that larger networks require a larger sample size of simulated networks for the MCMLE to outperform the MPLE. While the fairly small Faux Mesa High network only requires a sample size of about $50 - 100$ networks, the larger Faux Magnolia High network requires a sample size of at least 1,500 networks for the MCMLE to surpass the MPLE. For especially large networks (e.g., social media data) the sample size has to be set in order to justify the approximately exact, but computationally expensive and potentially prohibitive MCMLE method.

4.4 Bootstrapped MPLE

As discussed in the previous section, the MPLE converges to the MLE as the size of the network increases. Moreover, the MPLE is able to outperform the MCMLE if the sample size used in MCMLE is not large enough. The main reason why the MCMLE is

still widely preferred is that, in contrast to the MPLE, it does not underestimate the standard errors (van Duijn et al., 2009). By the definition of the ERGM it is obvious that this model is an exponential family distribution where θ is the natural parameter and $\Gamma(G)$ is the sufficient statistic. For exponential family distributions, a covariance matrix can be estimated by the inverse of the negative Hessian matrix $[-H]^{-1}$ of the likelihood function at the MLE. The problem with the MPLE is that calculating $[-H]^{-1}$ by the pseudolikelihood function will underestimate the variance of the MPLE (van Duijn et al., 2009), resulting in an underestimate of the width of the confidence intervals. van Duijn et al. show that constructing 95% MPLE confidence intervals can result in intervals that comprise the true value in less than 75% instead of the nominal 95%. In this paper, we are going to refer to the MPLE confidence intervals as *logistic regression confidence intervals* simply because the MPLE is calculated using logistic regression methods that also use the inverse of the negative Hessian matrix as an estimate for the covariance matrix.

Since the MPLE has the advantage of being approximately exact and computationally inexpensive, but has the disadvantage of underestimating corresponding confidence intervals, we apply a technique referred to as bootstrap resampling (Efron, 1982). Bootstrap resampling refers to constructing a sampling distribution for the parameter estimate by resampling the data with replacement, and re-estimating the model on the resampled data. Under non-parametric bootstrap resampling, the data are resampled directly from the dataset. Under the parametric boostrap, the data is resampled from the estimated model. The idea of using boostrap resampling with MPLE for ERGM was first introduced by Desmarais and Cranmer (2012) and provides a consistent estimate of MPLE confidence intervals. However, the methods introduced by Desmarais and Cranmer (2012) only applied to cases in which the researcher had a large sample of networks (e.g., a time series of networks). For the case in which there is just a single network to be studied, we propose the use of a parametric bootstrap. Under the parametric bootstrap, the sampling distribution of the MPLE is derived by re-estimating the MPLE on a sample of networks simulated from the MPLE estimated on the observed network. We verify the consistency of the bootstrapped MPLE by conducting a simulation study on the same two networks with the same parametrization as in the previous chapter: The Faux Mesa High friendship network and the Faux Magnolia High friendship network.

For the simulation study, we determine the MPLE for the model and treat these estimates as the networks' 'true' parameter values. We use these parameter values to simulate a sample of 1000 networks from the distribution of G. For each of the 1000

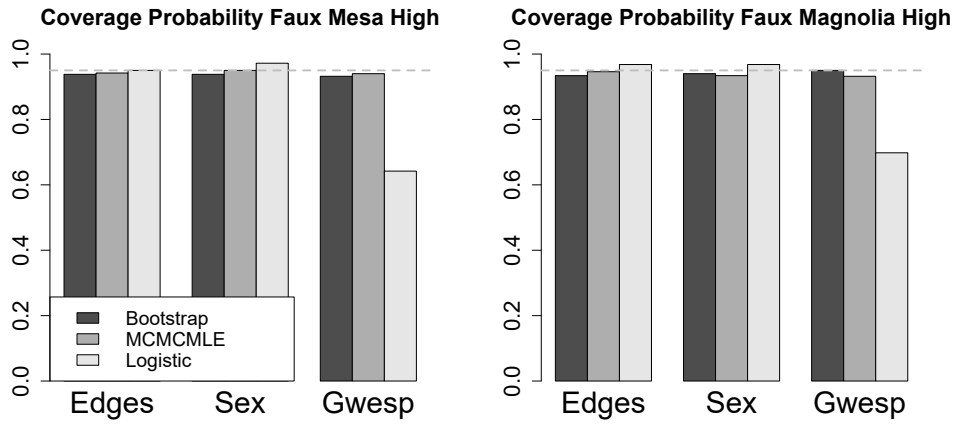

Figure 4.2. The Coverage Probability results of the Faux Mesa High network (left) and of the Faux Magnolia High network (right) for bootstrapped MPLE, MCMLE and logistic regression

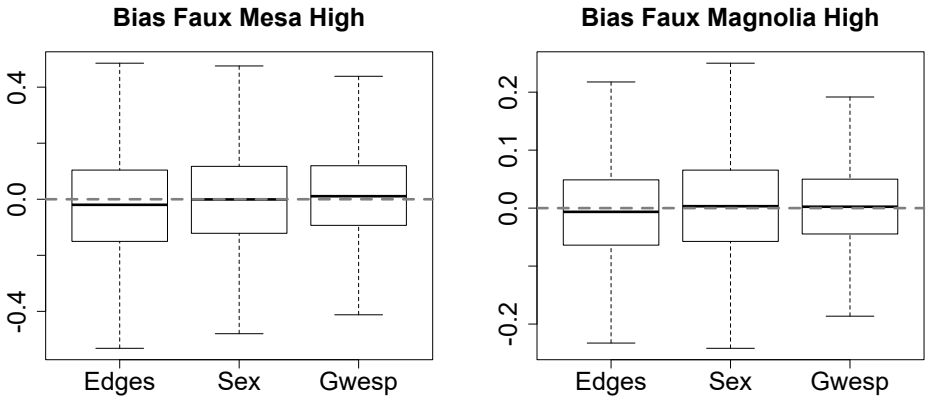

Figure 4.3. The boxplots visualize the bias ($\hat{\theta} - \theta$) over the 500 iterations for the Faux Mesa High network (left) and the Faux Magnolia High network (right)

networks, we calculate 95% confidence intervals based on the MCMLE and the logistic regression and examine whether the 'true' parameter values lie in these intervals. In addition, we determine the bootstrapped MPLE confidence intervals by sampling 500 networks for each of the originally sampled 1000 networks, by using the respective MPLE as parameter values. For every newly sampled network, we again determine the MPLE and then take the 2.5th and 97.5th percentile of the 500 MPLE estimates to obtain 95% bootstrap confidence intervals. We verify whether the 'true' parameter value can be found in the bootstrapped confidence interval.

Figure 4.2 visualizes the coverage percentages for each of the three methods for both

	MCMLE		Logistic Regression		bootstrapped MPLE	
	Estimate	St. Error	Estimate	St. Error	Lower Bound	Upper Bound
Edges	-5.884	0.065	-5.869	0.015	-6.007	-5.751
Sponsor Party	1.440	0.015	1.440	0.015	1.411	1.467
Altern. k-star	0.124	0.064	0.108	0.006	-0.011	0.2379

Table 4.1. Estimation results for the Cosponsorship network using MCMLE, logistic regression and bootstrapped MPLE

networks. The dashed line is set at 0.95 and represents the optimal value. It is evident that the bootstrapped MPLE performed equally well as the MCMLE, achieving results that obtain the true parameter values in approximately 95% of the cases. Additionally, a difference in the results between the smaller Faux Mesa High network and larger Faux Magnolia High network is not identifiable. Similar to the results of van Duijn et al. (2009) our results for the logistic regression differ distinctively from the anticipated 95%, confirming that the MPLE underestimates the variance of its estimates. Figure 4.3 illustrates the bias between the 'true' network coefficients θ and the MPLE estimates. The median MPLE estimates approximate the true parameters. It is especially worthwhile to mention that the bias of the larger Faux Magnolia High network is smaller than the bias of the Faux Mesa High network, supporting the fact that the MPLE converges to the MLE as the network size increases.

4.5 Cosponsorship Network Data

To illustrate the performance of MCMLE relative to that of the bootstrapped MPLE we apply both approaches to the data on cosponsorship of bills in the U.S. House of Representatives for the 108th Congress (2003–2004), developed by Fowler (2006). The cosponsorship network consists of 2,635 nodes, which we define as pieces of legislation (i.e., bills), considered by the Senate during the 108th Congress. In this undirected network bills are tied together based on the similarity of the sets of legislators who cosponsor them. Specifically, we include an edge between bills i and j if the correlation coefficient between the indicator vectors indicating whether i and j were sponsored each legislator is greater than a random uniform draw. This results in an undirected network with 28060 edges.

We build an ERGM specification that extends the work of Zhang et al. (2008) in

exploring the structure of cosponsorship ties. They find that congressional cosponsorship is primarily characterized by intra-party ties—among Republicans and among Democrats, but few cross-party ties. We test for this party-based clustering (i.e., homophily) in our ERGM. This is done through a term that accounts for the party of the senators who sponsored the two bills in the pair. A positive parameter value for this statistic indicates that ties tend to be formed between bills sponsored by the same political party.

We extend the homophily-based model to account for a network dynamic that is commonly found in the study of networks—that of popularity or preferential attachment (Barabási and Albert, 1999). The alternating k-star statistic was introduced by Snijders et al. (2006) and modified by Hunter and Handcock (2006). A positive parameter estimate associated with the alternating k-star statistic indicates that tie formation follows a form of preferential attachment (Snijders et al., 2006). This could arise in a network of bill-to-bill ties if the majority party in power was particularly disciplined at rallying its partisans to pile on to the bills that its members proposes, thus producing a large set of very similar bills. We estimate the ERGM using MCMLE and the bootstrapped MPLE. The MCMLE requires a sample size of at least 1000 networks to converge. The bootstrapped MPLE was estimated by using 500 simulated networks. As we described in the section *Estimation*, only one edge at a time is changed when simulating networks. The results can be found in table 4.1.

The MPLE estimate is equivalent to the logistic regression estimate, but the bootstrap confidence intervals, especially for the alternating k-star statistic, are much wider than would be calculated using the logistic regression standard errors. An estimate is generally considered statistically different from zero (i.e., statistically significant) if the confidence interval does not contain zero, or if the ratio of the estimate to the standard error exceeds 1.96 in magnitude. This cosponsorship network example perfectly illustrates the inferential problems that can arise with the conventional logistic regression standard errors when using MPLE. All of the parameter estimates are statistically significant according to the logistic regression estimates. However, the alternating k-star statistic is not significant according to either the MCMLE or the bootstrapped MPLE.

4.6 Parallel Computing with MPLE

The bootstrapped MPLE is not only simple and fast, it is highly parallel. Once the networks on which to estimate the bootstrap replicates are simulated, each re-estimate can be run in parallel. By using multiple cores, the computing time for estimating

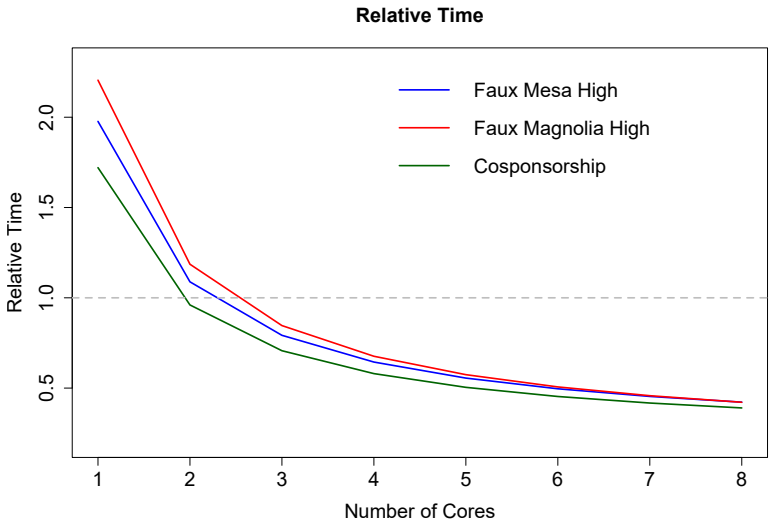

Figure 4.4. The y-axis gives the ratio of the bootstrapped MPLE time to that of the MCMLE time. Values below 1 indicate that the bootstrapped MPLE requires a shorter computing time.

bootstrapped MPLE confidence intervals can be reduced substantially. Figure 4.4 illustrates the relative computing time of the bootstrapped MPLE using 500 simulated networks and the MCMLE for the three networks Faux Mesa High (205 nodes), Faux Magnolia High (1461 nodes) and Cosponsorship (2635 nodes) for an increasing number of computing cores. For the small network we simulate 2000 networks using a MCMC interval of 2000 steps, for the medium network we simulate 8000 networks using a MCMC interval of 5000 steps and for the large network we simulate 10000 networks using $30,000$ MCMC steps in order to approximate the likelihood appropriately. The chosen sample sizes and MCMC steps are necessary to guarantee a good model fit. The small network took 14 seconds, the medium network took 123 seconds and the large network took 986 seconds to run. We define the simulation time of the bootstrapped MPLE as a function of the number of available computing cores x:

$$\text{MPLE time} = \text{network simulation time} + \frac{500 \cdot \text{MPLE estimation time}}{x}$$

Based on this, we define the relative computing time as $\frac{\text{MPLE time}}{\text{MCMLE time}}$. This means that a relative computing time greater than 1 indicates that the MCMLE computing time is shorter, while a relative computing time smaller than 1 indicates that the bootstrapped MPLE provides faster results.

