UNIVERSITY OF CALIFORNIA

Los Angeles

Novel Approaches to Degeneracy in Network Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Timothy Blackburn

2021

ABSTRACT OF THE DISSERTATION

Novel Approaches to Degeneracy in Network Models

by

Timothy Blackburn
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2021
Professor Mark Stephen Handcock, Chair

As technology advances, the manner in which humans communicate and collaborate becomes increasingly intricate and the study of complex networks becomes ever important. Exponential-family Random Graph Models (ERGMs) have long been at the forefront of the analysis of relational data due to their interpretability, flexibility, and strong theoretical foundation. However, ERGMs sometimes suffer from a serious condition known as degeneracy, in which the model exhibits unrealistic behavior or a severe lack-of-fit to the observed data if it can even be fit at all.

In an effort to overcome the issue of degeneracy, we take a variety of new approaches to network modeling and also build on the existing work of Fellows and Handcock (2017). We first consider several alternative models motivated by the maximization of various entropies or minimization of different divergence measures, as well as hierarchical models which expand on the ERG model. We derive their degenerate-inhibiting properties, but our analysis shows them ineffective for many purposes. Next, we provide an in-depth investigation of the Tapered ERGM and offer solutions to some related open questions, such as which terms to taper and how much. Several case studies are presented showing the effectiveness of the Tapered model and we derive some theoretical results. We also show, both theoretically and empirically, that the natural parameter estimates are largely numerically unaffected by the

amount of tapering. We then extend the ideas behind the Tapered ERGM and generalize it as member of a class of so-called Restorative Force Models (RFMs), which disallow degeneracy through various constraints. A treatment of this general class of models is conducted, and we examine the properties of several types of RFMs such as the Stereo ERGM, MAD ERGM, and LogCosh ERGM.

The dissertation of Timothy Blackburn is approved.

Jacob Gates Foster

Arash Ali Amini

Frederic R Paik Schoenberg

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2021

This work is dedicated to you, dear reader,

for taking the time to read this.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

"There are no bad ideas; only bad ideas in context." Without question, the author of this quote, my advisor Mark S. Handcock, deserves the most thanks in helping me get to where I am today. As the quote suggests, Professor Handcock allowed me the freedom to fail time and time again in my exploration of ideas. It was ultimately those failures, with my advisor's ever-present guiding hand, that eventually paved the way to the successes. Along my winding journey, Professor Handcock showed great patience as well as encouragement in my travails both professional and personal, and for that I am truly grateful.

I would also like to acknowledge the support of several of my cohorts, without which I may never have completed this PhD program: Seunghyun Min, for his constant help; Joshua Gordon for his technical prowess, and Aaron Danielson for his friendship. Credit is owed to Bryan Galvin for putting the idea of graduate school in my head. B3 was not the inspiration for beginning this endeavor, but he was the inspiration for finishing it, and for him I am eternally thankful.

VITA

2002–2006	Endowed Scholar, Dartmouth College.
2004–2006	Archivist for the Euler Archive, a project funded by the Swiss Consulate aimed at compiling everything created by the prolific Swiss mathematician Leonhard Euler, Dartmouth College.
2006	Gold Star award received for meritorious service as a volunteer in the Children's Hospital at Dartmouth (CHaD), a unit of the Dartmouth-Hitchcock New England Regional Medical Center.
2006	B.A. Mathematics *cum Laude*, Dartmouth College.
2016–2021	Teaching Assistant, Statistics Department, UCLA. Courses include Introduction to Probability and Introduction to Statistical Programming with R.
2017–2019	Co-Founder and Chief Scientist, Priceflow, Los Angeles, CA. Priceflow was an automotive tech start-up that was acquired by TrueCar in 2019.
2020	Presenter at Joint Statistical Meeting (JSM).
2020	Presenter at International Network for Social Network Analysis (INSNA Sunbelt).
2020–Present	Statistical Consultant, World Health Organization.

CHAPTER 1

Introduction

Given a set of N entities, how each entity relates to others is a question of interest. Such entities could be individuals in the workplace, countries within global markets, satellites in space, et cetera. We can refer to each entity as simply a *node*, and to each connection between nodes as an *edge*. This intuitive conceptualization of a network, the nodes together with edges, invokes its representation as a graph.

We formally define a graph G as a pairing of a node set V and an edge set E, so that $G = (V, E)$. Each node is given a unique label, and for simplicity we disallow multiple edges between nodes or any self-loops. Edges may be directed or undirected, and while methods exist to handle weighted values, for this work we only consider edges that take binary values indicating whether a relation between nodes exists or does not. Most often the number of nodes is fixed at N, and in the undirected case there are therefore $2^{\binom{N}{2}}$ possible graphs. Figure 1.1 shows an example graph.

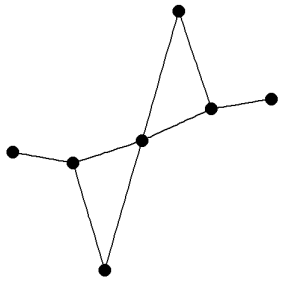

Figure 1.1: Simple random graph. This graph has seven nodes, eight edges, and two triangles.

Applying this simple yet abstract framework to the real world is where the difficulty arises.

Experience tells us that certain graphs (representative of their networks) are more likely to appear than others. We would not expect to encounter the empty graph, which contains no edges, or the complete graph, in which every node shares an edge with every other node, very often. We may have empirical data informing us of several network configurations, and would therefore expect there is a good chance of seeing such configurations again. Thus, we need a manner in which to place a probability distribution over all graphs. The most popular way of doing so is to use an Exponential-family Random Graph Model (ERGM).

An ERGM is a probability mass function

$$p_\theta(Y = y) = \frac{\exp(\langle \theta, t(y) \rangle)}{Z(\theta)} \tag{1.1}$$

where Y is a random graph whose realization is $y \in G_N$, the set of all possible graphs on N nodes; $t(y)$ is the vector of d sufficient statistics in $\mathbb{R}^d$; $\theta \in \mathbb{R}^d$ is the vector of parameters; and $Z(\theta)$ the normalizing constant. The sufficient statistics used are chosen by the researcher, depending on which characteristics of the network are of interest. In this way, ERGMs actually constitute a Family across different choices of $t()$. Regardless of which sufficient statistics are used, the ERGM will have the maximal entropy of any distribution satisfying the mean constraints put on said statistics. In other words, given some network data and a set of relevant constraints put on features of the network, the ERGM is the model with the least bias built in. This property is a leading justification for its use.

Yet, what is the ERGM's greatest strength may also be its biggest weakness. In an effort to maximize entropy, the ERGM can be thought of as "spreading out" mass across the graph space as much as possible while still maintaining the mean constraints. This sometimes leads to a large amount of mass being placed on extremal configurations (such as the empty and complete graphs) and very little mass being placed on the observed graph. This problem is referred to as *degeneracy*, and when a model is degenerate predictions from it will be invalid.

Figure 1.2 shows an example of degeneracy. This ERGM uses the edge count and triangle count as sufficient statistics, both of which are extremely common and useful choices amongst researchers. Here we have used the exact enumeration of all labeled graphs on $N = 7$ nodes

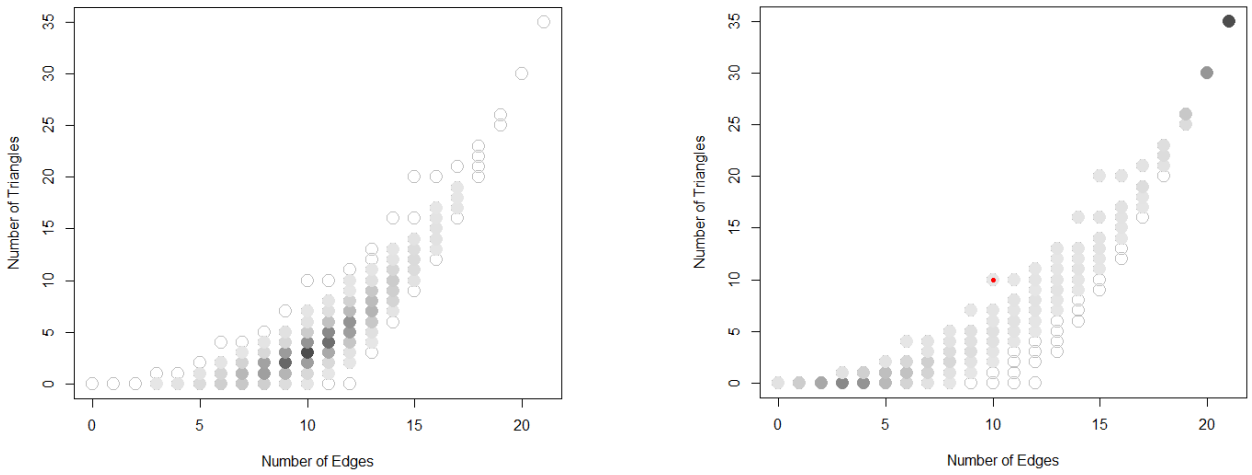

Figure 1.2: Degenerate ERGM. Each class of graphs, identified by the number of edges and triangles, is represented by a circle. *LEFT*: The number of graphs within each class, where the intensity of the shading is proportional to number of graphs. The darker the shading, the larger the number of graphs. *RIGHT*: The ERGM for mean edge and triangle constraints of 10 and 10, where the red dot denotes the class with these mean counts. The darker the shading, the more mass the ERGM places on that class. Note the mass placed on the extremes of the space.

as the data set. Using the edge count and triangle count to classify each graph, we end up with 110 distinct classes. The left panel depicts the number of graphs possible for each class within the graph space, with darker colors indicating relatively higher numbers. We see that most configurations lie within the center of the graph space. The right panel shows the ERGM with maximum likelihood parameter values corresponding to mean constraints 10 and 10 for the edge and triangle counts, respectively. Even though these constraints are realizable by a specific class, as indicated by the red dot, very little mass is placed on this observed class or the surrounding classes. Instead, the degeneracy of the model puts a large amount of mass toward the extremal configurations, especially the complete graph in the upper right hand corner. As a result, simulations from this ERGM yield graphs that are very dense (near or at the complete graph) or very sparse (near or at the empty graph), but very few similar to the observed class of graphs containing 10 edges and 10 triangles, despite the fact that those averages are met over the entire distribution.

1.1 Overview of the Dissertation

In an attempt to overcome the issue of degeneracy, a variety of new approaches toward network modeling have been taken. Chapter 2 details several models with solid theoretical foundations but which turned out to be fruitless. These include q-exponentials, reverse entropy models, and symmetric divergence models. Chapter 3 explores random effects models with both logistic and conjugate priors, treating the ERGM in a hierarchical fashion. An analysis into the limitations of this framework is presented as well as a key insight into why degeneracy occurs.

Chapter 4 dives into the Tapered ERGM of Fellows and Handcock (2017). We tackle the open question of how to choose the tapering parameters by showing that tapering all terms is not always necessary and that natural parameter estimates are largely unaffected by the amount of tapering. We prove several key theorems, including the non-degeneracy of the model and computation of the standard errors, among others. We provide an algorithm

for setting the level of tapering in different scenarios. Along the way, we also derive a bias-corrected measure of the sample kurtosis for use in measuring the bimodality of a distribution and assessing degeneracy.

Chapter 5 extends the ideas behind the Tapered ERGM beyond Gaussian tapering. We generalize the Tapered ERGM into a broader class of Restorative Force Models (RFMs) in which degeneracy is disallowed through varying constraints. Each constraint gives rise to a new model, and we examine the resulting maximum entropy distributions. We introduce and prove several key theorems for each of the Stereo ERGM, MAD ERGM, and LogCosh ERGM. A thorough comparison of the properties of these models is given.

Chapter 6 presents three case studies in which the Restorative Force Models and the methods outlined in the previous chapter are applied to published networks. We show that the Tapered ERGM and other RFMs can handle large networks on the order of thousands of nodes, compare the results from using different RFMs, and show that desirable models heretofore impossible to fit with the standard ERGM are achievable with RFMs and can lead to drastically different inferences.

Finally, an appendix is included at the end in which several derivations are worked out.

CHAPTER 2

Using Alternative Entropy and Divergence Measures

The Maximum Entropy Principle maintains that the best probability distribution for a given set of constraints and/or data is the one with the largest entropy. It is most common to consider the Shannon Entropy,

$$H(P) \equiv \mathbb{E}_P[-\log p(x)] = -\sum_{x \in \chi} p(x) \log p(x)$$

in which case it is well known that the distribution of maximal entropy will be a member of the Exponential Family (which includes ERGMs). However, there are other ways to measure entropy which lead to different maximum entropy models. Furthermore, there is a direct relation between entropy and divergence measures, and considering alternative divergence measures can also lead to new classes of models. These ideas are explored in this chapter.

2.1 q-Exponentials

In Tsallis (1988), the physicist Constantino Tsallis introduced a deformation of the Shannon Entropy which would come to be known as *q-Entropy*

$$H_q(p) \equiv \frac{\sum_{y \in G_N} p(y)^q - 1}{1 - q} \tag{2.1}$$

where $q \in \mathbb{R}$. Note that $H_q(p) \to H(p)$ as $q \to 1$. Tsallis later went on to publish his *q-Exponential* function (Tsallis, 1994), a deformation of the exponential function, defined as

$$\exp_q(x) \equiv [1 + (1 - q)x]^{\frac{1}{1-q}} \tag{2.2}$$

where $\exp_q(x) \to \exp(x)$ as $q \to 1$. Although not explicitly stated until much later (Umarov et al., 2008), Tsallis would modify this definition slightly to be

$$\exp_q(x) \equiv [1 + (1-q)x]_+^{\frac{1}{1-q}} \tag{2.3}$$

where $[s]_+ = \max\{s, 0\}$. It turns out, however, that the q-Exponential Family is connected to the α-Family of Information Geometry (Amari, 2012). From this perspective x can be any real number, but $\exp_q(x)$ is only defined for $1 + (1-q)x > 0$ (Amari and Ohara, 2011). Outside of Information Geometry, the q-Exponential Family is most often defined using (2.3).

Whether we use (2.2) or (2.3), any probability mass function using q-Exponentials has the potential to nullify certain classes of graphs, effectively removing them from the support. This potential restriction in the graph space was further motivation for investigating q-Exponentials as a solution to the degeneracy problem, and the idea of restricting the support has been explored elsewhere recently (Karwa et al., 2016).

However, maximizing $H_q(p)$ is not quite so straight-forward. Using the typical mean constraints on the sufficient statistics, we cannot arrive at a model containing the q-Exponential. Instead we arrive at the distribution

$$p_{*q,\theta}(Y = y) = [(1 - \frac{1}{q})(\langle \theta, t(y) \rangle - Z_q(\theta))]_+^{\frac{1}{q-1}} \tag{2.4}$$

See the appendix for derivation. If we use a set of modified constraints (see the note in the appendix), others (Amari and Ohara, 2011; Naudts, 2009) have arrived at a model involving the q-Exponential given by

$$p_{q,\theta}(Y = y) = \exp_q(\langle \theta, t(y) \rangle - Z_q(\theta)) \tag{2.5}$$

We will refer to (2.4) as the p_{*q} model and to (2.5) as the p_q model. These two models are very similar, but some subtle differences should be noted.

In the p_{*q} model, $p_{*q,\theta}(y) \equiv 0$ whenever $(1 - \frac{1}{q})(\langle \theta, t(y) \rangle - Z_q(\theta)) \leq 0$. This implies the parameter space is simply $\mathbb{R}^d$ for a model with d sufficient statistics; i.e. $\theta \in \mathbb{R}^d$. For the p_q

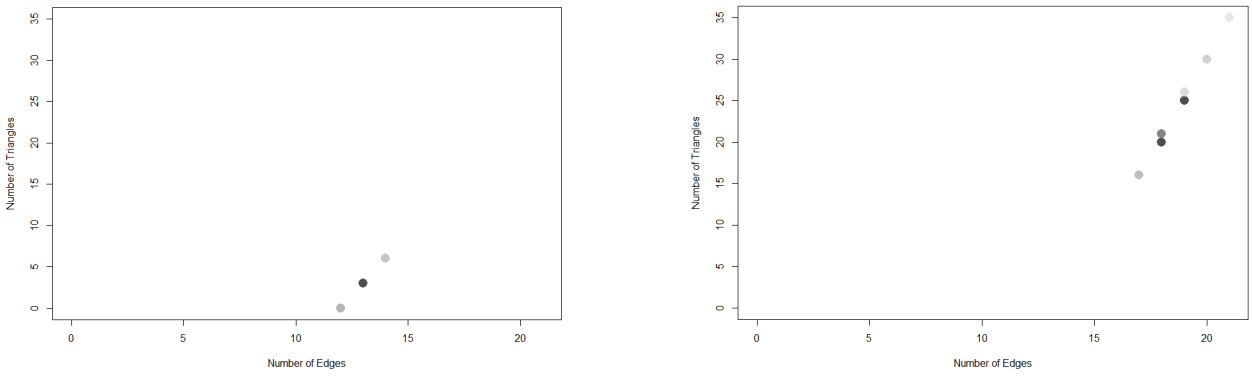

Figure 2.1: Degenerate p_{*q} model, $q = 1.5$, on the set of all graphs with seven nodes. The graph space has been severely restricted, with almost all classes receiving zero mass. The extreme sensitivity of the model is also demonstrated here. In the left panel, the distribution is fit with parameter $\theta = (0.5, -0.2)$. In the right panel, the distribution is fit with parameter $\theta = (0.5, -0.1)$. This slight perturbation results in dramatically different outcomes.

model, if we accept definition (2.2) then $p_{q,\theta}(y)$ is undefined whenever $1 + (1-q)(\langle \theta, t(y) \rangle - Z_q(\theta)) \leq 0$. This makes the parameter space for the p_q model much more complicated, as noted in Amari and Ohara (2011). In general, $\theta \in \Omega(q, t, G_N)$, where the parameter space Ω is dependent on the choice of q, the sufficient statistics vector t, and graph space G_N.

In the following section, we will explore the shortcomings of both of these models.

2.1.1 q-Exponential Network Models

Both models present challenges. With ERGMs, MCMC routines can be used for parameter estimation and explicit evaluation of the normalizing constant $Z(\theta)$ can be avoided. However, in our alternative models computation of the normalizing constant $Z_q(\theta)$ cannot be sidestepped due to its location within the expressions. Because we used the exact enumeration of all labeled graphs on $N = 7$ nodes, we were able to calculate $Z_q(\theta)$ in this case, but computing the normalizing constant would be a serious problem in larger networks.

The choice of q is also not obvious, nor are the acceptable values of q necessarily agreed upon. Tsallis placed no restriction on q other than $q \neq 1$ (Tsallis, 1994), while others have specified $q > 0$ (Amari and Ohara, 2011; Naudts, 2008). In practice, the choice of q will be reliant upon what mean constraints are imposed, as certain values of q do not afford certain distributions. As we will see below, for most values of q, the resulting distribution becomes severely degenerate.

Let us begin with the p_{*q} model. While the model does succeed in reducing the graph space, it does so in entirely the wrong way. Instead of removing extremal configurations, it tends to remove classes from the center of the graph space and keep *only* the extremal configurations. What's worse, the model is also extremely sensitive to changes in parameter values. Figure 2.1 shows the p_{*q} model with $q = 1.5$, for two different parameter values. As before, we use the edge and triangle counts as our sufficient statistics. The left panel shows the fit with $\theta = (0.5, -0.2)$, while the right panel shows the fit with $\theta = (0.5, -0.1)$. Two things immediately jump out at us. Firstly, most of the classes have vanished having been zeroed out by the model. Secondly, the small subset of the graph space that does receive positive mass changes drastically. This erratic behavior holds for most all values of q, making the p_{*q} model completely ineffective.

Turning our focus now to the p_q model, the situation is only slightly better. In general, this model is not as sensitive to small deviations in the parameters, but it also tends to exclude the wrong subset of graphs. The p_q model will most often place zero mass on the classes near the center of the graph space and more mass on the extremal configurations, exactly the opposite of what is desired. Furthermore, even if the q-Exponential model does manage to retain most of the classes, it also exhibits a "spreading" of the mass toward the extremes just as its exponential counterpart does, as shown in Figure 2.2. The q-Exponential model offers no respite from the degeneracy issue.

We examined two models: p_{*q} which stems directly from maximizing the q-Entropy, and p_q which belongs to the q-Exponential Family. These models are alluring in the sense that

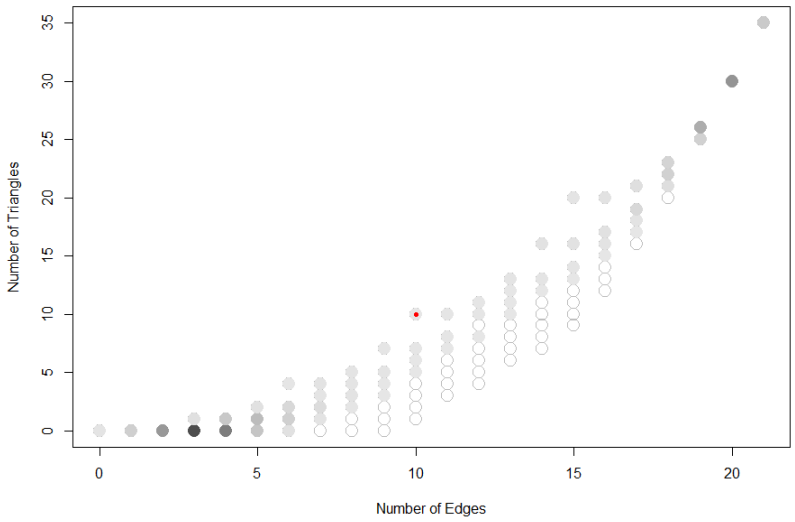

Figure 2.2: Degenerate q-Exponential model, $q = .9$. Here only 99 of the original 110 classes remain, as 11 classes vanish with zero mass (not visible in the graph above, though they are near the bottom of the convex hull). Unfortunately, the model is still degenerate as it places very little mass near the class with the mean edge and triangle counts (indicated by the red dot) and a great deal of mass toward the extremes of the space.

they provide a way of restricting the support of the model, but unfortunately this leads to more degenerate behavior and not less. Indeed, the degeneracy problems of both the p_{*q} and p_q models discussed above are even more severe than those encountered in ERGMs. Models which use q-Exponentials may be advantageous in Physics and other sciences, but they do not remedy the degeneracy issues of network science.

2.2 Maximizing Entropy, Minimizing Divergence

In this section, we look at the connection between entropy and divergence measures and again try to exploit these as possible solutions to degeneracy.

Let P and Q be two probability distributions over a space Ω. Let $f(t)$ be a real-valued convex function, defined for $t > 0$, such that $f(1) = 0$. The f-divergence of P from Q is defined as

$$D_f(P \parallel Q) \equiv \sum_{\Omega} f\left(\frac{p(x)}{q(x)}\right)\, q(x)$$

The Kullback-Leibler divergence is an f-divergence induced by the function

$$f(t) = t \log t$$

If we work through it, we arrive at the familiar expression for the KL-divergence of P from Q:

$$\begin{aligned} D_f(P \parallel Q) &= \sum_{\Omega} f\left(\frac{p(x)}{q(x)}\right)\, q(x) \\ &= \sum_{\Omega} \frac{p(x)}{q(x)} \log\left(\frac{p(x)}{q(x)}\right) q(x) \\ &= \sum_{\Omega} p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ &= D_{KL}(P \parallel Q) \end{aligned}$$

If we let $q(x) = 1$, we can measure the divergence of P from the uniform distribution as

$$\begin{aligned} &= \sum_{\Omega} p(x) \log\left(p(x)\right) \\ &= -H(P) \end{aligned}$$

where $H(P) = -\sum_{\Omega} p(x) \log\left(p(x)\right)$ is the entropy of P. Thus, maximizing entropy is equivalent to minimizing the KL-divergence from the uniform.

A key feature of the KL-divergence is its asymmetry: $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. As we've seen, entropy measures the divergence of a distribution from uniform. But what if we measured the divergence of the uniform from a distribution? This is referred to as *reverse KL-divergence*, and it is induced by:

$$f^*(t) = -\log t$$

In other words, $D_{f*}(P \parallel Q) = D_{KL}(Q \parallel P)$. Interestingly enough, forward-KL and reverse-KL are both types of α-divergences, with forward-KL corresponding to $\alpha = 1$ and reverse-KL corresponding to $\alpha = -1$.

2.2.1 Minimizing the f^*-divergence

Let our divergence measure be defined via the convex function

$$f^*(t) = -\log t$$

The corresponding f-divergence is then

$$D_{f^*}(P \parallel Q) = \sum_{\Omega} -\log\left(\frac{p(x)}{q(x)}\right) q(x)$$

If we let $q(x) = 1$, we can measure the divergence of P from the uniform distribution as

$$= \sum_{\Omega} -\log\left(p(x)\right)$$

Now consider Ω to be the set of all graphs x on a given number of nodes. Additionally, let $t(x) = (t_1(x), ..., t_d(x))$ be the vector of d sufficient statistics for the graph x. For example, for $d = 2$ we could have $t_1(x) = \{\text{number of edges of } x\}$ and $t_2(x) = \{\text{number of triangles of } x\}$.

We now seek to find P that minimizes the divergence (as measured above) from the uniform distribution subject to the following constraints:

$$\sum_{\Omega} p(x) = 1$$

$$p(x) \geq 0 \, \forall x \in \Omega$$

$$\sum_{x \in \chi} p(x) t_i(x) = \mu_i \; \forall i \in \{1, ..., d\}$$

Where the last equation represents $i = 1, ..., d$ arbitrary mean constraints. Using the method of Lagrange Multipliers, we can minimize f^* and obtain the following distribution (see the appendix for derivation):

$$p_\theta(Y = y) = \frac{1}{C(\theta)\left[1 + \langle \theta, t(y) \rangle\right]} \tag{2.6}$$

Here $\theta = \{\theta_i\}$ are the mean parameters and $C(\theta)$ is the normalizing constant. While this model is elegant in its simplicity, it too has problems. The most noticeable is that in order for the second constraint (all probabilities are non-negative) to be achieved, the values of θ_i must be such that $\langle \theta, \phi(y) \rangle > -1$. This has a rather limiting effect on the diversity of distributions attainable in that only relatively few means are achievable. Figure 2.3 illustrates these limitations. No matter what values of θ_i are chosen (such that $p_\theta(y)$ is non-negative), only distributions with means at the center of the convex hull are possible. This means that any mean-constraints outside of the center of the convex hull (the area in black) are not realized under the reverse entropy model. Unfortunately, most observable networks lie at or near the relative boundary away from the center, precisely where the reverse entropy model cannot reach. Moreover, it is only in the regions outside the center where ERGMs fail and degeneracy becomes an issue. In the following section, we discuss why the severe attraction to the center of parameter space is inherent in the nature of the reverse entropy model.

2.3 KL-divergence, Reverse KL-divergence, and Degeneracy

Denote the (forward) KL-divergence, the divergence of P from Q, as $D_{KL}(P \parallel Q) \equiv \sum_\Omega p(x) \log\left(\frac{p(x)}{q(x)}\right)$ and reverse KL-divergence, the divergence of Q from P, as $D_{KL}(Q \parallel P) \equiv \sum_\Omega q(x) \log\left(\frac{q(x)}{p(x)}\right)$.

It's easy to see that in both cases, the the KL-divergence is a weighted sum of the difference in log-likelihoods. The difference between them is that in forward KL, the weight is $p(x)$, the distribution we are taking as our solution to the Lagrangian, and in reverse KL, the weight is $q(x)$ - the mass given by the uniform distribution. While the uniform obviously gives the same mass to each graph, it does not give the same mass to each *class* of graphs, and these classes are often what we are concerned about.

