Title
Practical Network Modeling via Tapered Exponential-Family Random Graph Models

Journal
Journal of Computational and Graphical Statistics, 32(2)

ISSN
1061-8600

Authors
Blackburn, Bart
Handcock, Mark S

Publication Date
2023-04-03

DOI
10.1080/10618600.2022.2116444

Peer reviewed

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/10618600.2022.2116444

Practical Network Modeling via Tapered Exponential-Family Random Graph Models

Bart Blackburn and Mark S. Handcock

Department of Statistics, University of California, Los Angeles, Los Angeles, CA

ABSTRACT

Exponential-family Random Graph Models (ERGMs) have long been at the forefront of the analysis of relational data. The exponential-family form allows complex network dependencies to be represented. Models in this class are interpretable, flexible and have a strong theoretical foundation. The availability of powerful user-friendly open-source software allows broad accessibility and use. However, ERGMs sometimes suffer from a serious condition known as near-degeneracy, in which the model exhibits unrealistic probabilistic behavior or a severe lack-of-fit to real network data. Recently, Fellows and Handcock proposed a new model class, the Tapered ERGM, which circumvents the issue of near-degeneracy while maintaining the desirable features of ERGMs. However, the question of how to determine the proper amount of tapering needed for any model was heretofore left unanswered. This article develops a new methodology for how to determine the necessary level of tapering and as such provides a new approach to inference for the Tapered ERGM class. Noting that a Tapered ERGM can always be made nondegenerate, we offer data-driven approaches for determining the amount of tapering necessary. The mean-value parameter estimates are unaffected by tapering, and we show that the natural parameter estimates are numerically weakly varying by the level of tapering. We then apply the Tapered ERGM to two published networks to demonstrate its effectiveness in cases where typical ERGMs fail and present the case for Tapered ERGMs replacing ERGMs entirely.

ARTICLE HISTORY
Received August 2021
Accepted August 2022

KEYWORDS
Degeneracy; ERGM; Goodness of fit; Social network analysis

1. Introduction

Network models are widely used to represent relational information among interacting units and the structural implications of these relations. Social network studies have focused a great deal of attention on random graph models of networks whose nodes represent individual social entities and whose edges represent a specified relationship between the entities. Such entities could be individuals in the workplace, countries within global markets, satellites in space, or from a wide range of social or natural phenomena. We 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 (Krivitsky 2012), for this work we focus on edges that take binary values indicating whether a relation between nodes exists or does not. Most often the number of nodes is fixed and known ($N = |V|$) and in the undirected case there are therefore $|G_N| = 2^{\binom{N}{2}}$ possible graphs. In addition to the graph, it is common to have covariate data on the nodes and edges. Here we represent it by X and define a network as the union of the covariate and the graph structure (i.e., $\{X, G\}$). We focus on the situation where the covariate information is exogenous and suppress reference to the covariates for notational simplicity. For the more general case, see Fellows (2012).

Real-world networks reflect the complex social systems that are their source. As such, statistical models for network data should be able to represent complex dependencies. Exponential-family Random Graph Models (ERGMs) have shown themselves to be a useful class of models for representing complex social phenomena in this domain (Strauss (1986); Goodreau (2007); Handcock et al. (2008); Goodreau, Kitts, and Morris (2009)). An ERGM for the network can be expressed as

$$p_\theta(Y = y|X) = \frac{\exp(\theta \cdot t(y, X))}{c(\theta, X)} \qquad y \in G_N(X) \tag{1}$$

where Y is a random graph whose realization is $y \in G_N(X)$, the set of all possible graphs on N nodes with covariates X; $t(y, X)$ is a d-vector valued function defining a set of sufficient statistics; $\theta \in \mathbb{R}^d$ is a vector of parameters; and $c(\theta, X)$ the normalizing constant. Each ERGM is defined by the choice of sufficient statistics. These are chosen by the researcher, depending on domain knowledge, to specify the generating social processes. They can be any statistical summary of network properties and are typically motivated by social theory (Goodreau, Kitts, and Morris 2009) or symmetry arguments (Strauss 1986). In this way, ERGMs constitute a family of models across different choices of the sufficient statistics. Regardless of which sufficient statistics are used, the ERGM will have the maximal entropy of

CONTACT Bart Blackburn bart.blackburn@gmail.com Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095-1554.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

any distribution satisfying the d-dimensional mean constraints placed on $t(y, X), E[t(y, X)] = \mu$.