Figure 4.4 demonstrates that all three networks only require three cores for the bootstrapped MPLE to outperform the computing time of the MCMLE and that the

computing time can further be reduced if more computing cores are available. If exactly 500 computing cores are being used the ratio of the bootstrapped MPLE time to the MCMLE time levels off at 0.20 for the small and large network and 0.17 for the medium network, meaning that the computing time can be quintupled using the bootstrapped MPLE. This figure also depicts that larger network in general require a longer computation time and will benefit more if the bootstrapped MPLE is used.

One of the major disadvantages of MPLE over MCMLE is that degeneracy is not assessed while the model is being estimated. The bootstrapped MPLE, however, allows assessing degenerate models as well since the method requires simulating from the estimated parameters. In order to verify whether a model is degenerate or not, one can take a look at density and trace plots as visualized in figure 4.5. The trace plots on the left side depict the the attained values via MCMC simulated networks for every single statistic included into the model, centered on the statistic values of the observed network. The plots on the right side visualize the empirical density function of the respective statistic, based on the simulated networks (Hunter and Handcock, 2006). For a non-degenerated model the empirical density function should be approximately symmetrical around zero for every included centered statistic, since this corresponds with the expected value of a centered statistic. Otherwise, the values of the simulated networks systematically differ from the corresponding statistics in the observed network, making it unreasonable to assume that the simulated networks originate from the same distribution as the observed network. Furthermore, the trajectories in the trace plot should not indicate a dependence structure. This would be a signal that the constructed stochastic process violates the Markov properties.

4.7 Conclusion

In this paper we introduced the bootstrapped MPLE as an alternative method of statistical inference for ERGMs and compared the performance to the commonly applied MCMLE. Based on a simulation study we demonstrated that the larger the size of a network is the larger the MCMC sample size has to be in order for the MCMLE to outperform the fast and simple MPLE. However, the big disadvantage of the MPLE is that, even though it is an approximately exact estimator, it underestimates the standard error. For this reason, we propose a parametric bootstrap method of evaluating model uncertainty. On the basis of another simulation study on two different networks, we demonstrate that the bootstrapped MPLE covers the true coefficients just as well as the MCMLE, while

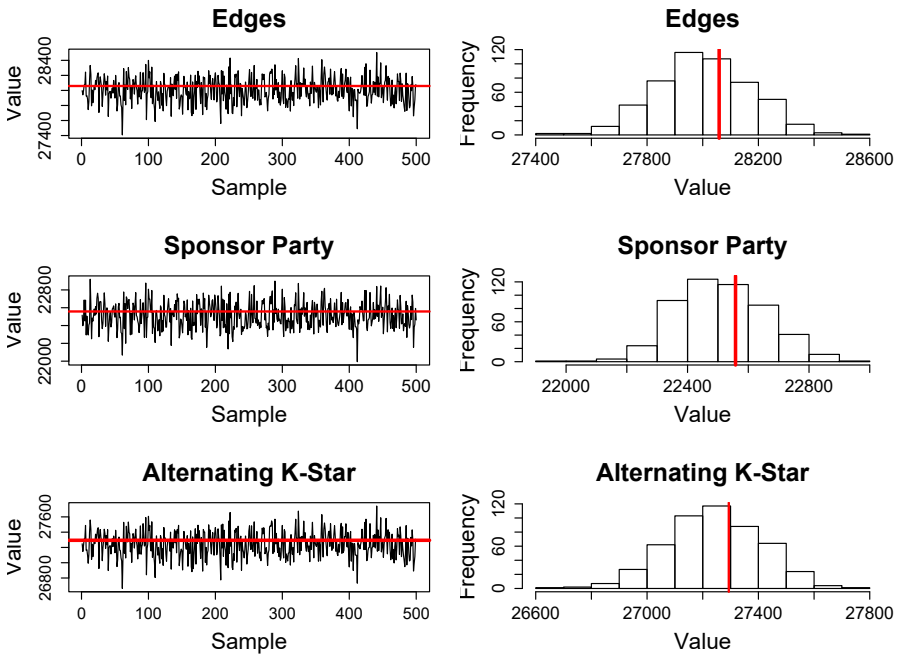

Figure 4.5. Network statistics of the 500 boootstrap samples for the cosponsorship network. The thick line in both, the traceplots and the histograms, represents the network statistics of the observed network.

the simple MPLE performs clearly poorer. This means that the bootstrapped MPLE combines the advantages of both methods, the MPLE and the MCMLE, because it is still simple and fast, and provides approximately exact results, but also accurately estimates model uncertainty. We conclude that the bootstrapped MPLE should be regarded as an alternative to the MCMLE. It also has the advantage of being parallel, which leads to a rapid speed-up of the calculation if multiple computing cores are used.

Acknowledgements

This work was supported in part by NSF grants SES-1558661, SES-1637089, SES-1619644, and CISE-1320219.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

Chapter 5
Accounting for Model Misspecification When Using Pseudolikelihood for ERGMs

This paper was written in collaboration with David R. Hunter and has not been submitted for publication at the time this dissertation was submitted. The content and ideas in this paper were developed together. All simulations and visualizations were created by myself. I also did the majority of the writing for this paper and have implemented the proposed approximate Godambe method into the ***ergm*** *R-package.*

The reputation of the maximum pseudolikelihood estimator (MPLE) for Exponential Random Graph Models (ERGM) has undergone a drastic change over the past 30 years. While first receiving broad support, mainly due to its computational feasibility and the lack of alternatives, the general opinion started to change with the introduction of approximate maximum likelihood estimator (MLE) methods that became practicable due to increasing computing power and the introduction of MCMC methods. The eventuating comparison studies appear to yield contradicting results regarding the preference of these two point estimators, however, there is consensus that the prevailing method to obtain an MPLE's standard error by the inverse Hessian matrix generally underestimates standard errors. We propose replacing the inverse Hessian matrix by an approximation of the Godambe matrix that results in confidence intervals with appropriate coverage rates and that, in addition, enables examining for model degeneracy.

5.1 Introduction

The methods discussed in this article have been implemented into the **ergm**-package, a package to fit, simulate and diagnose exponential-family models for networks (Handcock et al., 2019) in **R** (R Core Team, 2020). The updates discussed, provide more accurate standard errors and confidence intervals for maximum pseudolikelihood estimates.

The probabilistic modeling of networks goes back to Erdös and Renyi (1959) and Gilbert (1959), who introduced a model where each tie occurs independently with probability p. Holland and Leinhardt (1981) were the first to consider tie dependence within dyads in their p_1 model, a model that was later generalized by the Markov random graph model of Frank and Strauss (1986). This model assumes that only dyads that share a common node can depend on each other. The exponential random graph model (ERGM), or p^* model as it was called by Wasserman and Pattison (1996), generalizes the Markov random graph model and is to this day a popular way to model complex dependency structures of networks.

Let a network be represented as an adjacency matrix $A \in \mathbb{R}^{N\times N}$ with $A_{ij} = 1$ if there is an edge between i and j, $i \neq j, i, j \in \mathcal{N} = \{1, ..., N\}$ and $A_{ij} = 0$ otherwise. In an undirected network, i.e. a network where $A_{ij} = A_{ji}$, a dyad, or a pair of nodes along with the status of its ties, can either contain a tie or not. For the sake of simplicity, we will confine ourselves to undirected networks and disallow self-edges where $A_{ii} \neq 0$. An extension to directed networks is straightforward. The ERGM assumes that an observed network A^{obs} is a realization of matrix-like random variable Y with an underlying probability distribution defined over all possible networks on N nodes. The ERGM takes the form

$$P_\theta(Y = A) = \frac{\exp(\theta^\top \cdot T(A))}{k(\theta)} \tag{5.1}$$

for $A \in \mathcal{A}(\mathcal{N})$, where $\mathcal{A}(\mathcal{N})$ is the sample space of allowable networks and $\theta \in \mathbb{R}^q$ is a vector of parameters. In many applications, $\mathcal{A}(\mathcal{N})$ denotes the entire set $\{A \in \mathbb{R}^{N\times N}, A_{ij} = A_{ji} \in \{0, 1\}, A_{ii} = 0\}$ of possible networks on N nodes, while in other applications $\mathcal{A}(\mathcal{N})$ may be constrained to be a proper subset of this set. The sufficient statistics $T : \mathcal{A}(\mathcal{N}) \to \mathbb{R}^q$, $A \mapsto (g_1(A), \ldots, g_q(A))$ play a central role in the model, since they enable the inclusion of traditional *exogenous* covariates like a node's age as well as *endogenous* statistics, i.e., statistics that allow for inference on the structure of the network. Popular endogenous statistics are a network's number of triangles or the

number of ties that share one common node (two-stars). The normalizing constant

$$k(\theta) := \sum_{A^* \in \mathcal{A}(\mathcal{N})} \exp(\theta^\top \cdot T(A^*)), \tag{5.2}$$

a weighted sum over all possible networks on N nodes, assures that (5.1) defines a probability model.

5.2 Maximum Pseudolikelihood Estimation

The estimation of the parameter vector θ has been a major focus in ERGM literature. The challenge lies in the normalizing factor $k(\theta)$ that appears in the likelihood function and requires the calculation of a weighted sum with $2^{N(N-1)/2}$ summands for undirected networks. This number is very large for even relatively small networks, making straightforward calculation and therefore the computation of the maximum likelihood estimator (MLE) in most cases infeasible.

Frank and Strauss (1986) propose, and Strauss and Ikeda (1990) fully develop, an estimation approach based on the maximum pseudolikelihood estimator (MPLE) first introduced for lattice models by Besag (1974). The pseudolikelihood is a special form of a composite likelihood (Lindsay, 1988), an inference function where conditional or marginal densities are multiplied with one another, irrespective of whether these components are independent of each other or not. If independence does not hold, the inference function has the characteristics of a misspecified model's likelihood (White, 1982). For a detailed review of composite likelihood methods, we refer to Varin et al. (2011).

Some additional notation is necessary for introducing the pseudolikelihood function. Define A^c_{ij} as the network A without dyad ij, often also referred to as the rest of the network, and let A^+_{ij} and A^-_{ij} be network A where dyad ij is forced to be 1 and 0, respectively. Then based on Equation (5.1), the conditional probability of a tie given the rest of the network satisfies

$$\text{logit}(P_\theta(Y_{ij} = 1 | Y^c_{ij} = A^c_{ij})) = \theta^\top \cdot (T(A^+_{ij}) - T(A^-_{ij})), \tag{5.3}$$

where $\text{logit}(p) := \log p - \log(1-p)$. We refer to $\Delta_{ij} := T(A^+_{ij}) - T(A^-_{ij})$ as the vector of change statistics. Note that Equation (5.3) corresponds to a logistic regression model, where we assume a linear relationship between the predictor variables (here the change statistics) and the log-odds that $Y_{ij} = 1$. In the network setting the assumption of the

independence of observations translates to dyadic independence, i.e., the Y_{ij} are mutually independent,

$$P_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c) = P_\theta(Y_{ij} = 1).$$

If this is the case, the loglikelihood function is

$$\begin{aligned} p\ell(\theta) &= \log\Bigg(\prod_{ij \in \Omega(\mathcal{N})} P_\theta(Y_{ij} = 1)^{A_{ij}} P_\theta(Y_{ij} = 0)^{1 - A_{ij}} \Bigg) \\ &= \sum_{ij \in \Omega(\mathcal{N})} \Big[A_{ij} \cdot (\theta \Delta_{ij}) - \log\Big(1 + \exp(\theta \cdot \Delta_{ij})\Big) \Big], \end{aligned} \tag{5.4}$$

where $\Omega(\mathcal{N}) = \{ij \mid i, j \in \{1, \ldots, N\},\ i < j\}$ is defined as the set of all undirected dyads and the maximizer of (5.4) is equivalent to the MLE of a logistic regression fit of (5.3). This means that the maximizer of (5.4) can be obtained with any canned logistic regression software. In many ERGMs however, this independence assumption does not hold, making (5.4) an incorrect loglikelihood function. In this context, (5.4) is called the log-pseudolikelihood function and maximizing it results in what we refer to as the MPLE. Obtaining the MPLE is simple and fast, but this estimator can be imprecise, since a network's dependency structure is for the sake of simplicity deliberately ignored. In addition, even though the MPLE can be obtained using logistic regression software, the software's output has to be treated with caution. We will demonstrate in the two following sections that the standard errors obtained from logistic regression software is not appropriate if the independence assumption does not hold.