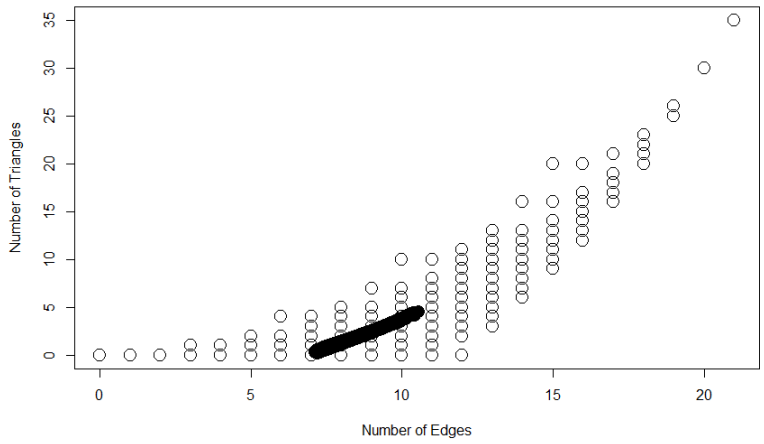

Figure 2.3: Limitations of the reverse entropy model. For any value of $\theta = (\theta_1, \theta_2)$ which keeps $p_\theta(y)$ non-negative for all graphs y, the mean number of edges and triangles is plotted in black. We see that only distributions with means in the center (the black region) of the parameter space are realizable. Any empty circle represents a class of graphs for which no distribution is possible having those means, even though graphs with those means do exist. Note that ERGMs can produce distributions with means at any location within the convex hull of the parameter space, but many of these distributions will be degenerate. Reverse entropy models will not be degenerate, but they can only produce distributions with means in the black region. This drawback effectively renders reverse entropy models useless.

In either case, the KL-divergence is an expectation of the difference in log-likelihoods. The distribution over which this expectation is taken is what accounts for the asymmetry of $D_{KL}(P \parallel Q)$ and $D_{KL}(Q \parallel P)$. The expression $w \log(w)$ is the product of two terms with opposite behaviors as $w \to 0$. However, $\lim_{w \to 0} w \log(w) = 0$, which shows that w, the expectation weights, are what dominate.

Consider first the forward KL-divergence. Here the weighting is given by $p(x)$. Since we are minimizing the divergence, the *best* case scenario occurs when $p(x)$ gives little mass to a class of graphs where $q(x)$ puts large mass. In this scenario, although the difference in log-likelihoods will tend toward $-\infty$, the divergence will be very small because of the

dominating $p(x)$ which is small. Since we are minimizing the divergence, this situation is sometimes favored. Thus, this method sometimes *encourages degeneracy*.

Now consider the reverse KL-divergence. Here the weighting is given by the uniform, $q(x)$. The *worst* case scenario here is when $q(x)$ gives a large amount of mass to a class of graphs that $p(x)$ gives little mass to. In this case, the difference in log-likelihoods will tend toward $+\infty$, and since the dominating $q(x)$ is large the divergence would be very large. Since we are minimizing the reverse KL, this situation is discouraged. This method thus *discourages degeneracy*.

This degeneracy-inhibiting quality of the reverse KL is a desirable trait, but it is often too strong. Figure 2.4 below is from Murphy (2012). It illustrates how the forward KL wants to "spread out" over the entire distribution, whereas the reverse KL wants to "lock onto" one of the modes.

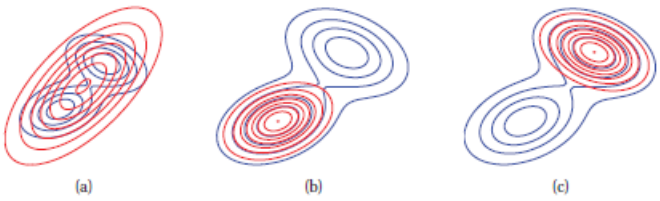

Figure 21.1 Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution p. The red curves are the contours of the unimodal approximation q. (a) Minimizing forwards KL: q tends to "cover" p. (b-c) Minimizing reverse KL: q locks on to one of the two modes. Based on Figure 10.3 of (Bishop 2006b). Figure generated by `KLfwdReverseMixGauss`.

Figure 2.4: Photo Credit: Murphy (2012)

If we minimize the f^*-divergence (the reverse KL-divergence) with only mean constraints, this tendency to lock onto the mode is too strong and yields a distribution too close to the uniform distribution in many scenarios. Similarly, in many scenarios minimizing the forward KL (maximizing Entropy) has too strong a tendency to do the opposite and spread mass too far, leading to degeneracy. The ERGM is the result of minimizing the forward KL-divergence from uniform, and the model in (2.6) is the result of minimizing the reverse KL-divergence from uniform. Under model (2.6) we no longer need to worry about degeneracy, but it comes

at a drastic cost. ERGMs and reverse entropy models are seemingly opposite sides of the same record, neither of which play a pleasant tune.

2.3.1 Minimizing Symmetrized Divergence

The previous section showed that both the KL- and reverse KL-divergence lead to unwanted extremes. A natural idea, then, is to consider what happens when both divergences are used together. The symmetrized divergence, also sometimes referred to as *Jeffrey's Divergence*, is

$$D(P,Q) \equiv D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)$$

which is clearly symmetric and non-negative.

Let Q be the uniform distribution. With $q(x) = 1$, we have

$$D(P,Q) = \sum_{\Omega} p(x) \log\left(p(x)\right) - \sum_{\Omega} \log\left(p(x)\right)$$

We seek to find P that minimizes $D(P,Q)$ subject to the following basic constraints, once more:

$$\sum_{\Omega} p(x) = 1$$

$$p(x) \geq 0 \, \forall x \in \Omega$$

$$\sum_{\Omega} p(x) t_i(x) = \mu_i \; \forall i \in \{1, ..., d\}$$

The method of Lagrange Multipliers is again used, with details provided in the appendix, and we obtain

$$p(Y = y) = \frac{1}{W\left(C(\theta) \exp\left(\langle \theta, t(y) \rangle\right)\right)}$$

where θ is the vector of parameters, $C(\theta)$ is the normalizing constant, and $W()$ is the *Lambert-W function*, also known as the *product logarithm*.

Unfortunately, this class of models is unwieldy for two reasons. Firstly, the normalizing constant $C(\theta)$ is hopelessly trapped within the W-function, making typical estimation routines impotent. Typically, evaluation of the normalizing constant can be avoided in finding the maximum likelihood estimate of θ, or in the case of ERGMs a clever method exists to estimate it. With the symmetrized divergence model, side-stepping the computation of the normalizing constant is not possible, and there no feasible technique to do so. Secondly, the W-function has such a small gradient that it is relatively insensitive to changes in it's argument. For example, $W(10) = 1.75$ while $W(1000) = 5.25$. This means that changes in network statistics commonly used, such as the number of edges or triangles, would not have much effect on the likelihood.

2.4 Discussion

In this chapter we have shown that entropy measures and divergence measures are really two sides of the same coin. Given that ERGMs, like so many others, are maximum entropy distributions, it makes sense to explore the other side of the coin. The Shannon Entropy and Kullback-Leibler divergence are the bedrock of most statistical formulations, but entropy and divergence measures come in many flavors. Limiting ourselves to just one type of measure would be myopic.

The q-Exponentials introduced by Tsallis and explored further by Amari provide interesting analogs to the typical models belonging to the Exponential Family. The use of q as a tuning parameter may have many advantages for modeling in physical sciences when the observed phenomena deviate from predictions of traditional models. In many ways, this is exactly what happens when degeneracy strikes in ERG models, yet the models employing q-Exponentials unfortunately do not relieve the problems of degeneracy. It is concluded that q-Exponentials should no longer be investigated in connection to degeneracy in network models.

Alternative divergence measures, on the other hand, perhaps do warrant further inves-

tigation. The reverse-entropy model developed in section 2.2 does not solve the problem of degeneracy because of its inability to reach the extremal configurations often observed in real-life networks. However, the marked stability and elegant form of the model may lend itself useful in other situations. Similarly, the symmetric-divergence model may prove valuable in the future if other computational techniques are developed to aid in estimation of the normalizing constant.

Finally, the exploration into alternative entropy and divergence measures has paid dividends in that it has led to a deeper understanding of *why* degeneracy occurs. We now know that ERGMs and, in general, all members of the Exponential Family encourage degeneracy in that any maximum entropy distribution (taken with respect to Shannon Entropy) will seek to spread probability mass as far as possible over the support. The reason we encounter degeneracy in ERGMs and not other members, such as the Normal distribution, is because the support of ERGMs is discrete and finite. Now that we understand that these degenerate tendencies are inherent in the ERGM, we know that we must go beyond ERGMs to remedy it. This is precisely what we do in the following chapters.

CHAPTER 3

Random Effects Models

In this section we take altogether different approaches by using hierarchical linear models. The following sections both specify the lowest level as an expanded or "full" ERGM. The Full-ERGM differs from the ERGM only in that it does not make the *homogeneity* assumption of Frank and Strauss (1986) that all vertices are indistinguishable. In the ERG model, the probability of an edge between two nodes is the same for all dyads, implying that isomorphic graphs (i.e., graphs with the same value of sufficient statistics) must all have the same probability. None of this is the case in the Full-ERG model. Under the Full-ERGM the probability of an edge between nodes is specific to which nodes are involved, and thus the model has a different parameter for each dyad. The model can be extended to include separate parameters for each 2-star, triangle, etc.

For example, if we fix the number of nodes at $n = 7$, there are 21 dyads and 35 triads. Allowing distinct parameters for each edge and triangle, the Full-ERGM likelihood is

$$p(Y = y|\delta, \tau) = \frac{\exp(\delta_1 d_{y_1} + ... + \delta_{21} d_{y_{21}} + \tau_1 t_{y_1} + ... + \tau_{35} t_{y_{35}})}{Z(\delta, \tau)} \tag{3.1}$$

where

- $d_{y_i} \in \{0, 1\}$, $i = 1, ..., 21$ are the edge indicators of graph $Y = y$
- δ_i , $i = 1, ..., 21$ are the corresponding parameters for each dyad
- $t_{y_i} \in \{0, 1\}$, $i = 1, ..., 35$ are the triangle indicators of graph $Y = y$
- τ_i , $i = 1, ..., 35$ are the corresponding parameters for each triad

- $Z(\delta, \tau) = \sum_y \exp(\delta_1 d_{y_1} + ... + \delta_{21} d_{y_{21}} + \tau_1 t_{y_1} + ... + \tau_{35} t_{y_{35}})$ is the normalizing constant

As in the standard ERGM, the parameters δ_i and τ_j can be interpreted as the log-odds for the edge d_i and triangle t_j, respectively. Since sparsity is a key feature of many networks, we would expect most δ_i and τ_i to take negative values. Thus we an induce sparsity by taking each δ_i, τ_i as a random draw from their own distributions.

3.1 Using Logistic Priors

Given that the parameters δ_i and τ_j are the log-odds for the edge d_i and triangle t_j, respectively, a natural choice for their priors is the Logistic Distribution indexed by location parameter μ and scale parameter s. The Logistic(μ, s) has pdf

$$f(x; \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s\left(1 + e^{-\frac{x-\mu}{s}}\right)^2}$$

Diaconis and Ylvisaker (1979) have shown that conjugate priors exist for any member of the exponential family, though the conjugate priors they specify are priors on the *mean parameters*. The Logistic(μ, s) priors we will use here are priors on the natural parameters themselves. Although the effect of either prior is mathematically the same, a prior on the mean parameters represents information about distributional means, whereas priors directly on the natural parameters encapsulate part of the data generating process. Nevertheless, we will now show that the Logistic(μ, s) prior is in fact conjugate to our Full-ERG model.

For simplicity, consider an ERGM with only a single edge and hence a single parameter, δ. The following generalizes to any number of edges or any other statistic. The single-edge model likelihood is

$$p(Y = y|\delta) \propto \exp(\delta d_y);$$

and the prior on δ is

$$f(\delta|\mu, s) \propto \exp\left(-\frac{\delta - \mu}{s}\right);$$

therefore the posterior is

$$\begin{aligned} p(\delta|\mu, s, Y) &\propto p(Y|\delta) \times f(\delta|\mu, s) \\ &\propto \exp\left(\delta d_y - \frac{\delta - \mu}{s}\right) \\ &\propto \exp\left(-\frac{1 - sd_y}{s}\left(\delta - \frac{\mu}{1 - sd_y}\right)\right) \\ &\sim \text{Logistic}\left(\frac{\mu}{1 - sd_y}, \frac{s}{1 - sd_y}\right) \end{aligned}$$

Thus, the Logistic(μ, s) is a conjugate prior of the Full-ERGM.

Since the Full-ERGM is a subsumed by the Exponential Family, there is a one-to-one correspondence between the natural parameters and mean parameters. Hence, our Logistic(μ, s) prior on the natural parameters implies a prior on the mean parameters. But what is this prior? We can work it out by again considering a single-edge model for simplicity. In this case, the mean parameter π is simply

$$\pi \equiv \mathbb{E}_\delta[d_y] = p(d_y = 1|\delta) = \frac{e^\delta}{e^\delta + 1}$$

Now considering the cdf of π and using $f(\delta; \mu, s)$ for the Logistic(μ, s) pdf of δ, we have

$$\begin{aligned} \Pr(\pi < x) &= \Pr\left(\frac{e^\delta}{e^\delta + 1} < x\right) \\ &= \Pr\left(\delta < \log\left(\frac{x}{1 - x}\right)\right) \\ &= \int_{-\infty}^{\log\left(\frac{x}{1-x}\right)} f(\delta; \mu, s) d\delta \\ &= \left.\frac{1}{1 + e^{-\frac{\delta - \mu}{s}}}\right|_{-\infty}^{\log\left(\frac{x}{1-x}\right)} \\ &= \frac{1}{1 + e^{\frac{\mu}{s}}\left(\frac{x}{1-x}\right)^{\frac{-1}{s}}} \end{aligned}$$

Differentiating this expression reveals the corresponding prior on π, $g(\pi; \mu, s)$ as

$$g(\pi; \mu, s) = \frac{e^{\frac{\mu}{s}}\left(\frac{\pi}{1-\pi}\right)^{1-\frac{1}{s}}}{s\pi^2\left(1 + e^{\frac{\mu}{s}}\left(\frac{\pi}{1-\pi}\right)^{\frac{-1}{s}}\right)^2}$$

While $g(\pi; \mu, s)$ is not exactly familiar, if we let $\mu = 0$ and $s = 1$, we have $g(\pi; \mu = 0, s = 1) = 1$. Thus, we at least have the reassuring fact that the Logistic$(0, 1)$ prior on the natural parameter δ corresponds to the Uniform prior on the mean parameter π.

It is worth exploring the other direction; that is, finding the resulting prior on the natural parameter δ if we place a prior on the mean parameter π. A common choice of prior on π is the Beta Distribution. Given that $\pi \sim \text{Beta}(\alpha, \beta)$, we again consider the single-edge case for simplicity and proceed as before:

$$\begin{aligned}
\Pr(\delta < x) &= \Pr\left(\log\left(\frac{\pi}{1-\pi}\right) < x\right) \\
&= \Pr\left(\pi < \frac{e^x}{1+e^x}\right) \\
&= \frac{1}{\text{B}(\alpha, \beta)} \int_0^{\frac{e^x}{1+e^x}} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi
\end{aligned}$$

Differentiating this expression to retrieve the pdf yields the corresponding prior on δ, $h(\delta; \alpha, \beta)$ as

$$h(\delta; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} \frac{(e^\delta)^{\alpha-1}}{(1+e^\delta)^{\alpha+\beta-2}} \frac{e^{-\delta}}{(e^{-\delta}+1)^2}$$

Note the logistic expression on the right. Indeed, letting $\alpha = 1$ and $\beta = 1$ gives

$$h(\delta; \alpha = 1, \beta = 1) = \frac{e^{-\delta}}{(e^{-\delta}+1)^2} = \text{Logistic}(0, 1)$$

Thus confirming that the Uniform or Beta$(1, 1)$ prior on the mean parameter π corresponds the Logistic$(0, 1)$ prior on the natural parameter δ. The fact the logistic expression is intrinsically a part of $h(\cdot)$ further justifies the use of the Logistic prior on δ and τ.

3.1.1 Hierarchical Model with Logistic Priors

One specification of the model is to take each δ_i as iid sampled from a Logistic distribution, as with each τ_j from a separate Logistic distribution. That is,

$$\delta_i \overset{\text{iid}}{\sim} \text{Logistic}(\mu_\text{d}, \text{s}_\text{d})$$

$$\tau_j \overset{\text{iid}}{\sim} \text{Logistic}(\mu_\text{t}, \text{s}_\text{t})$$

This leads us to the following priors:

$$\pi_\delta(\delta|\mu_d, s_d) \propto \prod_{i=1}^{n} e^{-\frac{\delta_i - \mu_d}{s_d}} \propto \exp\left(-\frac{n}{s_d}(\bar{\delta} - \mu_d)\right)$$

$$\pi_\tau(\tau|\mu_t, s_t) \propto \prod_{j=1}^{m} e^{-\frac{\tau_j - \mu_t}{s_t}} \propto \exp\left(-\frac{m}{s_t}(\bar{\tau} - \mu_t)\right)$$

The complete Hierarchical Full-ERG model is thus

$$\delta \overset{\text{iid}}{\sim} \pi_\delta(\delta|\mu_d, s_d)$$

$$\tau \overset{\text{iid}}{\sim} \pi_\tau(\tau|\mu_t, s_t)$$

$$Y \sim p(Y|\delta, \tau)$$

This main idea is to average over δ and τ, the random effects, to obtain a distribution for Y that is dependent only on the hyperparameters μ and s. The mechanics of this are fleshed out using conjugate priors in the following section.

3.2 Using Conjugate Priors

Following Diaconis and Ylvisaker (1979), we can place a conjugate prior on the likelihood of the form

$$\pi(\delta, \tau|a, b, c) = \frac{\exp\left(c\left[\sum_{i=1}^{n} a_i\delta_i + \sum_{j=1}^{m} b_j\tau_j - \kappa(\delta, \tau)\right]\right)}{Z(a, b, c)} \tag{3.2}$$

where $Z(a, b, c)$ is the normalizing constant. Note that a is the vector of a_i parameters, b is the vector of b_j parameters, and $c \in \mathbb{R}$ is a constant.

We wish the marginalize out δ and τ to obtain $p(Y = y|a, b, c)$; that is,

$$p(Y = y|a, b, c) = \int\int p(Y = y|\delta, \tau)\pi(\delta, \tau|a, b, c)d\delta d\tau$$

Perhaps the best route to the marginal distribution $p(Y)$ is through Baye's Rule. Because we are using a conjugate prior $\pi(\delta, \tau)$, we know the posterior $p(\delta, \tau|Y)$ must be of the form

$$p(\delta, \tau|Y) = \frac{\exp\left(c'\left[\sum_{i=1}^{n} a_i'\delta_i + \sum_{j=1}^{m} b_j'\tau_j - \kappa(\delta, \tau)\right]\right)}{Z(a', b', c')}$$

for some vector of parameters a', b' and real number c'. Then, using Bayes' Rule we have

$$\begin{aligned} p_{a,b,c}(Y) &= \frac{p(Y|\delta, \tau)\pi(\delta, \tau)}{p(\delta, \tau|Y)} \\ &= \exp\left(\sum_{i=1}^{n}(d_{y_i} + ca_i - c'a_i')\delta_i + \sum_{j=1}^{m}(t_{y_i} + cb_j - c'b_j')\tau_j - (1 + c - c')\kappa(\delta, \tau)\right) \times \frac{Z(a', b', c')}{Z(a, b, c)} \end{aligned}$$

Since the left side does not depend on δ or τ, the right side must not either. Hence we have the following relations for all i, j:

$$\begin{aligned} c' &= c + 1 \\ a_i' &= \frac{ca_i + d_i}{c + 1} \\ b_j' &= \frac{cb_j + t_j}{c + 1} \end{aligned}$$

This leaves us with only the ratio of normalizing constants:

$$\begin{aligned} p(Y = y|a, b, c) &= \frac{Z(a', b', c')}{Z(a, b, c)} \\ &= \frac{\int\int \exp\left(c'\left[\sum_{i=1}^{n} a_i'\delta_i + \sum_{j=1}^{m} b_j'\tau_j - \kappa(\delta, \tau)\right]\right) d\delta d\tau}{\int\int \exp\left(c\left[\sum_{i=1}^{n} a_i\delta_i + \sum_{j=1}^{m} b_j\tau_j - \kappa(\delta, \tau)\right]\right) d\delta d\tau} \end{aligned} \tag{3.3}$$

3.2.1 Prior parameters $a, b,$ and c

After some inspection of the form of $p(Y = y|a, b, c)$, some constraints on the parameters are apparent. Because both the numerator and denominator must be integrated over all real

numbers, we must take care that the integrals do not become divergent. Consider just the denominator of of our marginal distribution,

$$\int\int \exp\left(c\left[\sum_{i=1}^{n} a_i\delta_i + \sum_{j=1}^{m} b_j\tau_j - \kappa(\delta,\tau)\right]\right) d\delta d\tau$$

$$= \int\int \frac{\exp\left(c\sum_{i=1}^{n} a_i\delta_i + c\sum_{j=1}^{m} b_j\tau_j\right)}{\left(\sum_y \exp(\delta_1 d_{y_1} + ... + \delta_n d_{y_n} + \tau_1 t_{y_1} + ... + \tau_m t_{y_m})\right)^c} d\delta d\tau$$

Since δ and τ can approach $-\infty$, we cannot let $a_i < 0$ or $b_j < 0$ lest the numerator tends to $+\infty$. Thus we have $a_i, b_j \geq 0$ as a hard constraint. Similarly, as $\delta, \tau \to +\infty$ the denominator approaches $\exp\left(c\sum_{i=1}^{n} \delta_i + c\sum_{j=1}^{n} \tau_j\right)$. In order for the above integral to converge, we need

$$\exp\left(c\sum_{i=1}^{n} a_i\delta_i + c\sum_{j=1}^{m} b_j\tau_j\right) \leq \exp\left(c\sum_{i=1}^{n} \delta_i + c\sum_{j=1}^{n} \tau_j\right)$$

and hence, for all i, j

$$a_i c - c \leq 0 \implies a_i \leq 1$$

$$b_j c - c \leq 0 \implies b_j \leq 1$$

Furthermore, it is required that $c > 0$ otherwise the denominator will cause the integral to diverge.

Thus, for all i, j we have the set of constraints

$$0 \leq a_i, b_j \leq 1$$

$$c > 0$$

Note that this is entirely consistent with the interpretation of the parameters for conjugate priors of the exponential family: a and b are the prior means for the parameters d_i and t_i. Since d_i and t_i are indicator variables, they must be bounded between 0 and 1. Furthermore, c can be thought of as the effective number of graphs worth of prior information, and with this we have the constraint $c > 0$.

3.3 Empirical Results

Look again at the right panel of Figure 1.2. This shows a degenerate ERGM over the set of all graphs of $N = 7$ nodes, using the edge and triangle counts as sufficient statistics. It is degenerate in the sense that although the means of the distribution correspond to 10 edges and 10 triangles, the class indicated by the red dot, most of the mass is placed toward the extremes of the graph space, far away from the average class. We will use this degenerate point in the graph space, the class with 10 edges and 10 triangles, as our test case.

The core idea is to average over the parameters of the Full-ERGM, the random effects, to obtain a distribution for Y that is dependent only on the hyperparameters. The goal is to find hyperparameters (whether those be for logistic priors or conjugate priors) which also produce distributions with means at this desired class (our test case of 10 edges and 10 triangles). The hope is that these distributions would not be degenerate in the above sense, and that we could potentially control degenerate behavior by tuning the hyperparameters.

Unfortunately, results indicate that the Hierarchical Full-ERGM appears to behave much like the ERGM. Neither the logistic prior nor conjugate prior succeed in producing distributions with means in the test class that are not degenerate. We will demonstrate this with the random effects model using the logistic prior first.

3.3.1 Random Effects Model with Logistic Priors

A word about the methodology first. We have used the exact enumeration of all labeled graphs on $N = 7$ nodes as the data set. Because we have all graphs readily available to us, we also have exact computation of the normalizing constant, hence MCMC routines were not necessary. Our hierarchical model essentially has two levels. At the bottom level, we have the random effects δ_i and τ_j for each dyad i and triad j, respectively, as in (3.1). At the top level we have the logistic priors, one for the dyads and one for the triads, parametrized with means and scales μ_d, s_d and μ_t, s_t, respectively.

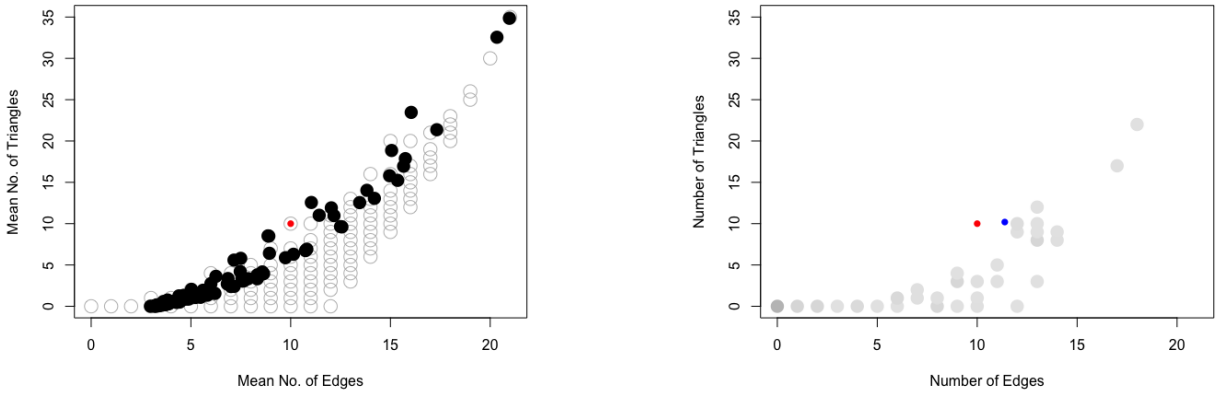

Figure 3.1: Random effects model with logistic priors parametrized by $\mu_d = -1.6$, $s_d = .1$, $\mu_t = .6$, $s_t = 1.1$. Each class of graphs, identified by the number of edges and triangles, is represented by a circle. *LEFT*: The black points represent means from 100 Full-ERGMs generated from the above hyperparameters. There is a large amount of diversity, though no distributions here land on the target class of 10 edges, 10 triangles (red dot). The average of these black points, however, does land in the target class. *RIGHT*: The distribution obtained by averaging 1000 Full-ERGMs generated from the above hyperparameters. The darker the shading, the more mass the model places on that class. The target class is indicated by the red dot, and the actual means of the distribution (11.38 edges, 10.19 triangles) are indicated by the blue dot. The slight discrepancy is due to the inexact nature of our scheme and finite number of distributions sampled, but it is clear that the resulting distribution is not what we had hoped for. We would like more mass to be concentrated near the target class.

Typically we would find the maximum likelihood estimates of the parameters of our likelihood (3.1), but we are instead interested in the hyperparameters for the logistic priors. Specifically we wish to find which set of μ_d, s_d, μ_t, s_t will yield a distribution with means at 10 edges and 10 triangles once all the random effects are averaged over. To achieve this, we do a search over the parameter space, allowing μ_d, μ_t to range over [-2.5, 2.5] and s_d, s_t to range over (0, 2.5]. With each set of four hyperparameters, the parameters δ_i and τ_j are generated to make a Full-ERGM, and the mean number of edges and triangles are calculated for the distribution. This is done 100 times for each quadruplet of hyperparameters, and the average of the means is recorded. From this scheme, we found that the set of hyperparameters which generate an average distribution closet to means 10 edges and 10 triangles is $\mu_d = -1.6$, $s_d = .1$, $\mu_t = .6$, $s_t = 1.1$. These values are not surprising. Our test class of 10 edges, 10 triangles is unique in the sense that it has a small number of edges but a relatively large number of triangles. These hyperparameters keep the number of edges low while still providing a sizable number of triangles. Furthermore, a class such as this is not unlike real-world networks which show the same clustering tendency; i.e., more triangles than would be expected given the number of edges.

Once the quadruplet of hyperparameters was found, we then sampled the random effects in order to make the Full-ERGM. With the Full-ERGM specified, we evaluated it over every graph in our dataset to see the distribution. This was done 100 times to get a sense of the diversity of distributions attainable from the hyperparameters. The left panel of Figure 3.1 shows the mean number of edges and triangles for each of these 100 distributions. We notice two things right away: firstly, the averages vary quite a bit over the convex hull; secondly, none of these means hit our target class of 10 edges, 10 triangles. However, the average of these means does place us very close to the target class. Thus, we see a degenerate behavior just one level higher. This may not be an issue though, since the idea was always to average over every distribution generated from the hyperparameters, which is the next step in the process.