Properties of exponential-family models have received extensive attention in the statistical literature (Barndorff-Nielsen 1978) and their application to networks has a long history (Holland and Leinhardt 1981; Strauss 1986). Schweinberger et al. (2020) review random graph models for complex random graphs. They emphasize the value of the exponential-family framework and address two issues that have arisen in modeling using ERGMs. One is that most ERGMs are not projective (Shalizi and Rinaldo 2013). For ERGMs, projectivity is a form of closure under marginalization and implies that the same parameters govern the marginal distributions of all subgraphs. While projectivity may be statistically convenient, it may not be realistic as it implies the subgraph distributions are unaffected by embeddedness within the overall graph. It does, however, emphasize the importance of likelihood-based inference which naturally deals with the lack of projectivity (Handcock and Gile 2010).

The second concern is that ERGMs with nontrivial dependence structure can be ill-behaved. In an effort to maximize entropy, the ERGM can be thought of as "spreading out" mass across the graph space $G_N(X)$ 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 in the region around the observed graph. This problem is referred to as *near-degeneracy*: despite having realistic mean values, no choice of the parameters places significant probability mass on graphs that are realistic.

Figure 1 shows an example of near-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 as the context. 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 near-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. The issue of near-degeneracy in ERGMs is well-documented but unresolved (Handcock 2003; Snijders et al. 2006; Schweinberger 2011; Rinaldo, Fienberg, and Zhou 2009).

However, recently there has been a breakthrough with the *Tapered ERGM* (Fellows and Handcock 2017). Fellows and Handcock (2017) propose an extension of the standard ERGM which disallows near-degeneracy through additional constraints on the sufficient statistics. This article further develops the ideas behind the Tapered ERGM and demonstrates the usefulness of this class of models.

In Section 2, we provide a development of the Tapered ERGM model and why tapering is effective in reducing the impact of degeneracy. Section 3 motivates the use of bimodality and kurtosis as numerical measures of near-degeneracy. In Section 4, we develop the methodology for Tapered ERGMs and the incorporation of kurtosis in automatic selection of the degree of tapering. In Section 5 we consider two network modeling situations where standard ERGMs would naturally be used, comparing the ERGM fits to those of Tapered ERGMS. We conclude in Section 6 with a discussion of the results and implications for practical modeling of complex social networks.