Frank and Strauss (1986) and Strauss and Ikeda (1990) were the first to attempt comparing the performance of MPLE to MLE. Due to the extreme difficulty of obtaining maximum likelihood estimators, these authors focused their comparisons on models with a univariate sufficient statistic and constrained the set of possible networks to those with a fixed number of edges. In these papers, the univariate sufficient statistics that were the attention of this investigation were the number of two-stars and triangles in a network, respectively. In an undirected network a two-star is defined as any set of edges (i, j) and (i, k), and a triangle is defined as any set of edges (i, j), (j, k) and (k, i), where $i, j, k \in \{1, \ldots, N\}$. The two-star model was referred to as the cluster model, while the triangle model was called the triad model. An approximate MLE was achieved by a trial and error method that required the simulation of networks from a model defined by a given parameter value and comparing the simulated network's sufficient statistic to the sufficient statistic in the original network. An estimator was defined as the approximate MLE once the simulated networks yielded the same sufficient statistic as the observed

network. The justification of this approach is rooted in the fact that in exponential family distributions the expectation of the sufficient statistics with respect to the MLE equals the sufficient statistic in the observation (Barndorff-Nielsen, 1978). While Frank and Strauss (1986) solely based their conclusion on the comparison of the MLE and MPLE point estimators, Strauss and Ikeda (1990) took the study one step further by also investigating the mean square error. The conclusion of this first comparison was that "the two methods appear to give estimators that are about equally good" (Strauss and Ikeda, 1990, p. 207).

Dahmström and Dahmström (1993) compared the MPLE to the MLE on the cluster and triad models for networks with 12 edges and $N = 7$ nodes, a sample space small enough to allow computation of the exact MLE. Comparing only the point estimates, these authors concluded that the MPLE and MLE can differ significantly.

Corander et al. (1998) compared the two estimation methods for the cluster and triad models based on mean squared error for networks of 40 to 100 nodes, approximating the MLE in a fashion similar to Strauss and Ikeda (1990). The authors conclude that the MLE has a smaller MSE for networks up to size $N = 40$, but that for larger networks both methods perform nearly equivalently well. Corander et al. (1998) also mentioned that unlike the MLE, the MPLE is not a function of the sufficient statistics, which means that different networks with the same sufficient statistics yield the same MLE but may have different MPLEs. This in turn implies that the MPLE violates the likelihood principle (Barnard et al., 1962; Birnbaum, 1962) according to which two networks with the same sufficient statistics should yield the same estimator.

All potential sufficient statistics as well as the corresponding convex hull can be obtained in **R** using *ergm.allstats()* in the **ergm**-package (Handcock et al., 2019) and *gConvexHull()* in the **rgeos**-package (Bivand and Rundel, 2020). Note that the calculation of all possible sufficient statistic combinations took about 6 hours for a network on 9 nodes. The computation time is increasing exponentially for each additional node.

```
R> library(ergm)
R> library(rgeos)
# create a network on 9 nodes
R> net <- network(n=9, directed=FALSE)
# calculate all possible vectors of statistics on a network
R> stats <- ergm.allstats(net~edges+ triangles , zeroobs=FALSE,
force=TRUE)
# produce the convex hull
```

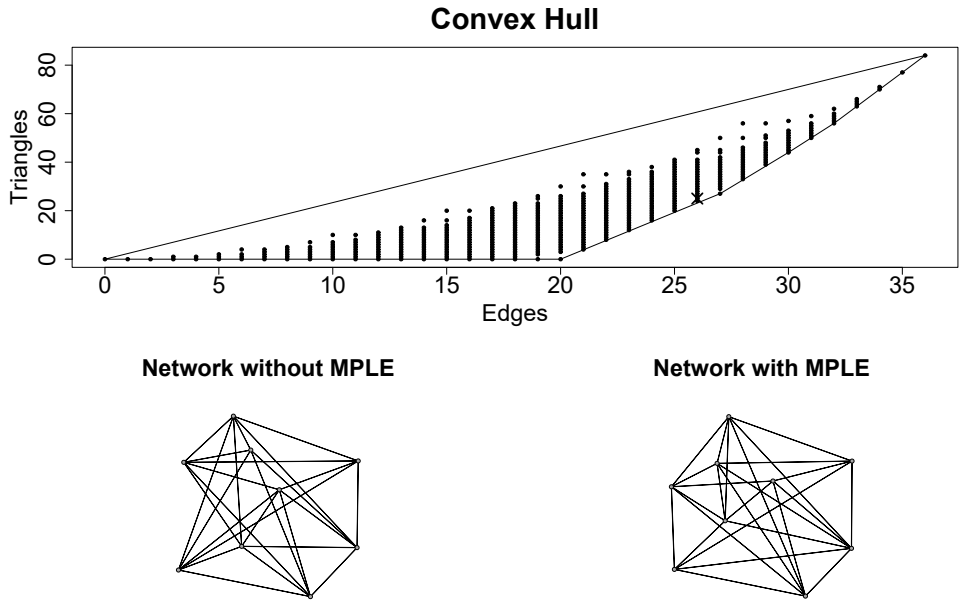

Figure 5.1. Top: Convex hull and potential sufficient statistics for a network of size $N = 9$ accounting for the number of edges and triangles. Bottom left: A network with 26 edges and 25 triangles that has no MPLE. Bottom right: A network with 26 edges and 25 triangles that has an MPLE.

```
R> ch <- gConvexHull(SpatialPoints(stats$statmat))
# plot sufficient statistics and convex hull
R> plot(a$statmat, pch=20, main="Convex Hull", xlab="Edges",
ylab="Triangles")
R> plot(ch, add=TRUE)
```

The top panel of Figure 5.1 visualizes every potential sufficient statistic of an ERGM that accounts for the number of edges and triangles of a network on $N = 9$ nodes and the resulting convex hull. Standard exponential family theory shows that the MLE exists for all statistics, except those on the boundary of the convex hull. This theory does not translate for the MPLE. Having 26 edges and 25 triangles, the network depicted in the bottom left panel of Figure 5.1 has statistics lying inside the convex hull, which guarantees the existence of an MLE, but interestingly one can show that the MPLE for this network does not exist. Konis (2007) shows that if the data may be separated, in

the sense that there exists a vector β such that

$$\beta^\top \cdot (T(A_{ij}^+) - T(A_{ij}^-)) \begin{cases} < 0 & \text{when } A_{ij} = 0, \\ > 0 & \text{when } A_{ij} = 1, \end{cases}$$

the MPLE does not exist. Konis (2007) argues that finding such β can be posed as a linear programming problem. In particular,

$$\begin{aligned} \text{maximize} \quad & (e^\top \bar{X})\beta \\ \text{subject to} \quad & \bar{X}\beta \geq 0 \end{aligned} \tag{5.5}$$

where $e^\top$ is a vector of ones, and $\bar{X}$ is the design matrix $(T(A_{ij}^+) - T(A_{ij}^-))$, where each element in a row that corresponds to a dyad with no tie, i.e., $A_{ij} = 0$, is being multiplied by -1. If there exist a β such that (5.5) has a solution, then the data is separable and the MPLE does not exist.

The network depicted in the bottom left of Figure 5.1 results in separated data, and hence does not yield an MPLE. However, this does not consequently mean that all networks with the same sufficient statistics do not have an MPLE. On the contrary, the network depicted in the bottom right of Figure 5.1 has 26 edges and 25 triangles just as the network to its left, with the difference of having an MPLE.

From version 4.0-5974 on the **ergm**-package automatically test for the existence of the MPLE through the **rcdd**-package (Geyer and Meeden, 2019). The following code creates the network depicted in the bottom left panel of Figure 5.1 and tests it for the existence of the MPLE.

```
R> set.seed(123)
# create an initial network on N=9 nodes
R> A <- matrix(rbinom(81, 1,0.3), 9,9)
R> diag(A)<- 0 # set diagonal elements to 0
# turn matrix into a network object
R> A <- as.network.matrix(A, directed=FALSE)
# simulate a network with 26 edges and 25 triangles using
# simulated annealing
R> san.net <- san(A ~ edges+triangles, target.stats=c(26,25))
# verify that the new network has 26 edges and 25 triangles
R> summary(san.net~edges+triangles)
```

```
   edges triangle
      26       25

R> model <- ergm(san.net ~ edges + triangles, estimate="MPLE")

Warning message:
In ergm.mple(nw, fd, m, MPLEtype = MPLEtype, init = init:
  The MPLE does not exist!
```

Another popular estimation approach for models with intractable normalizing constants is the Markov Chain maximum likelihood estimator (MCMLE), first proposed by Geyer and Thompson (1992) and then adapted to the ERGM framework by Snijders (2002) and Hunter and Handcock (2006). This family of estimation techniques attempts to approximate the MLE by estimating the normalizing constant through networks that were sampled using MCMC methods. The idea is that for any chosen $\theta_0 \in \mathbb{R}^q$ one can approximate the log-likelihood function by

$$\ell(\theta) - \ell(\theta_0) \approx (\theta - \theta_0)^\top \cdot T(A) - \log\left(\frac{1}{L}\sum_{i=1}^{L} \exp\Big((\theta - \theta_0)^\top \cdot T(A_i)\Big)\right),$$

where $A_1, \ldots, A_L$ are networks sampled from the distribution defined by θ_0.

The introduction of the MCMLE opened further possibilities for the evaluation of the MPLE. Analyzing two data sets (Sampson, 1968; Krackhardt, 1987) and comparing MCMLE and MPLE and the resulting standard errors, Snijders (2002) concluded that the MPLE has a tendency to underestimate standard errors and should therefore not be trusted. Later, Robins et al. (2007) and Lubbers and Snijders (2007) came to the same conclusion. This result, however, was to be expected, especially since Strauss and Ikeda (1990) had already clarified that "the quoted standard errors of the estimated parameters do not apply, because the [...] observations in the regression are certainly not independent" (p. 207). Despite the warning, Snijders (2002), Robins et al. (2007), and Lubbers and Snijders (2007) all appeared to draw their conclusions from the standard output from logistic regression-based estimates of the standard errors, which are based on an incorrect model.

van Duijn et al. (2009) investigated the efficiency, bias, standard errors, and confidence interval coverage rates of MPLE and of MCMLE, as an approximation of the MLE, for an undirected network on 36 nodes (Lazega, 2001) in natural and mean value

parameter space. The mean value parameter space is defined by the bijective mapping $\mu : \mathbb{R}^q \to \mathcal{C}$, $\mu(\theta) = \mathbb{E}_\theta[T(Y)]$, with $\mathcal{C}$ denoting the interior of the convex hull of the sufficient statistic's sample space. The convex hull $\overline{\mathcal{C}}$ is defined as the unique intersection of all convex sets containing $\{T(A) : A \in \mathcal{A}(\mathcal{N})\}$. For their simulation studies, van Duijn et al. (2009) treated a model's estimated MLE as the true parameter value and simulated networks from the corresponding probability distribution. Regarding the estimators' efficiency, the authors concluded that "the MLE is substantially more efficient than the MPLE" and that the difference is even more pronounced in mean value parameter space. In their studies, the MLE had larger bias in natural parameter space while the MPLE had larger bias in mean value parameter space. The MPLE standard errors obtained from the logistic regression output led to MPLE-based confidence intervals with coverage rates far below the nominal confidence level, confirming the conclusion made by Snijders (2002) that they are in general underestimated.