We sampled from the logistic priors 1000 times, generating 1000 Full-ERGMs and evaluating each model over every graph in the dataset, recording the entire distribution. We then averaged over the 1000 distributions to get our final distribution with the random effects marginalized out. Our final distribution yielded means of 11.38 edges and 10.19 triangles, fairly close to our target class. The results can be seen in the right panel of Figure 3.1. As you can see, unfortunately our distribution is still degenerate. There is less mass near the extremes compared to the typical ERGM, but there is still far too little mass near the target class of 10 edges, 10 triangles. We were unsuccessful in marginalizing out the degenerate behavior through the use of logistic priors. In the next section, we will demonstrate that conjugate priors fare no better.

3.3.2 Random Effects Models with Conjugate Priors

Our hierarchical model again has two levels. At the bottom level, as before, we have the random effects δ_i and τ_j for each dyad i and triad j, respectively, as in (3.1). This time for the top level we use a conjugate prior to the Full-ERGM given by (3.2). We generate samples of δ_i and τ_j from the prior using a Metropolis Hastings MCMC routine to create the Full-ERGMs, and then marginalize over the random effects to obtain our pmf (3.3). The objective is to find hyperparameters a, b, c for the conjugate prior that will yield a distribution with means near our test class of 10 edges, 10 triangles and hopefully not be degenerate.

This time there is no need to do a grid search for the desired values of a, b, c. Based on our analysis above, we have the constraints $0 \leq a_i, b_j \leq 1$ and $c > 0$, and recall that each a_i can be thought of as the mean of δ_i, and likewise each b_j as the mean of τ_j. Note that in our dataset there are 21 possible edges and 35 possible triangles. Thus, if we take $a_i = .5$ for all i and $b_j = .3$ for all j, we would expect $.5 \times 21 = 10.5$ edges and $.3 \times 35 = 10.5$ triangles, very close to our test class. The value of c can be experimented with, and we show results with two distinct values: $c = 1$ and $c = 10$.

Figure 3.2 shows these results. In both cases, we end up with distribution means very

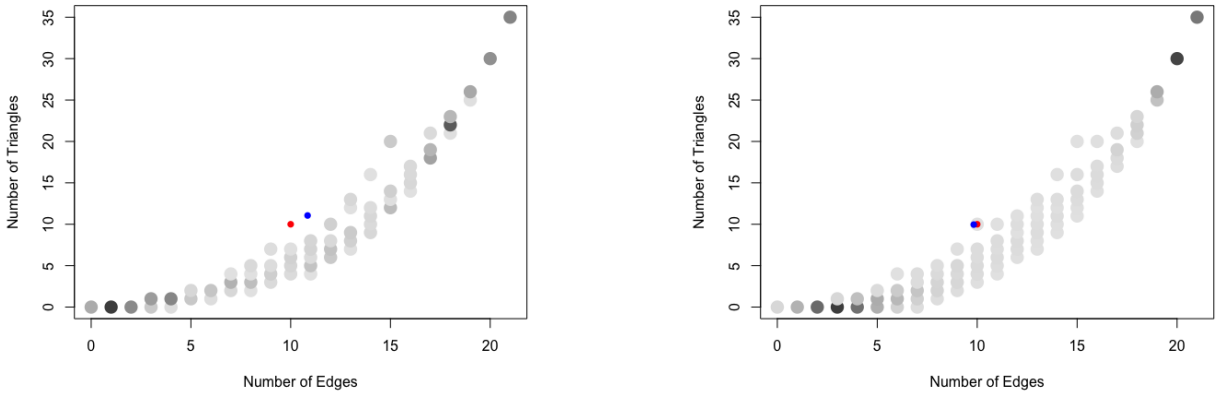

Figure 3.2: Random effects model with conjugate priors. Each class of graphs, identified by the number of edges and triangles, is represented by a circle. Darker shading indicates more mass is placed on the class. *LEFT*: Results using a conjugate prior with $a = .5$, $b = .3$, and $c = 1$. The red dot marks the target class of 10 edges, 10 triangles, and the blue dot marks the actual means of 10.85 edges, 11.06 triangles. Degenerate behavior appears again as little mass is placed near either of these markers. *RIGHT*: Results using a conjugate prior with $a = .5$, $b = .3$, and $c = 10$. The averages (blue dot) of 9.83 edges, 9.96 triangles coincide almost perfectly with the target class (red dot). Unfortunately, this distribution looks just as degenerate as the standard ERGM.

close to our target class of 10 edges, 10 triangles. Unfortunately, in both cases the models show degenerate behavior, placing most mass near the extremes of the graph space. However, we notice that as c increases, the model behaves more and more like the typical ERGM. In the following section we explore a simpler model to better understand the importance of the c parameter and discover an additional constraint that $c > \max(a, b)$. This effectively keeps us from lowering c all the way toward zero as an attempt to rein in the degenerate behavior.

Given that neither the logistic priors nor the conjugate prior produced any results strikingly different from the ERGM, the random effects models cannot be regarded as a solution to the degeneracy problem.

3.4 Elemental Model

Sometimes we can glean insight into difficult areas by first examining simple ones. Here we consider the most basic of situations - a graph with only two nodes - in order to better understand the effects of the parameters. This "elemental" graph Y may take only two states: $Y = 1$ in which there is an edge present between the two nodes, or $Y = 0$ in which no edge is present.

If we allow the edge to be present with probability p, the Erdos-Renyi Model is simply

$$\begin{aligned} Pr(Y = y|p) &= p^y(1-p)^{1-y} \\ &= \frac{e^{\delta y}}{1 + e^{\delta y}} \\ &= \exp(\delta y - \kappa(\delta)) \end{aligned}$$

where $\delta = \log\left(\frac{p}{1-p}\right)$ is the log-odds of the edge being present and $\kappa(\delta) = \log(1 + e^{\delta})$.

We wish to examine what happens if we place a prior over the log-odds δ.

A conjugate prior to the likelihood $p(Y = y|\delta)$ has the form

$$\pi(\delta|a, c) = \frac{\exp(a\delta - c\kappa(\delta))}{Z(a, c)}$$

where $Z(a,c) = \int \exp(a\delta - c\kappa(\delta))d\delta$ is the normalizing constant.

We wish the marginalize out δ to obtain $p(Y = y|a,c)$. We can arrive at an expression for the marginal distribution by first noting that, due to conjugacy, the posterior must have the form

$$p(\delta|Y) = \frac{\exp(a'\delta - c'\kappa(\delta))}{Z(a',c')}$$

for some real numbers a', c'. Using Baye's Rule, we then have

$$\begin{aligned} p(Y) &= \frac{p(Y|\delta)\pi(\delta)}{p(\delta|Y)} \\ &= \exp\left((y + a - a')\delta - (1 + c - c')\kappa(\delta)\right) \times \frac{Z(a',c')}{Z(a,c)} \end{aligned}$$

Since the left side does not depend on δ, the right side must not either. Hence we have the following relations: $a' = a + y$ and $c' = c + 1$. This leaves us with only the ratio of normalizing constants:

$$\begin{aligned} p(Y = y|a,c) &= \frac{Z(a',c')}{Z(a,c)} \\ &= \frac{\int \exp\left((a+y)\delta - (c+1)\kappa(\delta)\right) d\delta}{\int \exp\left(a\delta - c\kappa(\delta)\right) d\delta} \end{aligned}$$

3.4.1 Special Case for the Marginal Distribution

There is a surprisingly nice closed form for the marginal distribution $p(Y = y|a,c)$, if we assume that a, c are positive integers and $c > a$ Consider the denominator of the above expression,

$$\int \exp\left(a\delta - c\kappa(\delta)\right) d\delta = \int \frac{e^{a\delta}}{(1+e^{\delta})^c} d\delta$$

Integrating over all real values of δ, this evaluates simply to $B(a, c-a)$, where $B(p,q)$ is the beta function. To see this, first note that for positive integers p and q,

$$B(p,q) = \int_0^{\infty} \frac{t^{p-1}}{(1+t)^{p+q}} dt$$

Substituting $t = e^{\delta}$ gives

$$= \int_{-\infty}^{\infty} \frac{(e^{\delta})^{p-1}}{(1+e^{\delta})^{p+q}} e^{\delta} d\delta$$
$$= \int_{-\infty}^{\infty} \frac{e^{a\delta}}{(1+e^{\delta})^{c}} d\delta$$

if we let $p = a$ and $q = c - a$.

Thus, when a, c are positive integers with $c > a$ we can express the marginal distribution as the ratio of two beta functions:

$$p(Y = y|a, c) = \frac{B(a + y, c + 1 - a - y)}{B(a, c - a)}$$

3.4.2 Constraints on Parameters a and c

We would like to know the effects of the two parameters a and c on the probability of forming an edge, p. After some inspection of the form of $p(Y = y|a, c)$, some constraints on the parameters are apparent. Because both the numerator and denominator must be integrated over all real numbers, we must take care that the integrals do not become divergent. Consider again the denominator of $p(Y = y|a, c)$,

$$\int \exp\left(a\delta - c\kappa(\delta)\right) d\delta = \int \frac{e^{a\delta}}{(1+e^{\delta})^{c}} d\delta$$

Since δ can approach $-\infty$, we cannot let $a < 0$ lest the numerator tends to $+\infty$. Thus we have $a > 0$ as a hard constraint. Similarly, as $\delta \to +\infty$ the quantity $(1 + e^{\delta})^{c} \to e^{c\delta}$. In order for the above integral to converge, we need $e^{a\delta} < e^{c\delta}$ and hence $a < c$. This leads us to a set of constraints

$$0 < a < c$$

How these parameters influence $p(Y = y|a, c)$ is best viewed through the lens of the log-odds, δ. The plot below shows the influence of c on δ given a fixed value of a. So long as the above constraints are satisfied, δ ranges over all reals. However, we see that for a fixed value of a increasing c will only decrease δ, making the formation of an edge less and less likely.

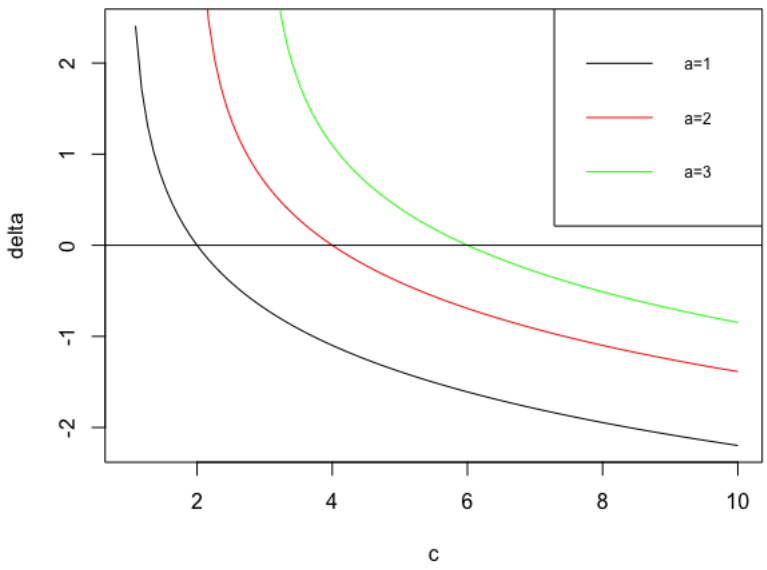

Figure 3.3: Effects of parameters a and c on the log-odds of an edge, δ. For a given value of a, δ is charted on the y-axis given the value of c on the x-axis. As $c \to +\infty$, $\delta \to -\infty$. As $c \to a$, $\delta \to +\infty$, reflecting the constraint that $c > a$.

3.5 Discussion

In this chapter we examined the expanded or "Full-ERGM", which abandons the homogeneity assumption and allows separate parameters for each dyad, triad, etc. within the network. We took a hierarchical approach where at the bottom level are the random effects for each dyad, etc. generated from the priors at the top level. We explored a logistic prior on the natural parameters themselves and conjugate priors on the mean parameters a la Diaconis and Ylvisaker (1979). Our objective was to find hyperparameters for these priors such that, when the random effects are averaged over, we arrive at distributions satisfying our constraints without being degenerate.

Our analysis shows that the Logistic prior, while being a natural choice for the natural parameters which represent the log-odds of a tie, etc., fail to achieve the desired result.

When marginalizing out the random effects, we can produce a distribution with means near the observed values but not one with considerable mass near the observed class.

Although also not successful at mitigating the effects of degeneracy, the analysis of the conjugate prior was fruitful in that we now have a better understanding of what each hyper-parameter controls. Notably, the discovery of the additional constraint that $c > \max(a, b)$ means that we will not be able to simply lower c in order to inhibit degeneracy by reigning in the "overspreading" of mass toward the extremal configurations. Note that this constraint would hold even for the basic ERGM and not only the Full-ERGM. This tells us that we must put in an additional constraint on the spread of the mass away from the observed class, and this is what is done, with stunning success, in the subsequent chapters.

CHAPTER 4

The Tapered ERGM

In this chapter we will take an in-depth look at the Tapered ERGM introduced by Fellows and Handcock (2017). We begin with a demonstration of what tapering looks like on an adolescent friendship network (the subsequent chapter provides several applications of the Tapered ERGM to real networks). We then prove several important properties of the Tapered ERGM, followed by a discussion of the kurtosis as a way to measure bimodality. Finally, we end with a general prescription for how and what to taper.

4.1 Why Tapering Works

$$p_{\theta,\tau}(Y = y) = \frac{\exp\left(\langle \theta, t(y) \rangle - \langle \tau, (\mu - t(y))^2 \rangle\right)}{Z(\theta, \tau)} \tag{4.1}$$

Equation 4.1 is the Tapered ERGM of Fellows and Handcock (2017), with a minor reparametrization. In the above, $\mu \equiv \mathrm{E}_{\theta,\tau}[t(y)]$. The form alone is enough to intuitively grasp why tapering works: we put a Gaussian penalty on graphs with statistics $t(y)$ that are too far away from the mean parameters μ. This stops the ERGM from putting mass on extremal configurations like the empty and complete graphs. The heavier the tapering, the less we allow the graph statistics to vary from the mean parameters; indeed, the Tapered ERGM is the maximal entropy distribution subject to constraints on the variance of the graph statistics (Fellows and Handcock, 2017). Later on we will generalize the Tapered ERGM using other forms of constraints, creating more models collectively known as Restorative Force Models.

Figure 4.1 shows tapering constraints applied to an adolescent friendship network (see

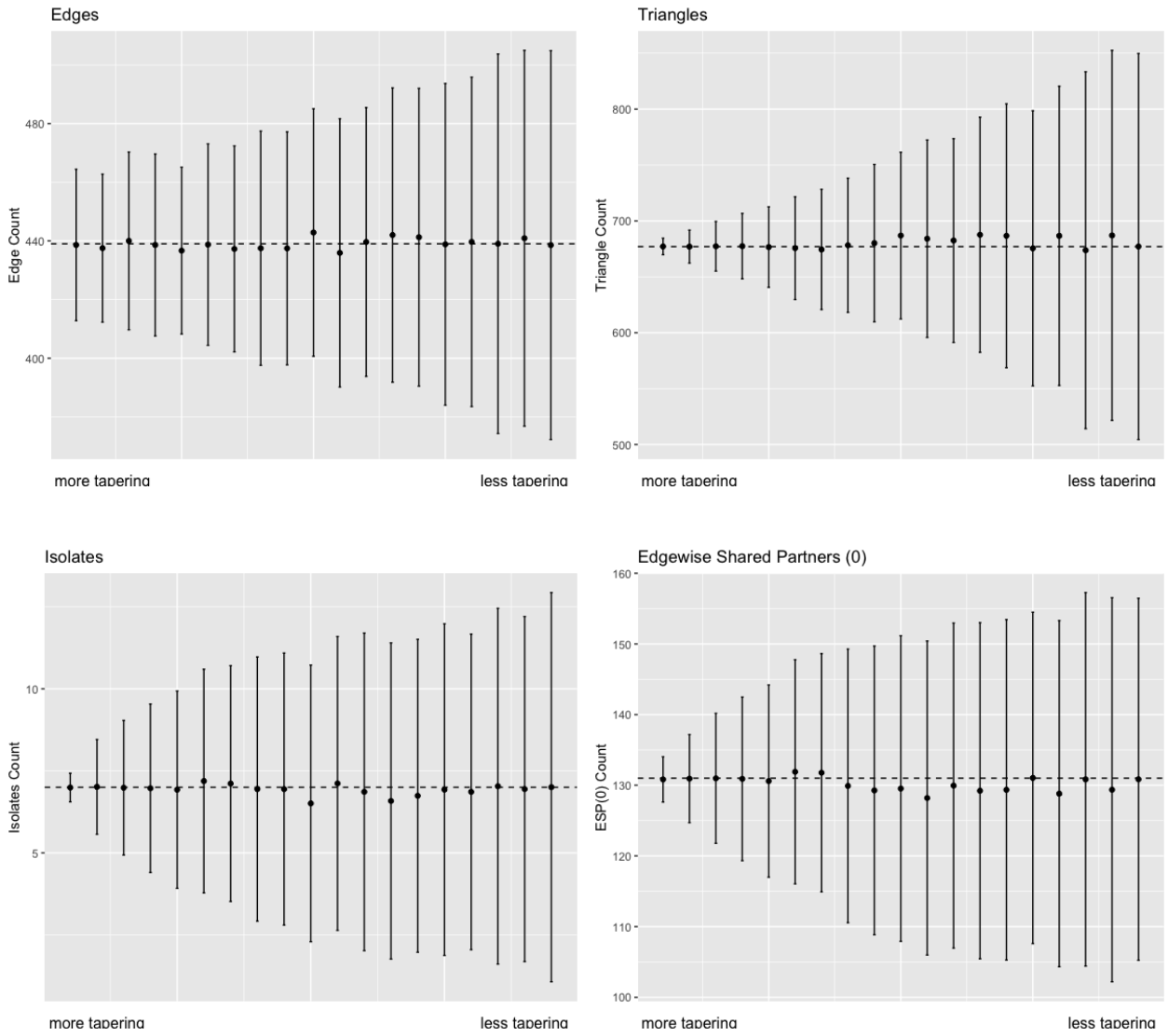

Figure 4.1: Variation in term counts across different levels of tapering. In each of the panels above, the dashed line indicates the term count in the observed network. Each point is the mean parameter at that level of tapering with corresponding error bars. We see that the mean parameters are consistently at the observed values. The isolates and ESP(0) plots do not show the effects of tapering until further left because the variance constraints are not realized until the tapering becomes heavier.

the case study of the Faux Desert High network in section 6.1). To the right of each panel, no tapering is applied and the mean parameter with error bars is plotted from the standard ERGM. As we move left within each panel, tapering is increased and the variance of the term is constrained more and more. Eventually those constraints become active, reducing the error bar of the mean parameter (i.e., the standard deviation of the term count). The mean parameters remain consistent at the observed values, we just allow them to vary less with increased tapering.

Of course, we need not rely on our conceptual intuition to see why the Tapered ERGM disallows degeneracy. We can prove that we can always find a parameter τ that will make $p_{\theta,\tau}(Y)$ non-degenerate, and we do so now.

In Horvát et al. (2015), the authors provide two critical results. When degeneracy occurs, the ERGM $p_\theta(Y)$ is plagued by mulitimodality. One way to ensure $p_\theta(Y)$ is unimodal is to require it does not have any local minima. The first result addresses this requirement.

Result 1. *Let $r(x) = h(x)\exp(\langle\theta, x\rangle)$, where x is a vector. $r(x)$ has no minima for all θ if and only if $h(x)$ is strictly log-concave.*

The next result involves $N(t(y))$, the counting function representing the number of graphs that have sufficient statistics $t(y)$. For example, if our vector of sufficient statistics for the graph y is $t(y)$ =(edge count,triangle count), then $N(0,0) = 1$ since there is only one graph with those statistics, namely the empty graph. It's worth pointing out that the standard ERGM is a pmf with respect to the counting measure. Furthermore, letting $t(y) \equiv t$, the probability a graph is sampled by the ERGM is

$$p(t|\theta) = \frac{N(t)}{Z(\theta)}\exp(\langle\theta, t\rangle)$$

where $p(t|\theta)$ is now a pmf with respect to the measure $N(t)$ due to the push-forward from the space of graphs Y to the space determined by $t(Y)$. From Result 1, Horvát et al. (2015) provide the following insight:

Result 2. *Let $\tilde{N}(t(y))$ be a smoothened, continuous version of $N(t(y))$. An ERGM is non-degenerate if and only if $\tilde{N}(t(y))$ is strictly log-concave.*

Because of it's discreteness, we need a continuous version of $N(t(y))$ in order to build on Result 1. Even with $\tilde{N}(t(y))$, the difficulty in utilizing this result is that computing $N(t)$ is in most cases impossible or at best extremely expensive. Under the Tapered ERGM, however, we have

$$p(t|\theta,\tau) = \frac{N(t)\exp(-\langle\tau,(\mu-t)^2\rangle)}{Z(\theta,\tau)}\exp(\langle\theta,t\rangle)$$

Now we don't necessarily have to worry about computing $N(t)$ and we can guarantee the Tapered ERGM is non-degenerate so long as $\tilde{N}(t)\exp(-\langle\tau,(\mu-t)^2\rangle)$ is strictly log-concave. Neither Horvát et al. (2015) nor Fellows and Handcock (2017) bother to define a smoothing function $\tilde{N}(t)$, but we will do so for completeness. Recall that t is the vector of sufficient statistics for a graph y. $N(t)$ is defined for all whole number-valued t that are in the support of t. For example, if t is the vector of edge and triangle counts, $t = (1,1)$ is not realizable. Thus, we need $\tilde{N}(t)$ such that it matches $N(t)$ if t is realizable yet also gives numerically similar values for any nearby vector in $\mathbb{R}^{k+}$. Then one possible choice for $\tilde{N}(t)$ that we can use is

$$\tilde{N}(t) = \begin{cases} N(t), \text{if } t \text{ is in the support} \\ \sum_i N(t_i)\exp(-||t-t_i||^2), \text{otherwise} \end{cases} \tag{4.2}$$

where the sum is taken over all values of t_i in the support.

Note that in Fellows and Handcock (2017), the authors prove the non-degeneracy for a larger class of models which subsumes the Tapered ERGM as we have defined it above. The larger class has the tapering center set to a general constant m instead of μ. We will now show a proof specific to the Tapered ERGM as defined here.

Theorem 3 (Theorem 2 in Fellows and Handcock (2017))**.** *For any vector μ of mean parameters, there exists a vector of tapering parameters τ such that the Tapered ERGM with tapering center μ is non-degenerate.*

Proof. We will use $\tilde{N}(t)$ as defined in equation 4.2 for our smoothing function. It suffices to show that $\tilde{N}(t)\exp(-\langle\tau,(\mu-t)^2\rangle)$ is strictly log-concave. Note that although $\mu=\mu(\theta,\tau)$ is dependent on parameters θ and τ, once those parameters are chosen $\mu(\theta,\tau)$ is a constant.

Let $r=\log(\tilde{N}(t))-h(t)$, where $h(t)=\langle\tau,(\mu-t)^2\rangle$. Then we have $\frac{\partial h}{\partial t_i}=-2\tau_i(\mu_i-t_i)$ and $\nabla^2 h$ a diagonal matrix

$$\nabla^2 h=\begin{bmatrix}2\tau_1 & & \\ & \ddots & \\ & & 2\tau_k\end{bmatrix}$$

Let $x=(x_1,...,x_k)$ be any nonzero column vector. Then $x^T\nabla^2 hx=\sum_i 2\tau_i x_i^2$. Thus, regardless of $\nabla^2\log(\tilde{N}(t))$, we can always choose τ large enough such that $x^T\nabla^2 rx<0$. Thus, r is concave and the Tapered ERGM is non-degenerate by Results 1 and 2. QED

Theorem 4 (Correction to Theorem 4 in Fellows and Handcock (2017))**.** *There exists a typo in Fellows and Handcock (2017) regarding the Hessian of the log-likelihood. In particular, the expression for* $\frac{\partial\mu_r(\theta,\tau)}{\partial\theta_i}$ *is incorrect. The correct expression is actually* $\frac{\partial\mu(\theta,\tau)}{\partial\theta_i}=(I-B)^{-1}c^i$. *The derivation below corrects for the error.*

Proof.

$$\begin{aligned}\frac{\partial\mu_r(\theta,\tau)}{\partial\theta_i}&=\operatorname{Cov}\left(t_r(y),t_i(y)-\sum_k 2\tau_k(\mu_k(\theta,\tau)-t_k(Y))\frac{\partial\mu_k(\theta,\tau)}{\partial\theta_i}\right)\\&=\operatorname{Cov}(t_r(Y),t_i(Y))+\sum_k 2\tau_k\frac{\partial\mu_k(\theta,\tau)}{\partial\theta_i}\operatorname{Cov}(t_r(Y),t_k(Y))\end{aligned}$$

Collecting all the partial derivatives on the left side, we have

$$\frac{\partial\mu_r(\theta,\tau)}{\partial\theta_i}-\sum_k 2\tau_k\frac{\partial\mu_k(\theta,\tau)}{\partial\theta_i}\operatorname{Cov}(t_r(Y),t_k(Y))=\operatorname{Cov}(t_r(Y),t_i(Y))$$

Which can be written as a system of linear equations

$$(I-B)\frac{\partial\mu(\theta,\tau)}{\partial\theta_i}=c^i$$

where, adopting the notation of Fellows and Handcock (2017), we define matrix B with $B_{rk} = 2\tau_k \operatorname{Cov}(t_r(Y), t_k(Y))$ and vector c^i with $c^i_r = \operatorname{Cov}(t_r(Y), t_i(Y))$. Thus, the correct expression is

$$\frac{\partial \mu(\theta, \tau)}{\partial \theta_i} = (I - B)^{-1} c^i$$

QED

Now that we have proven we can prevent multimodality, we need a way to measure it. This brings us to a discussion on kurtosis.

4.2 The Kurtosis and Bimodality

One of the hallmarks of degeneracy is bi/multimodality. When degeneracy strikes, often a large amount of mass is placed at or near the extremes of the space (empty and complete graphs) with very little mass placed near the observed graph. Figure 4.2 shows two bimodal marginal distributions taken from a degerate ERGM fit on the set of all graphs with seven nodes. In this degenerate fit, the mean parameters of 10 edges and 10 triangles are achieved, yet very little mass is put near those values. The Tapered ERGM allows us to rein in this bimodality by tapering sufficiently around the mean parameters until the distribution becomes unimodal. But the question remains as to how much tapering is sufficient in order to remove the bimodality. To answer this, we need an effective way to measure the bimodality of a distribution. Although not a magic bullet for measuring the bimodality of any distribution, the kurtosis is an effective instrument for our purposes.

Since its inception in 1905, the meaning and interpretation of the kurtosis statistic has been debated (Darlington, 1970; Moors, 1986; Westfall, 2014; DeCarlo, 1997; Chissom, 1970; Balanda and MacGillivray, 1988). For over a century, kurtosis has been at times rightly and wrongly associated with peakedness, heavy-tailedness, and modality. To make things clear, it is best to focus on the mathematics. The kurtosis of a random variable, X, is

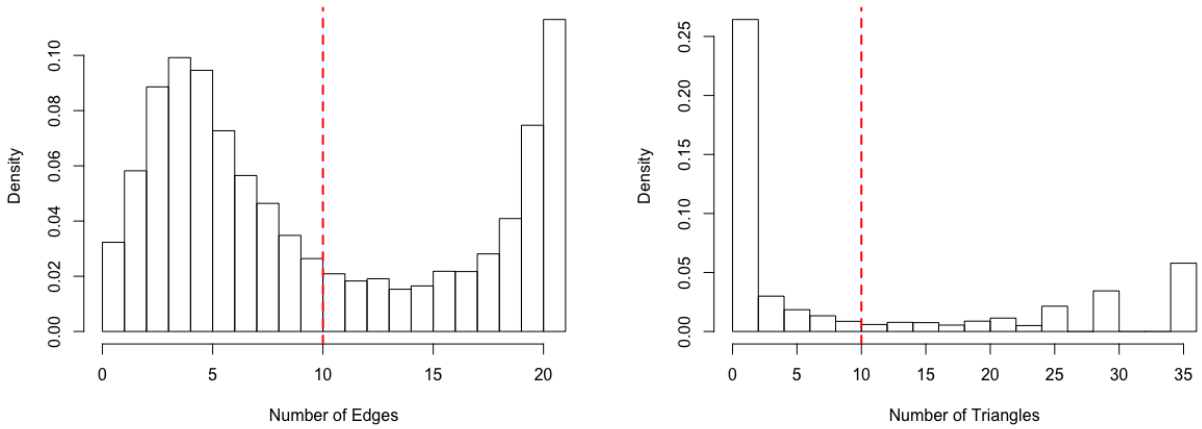

Figure 4.2: The marginal distributions of edges (left) and triangles (right) sampled from a degenerate ERGM. Much of the mass falls toward the empty and complete graphs with very little near the mean parameters (dashed red line).

$$\text{Kurt}[X] \equiv \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\text{E}[(X-\mu)^4]}{(\text{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\mu_2^2}$$

This can be equivalently stated as the expectation of Z^4, where Z is the standardized random variable. Using this framework, one can see immediately that only values with $|Z| > 1$ contribute non-negligibly to the kurtosis since raising a number less than 1 to the fourth power only brings that number closer to zero. Thus, as Westfall (2014) points out, the only unambiguous interpretation of the kurtosis is a measure of the tail extremity; i.e., the presence of outliers or the ability to produce outliers. We can make no assertion about the peakedness or even modality of the distributions if the peaks fall within one standard deviation of the mean.