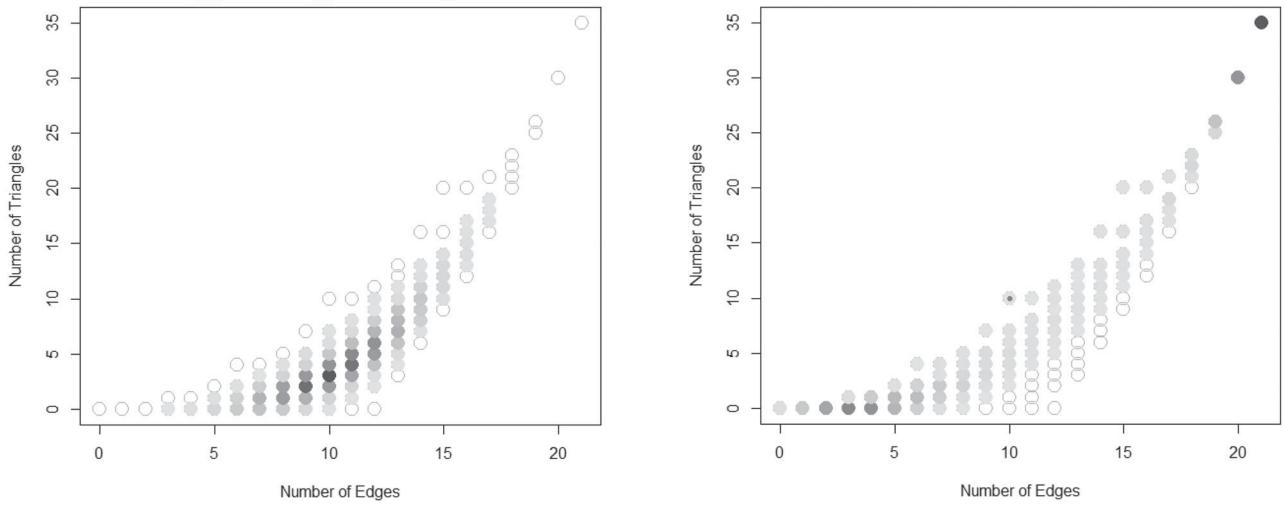

Figure 1. Near-degenerate ERGM. Each class of graphs, identified by the number of edges and triangles, is represented by a circle. *LEFT*: The number of graphs within each class, where the intensity of the shading is proportional to number of graphs. The darker the shading, the larger the number of graphs. *RIGHT*: The ERGM for mean edge and triangle constraints of 10 and 10, where the 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.