5.3 Estimating Standard Errors for MPLE

Although Strauss and Ikeda (1990) as well as van Duijn et al. (2009) acknowledge that MPLE standard errors obtained from logistic regression output are unsuitable, to the best of our knowledge, no one has yet formally introduced a correct way to specify standard errors for the MPLE in ERGMs.

Based on the log-pseudolikelihood (5.4), let us define $s(\theta)$ to be the vector of first derivatives,

$$s_k(\theta) = \frac{\partial}{\partial \theta_k} p\ell(\theta) = \sum_{ij} \left(A_{ij}\Delta_{ijk} - \frac{\exp(\theta\Delta_{ij})}{1+\exp(\theta\Delta_{ij})}\Delta_{ijk} \right), \tag{5.6}$$

and $J(\theta)$ the negative Hessian matrix,

$$J_{kl}(\theta) = -\frac{\partial}{\partial \theta_l} s_k(\theta) = \sum_{ij} \left(\frac{\exp(\theta\Delta_{ij})}{\left(1+\exp(\theta\Delta_{ij})\right)^2}\Delta_{ijk}\Delta_{ijl} \right), \tag{5.7}$$

where $k, l \in \{1, \ldots, q\}$. The Hessian matrix does not depend on the random variable A_{ij}, since the log-pseudolikelihood has the form of an exponential family model. Thus, $J(\theta) = -\nabla s(\theta) = -\mathbb{E}_\theta[\nabla s(\theta)]$, i.e., the negative Hessian matrix is both the Fisher information and the observed Fisher information. Furthermore, $J(\theta)^{-1}$ is the approximate covariance matrix used by logistic regression software to estimate standard errors.

One feature of a correctly specified likelihood $\ell(\theta)$ is that the Bartlett identities,

$$\begin{aligned} \mathbb{E}_\theta[\ell'(\theta)] &= 0, & (5.8) \\ \mathrm{Var}_\theta[\ell'(\theta)] &= -\mathbb{E}_\theta[\ell''(\theta)], & (5.9) \end{aligned}$$

hold, which justifies $J(\hat{\theta})^{-1}$ as the covariance matrix of $\hat{\theta}$. However, the pseudolikelihood is a form of misspecified likelihood, where (5.8) and (5.9) do not apply anymore, which consequently makes $J(\hat{\theta})^{-1}$ an incorrect covariance matrix for an estimator $\hat{\theta}$.

A more suitable method to estimate MPLE standard errors is by the calculation of the Godambe matrix (Godambe, 1960), also known as the sandwich information matrix, instead of the inverse Fisher information matrix, as for instance, was demonstrated for Potts models by Okabayashi et al. (2011). The Godambe matrix is defined as

$$G(\theta) = J(\theta)^{-1} V(\theta) J(\theta)^{-1}, \tag{5.10}$$

where $J(\theta)$ is referred to in this context as the sensitivity matrix and $V(\theta) = \mathrm{Var}_\theta[s(\theta)]$ is called the variability matrix. We can justify the Godambe matrix by usual Taylor approximation

$$s(\hat{\theta}) \approx s(\theta) + J(\theta)(\hat{\theta} - \theta),$$

where since $s(\hat{\theta}) = 0$ we obtain

$$(\hat{\theta} - \theta) \approx \left[-J(\theta)\right]^{-1} \left[s(\theta)\right]. \tag{5.11}$$

The usual derivation of the Godambe matrix relies on the multivariate central limit theorem. However, since $s(\theta)$ is not the sum of independent and identically distributed random vectors, nor is it generally possible to apply standard asymptotic arguments in the case of ERGMs (Shalizi and Rinaldo, 2013), it is not clear that we can employ this approach here. However, we may simply take the variance of Equation (5.11) to obtain

$$\mathrm{Var}(\hat{\theta}) \approx \left[-J(\theta)\right]^{-1} \mathrm{Var}\left[s(\theta)\right] \left[-J(\theta)\right]^{-1}.$$

Shalizi and Rinaldo (2013) show that many ERGMs are not consistent under sampling. In this paper we do not characterize consistency as the concept of an increasing sample size $N \to \infty$ as the number of observed networks $A_1^{obs}, \ldots, A_N^{obs}$ going to infinity, but as the number of nodes within a single observed network expanding. As a matter of fact, both MLE and MPLE are consistent and asymptotically normal for a sample of

n networks from a distribution P_θ as shown by Arnold and Strauss (1988), but it is not the prevailing situation that multiple networks are being sampled from a common distribution. It is more common to observe a subnetwork from a larger population network.

The variability matrix can in general not be directly computed for an ERGM. For this reason, we propose for the approximation of $V(\theta)$ to simulate R networks $A_1, \ldots, A_R$ from the distribution defined by the MPLE and then to calculate the vector of first derivatives of the pseudolikelihood function $s^1(\theta), \ldots, s^R(\theta)$ for each of the simulated networks. Here, the superscript indicates the vector of first derivatives of the rth simulated network. Let $\bar{s}(\theta) = R^{-1} \sum_{r=1}^{R} s^r(\theta)$ be the sample mean vector. Then

$$\hat{V}(\theta) = \frac{1}{R-1} \sum_{r=1}^{R} \left(\Big(s^r(\theta) - \bar{s}(\theta) \Big) \Big(s^r(\theta) - \bar{s}(\theta) \Big)^T \right). \tag{5.12}$$

If $\tilde{\theta}$ denotes the MPLE, then the Godambe matrix can be estimated as

$$\hat{G}(\tilde{\theta}) = J(\tilde{\theta})^{-1} \hat{V}(\tilde{\theta}) J(\tilde{\theta})^{-1}. \tag{5.13}$$

The estimation based on the Godambe matrix therefore requires the simulation of networks, which may appear to be a disadvantage. However, the simulation of networks serves the dual purpose of helping assess model degeneracy (Schmid and Desmarais, 2017): Simulated networks operate as a potential warning sign if the estimated probability distribution does not produce networks that appear to be sampled from the same distribution as the observed network. As we will illustrate later, the MPLE is especially prone to defining probability distributions that put most mass on networks that do not resemble the observed network.

The **ergm**-package can also handle an offset in the model, i.e., models of the form

$$\text{logit}(P_\theta(Y_{ij} = 1 | Y_{ij}^c = A_{ij}^c)) = \theta^\top \cdot (T(A_{ij}^+) - T(A_{ij}^-)) + \beta(t(A_{ij}^+) - t(A_{ij}^-))$$

where $\beta : \mathbb{R} \to \mathbb{R}$ is a function of $\delta_{ij} = t(A_{ij}^+) - t(A_{ij}^-)$, the change statistic of the offset variable. Common forms of the β-function are $\log(\delta_{ij})$ and $\gamma \cdot \delta_{ij}$, for a known $\gamma \in \mathbb{R}$. If an offset is present, then the calculation of the inverse Hessian matrix and the estimation of the Godambe matrix have to be adjusted. In particular, the score function (5.6)

changes to

$$s_k(\theta) = \frac{\partial}{\partial \theta_k} p\ell(\theta) = \sum_{ij} \left(A_{ij}\Delta_{ijk} - \frac{\exp(\theta\Delta_{ij} + \beta(\delta_{ij}))}{1 + \exp(\theta\Delta_{ij} + \beta(\delta_{ij}))}\Delta_{ijk} \right),$$

and the negative Hessian matrix (5.7) turns into

$$J_{kl}(\theta) = -\frac{\partial}{\partial \theta_l} s_k(\theta) = \sum_{ij} \left(\frac{\exp(\theta\Delta_{ij} + \beta(\delta_{ij}))}{\Big(1 + \exp(\theta\Delta_{ij} + \beta(\delta_{ij}))\Big)^2}\Delta_{ijk}\Delta_{ijl} \right).$$

In addition to offset models, the **ergm**-package can estimate models with a wide range of sample space constraints. One particular sample space constraint that applies to citation networks, is that a document that has been published before another document can not cite the latter (see Schmid et al. (2021a) for an application of a citation ERGM). Let $\Psi(\mathcal{N}) \subset \Omega(\mathcal{N})$, be the constraint sample space of a network on N nodes. Then, the pseudolikelihood simplifies to

$$p\ell_\Psi(\theta) = \sum_{ij \in \Psi(\mathcal{N})} \Big[A_{ij} \cdot (\theta\Delta_{ij}) - \log\Big(1 + \exp(\theta \cdot \Delta_{ij}\Big)\Big], \tag{5.14}$$

and the score function (5.6) and the negative Hessian matrix (5.7) are being defined as

$$s_{k,\Psi}(\theta) = \frac{\partial}{\partial \theta_k} p\ell_\Psi(\theta)$$

and

$$J_{kl,\Psi}(\theta) = -\frac{\partial}{\partial \theta_l} s_{k,\Psi}(\theta).$$

5.4 Simulation Studies

We calculate MPLE confidence intervals based on the Fisher information and the Godambe matrix and study the coverage rates of both methods. In addition, we also compare coverage rates of MCMLE confidence intervals (Hunter and Handcock, 2006) and 95% parametric bootstrap confidence intervals of the MPLE (Schmid and Desmarais, 2017; Hughes et al., 2011).

N=50	Edges	Two-stars	Triangles
MCMLE	0.952	0.956	0.964
Logistic	0.744	0.742	0.770
Godambe	0.952	0.948	0.964
Bootstrap	0.902	0.902	0.962

N=100	Edges	Two-stars	Triangles
MCMLE	0.940	0.948	0.940
Logistic	0.676	0.702	0.750
Godambe	0.954	0.952	0.936
Bootstrap	0.918	0.926	0.942

N=200	Edges	Two-stars	Triangles
MCMLE	0.960	0.948	0.946
Logistic	0.610	0.620	0.746
Godambe	0.952	0.954	0.946
Bootstrap	0.920	0.924	0.936

N=300	Edges	Two-stars	Triangles
MCMLE	0.950	0.946	0.956
Logistic	0.608	0.620	0.762
Godambe	0.950	0.950	0.942
Bootstrap	0.912	0.922	0.938

Table 5.1. Coverage rates of 95% confidence intervals for the (ρ, σ, τ)-model for four different network sizes $(N = 50, 100, 200, 300)$.

The parametric bootstrap confidence intervals are obtained by simulating 500 new networks from the distribution defined by the MPLE, and then estimating the MPLEs for each of the 500 simulated networks. The standard error obtained from the 500 MPLEs is being taking as the standard error that results in the bootstrap confidence interval.

It should be noted that in the following simulation study the true model is known. In this case the parametric bootstrap approach generates a legitimate sampling distribution and as a result, is a solid method of comparison. However, little is known about the parametric bootstrap distribution of a misspecified model which means that in such case the bootstrap distribution of the MPLE could be unsuitable.

We compare the coverage rates of all four methods in a simulation study and on a real-life network. For the simulation study, we follow Desmarais and Cranmer (2012) and consider the undirected (ρ, σ, τ)-model as introduced by Frank and Strauss (1986), where $\theta = (\rho, \sigma, \tau)$ represent the parameters for a network's number of edges, 2-stars and triangles, respectively. Recall that in an undirected network a two-star is defined as any set of edges (i, j) and (i, k), and a triangle is defined as any set of edges (i, j), (j, k) and (k, i), $i, j, k \in \{1, \ldots, N\}$. We set $\rho = -0.25, \sigma = -0.2$ and $\tau = 0.5$ and simulate 500 undirected networks for four different sizes $(N = 50, 100, 200, 300)$. For each simulated network we obtain the MCMLE with its corresponding 95% confidence interval as well as the MPLE with 95% confidence intervals

estimated by the Fisher matrix, the Godambe matrix, and parametric bootstrapping. The results are summarized in Table (5.1). Since estimating MPLE standard errors using Fisher information corresponds to logistic regression, we will denote coverage rates obtained by this method by *logistic*.