We can, however, extract more from the kurtosis in certain contexts. Darlington (1970)

makes the following argument for interpreting the kurtosis as a measure of bimodality.

$$\begin{aligned}\mathrm{Var}[Z^2] &= \mathrm{E}[Z^4] - (\mathrm{E}[Z^2])^2 \\ &= \mathrm{Kurt}[X] - 1\end{aligned}$$

From the above identity, Darlington argues the kurtosis can be interpretted as "a measure of the degree to which the values of Z^2 cluster around their mean of 1" and furthermore as "a measure of the degree to which a distribution's z-scores cluster around +1 and -1." From this identity we see that the lower bound on the kurtosis is 1, and that this can only be achieved in a symmetric two-point distribution, i.e., one that is completely bimodal.

It would appear then that a lower kurtosis would indicate bimodality, where several benchmarks could be used ($\mathrm{Kurt}[X] = 3$ for the normal distribution, $\mathrm{Kurt}[X] = 9/5$ for the uniform). However, others (Hildebrand, 1971; Westfall, 2014) were quick to demonstrate counterexamples where bimodal distributions still had a high (≈ 3) kurtosis value, such as a "two-tailed gamma" distribution or the so-called "slip-dress" distribution. In these contrived examples, the two modes are very close to one another about the mean, and heavy tails extend to infinity producing the high kurtosis value. Yet, these examples show us precisely why it is okay to interpret the kurtosis as a measure of bimodality in the context of network modeling. The bimodal scenarios we encounter with degeneracy occur when significant probability mass is placed at the extremal configurations, i.e., the empty and complete graphs (Horvát et al., 2015; Handcock et al., 2003). So we need not worry about a bimodal distribution slipping past us with a high kurtosis value for two reasons: (i) the separation of the modes is large; and (ii) there is no opportunity for heavy tails to cover up bimodal peaks since the pmfs have finite, bounded support over the space of possible graphs. Thus, we can use the kurtosis statistic to help us measure bimodality for our purposes of identifying degeneracy.

4.2.1 Using the Kurtosis

The kurtosis is bounded below by the square of the skewness plus one. This lower bound is achieved only in a completely bimodal distribution such as $Bernoulli(p = .5)$.

$$\frac{\mu_4}{\mu_2^2} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1$$

The above inequality suggests we can use the **bimodality coefficient** (Ellison, 1987), β, to measure bimodality:

$$\beta = \frac{\gamma_1^2 + 1}{\gamma_2}$$

where γ_1 is the skewness and γ_2 the kurtosis. β lies in (0,1] with 1 indicating complete bimodality. The uniform distribution has a bimodal coefficient of 5/9, and any value above this threshold can be considered bi/multimodal.

4.2.2 Bias Correction for Sample Kurtosis

Unfortunately, the kurtosis is often conflated with the *excess kurtosis*, which is just the kurtosis relative to the normal distribution or $\text{Kurt}[X] - 3$. In order for the bimodality coefficient β to be bounded between 0 and 1, γ_2 must not be the excess kurtosis.

The sample kurtosis is a biased measure of the population kurtosis. A correction exists, but this correction is for the excess kurtosis. Indeed, at the time of writing the only known kurtosis function in R only returns the excess/ corrected excess kurtosis. Even more unfortunate, an existing package to compute the bimodal coefficient makes use of the excess kurtosis function, and often returns values outside of (0,1]. Moreover, the incumbent excess correction is only unbiased for samples from a normal distribution, and in general unbiased sample kurtosis measures do not exist.

Nevertheless, a correction for the sample (non-excess) kurtosis would be nice. I propose a correction, K_C found via cumulants. Let m_r denote the rth sample moment; that is,

$$m_r = \frac{1}{n}\sum(x - \bar{x})^r$$

Letting $\kappa_2 = \mu_2$ and $\kappa_4 = \mu_4 - 3\mu_2^2$ be the second and fourth cumulants, respectively, we see that

$$\frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\mu_2^2} - 3$$

or that the ratio of the cumulants gives the excess kurtosis. As Joanes and Gill (1998) note, $\frac{\kappa_4}{\kappa_2^2}$ is often approximated by the ratio of the unbiased estimates for the cumulants, namely

$$K_2 = \frac{n}{n-1}m_2$$

$$K_4 = \frac{n^2}{(n-1)(n-2)(n-3)}\left((n+1)m_4 - 3(n-1)m_2^2\right)$$

We wish to estimate $\frac{\kappa_4}{\kappa_2^2} + 3$, and doing so gives the proposed correction for the sample kurtosis, K_C, as

$$K_C = \frac{K_4}{K_2^2} + 3 = \frac{(n-1)(n+1)}{(n-2)(n-3)}\frac{m_4}{m_2^2} - \frac{3n-5}{(n-2)(n-3)} \tag{4.3}$$

The proposed corrections make use of the ratio of expectations in place of the expectation of the ratio. Of course, these two things are not equal. We could attempt to improve this approximation through the second order approximation of the expectation of the ratio of two random variables A and B,

$$\mathrm{E}\left[\frac{A}{B}\right] \approx \frac{\mathrm{E}[A]}{\mathrm{E}[B]}\left(1 - \frac{\mathrm{Cov}[A,B]}{\mathrm{E}[A]\,\mathrm{E}[B]} + \frac{\mathrm{Var}[B]}{(\mathrm{E}[B])^2}\right)$$

but $\mathrm{Cov}[m_4, m_2^2]$ involves moments up to μ_6. Hence, K_C is itself still biased, but it does estimate the (non-excess) kurtosis and allows the bimodality coefficient to be bounded between (0,1]. If we do not wish to correct for the finite sample bias, we can simply estimate γ_2 with $\frac{m_4}{m_2^2}$. Note that for large values of n, there is virtually no difference between the uncorrected and corrected estimates.

4.2.3 Alternative Bimodality Coefficient

Using either the corrected sample estimate or just the sample estimate in the bimodality coefficient calculation yields a β value between 0 and 1 except in cases of very small n. However, even with this improvement the bimodality coefficient still suffers from a known disadvantage of being influenced by the skewness. It is possible to have $\beta > 5/9$ even in unimodal distributions if the distribution is highly skewed. An alternative to the bimodality coefficient as computed above is

$$\beta' = \frac{1}{\gamma_2}$$

which is clearly free from the influence of the skewness. The threshold of 5/9 can still be employed as the skewness of the uniform is 0 anyway. β' is in fact still bounded in (0,1] since the smallest kurtosis possible is 1. Moors (1986) actually suggested the use of $1/\gamma_2$ as a measure of concentration around $\mu \pm \sigma$.

Now that we have a way to measure bi/multimodality, we can use the bimodality coefficient as a measuring stick for what and how much to taper.

4.3 Tapering Methodology

There are two main concerns when using the Tapered ERGM: (i) will the level of tapering affect the parameter estimates; and (ii) what level of tapering should we use (and on which terms)? Fortunately, empirically we have seen that the answer to (i) is 'not really'. Figure 6.2 shows that estimates of θ are remarkably stable across a wide range of tapering levels. In other words, we shouldn't worry about being uncertain of what to set our tapering parameter, τ, to. In fact, we can show that as τ goes to zero, the Tapered ERGM is identically the ERGM.

Theorem 5. *Let $Q(Y)$ denote the standard ERGM and $P(Y)$ denote the Tapered ERGM. Then as $\tau \to 0$, $D_{KL}(P||Q) \to 0$, where $D_{KL}()$ is the Kullback-Leibler divergence.*

Proof. Let $Q(Y)$ be the standard ERGM and $P(Y)$ the Tapered ERGM. That is,

$$Q(Y = y) = \frac{\exp\left(\sum_i \theta_i t_i(y)\right)}{C(\theta)}$$

and

$$P(Y = y) = \frac{\exp\left(\sum_i \theta_i t_i(y) - \sum_k \tau_k(\mu_k - t_k(y))^2\right)}{C(\theta.\tau)}$$

The Kullback-Leibler Divergence from Q to P is

$$\begin{aligned}
D_{KL}(P||Q) &= \sum_y P(y) \log\left(\frac{P(y)}{Q(y)}\right). \\
&= \sum_y P(y) \log\left(\exp\left(-\sum_k \tau_k(\mu_k - t_k(y))^2 - \log(C(\theta,\tau)) + \log(C(\theta))\right)\right) \\
&= \sum_y P(y)\left(-\sum_k \tau_k(\mu_k - t_k(y))^2 - \log\left(\frac{C(\theta,\tau)}{C(\theta)}\right)\right) \\
&= -\sum_k \tau_k \sigma_k^2 - \mathbb{E}_p\left[\log\left(\frac{C(\theta,\tau)}{C(\theta)}\right)\right]
\end{aligned}$$

where $\sigma_k^2 = \mathbb{E}_p[(\mu_k - t_k(y))^2] = \mathrm{Var}_p[t_k(y)]$.

Clearly as $\tau \to 0$,

$$\exp\left(\sum_i \theta_i t_i(y) - \sum_k \tau_k(\mu_k - t_k(y))^2\right) \to \exp\left(\sum_i \theta_i t_i(y)\right)$$

Therefore, as $\tau \to 0$, $C(\theta,\tau) \to C(\theta)$ and $\log\left(\frac{C(\theta,\tau)}{C(\theta)}\right) \to \log(1)$.

Thus, $D_{KL}(P||Q) \to 0$ as $\tau \to 0$. QED

The answer to question (ii) is more nuanced. Just because the tapering parameter τ does not have much effect on θ does not mean we should be unconcerned with setting τ. Indeed, we should seek to taper as few terms, and as little on each term, as possible. The argument for this is as follows. We saw in Theorem 5 above that the smaller τ is the closer the Tapered ERGM is to the ERGM. Of course, in a non-degenerate scenario we would not need any tapering at all, but we most often cannot know if the ERGM will be degenerate a priori. So we should apply the minimum amount of tapering necessary in order to get the

model to fit. This can be done in the following manner, with greater explanation of each step to follow.

Algorithm 6. *Setting the Tapering Parameter*

1. *Choose only the dyad-dependent terms to taper.*
2. *Set a large value of τ_k in order to heavily taper each of the k terms.*
3. *If the MCMC estimation converges, proceed to the next step. If the model is still unable to be fit, go back to step 1 and taper all terms.*
4. *Relax the amount of tapering by decreasing each τ_k until the bimodality coefficient for each of the k statistics is no greater than 0.4.*

Let's work through this step by step. Step 1 advises us to taper only the dyad-dependent terms. It is often these terms, like the triangle count, that go off the rails when degeneracy strikes so it is natural to taper them. One may wonder why we don't simply taper all terms by default. The reason we do not is not only because Theorem 5 tells us we would like some $\tau_k = 0$ (i.e., untapered terms), but also because τ has an effect on the interpretation of the parameters. To see this, let $P(Y_{ij} = 1|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^+)$ and $P(Y_{ij} = 0|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^-)$. Then, under the Tapered ERGM the log-odds of a tie is

$$\begin{aligned}
\log\left(\frac{P(Y_{ij}^+)}{P(Y_{ij}^-)}\right) &= \log\left(\frac{\exp\left(\sum \theta_k t_k(Y_{ij}^+) - \sum \tau_k(\mu_k - t_k(Y_{ij}^+))^2\right)}{\exp\left(\sum \theta_k t_k(Y_{ij}^-) - \sum \tau_k(\mu_k - t_k(Y_{ij}^-))^2\right)}\right) \\
&= \sum \theta_k \Delta t_k(Y_{ij}) - \sum \tau_k[(\mu_k - t_k(Y_{ij}^+))^2 - (\mu_k - t_k(Y_{ij}^-))^2] \\
&= \sum \theta_k \Delta t_k(Y_{ij}) - \sum \tau_k((\mu_k - t_k(Y_{ij}^+)) + (\mu_k - t_k(Y_{ij}^-)))(-t_k(Y_{ij}^+) + t_k(Y_{ij}^-)) \\
&= \sum \theta_k \Delta t_k(Y_{ij}) + \sum \tau_k((\mu_k - t_k(Y_{ij}^+)) + (\mu_k - t_k(Y_{ij}^-)))\Delta t_k(Y_{ij}) \\
&= \sum \Delta t_k(Y_{ij})\left[\theta_k + \tau_k \delta_{kij}\right]
\end{aligned}$$

where $\Delta t_k(Y_{ij}) = t_k(Y_{ij}^+) - t_k(Y_{ij}^-)$ is the change statistic, and $\delta_{kij} = (\mu_k - t_k(Y_{ij}^+)) + (\mu_k - t_k(Y_{ij}^-))$ is the sum of the differences from the mean. Hence, the interpretation of

the Tapered ERGM is that the conditional log-odds of a tie is the sum of the (change in statistics) $\times$ (θ_k plus a penalty), where the penalty is determined by τ. We know empirically that θ is very stable across a wide range of τ, so we may as well make τ as small as possible to get as close as we can to the standard ERGM interpretation where θ_k is the conditional log-odds of a tie.

Step two tells us to set a *large* value of τ. This may seem to contradict everything we just discussed above about wanting τ close to zero. But it is in fact consistent because in Step 4 we then relax the tapering and dial back τ to smaller values. The reason we actually want to start by over-tapering is because at the MLE of θ, we know that $\mu = t_{obs}(y)$. Thus, the value of τ doesn't matter when we are in the vicinity of the observed graph where $t(y) \approx t_{obs}(y)$. The heavy tapering allows us to find an initial value of θ_{MLE} quickly during MCMC estimation. Once we have our initial estimate of θ_{MLE}, we can restart our MCMC routine at that initial value using smaller values of τ. Convergence of the Markov chain should still be quick since our initial estimate of θ_{MLE} is likely very close to θ_{MLE}. Usually it is enough to taper only the dependent terms, since in doing so the independent terms (like edge count, for example) end up being curtailed indirectly. However, sometimes it is too difficult for the MCMC routine to converge, and in this scenario it is wise to start over and taper all terms.

Once we have an initial estimate of θ_{MLE} set, step 4 tells us to decrease the tapering. We can decrease τ until one of two things happens: the MCMC fails to converge (we have relaxed too far and degeneracy may be occuring), or until the bimodality coefficient $\beta \geq 0.4$, where β will make use of the bias-corrected kurtosis of equation 4.3. The choice of 0.4 as the cut-off value for β is somewhat arbitrary but very reasonable. Recall that $\beta \in (0, 1]$ where 1 indicates complete bimodality. The normal distribution has a bimodality coefficient of $\beta = 3$, and the uniform distribution has $\beta = .55$. The threshold of 0.4 is a nice medium between these, so we should allow τ_k to be as small as possible such that it still produces $\beta \leq 0.4$.

Noticeably absent from the algorithm above is what constitutes a "large" value of τ_k. This is because each value of τ_k must be set relative to $\mu_k = \mathrm{E}[t_k(y)]$. In Fellows and Handcock (2017), the authors suggest $\tau_k = \frac{1}{r^2\mu_k}$, which ensures observations r standard deviations from the mean are tapered most. This also takes the standard deviation of $t_k(y)$ to be $\sqrt{\mu_k}$, an assumption of Poisson dispersion. In reality we do not know if the variance of $t_k(y)$ is over- or under-dispersed, and the tuning parameter r allows us to adjust for this. Using a default value of $r = 2$ stems from a rough use of the empirical rule in the normal distribution. Thus, setting a "large" value of τ_k might instead use $r < 2$; for example, very heavy tapering would use $r = .5$ which corresponds to $\tau_k = \frac{4}{\mu_k}$. We should point out that setting overly small values of r (i.e., excessively large tapering) is also a danger. Doing so will constrain the model too much and not allow the Markov chain to explore the graph space away from the observed graph. Using $r = 2$ as a starting point and then slowly lowering r to increase tapering is the way to proceed, since we must be careful not to immediately jump to r values so small that the model also cannot converge because it is overly constrained. If we find that lowering r (increasing tapering) still does not make the model converge, we should consider tapering all terms (not just the dyad-dependent terms) and starting again using $r = 2$.

The theorems proven in this chapter show that it is theoretically possible to fit any network using the Tapered ERGM, and the algorithm above shows that it is also practical.

4.4 Conclusion

We began this chapter by showing what tapering looks like on an adolescent friendship network. Figure 4.1 allows us to visualize the effects of increased tapering on the variance of feature counts within generated networks. We bolster this intuitive concept with Theorem 3, an alternative proof that the Tapered ERGM can always be made non-degenerate. We also include Theorem 4 which gives the corrected formulation of the standard errors.

We also discussed at length the concept of the Kurtosis and why it is appropriate in the context of ERGMs. Employing a novel bias-corrected measure of the Kurtosis, we can

use the bimodality coefficient threshhold of 0.4 to know if we have tapered enough. This is an integral part of Alogrithm 6 which lays out exactly how, what, and when to taper the terms of the Tapered ERGM. In the following chapter, we generalize the Tapered ERGM to a larger class of models which are also very effective in combating degeneracy.

CHAPTER 5

Restorative Force Models

The deviousness of ERGMs is that, when degenerate, they do everything asked of them but still not what you want. ERGMs are tractable, interpretable models that maintain whatever distributional constraints (most often mean constraints) placed on them. When degeneracy appears, the ERGM still has all of those qualities, yet somehow manages to give very little weight to classes near the observed network. Degenerate ERGMs 'run off' to the extremes, putting a great deal of mass on the configurations at or near the empty and complete graphs.

In real-world networks, we observe such phenomena very rarely. Carter Butts of UCI is quick to point out that there are *some* cases where we would expect to see a bimodal distribution of extremes, such as modeling political opinion (especially in the present atmosphere). However, the vast majority of networks do not exhibit such behavior. Rather than going to the fringes, there seems to be a restoring force that keeps networks from fracturing to the empty state or imploding to the complete. The classic Physics model of the action of a spring is a direct analogy. When the spring is stretched, a restoring force pulls the spring back to its equilibrium state. The spring does not stay completely stretched out nor entirely compacted. What is the restoring force, then, that keeps networks from devolving into either extreme state as predicted by a degenerate ERGM? Perhaps it is the conditioning of society that keeps individuals from extreme behavior. Perhaps it is our evolutionary biology that keep social networks as they are; we need others to survive, so we cannot live in an empty state, yet we also cannot form strong ties with every person we meet since there are limits to our time, memory, etc. We can posit many reasons for what the restoring force is and how it works, but that is not the point. What's important is that we find a mechanism to

model such a force. The restorative force models introduced here do just that by putting an additional set of constraints in the ERGM. These constraints limit the dispersion of the model in various ways so as to keep things near "equilibrium".

The Tapered ERGM is a member of the restorative force models, and in this chapter we generalize it to include other tapering/restoring mechanisms. The Tapered ERGM employs a Gaussian penalty as the restoring mechanism. We will introduce three other restorative force models with alternative mechanisms, the MAD ERGM, the Stereo ERGM, and the LogCosh ERGM, and we will offer a comparison between them. We begin with a theoretical justification for the general restorative force model (RFM).

5.1 Maximum Entropy Derivation

The aim of this section is to establish a justification for additional constraints that will temper degeneracy. The constraints may vary significantly, but regardless of what they are we can specify the form of the resulting maximum entropy model. We will assume only mean constraints on sufficient statistics $t_i(x)$ in addition to a mean constraint on a set of generic functions $G_i(x)$. Thus, we have the following constrained optimization problem:

maximize

$$-\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

subject to

$$p(x) \geq 0 \ \forall x \in \mathcal{X}$$

$$\sum_{x\in\mathcal{X}} p(x) = 1$$

$$\sum_{x\in\mathcal{X}} p(x)t_i(x) = \alpha_i$$

$$\sum_{x\in\mathcal{X}} p(x)G_i(x) \leq \eta_i$$

$$\forall i \in \{1, ..., d\}$$

Note that the generic constraints involving $G_i(x)$ may involve inequalities or equalities; in the case of inequalities the method of Lagrange Multipliers can be extended using the Karush-Kuhn-Tucker conditions. The Lagrangian is:

$$\mathcal{L}(p) = H(P)+\theta_0(\sum_{x\in\mathcal{X}} p(x)-1)+\sum_{x\in\mathcal{X}} \lambda(x)(p(x)-0)+\sum_{i=1}^{d} \theta_i(\sum_{x\in\mathcal{X}} p(x)t_i(x)-\alpha_i)-\sum_{i=1}^{d} \lambda_i(\sum_{x\in\mathcal{X}} p(x)G_i(x)-\eta_i)$$

In the case where $\mathcal{X}$ is finite, $p = [p(x)]_{x\in\mathcal{X}}$ is a finite dimensional vector indexed by x. If we take the derivative of the Lagrangian with respect to each vector element $p(x)$, we obtain:

$$\frac{\partial \mathcal{L}}{\partial p} = -1 - \log p(x) + \theta_0 + \lambda(x) + \sum_{i=1}^{d} \theta_i t_i(x) - \sum_{i=1}^{d} \lambda_i G_i(x)$$

We set $\frac{\partial \mathcal{L}}{\partial p} = 0$ and solve to find a maximum, noting that we can verify we are finding a maximum since $\frac{\partial^2 \mathcal{L}}{\partial p(x)^2} = -\frac{1}{p(x)} < 0 \ \forall x$. The result is

$$p(x) = \exp\left(-1 + \theta_0 + \lambda(x) + \sum_{i=1}^{d} \theta_i t_i(x) - \sum_{i=1}^{d} \lambda_i G_i(x)\right)$$

The terms $-1 + \theta_0$ can be absorbed into the normalizing constant. Furthermore, since $p(x) > 0 \ \forall x$ the constraint that p be non-negative can be removed by complementary slackness, and thus $\lambda(x) = 0$. We are left with the maximum entropy model of the form

$$p_{\theta,\lambda}(Y = y) = \frac{\exp\left(\langle \theta, t(y)\rangle - \langle \lambda, G(y)\rangle\right)}{Z(\theta, \lambda)} \tag{5.1}$$

The remainder of this chapter will introduce several different RFMs, each characterized by its particular dispersion constraints $G_i(x)$.

5.2 Introducing the MAD ERGM

The *MAD ERGM* is the maximum entropy distribution subject to constraints on the mean absolute deviation of each sufficient statistic:

$$\mathrm{E}_p(|\mu_i - t_i(Y)|) \leq \sigma_i \ \forall i \in \{1, ..., d\}$$

which has the form

$$p_{\theta,\nu}(Y = y) = \frac{\exp\left(\langle \theta, t(y)\rangle - \langle \nu, |\mu - t(y)|\rangle\right)}{Z(\theta, \nu)} \tag{5.2}$$

The Tapered ERGM uses a Gaussian penalty, whereas the MAD ERGM employs a Laplacian penalty and as such provides an alternative that does not taper as strongly.

5.2.1 Properties of the MAD ERGM

One known drawback of this model is its non-differentiability. Before getting into the limitations of this and the subsequent work-arounds, we can prove the following.

Theorem 7. *Given a vector of tapering parameters ν and an observed network y_{obs} with statistics $t(y_{obs})$, at the MLE of θ, $\hat{\theta}_{MLE}$, it holds that $\mu(\hat{\theta}_{MLE}, \nu) = t(y_{obs})$.*

Proof. The log-likelihood of (5.2) and its first derivative are

$$l(\theta|y, \nu) = \sum_k \theta_k t_k(y) - \sum_k \nu_k |\mu_k - t_k(y)| - \log(Z(\theta, \nu))$$

$$\begin{aligned}
\frac{\partial l}{\partial \theta_i} &= t_i(y) - \sum_k \nu_k \text{sgn}(\mu_k - t_k(y)) \frac{\partial \mu_k}{\partial \theta_i} - \mathbb{E}\left[t_i(y) - \sum_k \nu_k \text{sgn}(\mu_k - t_k(y)) \frac{\partial \mu_k}{\partial \theta_i}\right] \\
&= t_i(y) - \sum_k \nu_k \text{sgn}(\mu_k - t_k(y)) \frac{\partial \mu_k}{\partial \theta_i} - \mu_i + \sum_k \nu_k \frac{\partial \mu_k}{\partial \theta_i} \mathbb{E}\left[\text{sgn}(\mu_k - t_k(y))\right]
\end{aligned}$$

Note that $\text{sgn}(\mu_k - t_k(y))$ is undefined when $\mu_k = t_k(y)$. We find $\hat{\theta}_{MLE}$ by setting $\frac{\partial l}{\partial \theta_i} = 0$ or looking for where $\frac{\partial l}{\partial \theta_i}$ is undefined, which yields $\mu(\hat{\theta}_{MLE}, \nu) = t(y)$. QED

The fact that the log-likelihood is not differentiable at $\hat{\theta}_{MLE}$ has several consequences. Firstly, we will not be able to prove the MAD ERGM can be made non-degenerate as we did for the Tapered ERGM. However, although we cannot technically complete the proof,

we have no reason to believe otherwise; that is, we should always be able to find a vector of tapering parameters ν such that the MAD ERGM is non-degenerate.

The second consequence is that the Hessian is undefined at $\hat{\theta}_{MLE}$, and hence typical methods of computing the standard errors cannot be used. The work-around to this predicament is a simple bootstrap, which is outlined below.

Algorithm 8. *Bootstrapping the Standard Errors for the MAD ERGM*

1. *Fit the MAD ERGM to obtain* $\hat{\theta}_{MLE}$.
2. *Sample* N *graphs from the MAD ERGM parametrized by* $\hat{\theta}_{MLE}$, *where* N *is a sufficiently large number.*
3. *For each of the* N *graphs, fit a MAD ERGM (using the same level of tapering from the original fit) to obtain* N *new estimates of* θ_{MLE}. *The standard deviation of the* N *estimates is the standard error of* $\hat{\theta}_{MLE}$.

With the non-differentiability of the absolute deviation being a hindrance, one may ask why we don't redefine the distance function in a piece-wise manner to be smooth at zero. It turns out exactly this is achieved in more elegant way by using the log-cosh distance, which is explored in section 5.4.

5.2.2 Choosing the Tapering Parameters

For choosing tapering parameters ν, we can essentially follow alogrithm 6 with one minor adjustment. That algorithm is reproduced here with minor changes that we discuss below.

Algorithm 9. *Setting the Tapering Parameter*

1. *Choose only the dyad-dependent terms to taper.*
2. *Set a large value of* ν_k *in order to heavily taper each of the* k *terms.*

3. *If the MCMC estimation converges, proceed to the next step. If the model is still unable to be fit, go back to step 1 and taper all terms.*

4. *Relax the amount of tapering by decreasing each ν_k until the bimodality coefficient for each of the k statistics is no greater than 0.4.*

Tapering all terms is largely unnecessary since tapering only the dependent terms usually has spillover effects on the rest of the terms. Furthermore, we should seek to taper as few terms, and as little on each term, as possible given the effect of tapering on the interpretation of the parameters. To see this, consider the log-odds of a tie between two nodes. Let $P(Y_{ij} = 1|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^+)$ and $P(Y_{ij} = 0|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^-)$. Then, under the MAD ERGM the log-odds of a tie is

$$\begin{aligned}\log\left(\frac{P(Y_{ij}^+)}{P(Y_{ij}^-)}\right) &= \log\left(\frac{\exp\left(\sum\theta_k t_k(Y_{ij}^+) - \sum_k \nu_k|\mu_k - t_k(Y_{ij}^+)|\right)}{\exp\left(\sum\theta_k t_k(Y_{ij}^-) - \sum_k \nu_k|\mu_k - t_k(Y_{ij}^-)|\right)}\right) \\ &= \sum\theta_k\Delta t_k(Y_{ij}) + \sum\nu_k\left(|\mu_k - t_k(Y_{ij}^-)| - |\mu_k - t_k(Y_{ij}^+)|\right)\end{aligned}$$

where $\Delta t_k(Y_{ij}) = t_k(Y_{ij}^+) - t_k(Y_{ij}^-)$ is the change statistic. Hence, the interpretation of the MAD ERGM is that the conditional log-odds of a tie is the sum of the $\theta_k \times$ (change in kth statistic) plus a penalty, where the penalty is determined by ν. We should make ν as small as possible to get as close as we can to the standard ERGM interpretation where θ_k is the conditional log-odds of a tie.