The following R-code demonstrates how one can obtain confidence intervals for each of the four methods.

```
R> library(ergm)
# create an initial undirected network on N=50 nodes
R> init.net <- network(N=50, directed = FALSE, density = 0.1)
# set the parameters
R> truth <- c(-0.25, -0.2, 0.5)
# simulate a network from the ERGM defined by the true parameter
R> sim.net <- simulate(init.net~ edges+kstar(2)+ triangles,
R> nsim=1, coef=truth)
# get an MCMLE
R> mcmle <- ergm(sim.net ~ edges+kstar(2)+ triangles)
# get the MPLE with inverse Fisher standard errors
R> logistic <- ergm(sim.net ~ edges+kstar(2)+ triangles,
R> estimate="MPLE")
# get the MPLE with Godambe standard errors
R> godambe <- ergm(sim.net ~ edges+kstar(2)+ triangles,
R> estimate="MPLE",
R> control=control.ergm(MPLE.covariance.method="Godambe"))
# get the MPLE with bootstrap confidence intervals
R> bootstrap <- ergm(sim.net ~ edges+kstar(2)+ triangles,
R> estimate="MPLE",
R> control=control.ergm(MPLE.covariance.method="bootstrap"))
```

As expected MCMLE confidence intervals yield coverage rates close to the anticipated 95% regardless the network size. On the other hand, the logistic intervals are nowhere close to the desired coverage rate, with results ranging between 60% and 76%. These results, however, are not surprising, since an improper method to obtain standard errors was applied. It is also interesting to note that the overall coverage rate for this method appears to decrease as the network size increases. Calculating the MPLE standard errors based on the Godambe matrix, however, yields confidence intervals that perform just as well as MCMLE confidence intervals. The fourth method, confidence intervals by

	Structural		Nodal		Homophily		
	Edges	GWESP	Seniority	Practice	Practice	Gender	Office
MCMLE	0.944	0.920	0.937	0.940	0.958	0.954	0.957
Logistic	0.981	0.763	0.980	0.976	0.978	0.982	0.982
Godambe	0.942	0.922	0.944	0.933	0.939	0.917	0.922
Bootstrap	0.930	0.982	0.966	0.960	0.962	0.943	0.928

Table 5.2. Coverage Rates for the Lazega Law Firm Collaboration Network.

parametric bootstrap, clearly outperforms the logistic regression results, but also appears to not quite reach the anticipated coverage rates.

Next we apply the four methods on the same data as used by van Duijn et al. (2009), a collaboration network between 36 partners within a New England law firm (Lazega, 2001). We treat this network as an undirected network and only consider an edge between two partners if both sides indicate to have collaborated with each other. The model we use was initially specified by Hunter and Handcock (2006) and only slightly modified by van Duijn et al. (2009). The endogenous statistics consist of the number of edges, which is equivalent to the intercept in a logistic model, and the geometrically weighted edgewise shared partners statistic (GWESP), a statistic used to model the tendency towards triangles and clustering (Hunter and Handcock, 2006). The decay parameter for the GWESP-statistic has been fixed to its MLE (0.7781). Among the exogenous statistics, we include *seniority*, the rank number of partners divided by 36, as well as *practice* (corporate or litigation) as nodal attributes, and *practice*, *gender* and *office* as dyadic homophily attributes.

As usual for real data sets, the true parameter θ is unknown. Therefore, we obtain a MCMLE and treat this estimate as the truth. At this point we conduct another simulation study, where we simulate 1000 networks from the distribution defined by the MCMLE and calculate 95% confidence intervals for all four methods. Finally, we test whether the *true* value falls within a computed interval. Coverage rates for the four methods are reported in Table 5.2.

The coverage rates for the MCMLE as well as for the MPLE using the Fisher matrix align with the results of van Duijn et al. (2009). While the MCMLE yields coverage rates close to 95%, MPLE coverage rates appear to overestimate standard errors with the exception of the *GWESP*-statistic, which coverage rate is clearly too small. MPLE coverage rates that were obtained using an approximated Godambe matrix provide similarly satisfying results as the MCMLE rates for structural and nodal variables.

However, standard errors of variables accounting for homophily appear to be too small. The fourth method, parametric bootstrap intervals for the MPLE, provides adequate results.

The third simulation study investigates the performance of the inverse Hessian and Godambe matrix as estimates of the covariance matrix changes as the network size increases. We follow Krivitsky et al. (2011) by adding an offset to the model to adjust for network size ensuring that a node's mean degree remains the same as the network size increases. Since the number of edges translates to the intercept of an ERGM and therefore, converts to the density of a network, we depend the parameter for the number of edges statistic on the initially simulated network on the network size N. For a given network size N, we simulate an initial network from the ERGM consisting of the number of edges and the number of triangles with parameters set to 4 and -0.2, respectively. The offset is set to $\log(1/N)$. Next, we obtain the MPLE of the simulated network, treat the MPLE as the truth, and simulate 1000 networks. Since the model defined by the MPLE represents the true underlying model, we can calculate the actual inverse Hessian matrix $J(\theta)$ using equation (5.7). For the estimation of the variability matrix $V(\theta)$ we calculate the MPLE for each of the 1000 simulated networks and apply equation (5.12).

Figure 5.2 visualizes the 95% confidence ellipses calculated from covariance matrices for $N = 9, 50, 100$ and 500. The 'x' represents the true MPLE, while every grey dot represents the MPLE of one of the 1000 simulated networks. The dashed lines indicate the 95% confidence ellipses calculated from the inverse Hessian covariance estimate, while the solid lines indicate the 95% confidence ellipses obtained from the estimated Godambe matrix. The following code creates the Figure 5.2 panel for $N = 50$.

```
R> library(ergm)
R> library(ellipse)
R> set.seed(14392)
# create an initial matrix on N=50 nodes
R> N=50; A <- matrix(rbinom(N^2, 1,0.005), N,N)
R> diag(A)<- 0; A <- as.network.matrix(A, directed=FALSE)
# simulate a network from the true distribution
R> init.sim <- simulate(A ~ edges+triangles, nsim=1,
R> coef=c(log(1/N)+4,-0.2))
# calculate init.sim's MPLE and approximate Godambe matrix
R> m50 <- ergm(init.sim ~ edges+triangles, estimate="MPLE",
R> control=control.ergm(MPLE.covariance.method = "Godambe"))
```

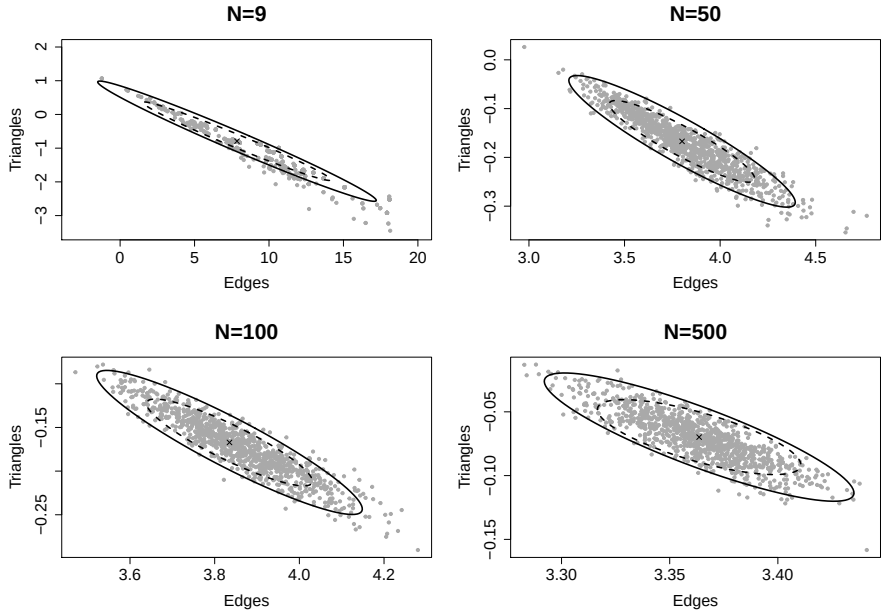

Figure 5.2. 95% confidence ellipses of the edges-triangle model with n nodes calculated using the inverse Hessian matrix (dashed) and Godambe matrix (solid). The 'x' indicates the parameters of the true model distribution. Every grey dot represents the MPLE of a network that was sampled from the true underlying distribution.

```
# calculate inverse Hessian
R> m50_2 <- ergm(init.sim ~ edges+triangles, estimate="MPLE")
# simulate 1000 networks
R> sim.net <- simulate(m50, nsim=1000)
# calculate MPLE of simulated networks,
R> sim.coef50 <- matrix(0, nrow=1000, ncol=2)
R> for(i in 1:1000){
R>    sim.mple<- ergm(sim.net[[i]] ~ edges+triangles,
R> estimate="MPLE")
R>    sim.coef50[i,]<- sim.mple$coef
R> }
# plot
R> plot(sim.coef50[,1], sim.coef50[,2], xlab="Edges",
R> ylab="Triangles", main="N=50", cex.axis=1.5, cex.lab=1.5,
R> cex.main=2,
R> pch=20, col="darkgrey" )
```

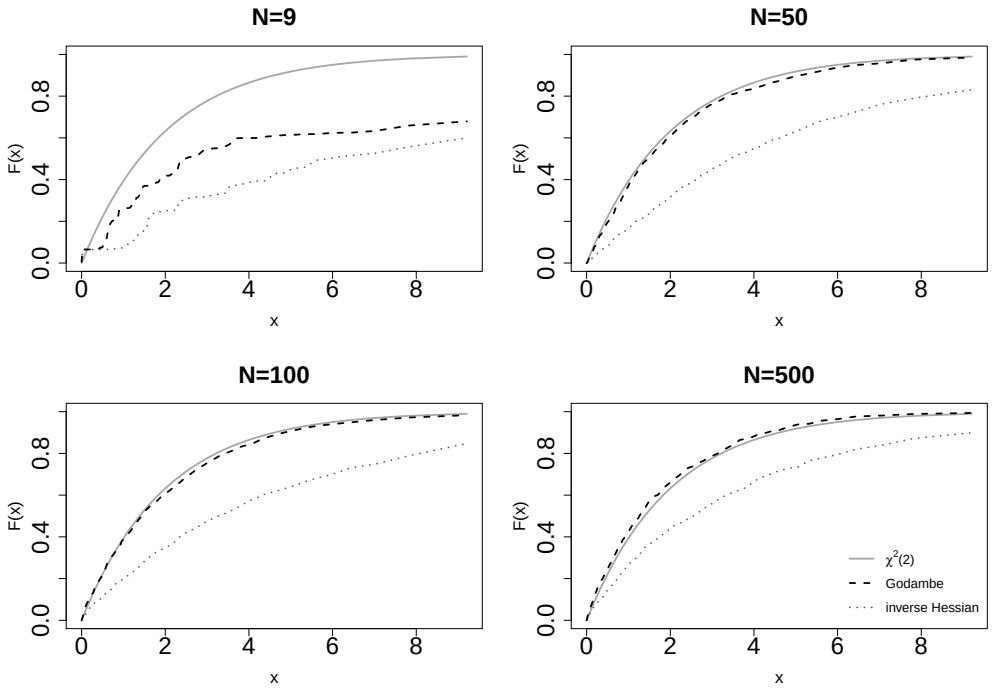

Figure 5.3. Empirical coverage rates of confidence ellipses obtained by the Godambe matrix, depicted by the dashed line (- - -), and the inverse Hessian matrix, depicted by the dotted line (···). The solid grey line shows the cdf of a χ^2-distribution with two degrees of freedom.

```
R> points(m50$coef[1], m50$coef[2], pch=4)
R> lines(ellipse(m50_2$covar, centre= m50$coef, level=0.95),
R> lty=2, lwd=2)
R> lines(ellipse(m50$covar, centre=m50$coef, level=0.95), lwd=2)
```

The multiplier to create confidence ellipses are obtained from a $\chi^2(2)$-distribution and the corresponding cdf is depicted as solid line in Figure 5.3. In addition, Figure 5.3 depicts the empirical coverage rates of the confidence ellipses that were obtained using the Godambe matrix (dashed line) and inverse Hessian matrix (dotted line). Based on this figure we can compare the expected coverage rates of a correctly specified confidence ellipse with the empirical coverage rates.