Step 2 instructs us to set a large value of ν_k, and hence we need to know a benchmark for what "large" is. As with the Tapered ERGM, each ν_k will be set relative to μ_k under an assumption of Poisson dispersion. And similar to the Tapered ERGM, we can set $\nu_k = \frac{1}{r\sqrt{\mu_k}}$ to ensure values of $t_k(y)$ more than r standard deviations from μ_k are tapered most heavily. In reality, we cannot know if the variance of $t_k(y)$ is over- or under-dispersed relative to the Poisson, but r can serve as a tuning parameter. The Tapered ERGM uses $r = 2$ (somewhat arbitrarily) as a benchmark for what is "far" away. If we wish for the tapering of the MAD ERGM to be on par with that of the Tapered ERGM, this implies we need $e^{-(\frac{x}{2})^2} = e^{-\frac{|x|}{s}}$

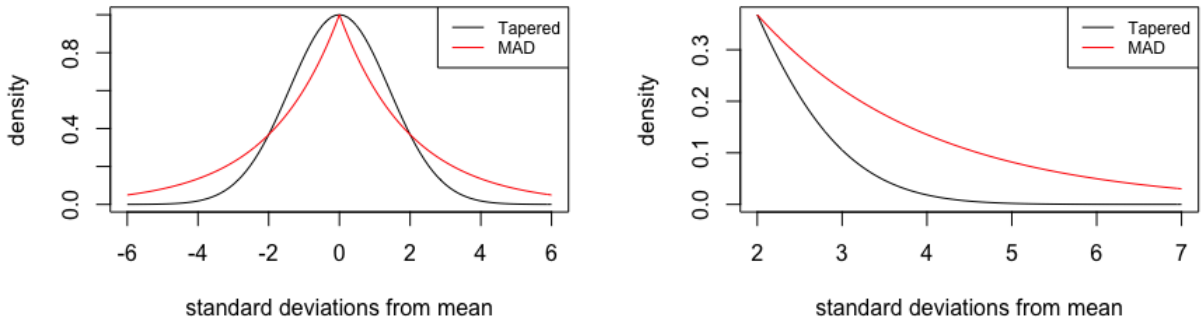

Figure 5.1: Tapered ERGM vs. MAD ERGM. *LEFT*: The MAD ERGM is scaled such that it down-weights statistics two standard deviations away just as much as the Tapered ERGM, hence the two curves intersect at ± 2. *RIGHT*: After 2 standard deviations, the MAD ERGM down-weights much less severely than the Tapered ERGM.

when $x = 2$, which of course yields $s = 2$. This is perhaps a roundabout way of saying that if we want the Tapered ERGM and MAD ERGM to penalize statistics two standard deviations away from the mean equally, we need to set $r = 2$ by default. Thus, a default setting of ν_k is $\nu_k = \frac{1}{2\sqrt{\mu_k}}$ and we can take a "large" value of ν_k to have $r < 2$. Note that even though those penalties will be precisely the same for statistics at two standard deviations, for values even further away the tapering under the MAD ERGM will be much less severe than that under the Tapered ERGM. Figure 5.1 shows this graphically.

The rest of algorithm 9 is straight-forward. The appeal of the MAD ERGM is that it down-weights graphs more mildly than the Tapered ERGM. The next RFM we discuss tapers even less.

5.3 Introducing the Stereo ERGM

The *Stereo ERGM* is the result of using an inverse stereographic projection onto a sphere of radius R in $\mathbb{R}^{d+1}$, where d is the number of statistics in the standard ERGM. It has the

form

$$p_{\theta,R}(Y = y) = \frac{\exp(\langle\theta, t(y)\rangle)}{Z(\theta, R)} \frac{1}{(R^2 + ||\mu - t(y)||^2)^2} \tag{5.3}$$

The Stereo ERGM is an RFM of equation 5.1 with a single constraint function $G(x) = \log(R^2 + ||\mu - t(x)||^2)$ and $\lambda = 2$, and it provides an even milder form of tapering. Unlike the exponential penalties of the Tapered and MAD ERGMs, this RFM utilizes a Cauchy-like polynomial penalty. Note that the parameter R is a single number and not a vector.

5.3.1 Geometric Interpretation

We can think of the probability mass assigned to each class of graphs as an area in $\mathbb{R}^d$, with d being the number of sufficient statistics. When an ERGM is degenerate, it will assign a large mass/area to classes far from the observed class. Figure 5.4 shows the same degenerate ERGM we have seen many times before, but in a different way illustrating the mass instead as an area in the plane.

Anyone who has ever looked at a map of the world has probably noticed that certain regions look distorted. While it is possible to find area-preserving mappings between the plane and sphere, it is exactly this area-distorting trait that we wish to take advantage of. By projecting the "areas" assigned by an ERGM onto a sphere, we can perhaps distort them in order to mitigate or eliminate degeneracy.

A mapping from the plane in $\mathbb{R}^d$ to the sphere in $\mathbb{R}^{d+1}$ can be accomplished using the *inverse stereographic projection*. Imagine that the plane intersects the sphere at its equator ($z = 0$). In a stereographic projection from the sphere to the plane, one can envision a light being shined downward from the north pole of the sphere. A light ray will intersect the sphere at some point $(y_1, ..., y_{n+1})$ and the plane at some point $(x_1, ..., x_n)$. The inverse of this projection allows us to map any point in the plane onto the sphere (see Figure 5.2).

As desired, the inverse stereographic projection does not preserve areas. In general, areas of regions in the plane farther from the origin will be deflated while areas of regions closer

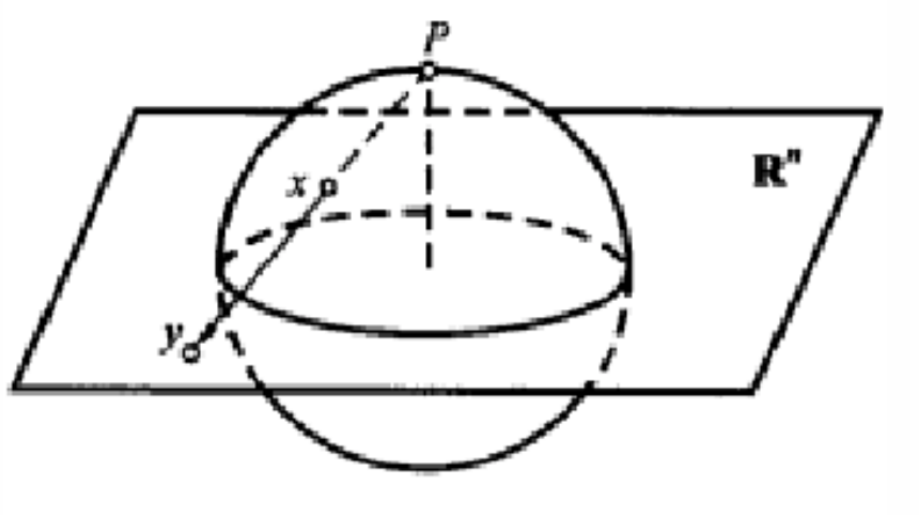

Figure 5.2: Stereographic Projection. A light ray from the north pole intersects the plane in $\mathbb{R}^n$ and the sphere in $\mathbb{R}^{n+1}$ in precisely one place, creating a map.

to the origin will be inflated (see Figure 5.3). In particular, the region at the origin of the plane will map to the south pole of the sphere and experience the largest increase in area. Let's now make these concepts mathematically precise.

Let $(x_1, ..., x_n)$ be a point on the plane in $\mathbb{R}^n$, and $(y_1, ..., y_{n+1})$ be a point on the sphere $y_1^2 + ... + y_{n+1}^2 = R^2$ in $\mathbb{R}^{n+1}$. To map $(x_1, ..., x_n) \mapsto (y_1, ..., y_{n+1})$, we will use the inverse stereographic projection $S(x_1, ..., x_n)$:

$$S(x_1, ..., x_n) = \left(\frac{2x_1}{R^2 + ||x||^2}, ..., \frac{2x_n}{R^2 + ||x||^2}, \frac{-R^2 + ||x||^2}{R^2 + ||x||^2} \right)$$

The Jacobian of this mapping is $J_{ij} = \frac{\partial S_i}{\partial x_j}$, and $g = J^T J$ is the metric tensor written in $\mathbb{R}^n$. The area element under this parametrization is given by

$$dA = \sqrt{|\det g|}\, dx_1 ... dx_n = \frac{4}{(R^2 + ||x||^2)^2}\, dx_1 ... dx_n$$

This explains the form and naming of the Stereo ERGM, taking the 'origin' of the plane (and center of the sphere) to be the vector of mean parameters $\mu = (\mu_1, ..., \mu_d)$.

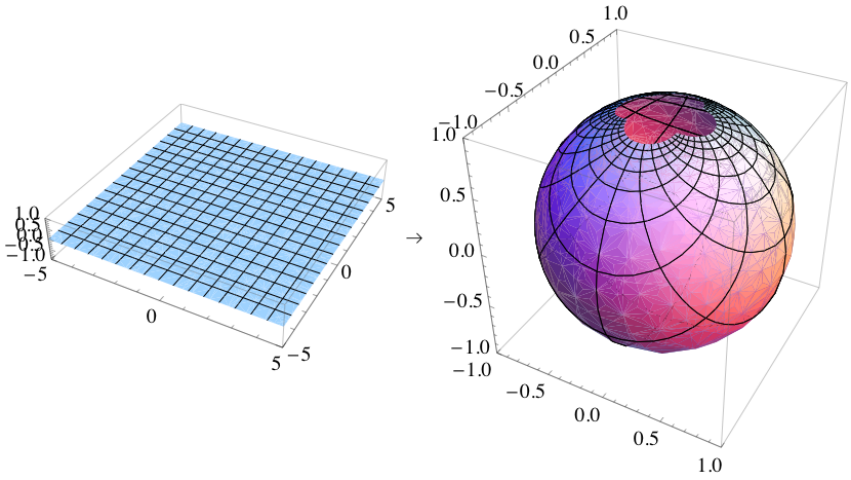

Figure 5.3: Distortion of areas under inverse stereographic projection. Areas of regions on the plane closer to the origin are expanded (mapped closer to the south pole), while areas of regions farther away are shrunk (mapped closer to the north pole). Photo Credit: Joshuardavis (public domain).

5.3.2 Properties of the Stereo ERGM

Revisiting equation 5.3 for the Stereo ERGM, we can see that as $R \to \infty$ the model approaches the standard ERGM. This is because if R is very large, $R^2 >> ||\mu - t(y)||^2$, making $\frac{1}{(R^2+||\mu-t(y)||^2)^2}$ essentially a constant that is absorbed into $Z(\theta, R)$. At the other extreme, as $R \to 0$ the model approaches a point mass. This is in line with our geometric intuition, as a sphere of infinite radius is essentially a plane (hence the standard ERG model) and a sphere of radius zero is a point. It is natural that the sphere we project onto be centered at the mean parameters $\mu = (\mu_1, ..., \mu_d)$, which at the MLE of θ are equal to the observed statistics by the following theorem.

Theorem 10. *Given a radius R and an observed network y_{obs} with statistics $t(y_{obs})$, at the MLE of θ, $\hat{\theta}_{MLE}$, it holds that $\mu(\hat{\theta}_{MLE}, R) = t(y_{obs})$.*

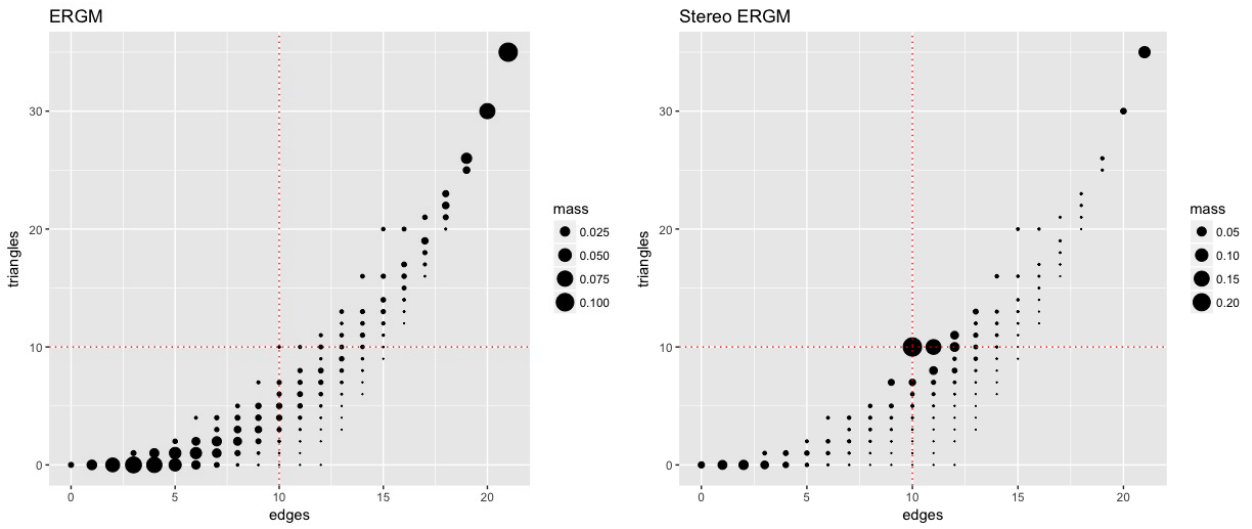

Figure 5.4: ERGM vs Stereo ERGM. Each class of graphs, identified by the number of edges and triangles, is represented by a circle with area proportionate to the probability mass assigned to the class. The sufficient statistics are the edge count and triangle count, with the mean parameters set at 10 edges and 10 triangles. *LEFT*: ERGM with degenerate behavior. Note that very little mass is placed near the means (indicated by the red dashed lines), and a great deal of mass is placed toward the extremes. *RIGHT*: Stereo ERGM with $R = 2$. A great deal of mass is now placed at the means and surrounding classes.

Proof. Let $Z(\theta, R)$ be the normalizing constant. Then

$$\frac{\partial}{\partial \theta_i} \log(Z(\theta, R)) = \mathrm{E}[t_i(y)] - 4\,\mathrm{E}\left[\frac{\sum_k (\mu_k(\theta, R) - t_k(y)) \frac{\partial \mu_k(\theta, R)}{\partial \theta_i}}{R^2 + ||\mu(\theta, R) - t(y)||^2}\right] = \mu_i(\theta, R)$$

since the second term is zero. The log-likelihood of the Stereo ERGM and its first derivative are

$$l(\theta|y, R) = \langle \theta, t(y) \rangle - \log(Z(\theta, R)) - 2\log(R^2 + ||\mu(\theta, R) - t(y)||^2)$$

$$\frac{\partial l}{\partial \theta_i} = t_i(y) - \mu_i(\theta, R) - \frac{4}{R^2 + ||\mu(\theta, R) - t(y)||^2} \sum_k (\mu_k(\theta, R) - t_k(y)) \frac{\partial \mu_k(\theta, R)}{\partial \theta_i} \tag{5.4}$$

We find $\hat{\theta}_{MLE}$ by setting $\frac{\partial l}{\partial \theta_i} = 0$ and solving, which yields $\mu(\hat{\theta}_{MLE}, R) = t(y)$. QED

The next theorem allows for calculation of the standard errors.

Theorem 11. *Given a radius R, the second derivative of the log-likelihood of the Stereo ERGM evaluated at $\hat{\theta}_{MLE}$ is*

$$\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\Big|_{\hat{\theta}_{MLE}} = -\frac{\partial \mu_i(\hat{\theta}_{MLE}, R)}{\partial \theta_j} - \frac{4}{R^2} \sum_k \frac{\partial \mu_k(\hat{\theta}_{MLE}, R)}{\partial \theta_i} \frac{\partial \mu_k(\hat{\theta}_{MLE}, R)}{\partial \theta_j}$$

Proof. Let $\mu_i(\theta, R) = \mathrm{E}[t_i(Y)]$. To ease notation, let $t_k(Y) \equiv t_k$ and $\mu_k(\theta, R) \equiv \mu_k$. Then we have

$$\begin{aligned}
\frac{\partial \mu_i(\theta, R)}{\partial \theta_j} &= \mathrm{Cov}[t_i, t_j] - 4\,\mathrm{Cov}\left[t_i, \frac{1}{R^2 + ||\mu - t||^2} \sum_k (\mu_k - t_k) \frac{\partial \mu_k}{\partial \theta_j}\right] \\
&= \mathrm{Cov}[t_i, t_j] - 4 \sum_k \frac{\partial \mu_k}{\partial \theta_j} \mathrm{Cov}\left[t_i, \frac{\mu_k}{R^2 + ||\mu - t||^2}\right] + 4 \sum_k \frac{\partial \mu_k}{\partial \theta_j} \mathrm{Cov}\left[t_i, \frac{t_k}{R^2 + ||\mu - t||^2}\right]
\end{aligned}$$

Note that near $\mu = (\mu_1, ..., \mu_d)$, $\phi(t) = \frac{1}{R^2 + ||\mu - t||^2}$ is concave. By Jensen's inequality,

$$\begin{aligned}
\phi(\mathrm{E}[t]) &\geq \mathrm{E}[\phi(t)] \\
\frac{1}{R^2} &\geq \mathrm{E}\left[\frac{1}{R^2 + ||\mu - t||^2}\right]
\end{aligned}$$

Thus, we can take a conservative overestimate of $\frac{\partial \mu_i(\theta, R)}{\partial \theta_j}$ as

$$\frac{\partial \mu_i(\theta, R)}{\partial \theta_j} = \mathrm{Cov}[t_i, t_j] + \frac{4}{R^2} \sum_k \frac{\partial \mu_k}{\partial \theta_j} \mathrm{Cov}[t_i, t_k]$$

Collecting all the partial derivatives on the left side, we have

$$\frac{\partial \mu_i(\theta, R)}{\partial \theta_j} - \frac{4}{R^2} \sum_k \frac{\partial \mu_k}{\partial \theta_j} \operatorname{Cov}[t_i, t_k] = \operatorname{Cov}[t_i, t_j]$$

Which can be written as a system of linear equations

$$(I - B) \frac{\partial \mu(\theta, R)}{\partial \theta_j} = c^j$$

where we define matrix B with $B_{ij} = \frac{4}{R^2} \operatorname{Cov}[t_i, t_j]$ and vector c^j with $c_i^j = \operatorname{Cov}[t_i, t_j]$. Thus, the vector of the derivatives of each mean parameter with respect to θ_j is

$$\frac{\partial \mu(\theta, R)}{\partial \theta_j} = (I - B)^{-1} c^j$$

This is entirely consistent with our geometric intuition: as $R \to \infty$, $B \to 0$ and $\frac{\partial \mu(\theta,R)}{\partial \theta_j} \to c^j$.

The fist derivative of the log-likelihood is given by equation 5.4. The second derivative of the log-likelihood is

$$\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} = -\frac{\partial \mu_i}{\partial \theta_j} - 4 \sum_k \left[\frac{(\mu_k - t_k)}{R^2 + ||\mu - t||^2} \frac{\partial^2 \mu_k}{\partial \theta_i \partial \theta_j} + \frac{\partial \mu_k}{\partial \theta_i} \left(\frac{(R^2 + ||\mu - t||^2) \frac{\partial \mu_k}{\partial \theta_j} - (\mu_k - t_k) \sum_k 2(\mu_k - t_k) \frac{\partial \mu_k}{\partial \theta_j}}{(R^2 + ||\mu - t||^2)^2} \right) \right]$$

which, simplified at the MLE, reduces to

$$\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\Big|_{\hat{\theta}_{MLE}} = -\frac{\partial \mu_i(\hat{\theta}_{MLE}, R)}{\partial \theta_j} - \frac{4}{R^2} \sum_k \frac{\partial \mu_k(\hat{\theta}_{MLE}, R)}{\partial \theta_i} \frac{\partial \mu_k(\hat{\theta}_{MLE}, R)}{\partial \theta_j}$$

QED

Unlike some of the other Restorative Force Models, we cannot always guarantee that the Stereo ERGM is non-degenerate. The next theorem makes clear the conditions under which the Stereo ERGM will be non-degenerate.

Theorem 12. *Let $\tilde{N}(t)$ be the smoothing function of equation 4.2. Define $h = 2 \log(R^2 + ||\mu - t||^2)$. If there exists a number R such that $R \geq \sqrt{2} \max_{k,t_k}(|u_k - t_k|)$ and $x^T \nabla^2 h x > x^T \nabla^2 \log(\tilde{N}(t)) x$ for any non-zero vector x, then we are guaranteed the Stereo ERGM is non-degenerate.*

Proof. From Results 1 and 2, the Stereo ERGM will be non-degenerate if $\tilde{N}(t)/(R^2 + ||\mu - t||^2)^2$ is log-concave. So let $r = \log(\tilde{N}(t)) - h(t)$, with $h(t)$ defined as above. Then we have the following partial derivatives of h:

$$\frac{\partial h}{\partial t_i} = \frac{4(t_i - \mu_i)}{R^2 + ||\mu - t||^2}$$

$$\frac{\partial^2 h}{\partial t_i^2} = \frac{4}{(R^2 + ||\mu - t||^2)^2}\left(R^2 + \sum_{k \neq i}(\mu_k - t_k)^2 - (\mu_i - t_i)^2\right)$$

$$\frac{\partial^2 h}{\partial t_i \partial t_j} = \frac{4}{(R^2 + ||\mu - t||^2)^2}\left(-2(\mu_i - t_i)(\mu_j - t_j)\right)$$

Clearly $\nabla^2 h$ is not diagonal. However, we can ensure $\nabla^2 h$ is positive-definite if it is diagonally dominant (and the diagonal entries are non-negative). Let us first note that

$$\left(\sum_k (\mu_k - t_k)\right)^2 > 0$$

$$\sum_k (\mu_k - t_k)^2 + 2\sum_{i \neq j}(\mu_i - t_i)(\mu_j - t_j) > 0$$

$$\sum_k (\mu_k - t_k)^2 > -2\sum_{i \neq j}(\mu_i - t_i)(\mu_j - t_j)$$

$$R^2 + \sum_k (\mu_k - t_k)^2 - 2(\mu_i - t_i)^2 > -2\sum_{i \neq j}(\mu_i - t_i)(\mu_j - t_j)$$

Where the last inequality holds so long as $R \geq \sqrt{2}|\mu_i - t_i|$. When the last inequality holds, $\nabla^2 h$ is diagonally dominant and therefore positive-definite.

Having $\nabla^2 h$ positive-definite is unfortunately not enough to guarantee $\nabla^2 r$ is negative-definite. But note that the condition $R \geq \sqrt{2}|\mu_i - t_i|$ is sufficient that $\nabla^2 h$ be positive-definite, but it is not necessary. The final inequality above may still hold even if $R < \sqrt{2}|\mu_i - t_i|$, and in general as $R \to 0$ the magnitude of $x^T \nabla^2 h x$ will become larger than that of $x^T \nabla^2 \log(\tilde{N}(t)) x$. Unfortunately, as $R \to 0$ we may also lose the positive-definiteness of $\nabla^2 h$ which is necessary. Thus, if there exists a number R such that $R \geq \sqrt{2} \max_{k,t_k}(|u_k - t_k|)$ and $x^T \nabla^2 h x > x^T \nabla^2 \log(\tilde{N}(t)) x$ for any non-zero vector x, we can be certain r is concave and hence the Stereo ERGM non-degenerate. Note that the maximum is taken over k to

make sure R is large enough for every statistic, but also over $t_k \equiv t_k(y)$ as R must also be larger than the furthest distance away from μ_k. QED

5.3.3 Choosing the Radius

While the last proof shows that we cannot always find a radius R to make the Stereo ERGM non-degenerate, it does give us some guidance on how to choose R. By the reasoning laid out above, we can begin with $R = \sqrt{2}\max_k(|u_k - t_k|)$. From there we can increase or decrease the value of R if necessary. But how to compute $|u_k - t_k|$? Of course our mean parameters μ_k are known, but $t_k \equiv t_k(Y)$ clearly depends on Y. We can, however, know a priori $t(Y)$ for two troublesome states of Y: the empty and complete graphs. It is therefore possible to find $|u_k - t_k|$ using either of those and the Euclidean metric, which is a natural choice since we are taking the standard ERGM to be in the plane. An even simpler calculation utilizes the fact that when y is the empty graph $t(y) = 0$, giving us $R = \sqrt{2}\max_k(u_k)$. This leads to the following prescription for choosing the radius, R.

Algorithm 13. *Choosing the Radius*

1. *Choose* $R = \sqrt{2}\max_k(u_k)$.
2. *If the MCMC estimation converges, proceed to the next step. If the model is still unable to be fit, go back to step 1 and decrease R.*
3. *Check that the bimodality coefficient for each of the k statistics is no greater than 0.4. If so, you can stop if the marginal distributions of the k statistics appear sufficiently spread out. If the marginal distributions appear too concentrated about the means, increase the value of R. If the bimodality coefficients are not less than 0.4, continue to decrease R until the model converges, if it does.*

Table 5.1: Summary of basic Stereo ERGM fit on Sampson's Monastery Network

Term	$\hat{\theta}$	$\hat{se}(\hat{\theta})$
edges	-1.73	0.289***
triangles	0.14	0.057*

$^*p < .05$ $^{**}p < .01$ $^{***}p < .001$

5.3.4 Illustration

In this section we provide a basic demonstration of the Stereo ERGM. We fit a simple edges + triangle model to the Sampson's Monastery Network (Sampson, 1968). This network is a time-aggregated network of three affectation networks. A tie from monk A to monk B exists if A nominated B as one of his three/four best friends at any of the three time points.

Using a standard ERGM, it is not possible to fit and edges + triangle model to this data set. However, with the Stereo ERGM we can fit such a model. Algorithm 13 was followed to choose a radius of $R = \sqrt{500}$. Table 5.1 gives the parameter estimates and figure 5.5 shows a decent fit for such a simple model.

Of course, a simple edges + triangle model is probably not the best model possible. Incorporating other information, such as the group affiliation of each monk ("Loyal", "Outcasts", or "Turks") results in a much better. As it turns out, when matching terms on group affiliation are added to the model it can be estimated using only a standard ERGM. This gives us an opportunity to fit the same model with a Stereo ERGM (again using $R = \sqrt{500}$) and compare the results to assess the impact of the tapering. Group specific triangle terms were also added to the model to examine the effects of triad closure versus homophily (note: the Outcasts triangle term was excluded because it could not be estimated simultaneously with an Outcasts matching term given the data). Table 5.2 shows that tapering under the Stereo ERGM does not impact the parameter estimates as the results are nearly identical. In this network, friendship ties tend to be driven my group homophily and not transitivity.

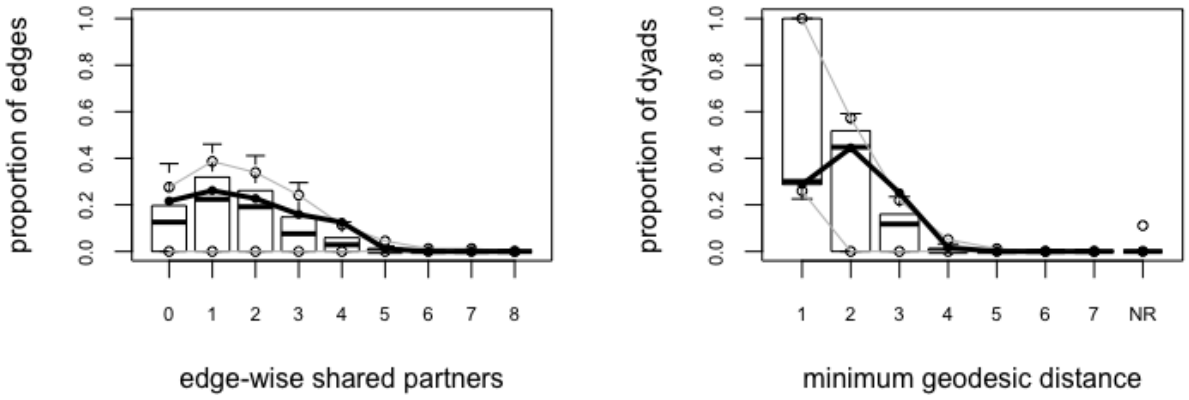

Figure 5.5: The edgewise shared partners distribution and minimum geodesic distance from networks simulated from the Stereo ERGM MLE compared to the observed network statistics (thick black line), where a simple edges + triangle Stereo ERGM was fit on Sampson's Monastery Network.