The results underline the conclusions made in the previous simulation studies that confidence intervals that are based on the inverse Hessian matrix are not reliable. For neither sample size the coverage rates of the inverse Hessian confidence ellipses are anywhere close to the intended coverage rates and it also doesn't appear to converge to the designated coverage rates as the network size increases. On the contrary, coverage

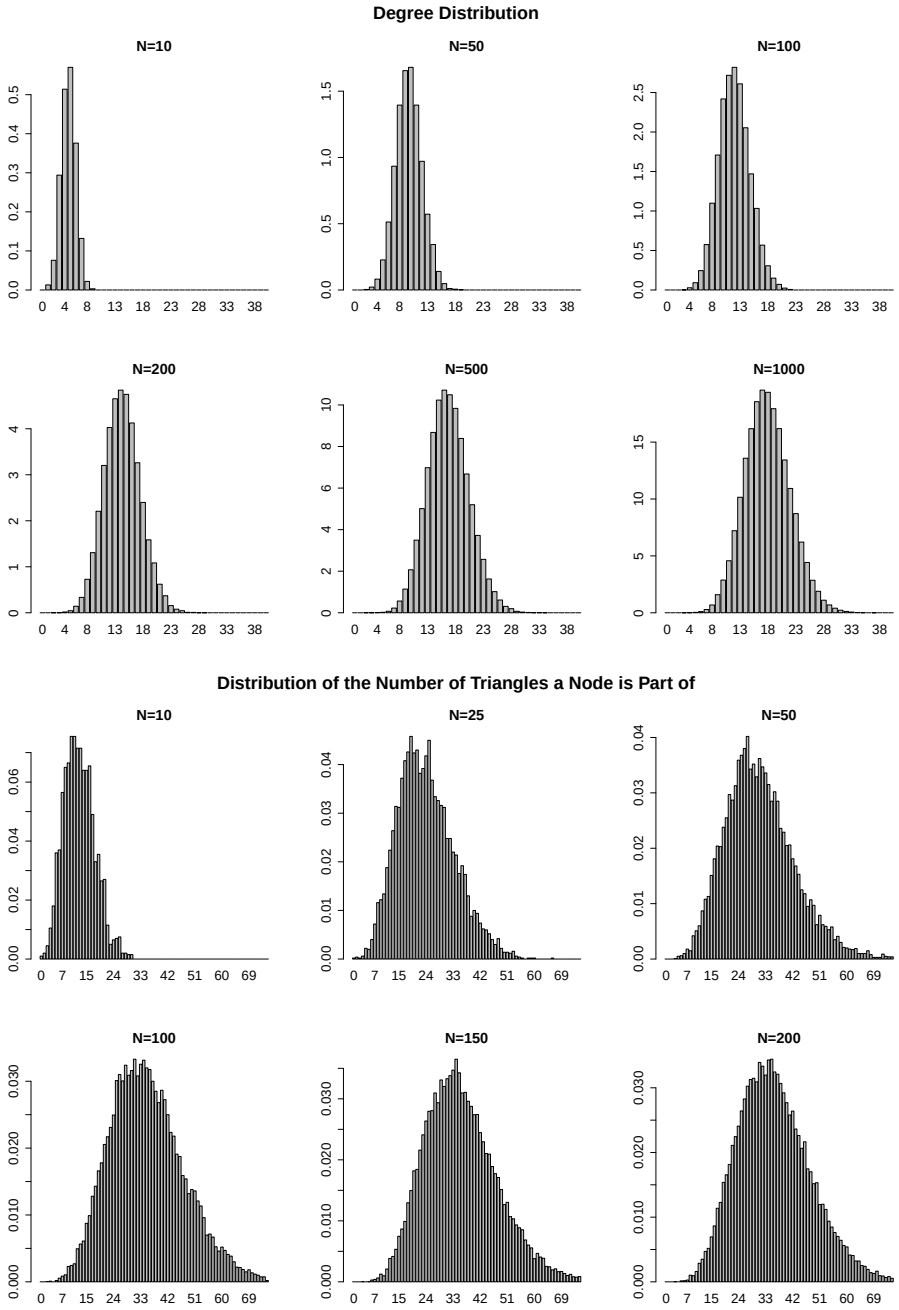

Figure 5.4. Degree distribution and distribution of the number of triangles a node is part of for an increasing network size.

rates of Godambe matrix ellipses are similar to the intended coverage rates for networks of already relatively small size N. While the empirical coverage rates for the network of size $N = 9$ is smaller than the expected coverage rates, the confidence ellipses for larger networks approximately cover the intended percentage of MPLEs.

This result, in particular, is interesting since Shalizi and Rinaldo (2013) conclude that most ERGMs cannot be embedded in a normally asymptotic framework, and yet we have evidence that the covariance does adhere to some sort of asymptotic.

So, what is the difference between the model that we have explored and the ERGMs that Shalizi and Rinaldo (2013) considered? The difference lies in the inclusion of the offset $\log(1/N)$. The work of Shalizi and Rinaldo (2013) only applies to a framework in which the number of nodes increases while the coefficients stay fixed. By including the offset these results do not apply anymore. Krivitsky et al. (2011) show that for the dyad-independent Erdös-Renyi-Gilbert model with this offset, i.e.,

$$P(Y = A|\theta) \propto \exp\Big((\log(1/N) + \theta) \cdot e(A)\Big),$$

the degree distribution converges to a Poisson distribution with mean $\exp(\theta)$. This means that a node's mean degree converges to $\exp(\theta)$ as the number of nodes N increases. Showing the asymptotic behavior of this dyad-independence model is not too surprising, especially since these results do not contradict the results of Shalizi and Rinaldo (2013). However, we have reason to believe that it is possible to extend this result to a Markov dependence model. A network is called Markov dependent if dyads that do not share a common node are conditionally independent (Frank and Strauss, 1986). The model in the latest simulation study accounts for the number of edges and triangles in a network, and we have empirical evidence that, in addition to the mean degree of a node converging as network size increases, the mean number of triangles a node is part of also converges as N increases. In other words, it appears to be possible to extend the work of Krivitsky et al. (2011) to dyad-dependent models, which would not only be novel but also establish a more general framework in which dyad-dependent asymptotics are achievable.

Figure 5.4 visualizes the degree distribution and the distribution of a node's number of triangles for an increasing network size. For each network size N, 200 networks were simulated from the ERGM

$$P(Y = A|\theta) \propto \exp\Big((\log(1/N) + \theta_1) \cdot e(A) + \theta_2 \cdot t(A)\Big),$$

where $\theta_1 = 4$ and $\theta_2 = -0.2$, and $e(A)$ and $t(A)$ refer to the network's number of nodes

and triangles, respectively. The degree distribution as well as the distribution of the number of triangles a node is part of was determined for each of the 200 networks and then averaged. Each bar in the lower part of Figure 5.4 depicts the average percentage of nodes in the simulated networks that are part of the given number of triangles.

As we can see from this figure, this distribution appears to converge as the network size increases. Interestingly, besides the offset proposed by Krivitsky et al. (2011), no additional offset term for the triangle statistic was necessary for the distribution to seemingly converge. Note that Frank and Strauss (1986) show that any Markov dependence model can be described by the ERGM that accounts for the number of edges, k-stars, and triangles in a network. Note that the k-star distribution of a node can be exactly determined by a node's degree, which makes us confident that Markov dependence graphs with offset can be embedded into an asymptotic framework.

5.5 Discussion

This paper proposes to estimate MPLE standard errors for ERGMs using an estimated Godambe matrix. Even though the exact calculation of the Godambe matrix is infeasible for most ERGMs, the approximated Godambe matrix performed exceptionally well in all simulation studies. In addition, the newly proposed method tests for model degeneracy, a check that can easily be forgotten when estimating the MPLE the conventional way.

There is still much to learn about the behavior of the MPLE. Despite the work of Shalizi and Rinaldo (2013), we have evidence that the covariance does adhere to some sort of asymptotic. This observation is reflected in the percentage of MPLEs that fall inside the $1 - \alpha\%$ Godambe confidence ellipses as depicted in Figure 5.2. It appears as if confidence ellipses that are based on the Godambe matrix cover the intended amount of MPLEs as the network size increases.

In addition, this paper provides empirical evidence that a type of consistency is possible even in ERGMs that lack dyadic independence. Confirming this assumption would undoubtedly be a very important result for statistical inference of ERGMs.

Chapter 6
Discussion and Future Work

In this thesis, I investigated and compared the MLE/MCMLE and the MPLE as different estimators of an ERGM. The MLE is computationally demanding, however, it possesses many favorable properties. One of these properties is that it satisfies the likelihood principle, which states that two networks of the same size with the same sufficient statistics have the same MLE. The MPLE, on the other hand, does not satisfy this property, which appears to be unfavorable at first, but can be turned into an advantage to improve starting values for the MCMLE algorithm. In Chapters 2 and 3, I have demonstrated that it is possible to search for networks with the same or at least similar sufficient statistics and to use these networks' MPLEs as alternative starting values. A naturally arising question is what distinguishes networks that lead to an MPLE close to the MLE from those networks with MPLEs further away. We suggest that it is preferable to search for networks that, despite having the same statistics as the observed one, can be accurately modeled by a dyad-independent model, since in this case MLE=MPLE. This, however, raises questions about the initial model specification. If the MLE of a dyad-dependent model equals the MPLE of an Erdös-Renyi-Gilbert generated network that has the very same network statistics, then this casts doubt on whether the observed network's dependency structure has been accurately captured in the model.

Exploring the space of possible MPLEs of networks with the same sufficient statistics and starting the simulated annealing algorithm with either the observed network or a dyad-independent network resulted in two distinct clouds of MPLE values. This leads to the question of whether there are any further clouds of MPLE values that have remained undetected and could potentially be of use as MCMLE starting values.

The MPLE for dyad-dependent models has mostly been regarded negatively in the ERGM literature. One reason was that MPLE standard errors appeared to be unreliable (van Duijn et al., 2009). In this thesis, I illustrate that until now MPLE standard errors

have been calculated incorrectly and introduce two approaches, the parametric bootstrap and the Godambe information matrix, to obtain standard errors with appropriate coverage rates. Both these methods require the simulation of networks, which 1) enables model degeneracy detection and 2) permits the comparison of MPLE and MLE. In the process of examining these techniques we found empirical evidence that, despite the work of Shalizi and Rinaldo (2013), some type of consistency for dyad-dependent ERGMs might be possible after all. Schweinberger and Stuart (2020) provide non-asymptotic consistency results for random graphs with additional block structure. Krivitsky et al. (2011) provide a theoretical proof that an Erdös-Renyi-Gilbert model with offset $\log(1/N)$ can be embedded into an asymptotic framework. Generalizing this proof to Markov dependence models with offset or dyad-dependence models with offset would be a potentially very important result for ERGM inference.

The decisive idea for the proof could be found in the composite likelihood literature. For instance, Lindsay et al. (2011) claim that for composite likelihoods, of which the pseudolikelihood is a special form, the Kullback-Leibler inequality applies, which implies that maximization of the likelihood leads to Fisher-consistent estimation of the parameter vector. Fisher consistency states that if an estimator is calculated using an entire population, the true value of the estimated parameter will be obtained. Lindsay et al. (2011) state that with added regularity conditions, Fisher consistency often implies consistency in probability.

Asymptotic normality is not the only aspect for which composite likelihoods could provide meaningful insight. As discussed in Chapter 1, both composite likelihoods and contrastive divergence put the MPLE and the MLE on two opposing ends of a spectrum. A natural question that arises is whether there is a connection between contrastive divergence and composite likelihoods. The answer is: Yes!

Asuncion et al. (2010) show that the connection lies in blocked contrastive divergence, which applies blocked Gibbs sampling in the MCMC. Blocked Gibbs sampling either randomly or systematically selects a block U_c of a network and updates the entire block collectively, while always conditioning on the rest of the network. In other words, a subset U_c is chosen and, conditional on the rest of the network $A^c_{U_c} = A_{V_c}$, the elements of A_{U_c} are jointly updated using the full conditional probability $P_\theta(Y_{U_c} = A_{U_c} \mid Y_{V_c} = A_{V_c})$.
To see the connection, consider the composite loglikelihood for a network A:

$$c\ell(\theta) \;=\; \sum_{c\in\Omega(N)} \log P_\theta(Y_{U_c} = A_{U_c} | Y_{V_c} = A_{V_c})$$

$$
\begin{aligned}
&= \sum_{c\in\Omega(N)} \log P_\theta(Y_{U_c} = A_{U_c}, Y_{V_c} = A_{V_c}) - \sum_{c\in\Omega(N)} \log P_\theta(Y_{V_c} = A_{V_c}) \\
&= m \cdot \log \frac{\exp(\theta' g(A))}{k(\theta)} - \sum_{c\in\Omega(N)} \log \sum_{A^*_{U_c}} \frac{\exp(\theta' \cdot g(A^*_{U_c}, A_{V_c}))}{k(\theta)} \\
&\propto \theta' \cdot g(A) - \frac{1}{m} \sum_{c\in\Omega(N)} \log \underbrace{\sum_{A^*_{U_c}} \exp(\theta' \cdot g(A^*_{U_c}, A_{V_c}))}_{:=k_{c\ell}(\theta)}, \qquad (6.1)
\end{aligned}
$$

where $\sum_{A^*_{U_c}}$ indicates the sum of all possible configurations of subgraph A_{U_c}.

The calculation of the composite likelihood normalizing constant requires the calculation of two sums, $\sum_{c\in\Omega(N)}$, the sum over all possible blocks of size m and $\sum_{A^*_{U_c}}$, the sum over all possible networks on m nodes. The first sum consists of $\binom{N}{m}$ summands while the second one requires 2^m summands.

After having defined $k_{c\ell}(\theta)$ in (6.1), we can perform a similar calculation as in (1.8) by choosing any $\theta_0 \in \Theta$:

$$
\begin{aligned}
\frac{k_{c\ell}(\theta)}{k_{c\ell}(\theta_0)} &= \frac{1}{k_{c\ell}(\theta_0)} \sum_{A^*_{U_c}} \exp(\theta' \cdot g(A^*_{U_c}, A_{V_c})) \\
&= \sum_{A^*_{U_c}} \exp((\theta - \theta_0)' g(A^*_{U_c}, A_{V_c})) \cdot \frac{\exp(\theta_0' \cdot g(A^*_{U_c}, A_{V_c}))}{k_{c\ell}(\theta_0)} \\
&= \sum_{A^*_{U_c}} \exp((\theta - \theta_0)' g(A^*_{U_c}, A_{V_c})) \cdot P_{\theta_0}(Y_{U_c} = A^*_{U_c}, Y_{V_c} = A_{V_c}) \\
&= \mathbb{E}^{U_c}_{\theta_0} \left[\exp((\theta - \theta_0)' \cdot g(Y))\right]. \qquad (6.2)
\end{aligned}
$$

Similar to (1.10), this expectation can be approximated by Monte Carlo with application of blocked Gibbs sampling.

I will adapt the notation used by Asuncion et al. (2010) and Hummel (2011) and denote Bm-CDn as the blocked CD that uses n Gibbs updates and blocks of fixed size m. Consequently, B1-CDn refers to block sizes of one, or blocks that only contain one dyad. In other words, this is CD-n.

Next, consider B3-CDn, blocked contrastive divergence with block size 3. The simplest way to choose blocks of size 3 is by randomly sampling three dyads: ij, kl, and rs. Then, the joint probability is defined as

$$
P_\theta(Y_{ij} = a_1, Y_{kl} = a_2, Y_{rs} = a_3 \mid A^c_{ij,kl,rs}), \qquad (6.3)
$$

where $a_1, a_2, a_3 \in \{0, 1\}$ indicate the presence or absence of a tie and $Y^c_{ij,kl,rs} = A^c_{ij,kl,rs}$

denotes the network without the three chosen dyads ij, kl, and rs. When assuming Markov dependence, then the joint probability (6.3) is equal to

$$P_\theta(Y_{ij} = a_1 \mid A^c_{ij,kl,rs}) \times P_\theta(Y_{kl} = a_2 \mid A^c_{ij,kl,rs}) \times P_\theta(Y_{rs} = a_3 \mid A^c_{ij,kl,rs}) \tag{6.4}$$

and not much is gained in this particular case by considering blocks.

A more useful case is to restrict the blocks by dyads of the form ij, jk, and ik (Hummel, 2011). This means that only blocks where the three dyads form a triad are considered. These specific blocks capture one of the most dominant dependencies in networks: dyads between nodes in a group of three depend on each other. The resulting joint probability is then

$$P_\theta(Y_{ij} = a_1, Y_{jk} = a_2, Y_{ik} = a_3 \mid A^c_{ij,jk,ki}). \tag{6.5}$$

The question that now arises is whether B3-CDn for triads improves the estimates of an ERGM, i.e., is the resulting MCLE expected to be more similar to the MLE than the MPLE? Furthermore, since B3-CDn is expected to be somewhat closer to the MLE than the MPLE, it is of interest to investigate whether the estimates of uncertainty improve as well.

Even though blocking nodes of $m = 3$ simplifies the calculation of the normalizing constant, it can still result in a weighted sum that cannot be calculated directly. For example, the normalizing constant of a network on $N = 500$ nodes with $m = 3$, results in 165 million summands. The question of whether there is any advantage in approximating a composite likelihood with blocks smaller than N instead of approximating the full likelihood function has yet to be answered.

Another common source of dependence in networks is that of the popularity of a node. The more connections a node has the more likely it is for this node to create new connections. In the ERGM framework, this tendency of the popular becoming more popular is commonly captured by the k-star statistic, a statistic that counts the number of k dyads that share a common node. This dependency can be accounted for by BN-CDn and restricting the blocks to dyads of the form $i1$, $i2$, …, iN, where N is the number of nodes in the network. Then, the joint probability is

$$P_\theta(Y_{i1} = a_1, Y_{i2} = a_2, \ldots, Y_{iN} = a_N \mid A^c_{i1,i2,\ldots iN}), \tag{6.6}$$

where $Y^c_{i1,i2,\ldots,iN} = A^c_{i1,i2,\ldots iN}$ is the network without dyads $i1$, $i2$, …, iN. Just as for

the triad case, one can now examine whether accounting for popularity dependencies improves the estimation of coefficients as well as the estimation of uncertainty.

Note that combining blocks that can account for Markov dependencies as well as blocks that can account for popularity dependencies would result in a block that contains all dyads of a network and hence, in the full likelihood.

Considering the MLE and MPLE as two ends of a spectrum of possible estimators, raises additional open questions. For instance, there is still little known about how standard errors of contrastive divergence estimators can be ideally estimated. Is a Godambe information matrix approach just as reliable for this as it is for pseudolikelihoods? It also would be interesting to examine how the standard errors perform as the fixed step size n increases. In conclusion, there are still plenty of unanswered questions in the theory of ERGM estimation that require further exploration.

Bibliography

Akcigit, U. and W. R. Kerr (2018). Growth through heterogeneous innovations. *Journal of Political Economy 126*(4), 1374–1443.

Arnold, B. C. and D. Strauss (1988). Pseudolikelhood estimation. Technical report, University of California, Riverside Department of Statistics.

Asuncion, A., Q. Liu, A. Ihler, and P. Smyth (2010). Learning with blocks: Composite likelihood and contrastive divergence. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics 9*, 33–40.

Atchison, A. L. (2017). Negating the gender citation advantage in political science. *PS: Political Science & Politics 50*(2), 448–455.

Bailey, M. A. and F. Maltzman (2008). Does legal doctrine matter? Unpacking law and policy preferences on the US Supreme Court. *American Political Science Review 102*(3), 369–384.

Bailey, M. A. and F. Maltzman (2011). *The constrained court: Law, politics, and the decisions justices make*. Princeton University Press.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Barnard, G. A., G. M. Jenkins, and C. B. Winsten (1962). Likelihood inference and time series. *Journal of the Royal Statistical Society. Series A (General) 125*(3), 321–372.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. Wiley.

Basiak Jr, J. F. (2006). The roberts court and the future of substantive due process: The demise of split-the-difference jurisprudence. *Whittier L. Rev. 28*, 861.

Benjamin, S. M. and B. A. Desmarais (2012). Standing the test of time: The breadth of majority coalitions and the fate of us supreme court precedents. *Journal of Legal Analysis 4*(2), 445–469.

Berlemann, M. and R. Christmann (2020). Disposition time and the utilization of prior judicial decisions: Evidence from a civil law country. *International Review of Law and Economics 62*, 105887.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(3), 259–302.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*(298), 269–306.

Bivand, R. and C. Rundel (2020). *rgeos: Interface to geometry engine - Open source ('GEOS')*. R package version 0.5-3.

Black, R. C. and J. F. Spriggs (2013). The citation and depreciation of us supreme court precedent. *Journal of Empirical Legal Studies 10*(2), 325–358.

Bommarito II, M. J., D. Katz, and J. Zelner (2009). Law as a seamless web? Comparison of various network representations of the united states supreme court corpus (1791-2005). In *Proceedings of the 12th international conference on artificial intelligence and law*, pp. 234–235. ACM.

Box-Steffensmeier, J. M. and D. P. Christenson (2014). The evolution and formation of amicus curiae networks. *Social Networks 36*, 82–96.

Bratton, K. A. and S. M. Rouse (2011). Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Studies Quarterly 36*(3), 423–460.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology 38*(1), 155–200.

Caldeira, G. A. (1988). Legal precedent: Structures of communication between state supreme courts. *Social networks 10*(1), 29–55.

Carreira-Perpiñán, M. A. and G. Hinton (2005, 06–08 Jan). On contrastive divergence learning. In R. G. Cowell and Z. Ghahramani (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Volume R5 of *Proceedings of Machine Learning Research*, pp. 33–40. PMLR. Reissued by PMLR on 30 March 2021.

Chayes, J. (2013). Mathematics of web science: structure, dynamics and incentives. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1987), 20120377.

Clark, T. S. and B. Lauderdale (2010). Locating supreme court opinions in doctrine space. *American Journal of Political Science 54*(4), 871–890.

Collins Jr, P. M. (2008). Amici curiae and dissensus on the US Supreme Court. *Journal of Empirical Legal Studies 5*(1), 143–170.

Corander, J., K. Dahmström, and P. Dahmström (1998). Maximum likelihood estimation for Markov graphs. Technical report, Department of Statistics, University of Stockholm.

Costa, M., B. A. Desmarais, and J. A. Hird (2016). Science use in regulatory impact analysis: The effects of political attention and controversy. *Review of Policy Research 33*(3), 251–269.

Cranmer, S., T. Heinrich, and B. Desmarais (2014). Reciprocity and the structural determinants of the international sanctions network. *Social Networks 36*, 5–22.

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

Cranmer, S. J. and B. A. Desmarais (2016). A critique of dyadic design. *International Studies Quarterly 60*(2), 355–362.

Cranmer, S. J., B. A. Desmarais, and E. J. Menninga (2012). Complex dependencies in the alliance network. *Conflict Management and Peace Science 29*(3), 279–313.

Cross, F. B. (2010). Determinants of citations to supreme court opinions (and the remarkable influence of justice scalia). *Supreme Court Economic Review 18*(1), 177–202.

Cross, F. B. and S. Lindquist (2005). The decisional significance of the chief justice. *U. Pa. L. Rev. 154*, 1665.

Dahmström, K. and P. Dahmström (1993). ML-estimation of the clustering parameter in a Markov graph model. Technical report, Department of Statistics, University of Stockholm.

Danelski, D. J. and A. Ward (2016). *The Chief Justice: Appointment and Influence*. University of Michigan Press.

Darroch, J. and R. Ratcliff (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics 43*, 1470–1480.

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

Desmarais, B. A. and S. J. Cranmer (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications 391*(4), 1865–1876.

Dey, C. J. and J. S. Quinn (2014). Individual attributes and self-organizational processes affect dominance network structure in pukeko. *Behavioral Ecology 25*(6), 1402–1408.

Dincer, O. (2019). Does corruption slow down innovation? Evidence from a cointegrated panel of US states. *European Journal of Political Economy 56*, 1–10.

Dion, M. L., J. L. Sumner, and S. M. Mitchell (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis 26*(3), 312–327.

Duque, M. G. (2018). Recognizing international status: A relational approach. *International Studies Quarterly 62*(3), 577–592.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. In *CBMS-NSF Regional conference series in applied mathematics. 25 cm. 92 p.* Society for industrial and applied mathematics, Philadelphia. US.

Erdös, P. and A. Renyi (1959). On random graphs. *Publicationes Mathematicae Debrecen 6*, 290–297.

Erikson, E. (2013). Formalist and relationalist theory in social network analysis. *Sociological Theory 31*(3), 219–242.

Ethayarajh, K., A. Green, and A. H. Yoon (2018). A rose by any other name: Understanding judicial decisions that do not cite precedent. *Journal of Empirical Legal Studies 15*(3), 563–596.

Fix, M. P. and B. R. Fairbanks (2019). The effect of opinion readability on the impact of US Supreme Court precedents in state high courts. *Social Science Quarterly 101*(2), 811–824.

Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis 14*(04), 456–487.

Fowler, J. H. and S. Jeon (2008). The authority of supreme court precedent. *Social networks 30*(1), 16–30.

Fowler, J. H., T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck (2007). Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court. *Political Analysis 15*(3), 324–346.

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

Geyer, C. J. and G. D. Meeden (2019). *rcdd: Computational Geometry*. R package version 1.2-2.

Geyer, C. J. and E. A. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological) 54*(3), 657–699.