Table 5.2: ERGM and Stereo ERGM fits on Sampson's Monastery Network

Term	ERGM	Stereo ERGM
edges	-1.66 (0.517)**	-1.67 (0.520)**
triangles	-0.102 (0.142)	-0.09 (0.155)
match(Loyal)	4.06 (1.065)***	4.07 (1.090)***
match(Outcasts)	3.98 (0.993)***	3.80 (1.032)***
match(Turks)	2.47 (1.570)	1.40 (1.699)
triangle(Loyal)	-0.27 (0.206)	-0.28 (0.220)
triangle(Turks)	0.12 (0.204)	0.22 (0.226)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

5.4 Introducing the LogCosh ERGM

The next RFM uses a somewhat more exotic distance measure but offers many of the advantages of the previous RFMs without any of the disadvantages. The *LogCosh ERGM* is the maximum entropy distribution subject to constraints on the mean log-cosh deviation of each sufficient statistic:

$$\mathrm{E}_p[\log(\cosh(\mu_i - t_i(Y)))] \leq \sigma_i \ \forall i \in \{1, ..., d\}$$

where $\cosh(x)$ is the hyperbolic cosine, $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$.

Under these constraints, we arrive at the form

$$p_{\theta,\lambda}(Y = y) = \frac{\exp\left(\langle \theta, t(y) \rangle\right)}{Z(\theta, \lambda)} \prod_k \operatorname{sech}(\mu_k - t_k(y))^{\lambda_k} \tag{5.5}$$

5.4.1 Justification

Let $H(x, y)$ denote the log-cosh distance between real numbers x and y, $H(x, y) = \log(\frac{1}{2}(e^{x-y} + e^{y-x}))$. H clearly satisfies (i) $H(x, y) \geq 0 \, \forall x, y$, (ii) $x = y \iff H(x, y) = 0$, and (iii) $H(x, y) = H(y, x)$. H is not a metric, however, because it does not satisfy the triangle inequality. Nonetheless, H is still a preferable distance measure because of its asymptotic and infinitesimal behavior.

Theorem 14. *As $x \to \infty$, $\log(\cosh(x)) \to |x| - \log 2$. As $x \to 0$, $\log(\cosh(x)) \approx \frac{x^2}{2}$.*

Proof. For the first claim, we can rewrite $\cosh(x)$ as $\cosh(x) = \frac{1}{2}(e^x + e^{-x}) = \frac{1}{2}e^x(1 + e^{-2x})$. This implies $\log(\cosh(x)) = x + \log(1 + e^{-2x}) - \log 2$. So as $x \to \infty$, $\log(\cosh(x)) \to x - \log 2 = |x| - \log 2$.

For the second claim, we make use of two well-known power series:

$$\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots$$

$$\log(1 + x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{n+1}}{n + 1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

where the second power series converges when $|x| < 1$. Using the first power series, we can say $\log(\cosh(x)) = \log(1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots) \approx \log(1 + \frac{x^2}{2})$. Using the second power series, we can say $\log(1 + \frac{x^2}{2}) = \frac{x^2}{2} - \frac{x^4}{8} + \frac{x^6}{24} \cdots \approx \frac{x^2}{2}$ when x is small. Thus, when $x \to 0$, $\log(\cosh(x)) \approx \frac{x^2}{2}$. QED

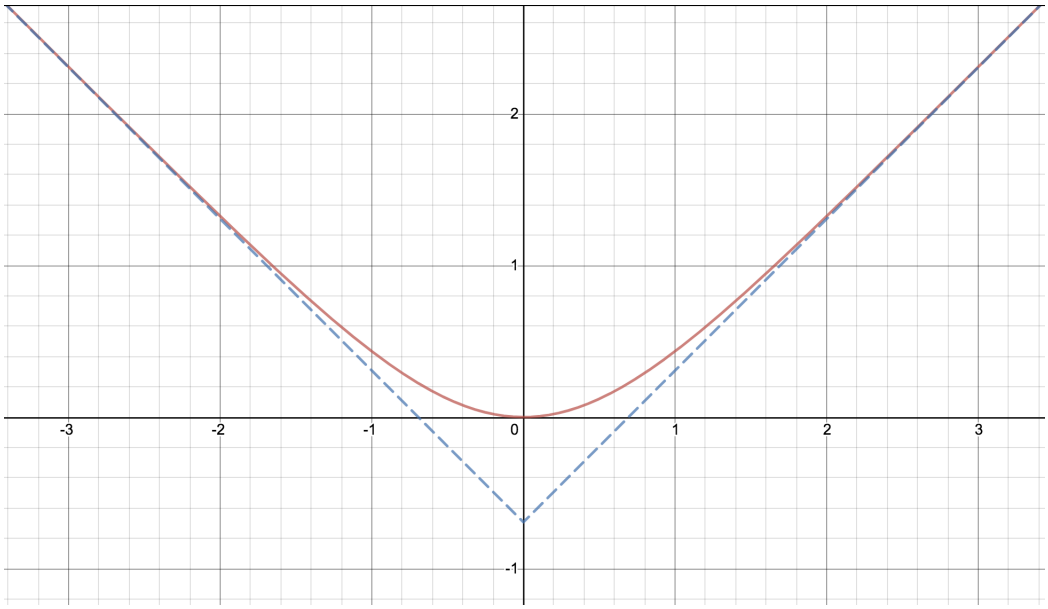

Figure 5.6: Behavior of the log-cosh penalty. $\log(\cosh(x))$ is the solid line in red, $|x| - \log 2$ is the dashed line in blue.

Thus, when $|x - y|$ is large, $H(x, y)$ behaves like the absolute deviation and when $|x - y|$ is small it behaves like the squared deviation (see Figure 5.6). In a sense, the LogCosh ERGM of equation 5.5 is the best of both worlds as it combines the mild tapering of the MAD ERGM with the differentiability of the Tapered ERGM. Moreover, unlike the Stereo ERGM, we can also guarantee the non-degeneracy of the LogCosh ERGM. The next section formally proves this assertion and others.

5.4.2 Properties of the LogCosh ERGM

In this section we will prove that for the LogCosh ERGM the mean parameters are equal to the observed values at the MLE, the LogCosh ERGM can always be made non-degenerate,

and we will find an expression for the Hessian.

Theorem 15. *Given a vector of tapering parameters λ and an observed network y_{obs} with statistics $t(y_{obs})$, at the MLE of θ, $\hat{\theta}_{MLE}$, it holds that $\mu(\hat{\theta}_{MLE}, \lambda) = t(y_{obs})$.*

Proof. Let $Z(\theta, \lambda)$ be the normalizing constant. Then

$$\frac{\partial}{\partial \theta_i} \log(Z(\theta, \lambda)) = \mathrm{E}[t_i(y)] - \sum_k \lambda_k \frac{\partial \mu_k}{\partial \theta_i} \mathrm{E}[\tanh(\mu_k - t_k(y))] = \mu_i(\theta, \lambda)$$

since the second term is zero. The log-likelihood of the LogCosh ERGM and its first derivative are

$$l(\theta|y, \lambda) = \langle \theta, t(y) \rangle - \log(Z(\theta, \lambda)) - \sum_k \lambda_k \log(\cosh(\mu_k - t_k(y)))$$

$$\frac{\partial l}{\partial \theta_i} = t_i(y) - \mu_i(\theta, \lambda) - \sum_k \lambda_k \frac{\partial \mu_k}{\partial \theta_i} \tanh(\mu_k - t_k(y)) \tag{5.6}$$

We find $\hat{\theta}_{MLE}$ by setting $\frac{\partial l}{\partial \theta_i} = 0$ and solving, which yields $\mu(\hat{\theta}_{MLE}, \lambda) = t(y)$. QED

Theorem 16. *For any vector μ of mean parameters, there exists a vector of tapering parameters λ such that the LogCosh ERGM with tapering center μ is non-degenerate.*

Proof. Let $t(y) \equiv t$ to ease notation. We will again use $\tilde{N}(t)$ as defined in equation 4.2 for our smoothing function. It suffices to show that $\tilde{N}(t) \prod_k \mathrm{sech}(\mu_k - t_k)^{\lambda_k}$ is strictly log-concave. Note that although $\mu = \mu(\theta, \lambda)$ is dependent on parameters θ and λ, once those parameters are chosen $\mu(\theta, \lambda)$ is a constant.

Let $r = \log(\tilde{N}(t)) - h(t)$, where $h(t) = \log\left(\prod_k \cosh(\mu_k - t_k)^{\lambda_k}\right) = \sum_k \lambda_k \log(\cosh(\mu_k - t_k))$. Then we have $\frac{\partial h}{\partial t_i} = -\lambda_i \tanh(\mu_i - t_i)$, $\frac{\partial^2 h}{\partial t_i^2} = \lambda_i \mathrm{sech}^2(\mu_i - t_i)$ and $\frac{\partial^2 h}{\partial t_i \partial t_j} = 0$. Evaluating at $\hat{\theta}_{MLE}$ gives $\lambda_i \mathrm{sech}^2(0) = \lambda_i$, and hence $\nabla^2 h$ is a diagonal matrix

$$\nabla^2 h = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix}$$

Let $x = (x_1, ..., x_k)$ be any nonzero column vector. Then $x^T \nabla^2 h x = \sum_i 2\lambda_i x_i^2$. Thus, regardless of $\nabla^2 \log(\tilde{N}(t))$, we can always choose λ large enough such that $x^T \nabla^2 r x < 0$. Thus, r is concave and the LogCosh ERGM is non-degenerate by Results 1 and 2. QED

Theorem 17. *Given a vector of tapering parameters λ, the second derivative of the log-likelihood of the LogCosh ERGM evaluated at $\hat{\theta}_{MLE}$ is*

$$\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\Big|_{\hat{\theta}_{MLE}} = -\frac{\partial\mu_i(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_j} - \sum_k \lambda_k \frac{\partial\mu_k(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_i} \frac{\partial\mu_k(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_j}$$

Proof. Let $\mu_i(\theta, \lambda) = \mathrm{E}[t_i(Y)]$. To ease notation, let $t_k(Y) \equiv t_k$ and $\mu_k(\theta, \lambda) \equiv \mu_k$. Then we have

$$\frac{\partial\mu_i(\theta, \lambda)}{\partial\theta_j} = \mathrm{Cov}[t_i, t_j] - \sum_k \lambda_k \frac{\partial\mu_k}{\partial\theta_j} \mathrm{Cov}[t_i, \tanh(\mu_k - t_k)]$$

Collecting all the partial derivatives on the left side, we have

$$\frac{\partial\mu_i(\theta, \lambda)}{\partial\theta_j} + \sum_k \lambda_k \frac{\partial\mu_k}{\partial\theta_j} \mathrm{Cov}[t_i, \tanh(\mu_k - t_k)] = \mathrm{Cov}[t_i, t_j]$$

Which can be written as a system of linear equations

$$(I + B)\frac{\partial\mu(\theta, \lambda)}{\partial\theta_j} = c^j$$

where we define matrix B with $B_{ij} = \lambda_j \mathrm{Cov}[t_i, \tanh(\mu_j - t_j)]$ and vector c^j with $c_i^j = \mathrm{Cov}[t_i, t_j]$. Thus, the vector of the derivatives of each mean parameter with respect to θ_j is

$$\frac{\partial\mu(\theta, \lambda)}{\partial\theta_j} = (I + B)^{-1} c^j$$

The fist derivative of the log-likelihood is given by equation 5.6. The second derivative of the log-likelihood is

$$\frac{\partial^2 l}{\partial\theta_i \partial\theta_j} = -\frac{\partial\mu_i}{\partial\theta_j} - \sum_k \left[\lambda_k \tanh(\mu_k - t_k)\frac{\partial^2\mu_k}{\partial\theta_i\partial\theta_j} + \lambda_k \mathrm{sech}(\mu_k - t_k)\frac{\partial\mu_k}{\partial\theta_i}\frac{\partial\mu_k}{\partial\theta_j}\right]$$

which, simplified at the MLE, reduces to

$$\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\Big|_{\hat{\theta}_{MLE}} = -\frac{\partial\mu_i(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_j} - \sum_k \lambda_k \frac{\partial\mu_k(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_i} \frac{\partial\mu_k(\hat{\theta}_{MLE}, \lambda)}{\partial\theta_j}$$

QED

5.4.3 Choosing the Tapering Parameter

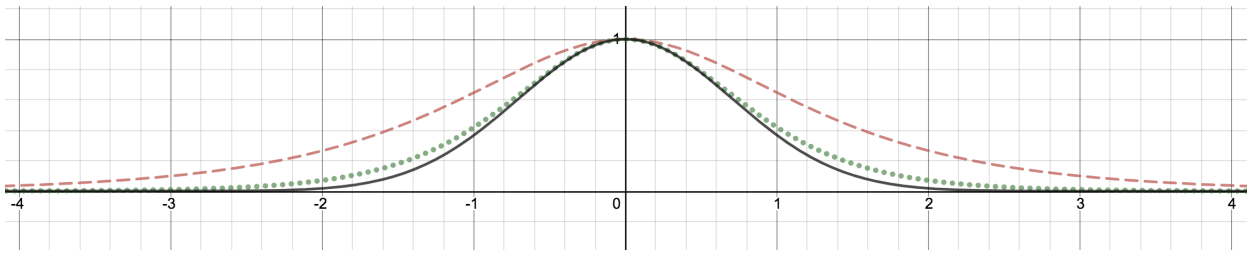

Figure 5.7: Contrasting levels of tapering. $\text{sech}(x)$ is the dashed red line, $\text{sech}^2(x)$ is the dotted green line, and e^{-x^2} is the solid black line.

We have seen that the LogCosh ERGM can always be made non-degenerate with sufficient tapering, which is controlled by the vector of tapering parameters λ. This section addresses how to choose λ. The methodology behind choosing λ is nearly identical to that of choosing τ for the Tapered ERGM. Indeed, we should follow algorithm 6, reproduced below, with only a few minor changes that we discuss afterwards.

Algorithm 18. *Setting the Tapering Parameter*

1. *Choose only the dyad-dependent terms to taper.*
2. *Set a large value of λ_k in order to heavily taper each of the k terms.*
3. *If the MCMC estimation converges, proceed to the next step. If the model is still unable to be fit, go back to step 1 and taper all terms.*
4. *Relax the amount of tapering by decreasing each λ_k until the bimodality coefficient for each of the k statistics is no greater than 0.4.*

In most cases, we need only taper the dependent terms because doing so indirectly tapers the other terms as well. It is best to taper as few terms as necessary since tapering does effect the overall interpretation of the parameters, which can be seen by considering the log-odds of a tie between nodes. Let $P(Y_{ij} = 1|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^+)$ and $P(Y_{ij} = 0|Y_{ij}^c = y_{ij}^c) \equiv P(Y_{ij}^-)$.

Then, under the LogCosh ERGM the log-odds of a tie is

$$\log\left(\frac{P(Y_{ij}^{+})}{P(Y_{ij}^{-})}\right) = \log\left(\frac{\exp\left(\sum\theta_k t_k(Y_{ij}^{+})\prod_k \operatorname{sech}(\mu_k - t_k(Y_{ij}^{+}))^{\lambda_k}\right)}{\exp\left(\sum\theta_k t_k(Y_{ij}^{-})\prod_k \operatorname{sech}(\mu_k - t_k(Y_{ij}^{-}))^{\lambda_k}\right)}\right)$$
$$= \sum\theta_k\Delta t_k(Y_{ij}) + \sum\lambda_k \log\left(\frac{\cosh(\mu_k - t_k(Y_{ij}^{-}))}{\cosh(\mu_k - t_k(Y_{ij}^{+}))}\right)$$

where $\Delta t_k(Y_{ij}) = t_k(Y_{ij}^{+}) - t_k(Y_{ij}^{-})$ is the change statistic. Hence, the interpretation of the LogCosh ERGM is that the conditional log-odds of a tie is the sum of the $\theta_k \times$ (change in kth statistic) plus a penalty, where the penalty is determined by λ. We should make λ as small as possible to get as close as we can to the standard ERGM interpretation where θ_k is the conditional log-odds of a tie.

Step two of the algorithm instructs us to assign a large value of λ_k to heavily taper each term, allowing the MCMC estimation to converge faster. Knowing what a "large" value of λ_k is is the main difference between setting λ of the LogCosh ERGM and τ of the Tapered ERGM. The same logic used with the Tapered ERGM also applies here: assume Poisson dispersion and set λ_k so that values more than r standard deviations away from μ_k are tapered heavily. By Theorem 14 we know that for sizable deviations from μ, $\log(\cosh(\mu - t(Y)))$ behaves like $|\mu - t(Y)| - \log 2$. Thus, we should set $\lambda_k = \frac{1}{r\sqrt{\mu_k - log2}}$. Note that r is a tuning parameter that ultimately controls whether or not the distribution of $t_k(Y)$ is under- or over-dispersed relative to the Poisson. A large value of λ_k may take $r < 4$, using the same logic that was applied to the MAD ERGM. Figure 5.7 clearly shows that as λ gets larger, the tapering becomes more severe, but notice that the $\operatorname{sech}(x)$ will always have fatter tails than the Gaussian. The rest of algorithm 18 is straight-forward.

5.5 Comparison of Restorative Force Models

Each Restorative Force Model comes with its own set of advantages and disadvantages. Here we compare the pros and cons of each in hopes of identifying under what circumstances which may be ideal for the researcher.

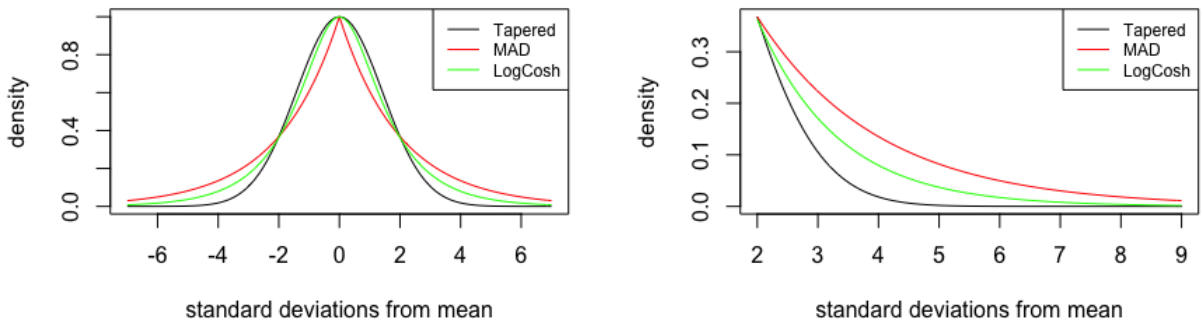

Figure 5.8: Comparison of tapering mechanisms. *LEFT*: The curves follow the default tapering recommendations such that they all down-weight statistics two standard deviations away from the mean equally, hence the two curves intersect at ± 2. *RIGHT*: After 2 standard deviations, the MAD and LogCosh ERGMs down-weight much less severely than the Tapered ERGM.

We begin with the Tapered ERGM. Like the MAD and LogCosh ERGMs, we can always guarantee that the Tapered model will be non-degenerate. The advantage the Tapered ERGM has above all others is that it is the easiest to fit. Finding the tapering parameters τ such that the model converges is relatively easy, whereas in the other RFMs the heavy tails can make this task more difficult. The flip side of this is that the Tapered model tapers the most heavily of all the RFMs, and indeed this is the model's biggest disadvantage. If we desire a wider range of networks differing from the observed network when simulating from our RFM, it is worth considering other RFMs.

Following that idea of easing the tapering to increase the variance in network statistics, the MAD ERGM offers the second least amount of tapering of the RFMs. This is of course due to the fact that the tapering is accomplished through a penalty on the absolute deviation instead of the squared deviation. However, this comes with the cost of having the MLE occur at a non-differentiable point which causes the Hessian to not exist. The standard errors can then only be calculated via bootstrap methods, which is the main disadvantage of the MAD

ERGM. Because of the heavy tails, the MAD ERGM can also sometimes be difficult to fit.

The LogCosh ERGM is a nice blend between the two aforementioned RFMs. Because the log-cosh penalty behaves like the squared deviation for small differences, but like the absolute deviation for large differences, we get the light tapering of the MAD ERGM with the stability of the Tapered ERGM. The standard errors can be computed in the typical manner using the Hessian as well. The LogCosh ERGM can be more difficult to fit than the Tapered ERGM due to its heavy tails, and this is the model's only disadvantage.

Finally, the Stereo ERGM offers the least amount of tapering amongst all of the RFMs. It also has the advantage of having only one tapering parameter to set, the radius R, as opposed to a vector of tapering parameters τ. However, it has a potentially crippling disadvantage that we cannot always guarantee the model will be non-degenerate. Because the Stereo ERGM has the heaviest tails out of all the RFMs, it is probably best to use the Stereo model only when dealing with small networks that require very little tapering.

Figure 5.8 shows the tapering effects from the Tapered, MAD, and LogCosh ERGMs (the Stereo is not included because that tapering mechanism is set in a different manner which makes for an uneven comparison). We can clearly see how the difference in the tails will allow for greater variation in simulated networks. The choice of which Restorative Force Model to use is really up to the researcher and context dependent, but the advantages and disadvantages laid out here should be taken into consideration when choosing.

CHAPTER 6

Case Studies

In this chapter we fit Restorative Force Models, including and especially the Tapered ERGM, to several real-world networks, each time noting the tapering methodology. These networks are a variety of sizes and come from various domains. We demonstrate the effectiveness of the RFMs in all scenarios and provide examples of how to get around any potential issues.

6.1 Faux Desert High Network

Derived from a national study on adolescent health (Resnick et al., 1997), the Faux Desert High Network is a simulated social network of middle and high school students. This is a medium-sized network comprised of 107 students, with 439 directed edges between them representing friendship nominations. We have information on the grade (7 through 12), sex, and race (with the vast majority identifying as White, but also including Black, Hispanic, Asian, and Other) of each student.

Additionally, we note there are 677 triangles in the network. We would like to know if these 3-cycles are a product of homophily ("birds of a feather flock together"), transitivity ("a friend of my friend is also my friend"), or some combination thereof. Typically, we cannot fit an ERGM with a triangle term, as the term nearly always induces degeneracy, and we are forced to use less than satisfying alternatives. However, this is an exceptional case, and we actually can fit such a model for this network using only a standard (untapered) ERGM. This gives a unique opportunity for a direct comparison between the ERGM and Tapered ERGM and for the effects of tapering to be explicitly measured. The ERGM can be

fit using relatively few terms, which are summarized in Table 6.1. We see that the triangle term is essentially zero, and there are strong effects of matching on grade at every level. In other words, under this model homophily on grade level is almost solely responsible for the observed clustering. This is unsurprising given most activities and classes within a school are segregated by grade. Figure 6.1 displays some graphical goodness-of-fit diagnostics showing that the model is indeed a good fit.

How might this fit change if instead we used a Tapered ERGM? We can consider two different scenarios here. First, consider the exact same model as the ERGM, but we instead decide to taper the dependent terms (as recommended by algorithm 6), which in this model are the triangles, isolates, and the zero edgewise shared partner (esp(0)) terms. The heavier the tapering, the smaller the standard deviation of the counts of each term. The left panel of Figure 6.2 shows what happens across different levels of tapering. On the far right of this plot are the ERGM parameter estimates. As we move left along the x-axis, the tapering increases and the standard deviation of the triangle count decreases (as do the standard deviations of the other terms, though not as much). We see that not only do the parameter estimates themselves remain basically unchanged, so too do their standard errors. Only under severe tapering (far left of the plot) do the standard errors grow significantly larger.

The second scenario to consider is a very practical one. Imagine we do not have any nodal attributes in our data. As such, we cannot match on grade level in our model. We would still like to fit a triangle term, but alas, without the nodal attributes the triangle term forces the ERGM to be degenerate and MCMC estimation fails. This is where the Tapered ERGM flexes its power. If we taper the dependent terms (triangles, isolates, and esp(0)), we can fit the model without problem. Moreover, we can also choose to taper only the triangle term and the results are nearly identical. The right panel of Figure 6.2 shows the parameter estimates and standard errors of the Tapered ERGM without nodal covariates. What is remarkable is how close these estimates are to that of the standard ERGM which did incorporate the nodal attributes. The Tapered ERGM not only allowed us to fit an

otherwise degenerate model, the results are very similar to that of the ERGM using more information. Note that the triangle term is statistically significant in this model, but the parameter estimate is still very close to zero. The key point to take away here is that the level of tapering essentially does not effect parameter estimates; in fact, tapering even gives reasonable estimates in models heretofore impossible to fit.

Tapering is always done relative to each term, specifically relative to each term's corresponding mean parameter. For example, we can control the level of tapering on the triangle term through $\tau_{tri} = 1/(r * \mu_{tri})$, where r is a user specified multiplier and μ_{tri} is the mean value parameter for triangles. Figure 4.1 shows what happens to the term counts as we vary r. The relation above shows that r is inversely proportional to the amount of tapering, τ; small values of r lead to heavy tapering (leftward) and tapering decreases as r increases (rightward). Because the Tapered ERGM centers tapering on the mean parameters, the mean parameters all lie near the observed values (dashed lines in the plot). As we move left, tapering increases and eventually the variance constraints for each term all become active. Certain terms like the triangle count exhibit tapering at nearly all levels of r (as expected since degeneracy often causes the triangle count to explode), whereas other terms like the number of isolates do not show the effects of tapering until large values of τ. It is worth noting that the edge count was not tapered in this model, yet it exhibits tapering because all of the dependent terms - triangles, isolates, esp(0) - were tapered. Because the mean parameters are consistent across levels of tapering, we should strive for as little tapering as necessary.

6.2 Last.fm Friendship Network

Last.fm is an online music service that allows users to create a community of "friends" in addition to streaming music. Users of the website fill out a profile and are encouraged to "find like-minded music fans" (Last.fm, 2020a), with each user's listening taste and history entirely recorded and on display. The site allows users to designate each other as 'friends'

Table 6.1: Summary of ERGM fit on Faux Desert High Network

Term	$\hat{\theta}$	$\hat{se}(\hat{\theta})$
edges	-3.48	0.10
triangles	-0.0069	0.038
isolates	1.16	0.47
esp(0)	-1.35	0.13
match.grade.7	2.22	0.24
match.grade.8	2.06	0.17
match.grade.9	1.99	0.16
match.grade.10	1.56	0.11
match.grade.11	1.77	0.14
match.grade.12	1.28	0.28

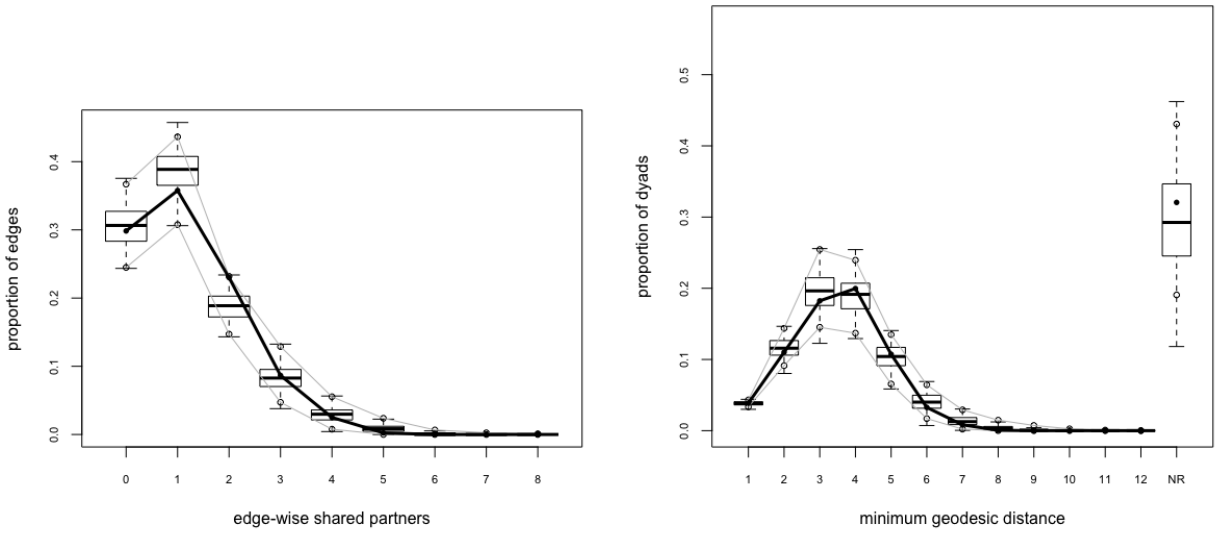

Figure 6.1: The edgewise shared partners distribution and minimum geodesic distance from 100 networks simulated from the Tapered ERGM MLE compared to the observed network statistics (thick black line), where the Tapered ERGM was fit to the Faux Desert High Network.