Gilbert, E. N. (1959, 12). Random graphs. *Ann. Math. Statist. 30*(4), 1141–1144.

Gillman, H. (2001). What's law got to do with it? Judicial behavioralists test the 'legal model' of judicial decision making. *Law & Social Inquiry 26*(2), 465–504.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics 31*(4), 1208–1211.

Goodreau, S. M., J. A. Kitts, and M. Morris (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography 46*(1), 103–125.

Hackney, J. K. and K. W. Axhausen (2006). An agent model of social network and travel behavior interdependence. In *Conference on Issues in Behavioral Demand Modeling and the Evaluation of Travel Time*.

Hallinan, M. T. and W. N. Kubitschek (1990). Sex and race effects of the response to intransitive sentiment relations. *Social psychology quarterly 53*(3), 252–263.

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Technical report, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, and M. Morris (2019). *ergm: Fit, simulate and diagnose exponential-family models for networks*. The Statnet Project (`https://statnet.org`). R package version 3.10.0-4837.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software 24*(1), 1548–7660.

Hansford, T. G. and J. F. Spriggs (2006). *The politics of precedent on the US Supreme Court*. Princeton University Press.

Hardt, H., H. J. Kim, A. E. Smith, and P. Meister (2019). The gender readings gap in political science graduate training. *The Journal of Politics 81*(4), 1528–1532.

Heaney, M. (2014). Multiplex networks and interest group influence reputation: An exponential random graph model. *Social Networks 36*, 66–81.

Heaney, M. and P. Leifeld (2018). Contributions by interest groups to lobbying coalitions. *The Journal of Politics 80*(2), 494 – 509.

Hinkle, R. K. and M. J. Nelson (2016). The transmission of legal precedent among state supreme courts in the twenty-first century. *State Politics & Policy Quarterly 16*(4), 391–410.

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

Holland, P. W. and S. Leinhardt (1971). Transitivity in structural models of small groups. *Small Group Research 2*(2), 107–124.

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

Hollway, J. and J. Koskinen (2016). Multilevel embeddedness: The case of the global fisheries governance complex. *Social Networks 44*, 281–294.

Hughes, J., M. Haran, and P. C. Caragea (2011). Autologistic models for binary data on a lattice. *Environmetrics 22*(7), 857–871.

Hummel, R. M. (2011). *Improving estimation for exponential-family random graph models.* Ph. D. thesis, The Pennsylvania State University.

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

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

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

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

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

Hyvärinen, A. (2006). Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation 18*, 2283–2292.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*(4598), 671–680.

Knight, J. and L. Epstein (1996). The norm of stare decisis. *American Journal of Political Science 40*(4), 1018–1035.

Konis, K. (2007). *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models (Ph.D. Thesis).* Ph. D. thesis, Worcester College, Oxford University.

Koontz, T. M. and C. W. Thomas (2018). Use of science in collaborative environmental management: Evidence from local watershed partnerships in the puget sound. *Environmental science & policy 88*, 17–23.

Krackhardt, D. (1987). Cognitive social structure. *Social Networks 9*, 109–134.

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

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

Krivitsky, P. N., D. R. Hunter, M. Morris, and C. Klumb (2021). ergm 4.0: New features and improvements.

Lazega, E. (2001). *The collegial phenomenon: The social mechanism and cooperation among peers in a corporate law partnership.* Oxford University Press.

Lehmann, E. and G. Casella (1998). *Theory of point estimation.* Springer Verlag.

Levy, M. (2016, 07). gwdegree: Improving interpretation of geometrically-weighted degree estimates in exponential random graph models. *The Journal of Open Source Software 1*(3), 36.

Liang, P. and M. I. Jordan (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, pp. 584–591. ACM.

Lindquist, S. A., W. L. Martinek, and V. A. Hettinger (2007). Splitting the difference: Modeling appellate court decisions with mixed outcomes. *Law & Society Review 41*(2), 429–456.

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

Lindsay, B. G., G. Y. Yi, and S. Jianping (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica 21*(1), 71–105.

Lissner, M. and B. W. Carver (2010). *CourtListener.com: A platform for researching and staying abreast of the latest law.*

Lomi, A. and F. Fonti (2012). Networks in markets and the propensity of companies to collaborate: An empirical test of three mechanisms. *Economics Letters 114*(2), 216–220.

Lu, Z., B. Savas, W. Tang, and I. S. Dhillon (2010). Supervised link prediction using multiple sources. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 923–928. IEEE.

Lubbers, M. J. and T. A. B. Snijders (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks 29*(2), 489–507.

Lupu, Y. and J. H. Fowler (2013). Strategic citations to precedent on the US Supreme Court. *The Journal of Legal Studies 42*(1), 151–186.

Lupu, Y. and E. Voeten (2012). Precedent in international courts: A network analysis of case citations by the european court of human rights. *British Journal of Political Science 42*(2), 413–439.

Maliniak, D., R. Powers, and B. F. Walter (2013). The gender citation gap in international relations. *International Organization 67*(4), 889–922.

Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis 10*(2), 134–153.

Mukherjee, S. et al. (2020). Degeneracy in sparse ergms with functions of degrees as sufficient statistics. *Bernoulli 26*(2), 1016–1043.

Okabayashi, S., L. Johnson, and C. Geyer (2011, 1). Extending pseudo-likelihood for Potts models. *Statistica Sinica 21*(1), 331–347.

Osei, A. (2018). Elite theory and political transitions: Networks of power in Ghana and Togo. *Comparative Politics 51*(1), 21–42.

Pang, X., B. Friedman, A. D. Martin, and K. M. Quinn (2012). Endogenous jurisprudential regimes. *Political Analysis 20*(4), 417–436.

Pattyn, V., A. Gouglas, and J. De Leeuwe (2020). The knowledge behind Brexit. A bibliographic analysis of ex-ante policy appraisals on Brexit in the United Kingdom and the European Union. *Journal of European Public Policy 17*, 1–19.

Pelc, K. J. (2014). The politics of precedent in international law: A social network application. *American Political Science Review 108*(3), 547–564.

Pomerance, B. (2018). Center of order: Chief justice john roberts and the coming struggle for a respected supreme court. *Alb. L. Rev. 82*, 449.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Resnick, M., P. Bearman, R. Blum, K. Bauman, K. Harris, J. Jones, J. Tabor, T. Beuhring, R. Sieving, M. Shew, M. Ireland, L. Bearinger, and J. Udry (1997, 9). Protecting adolescent's from harm: Findings from the national longitudinal study on adolescent health. *JAMA - Journal of the American Medical Association 278*(10), 823–832.

Richards, M. J. and H. M. Kritzer (2002). Jurisprudential regimes in supreme court decision making. *American Political Science Review 96*(2), 305–320.

Robins, G., P. Pattison, Y. Kalish, and D. Lusher (2007, May). An introduction to exponential random graph (p*) models for social networks. *Social Networks 29*(2), 173–191.

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

Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millán-Zárate, E. Díaz-Peredo, F. Sánchez-Solano, E. Pérez-Rueda, C. Bonavides-Martínez, and J. Collado-Vides (2001). RegulonDB (version 3.2): Transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic acids research 29*(1), 72–74.

Sampson, S. F. (1968). *A novitiate in a period of change: An experimental and case study of relationships*. Ph. D. thesis, Cornell University.

Saul, Z. and V. Filkov (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics (Oxford, England) 23*, 2604–11.

Schmid, C. S., T. H. Chen, and B. A. Desmarais (2020). *cERGM: A package to fit citation Exponential Random Graph Models*. R package version 1.0.0.

Schmid, C. S., T. H. Chen, and B. A. Desmarais (2021a). Generative dynamics of supreme court citations: Analysis with a new statistical network model.

Schmid, C. S., T. H. Chen, and B. A. Desmarais (2021b). Replication Data for Generative Dynamics of Supreme Court Citations: Analysis with a New Statistical Network Model by Schmid, Chen, and Desmarais.

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

Schmid, C. S. and D. R. Hunter (2021). Accounting for model misspecification when using pseudolikelihood for ERGMs. Technical report, The Pennsylvania State University, Department of Statistics.

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

Schweinberger, M. and J. R. Stuart (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *The Annals of Statistics 48*(1), 374–396.

Shalizi, C. R. and A. Rinaldo (2013, 04). Consistency under sampling of exponential random graph models. *Ann. Statist. 41*(2), 508–535.

Shen-Orr, S., R. Milo, S. Mangan, and U. Alon (2002, 06). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics 31*, 64–68.

Simpson, S. L., S. Hayasaka, and P. J. Laurienti (2011). Exponential random graph modeling for complex brain networks. *PloS one 6*(5), e20039.

Smith, S., F. Van Tubergen, I. Maas, and D. A. McFarland (2016). Ethnic composition and friendship segregation: Differential effects for adolescent natives and immigrants. *American Journal of Sociology 121*(4), 1223–1272.

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

Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology 37*(1), 131–153.

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

Spaeth, H., L. Epstein, T. Ruger, K. Whittington, J. Segal, and A. D. Martin (2014). Supreme court database code book.

Spriggs, J. F. and T. G. Hansford (2001). Explaining the overruling of US Supreme Court precedent. *The Journal of Politics 63*(4), 1091–1111.

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

Strogatz, S. H. (2001). Exploring complex networks. *Nature 410*(6825), 268.

Sunstein, C. R. (2008). Trimming. *Harv. L. Rev. 122*, 1049.

Thurner, P., C. Schmid, S. Cranmer, and G. Kauermann (2019). Network interdependencies and the evolution of the international arms trade. *Journal of Conflict Resolution 63-7*, 1736–1764.

Tieleman, T. and G. E. Hinton (2009). Using fast weights to improve persistent contrastive divergence. In *ICML*, pp. 1033–1040.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics 33*, 1–67.

US. Congress. Senate (2016). *Constitution of the United State of America: Analysis, and Interpretation – Centennial Edition*. 112th Cong., 2d sess. S. Doc. 112-9. `https://www.govinfo.gov/app/details/GPO-CONAN-2017/`.

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

Varin, C., N. Reid, and D. Firth (2011, 01). An overview of composite likelihood methods. *Statistica Sinica 21*, 5–42.

Wasserman, S. and P. E. Pattison (1996). Logit models and logistic regression for social networks: I. An introduction to markov graphs and p^*. *Electronic Journal of Statistics 61*(3), 401–425.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Wilkinson III, J. H. (2005). The Rehnquist court at twilight: The lures and perils of split-the-difference jurisprudence. *Stan. L. Rev. 58*, 1969.

Zhang, Y., A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha (2008). Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications 387*(7), 1705–1712.

Vita

Christian Song-Hyo Schmid

Education:

Doctor of Philosophy (Aug 2021) in Statistics with a graduate minor in Computational Science. The Pennsylvania State University, University Park, PA, USA. Thesis Title:

Master of Science (Feb 2015) in Statistics. Ludwig Maximilians University, Munich, Germany. Thesis Title: A Statistical Analysis of the International Arms Trade Data from 1950-2013

Bachelor of Science (Feb 2012) in Mathematics. Ludwig Maximilians University, Munich, Germany. Thesis Title: The Axiom of Choice and the Banach-Tarski Paradox.

Publications:

Christian S. Schmid, Ted H. Chen and Bruce A. Desmarais. Generative Dynamics of Supreme Court Citations: Analysis with a New Statistical Network Model. Accepted for publication at Political Analysis

Paul W. Thurner, Christian S. Schmid, Skyler J. Cranmer, and Göran Kauermann. Network Interdependencies and the Evolution of the Arms Trade Network. Journal of Conflict Resolution, Oct 2019.

Christian S. Schmid and Bruce A. Desmarais. Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. 2017 IEEE International Conference on Big Data (Big Data), pages 116–121, Dec 2017.

Christian S. Schmid and David R. Hunter. Improving ERGM Starting Values Using Simulated Annealing. (submitted)

Christian S. Schmid and David R. Hunter. Accounting for Model Misspecification When Using Pseudolikelihood for ERGMs. (in progress)

Academic Awards

Bruce Russett Award for the Best Paper Published in the Journal of Conflict Resolution in 2019. Paul W. Thurner, Christian S. Schmid, Skyler J. Cranmer, and Göran Kauermann. Network Interdependencies and the Evolution of the Arms Trade Network.

Teaching Award for Support of Pedagogy in Undergraduate Instruction 2019. Department of Statistics, Pennsylvania State University

Graduate Consulting Award 2015. Statistical Consulting Center, Ludwig Maximilians University