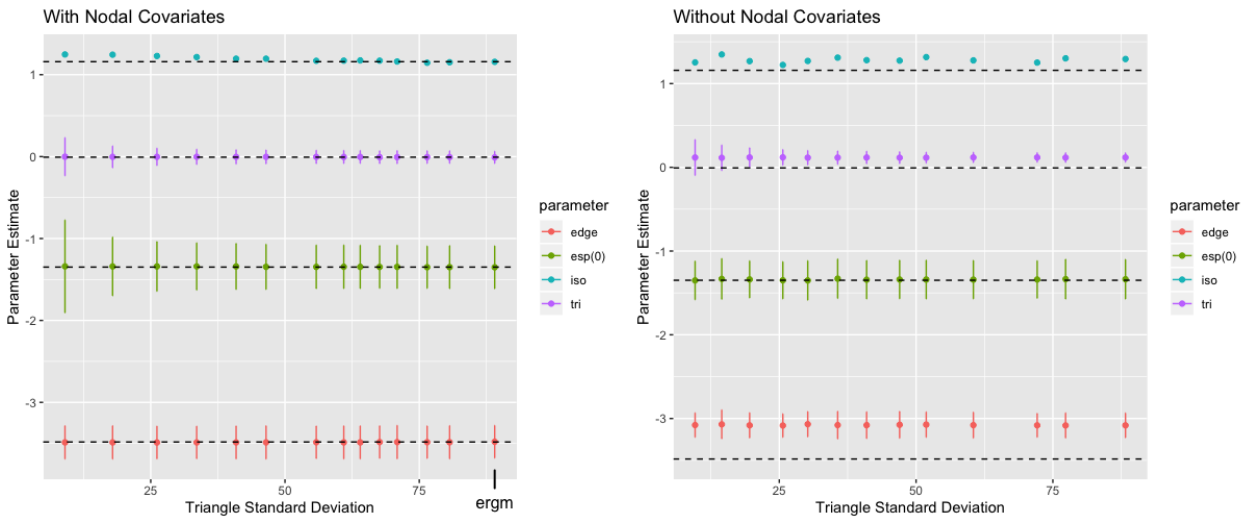

Figure 6.2: Similarity of parameter estimates across levels of tapering in the Faux Desert High Network. *LEFT*: Tapered models in which the nodal attribute 'grade' is included. The points on the far right of the plot are the estimates from the standard (untapered) ERGM, and the dashed line is set at those numerical values. We see that regardless of how much tapering we apply, the parameter estimates are spot on and the standard errors are comparable to that of the standard ERGM. *RIGHT*: Tapered models in which no nodal attributes are included. A standard ERGM with a triangle term cannot be fit in this case, but the parameter estimates from the standard ERGM which does include the 'grade' attribute are plotted as the dashed line for reference (exactly as in the left panel). We see that even without the nodal attributes, the Tapered ERGM is able to fit a triangle model and still arrive at stable estimates very similar to that of the ERGM including nodal attributes. Once again, the standard errors are comparable to that of the untapered ERGM. In both the left and right panel, the error bars have been omitted from the isolates term because the low number of isolates in the network lead to large standard errors which otherwise distort the graph.

and has various platforms for users to interact (discussion boards, private message, etc). Last.fm claims to have over 60 million users across the globe (Last.fm, 2020b).

This data set was collected by Toivonen et al. (2009) and used in their comparative study of social network models. Because Last.fm has an abundance of users, they chose only the largest connected component of users who self-identify Finland as their country. As such, the network is still a very large one with 8,003 nodes. A tie is counted only if both users acknowledge each other as friends, leaving an undirected network with 16,824 ties. This network contains only mutual friendship structure and does not include any nodal covariates or information on musical preference. While the authors do include ERGMs in their comparative study, they do not fit an ERGM to this data set because "generating networks of size 8,000 and fitting their parameters did not seem feasible for the ERG model" (Toivonen et al., 2009).

Whereas a standard ERGM may struggle with a network this size, the Tapered ERGM is able to fit the data with relative ease. However, before we state the results, it is worth mentioning the farcical objective here. To fit a network as large and complex as this one based on structure alone is to completely ignore the social processes at work. Experience tells us that many factors likely play a part in friendship formation, whether it be digitally or otherwise. Age, geography, gender, kinship, etc. all play a role in musical preference as well, which is assumed to drive the ties. We would not expect the musical taste of an adult male in a rural setting to match that of a young woman in an urban environment. So to make the assumption of homogeneity amongst these 8003 users, even if they all come from the same country, is completely absurd. At the very least, it would be good to know each individual's most listened to genre of music. Without any further information, we can know what to expect before we even fit the model: a fit so tight around the observed values that inference may become impossible.

Nonetheless, we include the Tapered ERGM fit here if only to show that following algorithm 6 works even for sufficiently large networks. Using said algorithm, we begin by

Table 6.2: Summary of Tapered ERGM fit on Last.fm Network

Term	$\hat{\theta}$	$\hat{se}(\hat{\theta})$	obs. count	Var(count)	τ (using $r = .06$)	bimod. coeff
edges	-4.71	0.1856	16,824	28.73	0.0165	0.332
triangles	0.26	0.2397	10,083	17.63	0.0275	0.314
gwdegree(decay=.25, fixed=TRUE)	1.60	0.2492	9,601.86	16.21	0.0289	0.328
esp(0)	-4.31	0.3309	5242	9.32	0.0530	0.336
esp(1)	-1.81	0.3575	4302	7.87	0.0646	0.316

attempting to taper only the dependent terms. However, no matter how severely we taper those terms, it is still not enough to get the model to converge. Thus when we arrive at step 3, we are forced to go back to step 1 and taper all of the terms. Going forward we find that we must set very large values of τ corresponding to $r = .02$, 100 times the default level of tapering. Recall that in the Tapered ERGM, $\tau_k = \frac{1}{r^2 \mu_k}$. Using these values, the model converges fairly quickly and we can move on to step 4 in which we begin to relax the tapering. We are only able to decrease the amount of tapering to levels where $r = .06$ and still achieve convergence, which is still a significant amount of tapering. At this level of tapering all of the marginal distributions of terms have a bimodality coefficient under 0.4 as desired. Table 6.2 gives the results of the model, and figures 6.3 and 6.4 show some goodness-of-fit plots along with the marginal distributions of the edges and triangle terms. The fit is a very tight one as we expected; this is not a failure of the tapering process, but rather a result of a lack of any additional information.

What do these results tell us? At first glance, table 6.2 would indicate that the triangle term is not significant. However, there is more going on here. The presence of the triangle term in the model greatly enhances the overall fit and in particular controls the amount of clustering, as shown by the second panel of figure 6.4. To see this, we can fit the same model without the triangle term and look at the marginal distribution of triangles. Figure 6.5 shows the marginal distribution of triangles from 1000 simulated networks sampled from the same model described in table 6.2, with the same level of tapering ($r = .06$), but with the triangle term removed. We can see that the triangle count is consistently 3000+ below

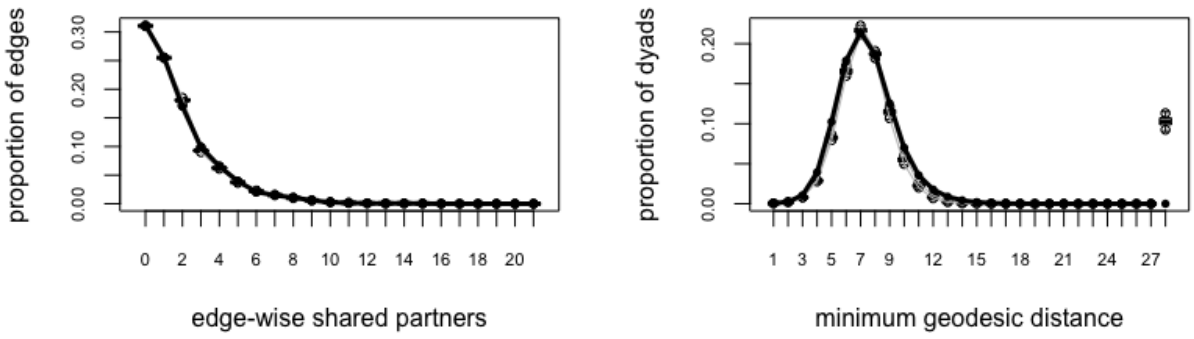

Figure 6.3: The edgewise shared partners distribution and minimum geodesic distance from 100 networks simulated from the Tapered ERGM MLE compared to the observed network statistics (thick black line), where the Tapered ERGM was fit to the Last.fm friendship network. When modeling only structure, we are forced to taper very heavily to get convergence, resulting in a very tight fit around the observed values.

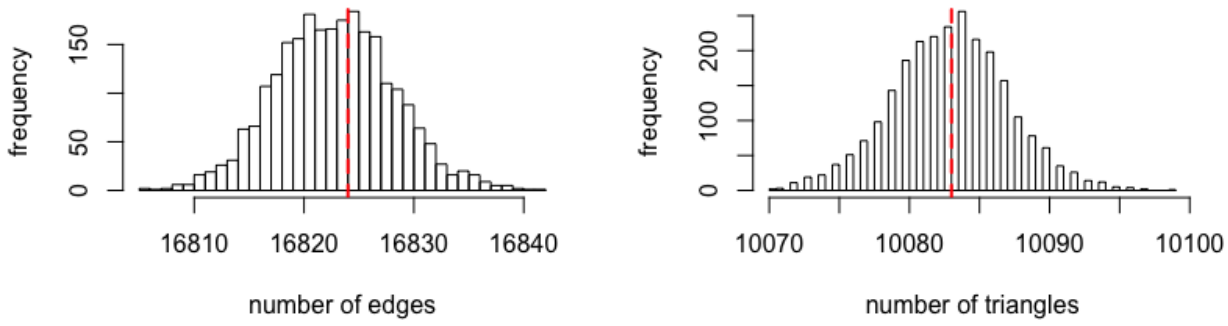

Figure 6.4: The marginal distributions of edges (left) and triangles (right) sampled from 2400 networks simulated from the Tapered ERGM MLE fit to the Last.fm network. Notice the scale on the x-axis indicating how tight the fit is around the observed values (dashed red line).

the observed number of triangles. Thus, we should believe that there is transitivity within the Last.fm network despite the non-significance of the triangle term, which is due to the homogeneity assumption we are forced to make.

There is yet another way to verify that transitivity is present within the network. We can instead fit a model using a gwesp term, an alternative to the triangle count which is also a measure of transitivity. What the gwesp term loses in interpretability it makes up for in numerical stability, and as such we can fit a model using said term by only tapering twice as much as the default, using a modest level of tapering at $r = 1$. Table 6.3 shows the summary of the model fit where the gwesp term is highly significant, in line with our conclusion above. Figure 6.6 shows some goodness-of-fit diagnostics for this model. Without having to taper so severely, the fit is not so overly tight as the previous one, albeit perhaps a worse fit overall. Note that this model also underestimates the triangle count and hence the amount of clustering, which further illustrates the importance of being able to include a

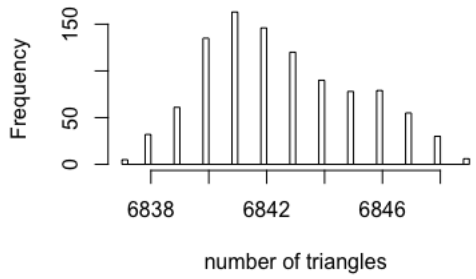

Figure 6.5: The marginal distribution of triangles from 1000 simulated networks sampled from the same model outlined in table 6.2, but with the triangle term removed. The observed number of triangles in the Last.fm network is 10,083. Removing the triangle term causes the model to severely underestimate the amount of clustering.

Table 6.3: Summary of Alternative Tapered ERGM fit on Last.fm Network

Term	$\hat{\theta}$	$\hat{se}(\hat{\theta})$	obs. count	Var(count)	τ (using $r = 1$)	bimod. coeff
edges	-9.25	0.0217	16,824	4426.05	0.000059	0.353
gwdegree(decay=.25, fixed=TRUE)	1.04	0.0522	9,601.86	363.35	0.00010	0.343
gwesp(decay=.25, fixed=TRUE)	3.35	0.0209	13,443.30	4029.10	0.000074	0.361

triangle term.

This case study demonstrates that RFMs are powerful tools that can be used to make inferences on very large networks even in suboptimal conditions (when no nodal information is present).

6.3 London Gang Network

The data for this network was gathered by two sociologists investigating the role of ethnicity within a London street gang (Grund and Densley, 2012). The gang was believed to have formed in 2005 and mainly operates in a low-income housing area of inner-city London. Using police arrest and conviction data, as well as fieldwork that involved interviewing some

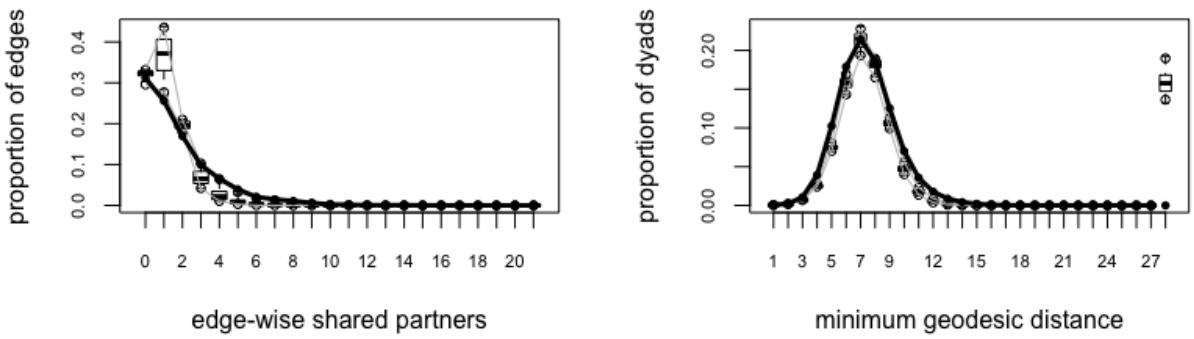

Figure 6.6: The edgewise shared partners distribution and minimum geodesic distance from 100 networks simulated from the Tapered ERGM MLE compared to the observed network statistics (thick black line), where the Tapered ERGM was fit to the Last.fm friendship network using a gwesp term instead of a triangle term. Only a modest level of tapering ($r = 1$) was needed to fit this model.

of the gang members, the authors of the study focus on 54 "confirmed" members of the gang who were known to be affiliated between 2006 and 2009. The data set contains a number covariates including the birthplace, age, number of arrests, number of convictions, incarcerations, and rankings of each gang member. A tie exists between two gang members if they *co-offended* (were arrested together for committing a crime) at least once. The network consists of 133 undirected ties. Figure 6.7 shows that there are six isolates within the network, though the authors later removed them and analyzed only the largest connected component using standard ERGMs (Grund and Densley, 2015). Though somewhat of a common practice, removing isolates is rarely justified and distorts the social processes at work in forming the network. Therefore, in the forthcoming treatment we analyze the network both ways, with and without the isolates.

Although every member of the gang would be racially defined as Black, they do not all share the same ethnicity. Grund and Densley (2015) use place of birth and national heritage

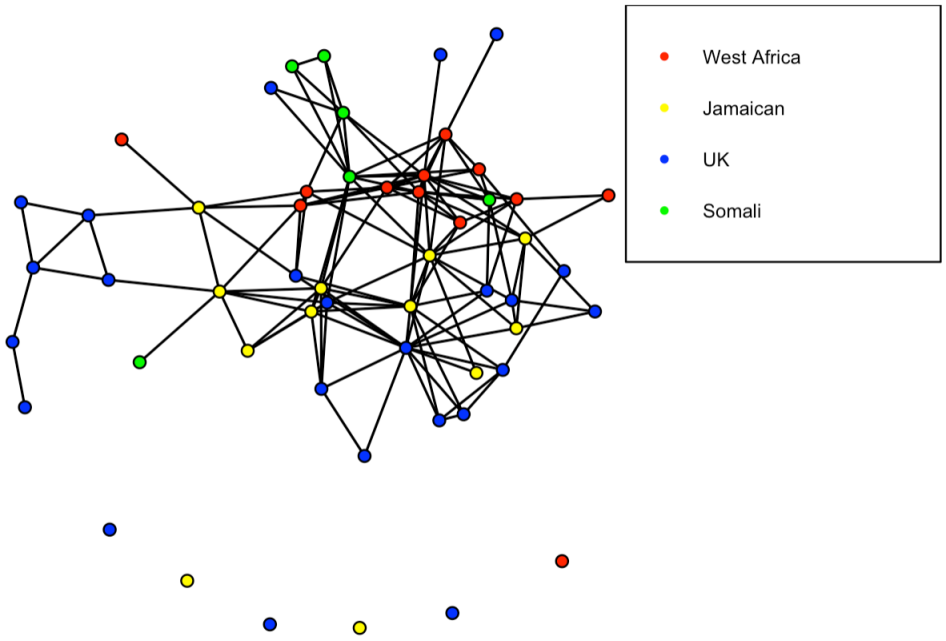

Figure 6.7: The London Gang Network. A tie exists between two gang members if they have committed at least one crime together. All gang members are Black but the gang is comprised of four distinct ethnicities, categorized by the authors as their countries of origin.

to serve as "a proxy measure for ethnic background." The authors are quick to admit that two individuals from the same region may not identify as the same ethnicity with regard to culture, language, etc., but their "fieldwork with the gang confirms the validity of this categorization." As such, they identify four distinct ethnic identities within the gang: (1) Somali (n = 6), (2) West African (Congo, Ghana, Ivory Coast, Nigeria, and Sierra Leone, n = 12, including two siblings), (3) Jamaican (n = 12), and (4) British (n = 24).

Grund and Densley (2015) posit that who co-offends with whom is driven by ethnic homophily, triad-closure, and potentially an interaction between the two. Specifically, they hypothesize that "gang members are even more likely to offend with each other when they have the same ethnic background AND share another co-offender from the same ethnic background" (Grund and Densley, 2015). To disentangle these effects, the authors fit ERGMs

to the data. Clearly, the most important term for these purposes would be the triangle, which can also be indexed by ethnic attribute. That is, including a separate triangle term for each of the four ethnicities, along with matching on ethnicity to measure homophily, would provide a conclusion to their hypothesis. Unfortunately, the authors note that counting triangles elicits degeneracy and cannot include such terms. As a workaround to measuring the effects of triad closure, they include a gwesp term and a customized gwesp term which only counts edgewise shared partners matching on the same ethnicity. With these and ethnic matching terms all significant, the authors conclude that their hypothesis is correct.

With Restorative Force Models, we do have the ability to measure the effect of triad closure directly by fitting triangle terms and our model provides clear answers to the questions of the researchers. ERGMs have the functionality to model triangles based on specific attributes, in this case ethnicity, but typically this presents the problem of degeneracy during maximum likelihood estimation of parameters. With RFMs this isn't so, and we can easily fit such terms. With the same objective of disentangling the effects of ethnic homophily and triad closure on who co-offends with whom, as well as any interaction, we fit a separate triangle term for each ethnicity as well as a matching term for each ethnicity. Because triangles can also be ethnically heterogeneous, we also fit a general triangle term to account for the effect of triad closure where gang members do not all share the same ethnicity. Looking at the data we see that for one particular ethnic group, Somalis, any homogenous ties also occur within homogenous triads. Thus, we cannot include both a Somali triangle term and a Somali matching term together in the model because it is not possible to estimate both simultaneously. We therefore make the decision to include the Somali triangle term but remove the Somali matching term. Table 6.4 shows the results of two Tapered ERGMs. Model 1 was fit to the largest connected component of the gang network as Grund and Densley (2015) did; Model 2 was fit to the entire network including the six isolates and hence also has an isolates term. The results of both models are expectedly very similar to each other. Models 1 and 2 were both fit with the least amount of tapering possible in order to achieve convergence.

Table 6.4: Summary of Tapered ERGMs fit on London Gang Network

Term	Model 1	Model 2
edges	-3.24 (0.152)***	-3.31 (0.157)***
triangles	0.68 (0.127)***	0.70 (0.128)***
triangles(West Africa)	0.12 (0.444)	0.16 (0.435)
triangles(Jamaican)	0.19 (0.710)	0.53 (0.662)
triangles(UK)	0.66 (0.507)	0.69 (0.492)
triangles(Somali)	2.41 (0.394)***	2.58 (0.368)***
match(West Africa)	1.01 (0.557)	0.89 (0.500)
match(Jamaican)	1.38 (0.608)*	0.93 (0.505)
match(UK)	0.26 (0.348)	0.24 (0.333)
isolates		0.82 (0.533)

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Unsurprisingly, the Somali triangle term is highly significant (as would be a Somali matching term had it been included instead of the Somali triangle term) since that ethnicity tends to cluster tightly together with regard to co-offending. What is surprising, however, is that the general triangle term is also highly significant while *nothing else is* (save the edge term). This tells us that outside of the Somali gang members, the most important thing driving who co-offends with whom is whether or not doing so would close a triad, regardless of the ethnicities of those in the triad. Neither ethnic homophily nor homogenous triad closure are significant for any ethnicity other than the Somalis (notwithstanding the slightly significant Jamaican matching term in Model 1). This leads us to conclusions almost entirely opposite of those made by Grund and Densley (2015): for this particular gang, gang members are more likely to offend with each other if doing so would close a triad; they are not more likely to offend with each other when they have the same ethnic background or if they share another co-offender from the same ethnic background (excepting Somali gang members).

Figures 6.8 and 6.9 show that both Model 1 and 2 are superior fits to the data, especially with regard to the edgewise shared partner distribution, further showing the importance of

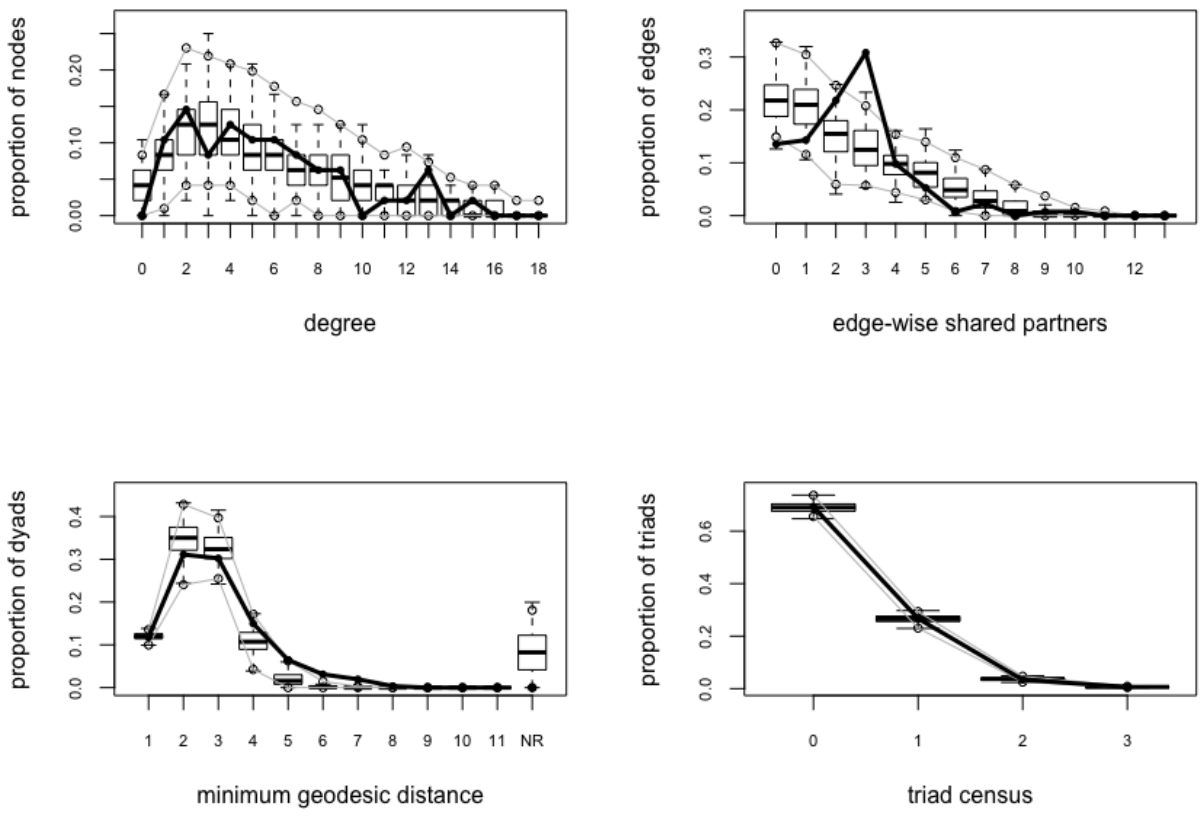

Figure 6.8: Goodness-of-fit diagnostic plots for the Tapered ERGM fit on the largest connected component of the London gang network (Model 1 in table 6.4).

the general triangle term. The excellent fit of Model 2 to the degree distribution underscores the wisdom of not removing the isolates from a network when modeling. It is worth noting that other models were fit including the other covariates (number of arrests, number of convictions, prison, age, ranking), but none improved the overall fit and none were significant. This example clearly demonstrates the fundamental need for RFMs, since without the ability to fit fundamental terms like the triangle it is very possible to make incorrect inferences.

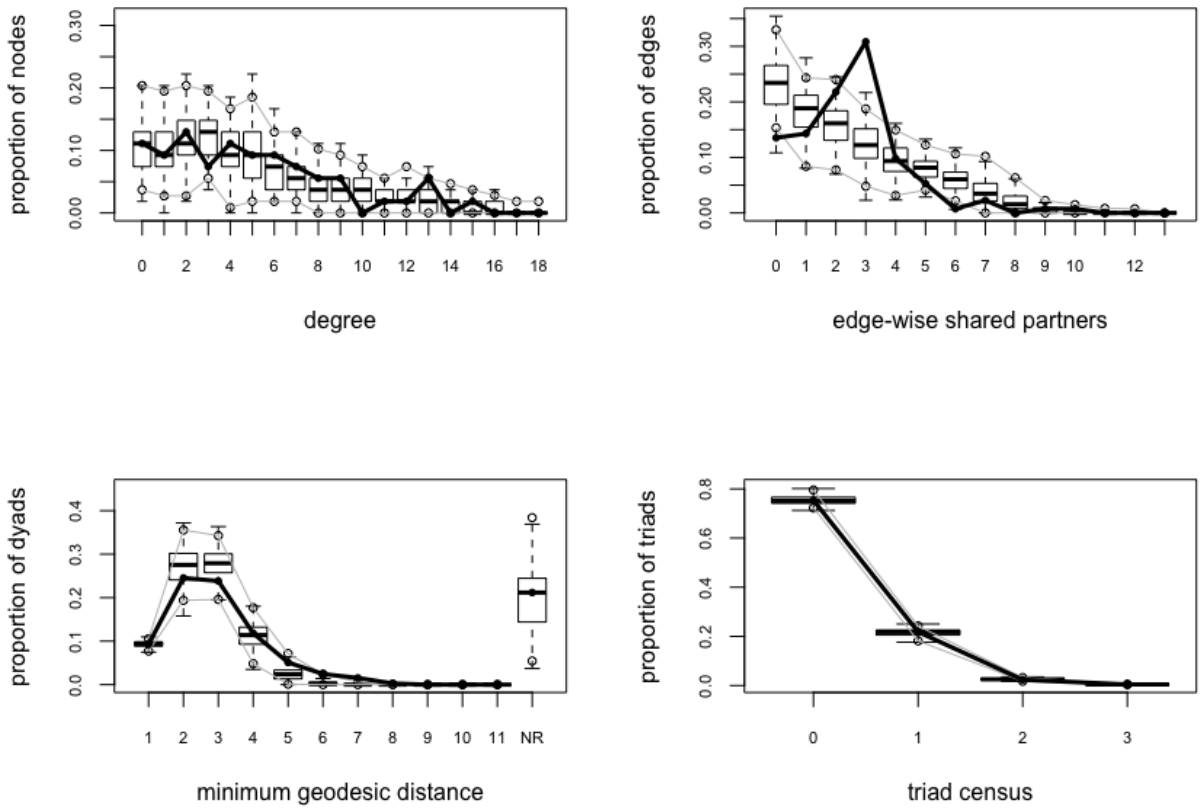

Figure 6.9: Goodness-of-fit diagnostic plots for the Tapered ERGM fit on the London gang network (Model 2 in table 6.4).

6.3.1 Using other Restorative Force Models

Heretofore, all analysis had been carried out using the Tapered ERGM. In this section we apply other RFMs, namely the MAD ERGM and the LogCosh ERGM, to the full London Gang Network and compare the results. The Stereo ERGM unfortunately could not be fit to this data set.

During the process of fitting the MAD and LogCosh ERGMs, the tapering coefficients τ required for the models to achieve convergence were much larger than expected. In every model only the dyad-dependent terms were tapered, namely triangles and the four ethnically homogenous triangles. Table 6.5 shows that each element of τ for the MAD and LogCosh ERGM was about an order of magnitude larger than its tapering counterpart of the Tapered ERGM (the same τ was used for both the MAD and LogCosh ERGM). Decreasing τ any further resulted in the models not converging. The heavy tails of the MAD and LogCosh ERGMs are what call for the larger tapering coefficients. The τ for the Tapered ERGM was also relaxed as much as possible until convergence was no longer possible.

Despite the larger tapering coefficients for the MAD and LogCosh ERGMs, the variance of the triangle term in those models is still much larger than the variance of the triangle term in the Tapered ERGM. Table 6.6 shows the variance of each tapered term for each RFM. The MAD and LogCosh models allow for greater variation in the general triangle count, as shown by Figure 6.10. Note that the apparent trade-off between increasing the general triangle variance and lowering the homogenous triangle variance, which makes sense if our simulates are to remain representative of the observed network.

The fact that the τ vectors of the MAD and LogCosh ERGMs are so much larger than that of the Tapered ERGM tells us two things. Firstly, the ability of the MAD and LogCosh models to 'reach farther' with their heavy tells means they are more susceptible to degeneracy and hence larger tapering coefficients must be employed. Secondly, for the Tapered ERGM we can see that setting a small τ vector perhaps does not imply as small a penalty as we might expect. Clearly, what's happening is that the squared deviation of the tapering penalty is

Table 6.5: Summary of RFM tapering coefficients fit on London Gang Network

Term	Tapered ERGM τ	MAD ERGM τ	LogCosh ERGM τ
triangles	0.0081	.17	.17
triangles(West Africa)	0.06	1.2	1.2
triangles(Jamaican)	0.15	1.8	1.8
triangles(UK)	0.1125	1.8	1.8
triangles(Somali)	0.225	1.8	1.8

Table 6.6: Variance of tapered terms across RFM fits on London Gang Network

Term	Tapered ERGM Variance	MAD ERGM Variance	LogCosh ERGM Variance
triangles	68.92	131.02	151.86
triangles(West Africa)	9.05	1.37	1.93
triangles(Jamaican)	4.73	0.55	1.10
triangles(UK)	6.05	0.50	1.04
triangles(Somali)	10.22	0.41	5.01

effectively more influential on the variance than the tapering coefficients themselves.

Table 6.7 shows the results of the Tapered, MAD, and LogCosh ERGM fits. We can see that the parameter estimates are extremely similar across every term. The standard errors are also very similar to one another, with the exception of two terms within the MAD ERGM. The MAD ERGM differs from the Tapered and LogCosh models in that it finds the homogoenous British triangle - not the homogenous Somali triangle - significant. This discrepancy could be due to the fact that a closed form for the standard errors of the MAD ERGM does not exist, and the standard errors had to be estimated using the bootstrap method. All of the RFMs find the triangle term highly significant ($p < .001$), and the conclusion is overwhelmingly clear: for this London gang, gang members are more likely to offend with each other if doing so would close a triad, regardless of the ethnicities of the gang members comprising the triad.

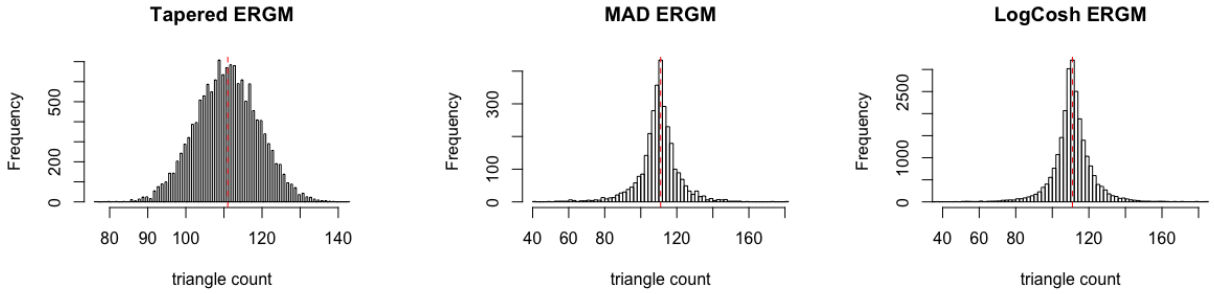

Figure 6.10: Triangle counts across different Restorative Force Models fit on the London Gang Network. The shape of the distributions reflect the tapering penalty used: squared deviation (Tapered), mean absolute deviation (MAD), and log-cosh deviation (LogCosh).

Table 6.7: Summary of Restorative Force Models fit on London Gang Network

Term	Tapered ERGM	MAD ERGM	LogCosh ERGM
edges	-3.31 (0.157)***	-3.35 (0.133)***	-3.33 (0.157)***
triangles	0.70 (0.128)***	0.72 (0.038)***	0.71 (0.096)***
triangles(West Africa)	0.16 (0.435)	0.13 (0.252)	0.16 (0.765)
triangles(Jamaican)	0.53 (0.662)	0.55 (0.478)	0.56 (1.052)
triangles(UK)	0.69 (0.492)	0.67 (0.250)**	0.67 (1.021)
triangles(Somali)	2.58 (0.368)***	2.08 (1.990)	2.54 (0.443)***
match(West Africa)	0.89 (0.500)	0.90 (0.471)	0.89 (0.501)
match(Jamaican)	0.93 (0.505)	0.89 (0.541)	0.90 (0.519)
match(UK)	0.24 (0.333)	0.27 (0.298)	0.25 (0.335)
isolates	0.82 (0.533)	0.78 (0.501)	0.80 (0.537)

*$p < .05$ **$p < .01$ ***$p < .001$

6.4 Discussion

The networks analyzed here provide a variety of insights. The Faux Desert High Network allowed us to empirically show that the choice of the tapering parameters τ does not critically effect the parameter estimates and thus has no effect on inference (within reason; severe tapering can cause overly small standard errors and misguided inference). The Last.fm case study showed us that RFMs can also handle larger networks upwards of at least 8,000 nodes. However, in doing so it may be necessary to taper all terms within the model and not just the dyad-dependent ones. The Last.fm analysis also underscored the foolishness of the homogeneity assumption for large networks when using only structural information.

Although not used in any of the case studies here, the Stereo ERGM is yet another option. Because the non-degeneracy of this model is not guaranteed, the Stereo ERGM is best used on smaller networks when only very light tapering may be needed. Lastly, the London Gang Network demonstrated how important the Restorative Force Models really are. Without them, desirable terms like the triangle cannot be fit and incorrect inferences may occur. The analysis on the London Gang Network resulted in conclusions nearly polar opposite of those reached by the authors of the original analysis done using standard ERGMs which were unable to use triangle terms. This case study also showed that RFMs other than the Tapered ERGM, especially the LogCosh ERGM, are viable options as well but may require more effort in setting the tapering parameters.

CHAPTER 7

Closing Thoughts

Often the path to discovery is a long and winding one, and this process was no different. While there surely is no roadmap to innovation, it is nonetheless worth reflecting on the journey in that it may provide guidance for future endeavors. A well-defined problem is key, and the problem of degeneracy was certainly that. The task was to find network models that do not suffer from degeneracy and develop them rigorously. How to go about doing so was much less clear.

Knowing that degeneracy is deeply embedded within ERGMs, it made sense to approach things from the very foundation of ERGMs, that is, the maximum entropy derivation. Alternative forms of entropy were played with and tested, and although no success was found through this approach, it did yield valuable insights into the nature of degeneracy and why it is inherent in ERGMs (see section 2.3).

Knowing that certain configurations (i.e., most social networks with sparsity and clustering) are more problematic than others, hierarchical models were a good choice to try to filter out the effects of degeneracy. Exploration into this area again proved fruitless, but it led to a better understanding of how priors can influence network models and why they fail to solve the problem of degeneracy.

The steps to the promise land are often laid down by many people, and the first major step toward solving degeneracy came from Horvát et al. (2015) when he showed the criteria necessary to guarantee non-degeneracy. It was then the work of Fellows and Handcock (2017) that built on top of those results to introduce the Tapered ERGM. The research done here

tied up any loose ends with the Tapered ERGM, including recommending how much to taper and empirically showing the small effects of tapering on paramter estimates. The Tapered ERGM opened the door to the more general Restorative Force Models, creating an arsenal of models with which to conquer degeneracy.

Perhaps most interestingly, in coming up with an algorithm for how much to taper in the Tapered ERGM, this led to a re-examination into the old idea of kurtosis. Incorporating the idea of using the kurtosis to measure degeneracy into the maximum entropy derivation led to the most promising model of all, the *kurstosis-constrained ERGM , or, KERGM.* Everything has come full-circle, going back to the fist ideas involving entropy. Although the KERGM did not make it into this dissertation, the KERGM is a direct result of the work done here.

So, perhaps the best advice on seeking discovery is to turn down many avenues, possibly even more than once.

CHAPTER 8

Appendix

8.1 Models Related to q-Entropy

8.1.1 The p_{*q} Model

Here we show the derivation of the p_{*q} model. We use the method of Lagrange Multipliers.

Let $t(y)$ be a vector in $\mathbb{R}^d$ of d sufficient statistics, indexed by functions $\phi_i : G_N \to \mathbb{R}$ and α be the vector in $\mathbb{R}^d$ of mean constraints, indexed by $\alpha_i \in \mathbb{R}$. We wish to maximize the q-Entropy, $H_q(p)$, subject to the linear constraints $\mathbb{E}_P[t_i(Y)] = \alpha_i$. That is, maximize

$$H_q(p) \equiv \frac{\sum_{y \in G_N} p(y)^q - 1}{1 - q}$$

subject to the constraints

$$\sum_{y \in G_N} p(y) t_i(y) = \alpha_i$$

$$p(y) \geq 0, \forall y \in G_N$$

$$\sum_{y \in G_N} p(y) = 1$$

We introduce the Lagrange multipliers $\lambda(y) \geq 0$ for the constraint $p(y) \geq 0$, $\theta_0 \in \mathbb{R}$ for the normalization constraint $p(G_N) = 1$, and θ_i for the constraints $\mathbb{E}_P[t_i(Y)] = \alpha_i$. The Lagrangian is $\mathcal{L}(p, \theta, \theta_0, \lambda) =$

$$-H_q(p) + \sum_{i=1}^{d} \theta_i(\alpha_i - \sum_{y \in G_N} p(y) t_i(y)) + \theta_0(\sum_{y \in G_N} p(y) - 1) + \sum_{y \in G_N} \lambda(y)(p(y) - 0)$$

or

$$\frac{\sum_{y\in G_N} p(y)^q - 1}{q-1} + \sum_{i=1}^{d} \theta_i(\alpha_i - \sum_{y\in G_N} p(y)t_i(y)) + \theta_0(\sum_{y\in G_N} p(y) - 1) + \sum_{y\in G_N} \lambda(y)p(y)$$

If we treat p as a finite dimensional vector indexed by y, we can differentiate the Lagrangian with respect to the vector element $p(y)$ and obtain:

$$\frac{\partial \mathcal{L}}{\partial p(y)} = \frac{q}{q-1}p(y)^{q-1} - \sum_{i=1}^{d} \theta_i t_i(y) + \theta_0 + \lambda(y)$$

$$\frac{\partial \mathcal{L}}{\partial p(y)} = \frac{q}{q-1}p(y)^{q-1} - \langle \theta, t(y)\rangle + \theta_0 + \lambda(y)$$

Setting $\frac{\partial \mathcal{L}}{\partial p(y)} = 0$ and solving gives

$$p(y) = \left[\frac{q-1}{q}(\langle \theta, t(y)\rangle - \theta_0 - \lambda(y))\right]^{\frac{1}{q-1}}$$

Since we define $p(y) \equiv 0$ when the argument is not positive, we are ensured $p(y) \geq 0$ $\forall y$ and can eliminate the constraint $\lambda(y)$. Letting θ_0 be the normalizing constant $Z_q(\theta)$, we arrive at equation (2.4):

$$p_{*q,\theta}(Y = y) = [(1 - \frac{1}{q})(\langle \theta, t(y)\rangle - Z_q(\theta))]_+^{\frac{1}{q-1}}$$

8.1.2 A Note on Models using the q-Exponential

Tsallis (1994) introduced the so-called *q-expectation*:

$$\mathbb{E}_q[\phi(x)] = \sum_{x\in\chi} p(x)^q \phi(x)$$

For what it's worth, Tsallis points out that the q-expectation is not a mean value of $\phi(x)$, but it is a mean value of $p(x)^{q-1}\phi(x)$.

Tsallis maximizes H_q given the normalization constraint $P(\chi) = 1$ and a new constraint using his q-expectation:

$$\sum_{x\in\chi} p(x)^q x = \alpha \tag{8.1}$$

where α is a known real number. He arrives at the distribution

$$p(x) = \frac{[1-\beta(1-q)x]^{\frac{1}{1-q}}}{Z_q} = \frac{\exp_q(-\beta x)}{Z_q} \tag{8.2}$$

where $Z_q = \sum_{x\in\chi}[1-\beta(1-q)x]^{\frac{1}{1-q}}$ and β is a Lagrange multiplier.

Although Tsallis does not provide any proof of his claim, we will work through it now. Using the constraint $p(x) \geq 0$ and the normalization constraint $P(\chi) = 1$, but now also using the q-expectation constraint $\sum_{x\in\chi} p(x)^q\phi_i(x) = \alpha_i$, maximizing H_q implies the Lagrangian is:

$$\mathcal{L}(p,\theta,\theta_0,\lambda) = -H_q + \sum_{i=1}^{d}\theta_i(\alpha_i - \sum_{x\in\chi} p(x)^q\phi_i(x)) + \theta_0(\sum_{x\in\chi} p(x) - 1) + \sum_{x\in\chi}\lambda(x)(p(x)-0)$$

$$\mathcal{L}(p,\theta,\theta_0,\lambda) = \frac{\sum_{x\in\chi} p(x)^q - 1}{q-1} + \sum_{i=1}^{d}\theta_i(\alpha_i - \sum_{x\in\chi} p(x)^q\phi_i(x)) + \theta_0(\sum_{x\in\chi} p(x) - 1) + \sum_{x\in\chi}\lambda(x)(p(x)-0)$$

Differentiating with respect to $p(x)$ gives:

$$\frac{\partial\mathcal{L}}{\partial p(x)} = \frac{q}{q-1}p(x)^{q-1} - qp(x)^{q-1}\sum_{i=1}^{d}\theta_i\phi_i(x) + \theta_0 + \lambda(x)$$

$$\frac{\partial\mathcal{L}}{\partial p(x)} = \frac{q}{q-1}p(x)^{q-1} - qp(x)^{q-1}\langle\theta,\phi(x)\rangle + \theta_0 + \lambda(x)$$

Setting $\frac{\partial\mathcal{L}}{\partial p(x)} = 0$ and solving gives

$$p(x) = (\frac{(q-1)(-\theta_0-\lambda(x))}{q[1-(q-1)\langle\theta,\phi(x)\rangle]})^{\frac{1}{q-1}}$$

$$p(x) = (\frac{q[1+(1-q)\langle\theta,\phi(x)\rangle]}{(q-1)(-\theta_0-\lambda(x))})^{\frac{1}{1-q}}$$

Letting $Z = (\frac{q}{(q-1)(-\theta_0-\lambda(x))})^{\frac{1}{q-1}}$ and using the q-exponential $\exp_q(x)$, we arrive at

$$p(x) = \frac{\exp_q(\langle\theta,\phi(x)\rangle)}{Z} \tag{8.3}$$

Note that in obtaining this form, our Z differs from Tsallis' Z_q given above. More importantly, a major problem with the q-expectation is that $\sum_{x\in\chi} p(x)^q \neq 1$. Tsallis himself would later call his use of the q-expectation "clearly unacceptable" (Tsallis et al., 1998). He instead suggests using the *escort probabilities*, which amount to using the constraint

$$\mathbb{E}_{P_e}[\phi_i(X)] = \frac{1}{h_q}\sum_{x\in\chi} p(x)^q \phi_i(x) = \alpha_i \tag{8.4}$$

where P_e is the *escort distribution*

$$P_e(x) = \frac{p(x)^q}{\sum_{x\in\chi} p(x)^q} = \frac{p(x)^q}{h_q}$$

However, maximizing H_q under escort constraints does not lead to a feasible model.

8.2 Reverse Kullback-Leibler Divergence

Let P and Q be two probability distributions over a space Ω. Every f-divergence measure between two distributions is induced by a real-valued convex function $f(t)$, defined for $t > 0$, such that $f(1) = 0$. The reverse Kullback-Leibler divergence of P from Q is defined via the function

$$f^*(t) = -\log t$$

The corresponding f-divergence is then

$$\begin{aligned} D_{f^*}(P \parallel Q) &= \sum_{\Omega} f^*\left(\frac{p(x)}{q(x)}\right) q(x) \\ &= \sum_{\Omega} -\log\left(\frac{p(x)}{q(x)}\right) q(x) \\ &= \sum_{\Omega} \log\left(\frac{q(x)}{p(x)}\right) q(x) \end{aligned}$$

If we let $q(x) = 1$, we can measure the divergence of P from the uniform distribution as

$$= \sum_{\Omega} \log\left(\frac{1}{p(x)}\right)$$
$$= \sum_{\Omega} -\log\left(p(x)\right)$$

Here we find the distribution P that minimizes the reverse KL-divergence from the uniform distribution subject to the following constraints:

$$\sum_{\Omega} p(x) = 1$$
$$p(x) \geq 0 \, \forall x \in \Omega$$
$$\sum_{x \in \chi} p(x) t_i(x) = \mu_i$$

Where the last equation represents $i = 1, ..., d$ arbitrary mean constraints for the $t_i(x)$ sufficient statistics.

We again use the method of Lagrange Multipliers. Using λ_0, $\lambda(x)$, and ϕ_i for the given constraints, the Lagrangian is

$$\mathcal{L}(p, \lambda, \theta) = -\sum_{\Omega} \log\left(p(x)\right) + \lambda_0 \left(\sum_{\Omega} p(x) - 1\right) + \lambda(x)\left(p(x) - 0\right) + \sum_{i=1}^{d} \phi_i \left(\sum_{\Omega} p(x) t_i(x) - \mu_i\right)$$

In the case where Ω is finite, $p = [p(x)]_{x \in \Omega}$ is a finite dimensional vector indexed by x. If we take the derivative of the Lagrangian with respect to each vector element $p(x)$, we obtain:

$$\frac{\partial \mathcal{L}}{\partial p(x)} = -\frac{1}{p(x)} + \lambda_0 + \lambda(x) + \sum_{i=1}^{d} \phi_i t_i(x)$$

We can verify that we will find a minimum by noting that

$$\frac{\partial^2 \mathcal{L}}{\partial p(x)^2} = \frac{1}{p(x)^2} > 0 \, \forall x$$

To find this minimum we set $\frac{\partial \mathcal{L}}{\partial p(x)} = 0$ and solve, which yields:

$$\begin{aligned} p(x) &= \frac{1}{\lambda_0 + \lambda(x) + \sum_{i=1}^{d} \phi_i t_i(x)} \\ &= \frac{1}{\lambda_0 \left[1 + \frac{\lambda(x)}{\lambda_0} + \frac{1}{\lambda_0} \sum_{i=1}^{d} \phi_i t_i(x)\right]} \end{aligned}$$

If we restrict ϕ_i such that the denominator is always positive, we can eliminate $\lambda(x)$ by complementary slackness. Then, setting $\frac{\phi_i}{\lambda_0} = \theta_i$ and $\lambda_0 = C(\theta)$, the normalizing constant, we obtain

$$p(Y = y) = \frac{1}{C(\theta)\left[1 + \langle \theta, t(y) \rangle\right]}$$

8.3 Symmetrized Divergence

The symmetric divergence between two probability distributions P and Q (over a space Ω) is given by

$$D(P, Q) \equiv D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)$$

Here we find the model class that minimizes the symmetric divergence from the uniform distribution. Let Q be the uniform distribution, so $q(x) = 1$. Then,

$$D(P, Q) = \sum_{\Omega} p(x) \log\left(p(x)\right) - \sum_{\Omega} \log\left(p(x)\right)$$

We seek to find P that minimizes $D(P, Q)$ subject to the following basic constraints:

$$\sum_{\Omega} p(x) = 1$$

$$p(x) \geq 0 \, \forall x \in \Omega$$

$$\sum_{\Omega} p(x) t_i(x) = \mu_i$$

Where the last equation represents $i = 1, ..., d$ arbitrary mean constraints for the $t_i(x)$ sufficient statistics. We denote the vector of sufficient statistics as $t(x) = t$.

Using λ_0, $\lambda(x)$, and θ_i for the given constraints, the Lagrangian is

$$\mathcal{L}(p,\lambda,\theta) = \sum_{\Omega} p(x)\log\left(p(x)\right) - \sum_{\Omega}\log\left(p(x)\right) + \lambda_0\left(\sum_{\Omega} p(x) - 1\right) + \lambda(x)\left(p(x) - 0\right) + \sum_{i=1}^{d}\theta_i\left(\sum_{\Omega} p(x)t_i(x) -\right.$$

In the case where Ω is finite, $p = [p(x)]_{x\in\Omega}$ is a finite dimensional vector indexed by x. If we take the derivative of the Lagrangian with respect to each vector element $p(x)$, we obtain:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial p(x)} &= 1 + \log\left(p(x)\right) - \frac{1}{p(x)} + \lambda_0 + \lambda(x) + \sum_{i=1}^{d}\theta_i t_i(x) \\ &= 1 + \log\left(p(x)\right) - \frac{1}{p(x)} + \lambda_0 + \lambda(x) + \theta^T t\end{aligned}$$

We can verify that we will find a minimum by noting that

$$\frac{\partial^2\mathcal{L}}{\partial p(x)^2} = \frac{1}{p(x)} + \frac{1}{p(x)^2} > 0\,\forall x$$

To find this minimum we set $\frac{\partial\mathcal{L}}{\partial p(x)} = 0$ and solve, which yields:

$$\begin{aligned}1 + \log\left(p(x)\right) - \frac{1}{p(x)} + \lambda_0 + \lambda(x) + \theta^T t &= 0 \\ 1 + \lambda_0 + \lambda(x) + \theta^T t &= \frac{1}{p(x)} + \log\left(\frac{1}{p(x)}\right) \\ \exp\left(1 + \lambda_0 + \lambda(x) + \theta^T t\right) &= \exp\left(\frac{1}{p(x)} + \log\left(\frac{1}{p(x)}\right)\right) \\ \exp\left(1 + \lambda_0 + \lambda(x) + \theta^T t\right) &= \frac{1}{p(x)}\exp\left(\frac{1}{p(x)}\right) \\ W\left(\exp\left(1 + \lambda_0 + \lambda(x) + \theta^T t\right)\right) &= \frac{1}{p(x)} \\ p(x) &= \frac{1}{W\left(\exp\left(1 + \lambda_0 + \lambda(x) + \theta^T t\right)\right)}\end{aligned}$$

where $W()$ is the *Lambert-W function*, also known as the *product logarithm*. Because W is positive whenever its argument is positive, we can eliminate $\lambda(x)$. Allowing the normalizing constant $C(\theta)$ to be $\exp(1 + \lambda_0)$, we have

$$p(x) = \frac{1}{W\left(C(\theta)\exp\left(\theta^T t\right)\right)}$$

BIBLIOGRAPHY

Amari, S.-i. (2012), *Differential-geometrical methods in statistics*, vol. 28, Springer Science & Business Media.

Amari, S.-i. and Ohara, A. (2011), "Geometry of q-exponential family of probability distributions," *Entropy*, 13, 1170–1185.

Balanda, K. P. and MacGillivray, H. (1988), "Kurtosis: a critical review," *The American Statistician*, 42, 111–119.

Chissom, B. S. (1970), "Interpretation of the kurtosis statistic," *The American Statistician*, 24, 19–22.

Darlington, R. B. (1970), "Is kurtosis really "peakedness?"," *The American Statistician*, 24, 19–22.

DeCarlo, L. T. (1997), "On the meaning and use of kurtosis." *Psychological methods*, 2, 292.

Diaconis, P. and Ylvisaker, D. (1979), "Conjugate priors for exponential families," *The Annals of statistics*, 269–281.

Ellison, A. M. (1987), "Effect of seed dimorphism on the density-dependent dynamics of experimental populations of Atriplex triangularis (Chenopodiaceae)," *American Journal of Botany*, 74, 1280–1288.

Fellows, I. and Handcock, M. (2017), "Removing phase transitions from Gibbs measures," in *Artificial Intelligence and Statistics*, pp. 289–297.

Frank, O. and Strauss, D. (1986), "Markov graphs," *Journal of the american Statistical association*, 81, 832–842.

Grund, T. U. and Densley, J. A. (2012), "Ethnic heterogeneity in the activity and structure of a Black street gang," *European Journal of Criminology*, 9, 388–406.

— (2015), "Ethnic homophily and triad closure: Mapping internal gang structure using exponential random graph models," *Journal of Contemporary Criminal Justice*, 31, 354–370.

Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. (2003), "Assessing degeneracy in statistical models of social networks," Tech. rep., Citeseer.

Hildebrand, D. K. (1971), "Kurtosis measures bimodality?" *The American Statistician*, 25, 42–43.

Horvát, S., Czabarka, É., and Toroczkai, Z. (2015), "Reducing degeneracy in maximum entropy models of networks," *Physical review letters*, 114, 158701.

Joanes, D. and Gill, C. (1998), "Comparing measures of sample skewness and kurtosis," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 183–189.

Karwa, V., Petrović, S., and Bajić, D. (2016), "DERGMs: Degeneracy-restricted exponential random graph models," *arXiv preprint arXiv:1612.03054*.

Last.fm (2020a), "About Last.fm," https://www.last.fm/about, accessed: 2020-03-10.

— (2020b), "About Us," https://store.last.fm/pages/about-us, accessed: 2020-03-10.

Moors, J. J. A. (1986), "The meaning of kurtosis: Darlington reexamined," *The American Statistician*, 40, 283–284.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Naudts, J. (2008), "Generalised exponential families and associated entropy functions," *Entropy*, 10, 131–149.

— (2009), "The q-exponential family in statistical physics," *Central European Journal of Physics*, 7, 405–413.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., et al. (1997), "Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health," *Jama*, 278, 823–832.

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

Toivonen, R., Kovanen, L., Kivelä, M., Onnela, J.-P., Saramäki, J., and Kaski, K. (2009), "A comparative study of social network models: Network evolution models and nodal attribute models," *Social networks*, 31, 240–254.

Tsallis, C. (1988), "Possible generalization of Boltzmann-Gibbs statistics," *Journal of statistical physics*, 52, 479–487.

— (1994), "What are the numbers that experiments provide," *Quimica Nova*, 17, 468–471.

Tsallis, C., Mendes, R., and Plastino, A. R. (1998), "The role of constraints within generalized nonextensive statistics," *Physica A: Statistical Mechanics and its Applications*, 261, 534–554.

Umarov, S., Tsallis, C., and Steinberg, S. (2008), "On a q-central limit theorem consistent with nonextensive statistical mechanics," *Milan journal of mathematics*, 76, 307–328.

Westfall, P. H. (2014), "Kurtosis as peakedness, 1905–2014. RIP," *The American Statistician*, 68, 191–195.