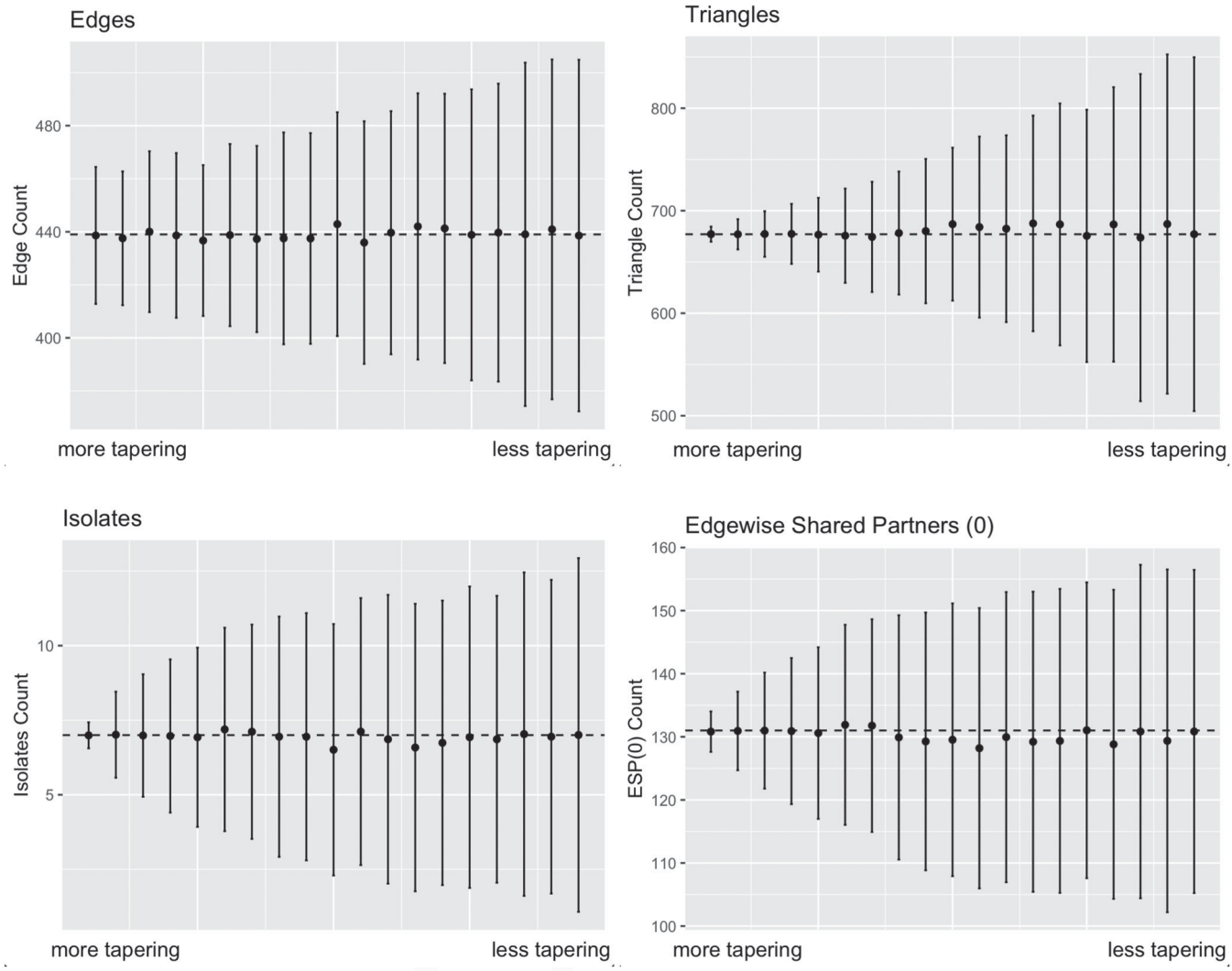

Figure 2. Variation in term counts across different levels of tapering. In each of the panels above, the dashed line indicates the term count in the observed network. Each point is the mean parameter at that level of tapering with corresponding variation bars (plus/minus two standard deviations). We 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.

2. The Tapered Version of ERGM

We start with a mechanistic explanation of the Tapered ERGM and follow it up with a conceptual and technical explanation. The Tapered ERGM of Fellows and Handcock (2017) is the maximum entropy distribution subject to additional upper bounds on the variance of the sufficient statistics. The solution is:

$$p_{\theta,\tau}(Y = y) = \frac{\exp\left(\theta \cdot t(y) - \tau \cdot (\mu(\theta,\tau) - t(y))^2\right)}{c(\theta,\tau)}$$
$$\times\ \tau \in \mathbb{R}^d \geq 0 \quad \theta \in \mathbb{R}^d \qquad (2)$$

where we have suppressed the expression of the covariates. In the above, $\mu(\theta,\tau) \equiv \mathrm{E}_{\theta,\tau}[t(Y)]$. The form alone is enough to intuitively grasp why tapering works: it adds additional terms to the standard ERGM that measure the deviation of the statistics from their mean. If the elements of τ are positive, graphs with statistics far from their central location have lower probability. This reduces the propensity of the model to place significant mass on extremal configurations such as the empty and complete graphs. The larger τ, the heavier the tapering and the less the graph statistics vary from their mean parameters. It is also possible to generalize the Tapered ERGM using other forms of additional constraints, creating classes of models collectively known as Restorative Force Models (Blackburn 2021).

2.1. Why Tapering Works

Figure 2 shows the effect of tapering applied to an adolescent friendship network from the National Longitudinal Study of Adolescent Health (Resnick et al. 1997). On the far right of each panel, the mean parameter with two standard deviation bars are plotted for the standard ERGM (no tapering). 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 variation of the mean parameter (i.e., the standard deviation of the term count). The mean parameters always remain consistent at the observed values; they just vary less with increased tapering.

We need not rely on our conceptual intuition to see why the Tapered ERGM reduces near-degeneracy. We can prove

that we can always find a parameter τ that will make $p_{\theta,\tau}(Y)$ nondegenerate, and we do so now. In Horvát, Czabarka, and Toroczkai (2015), the authors provide two critical results. When near-degeneracy occurs, the ERGM $p_\theta(Y)$ is plagued by multimodality. 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. Then $r(x)$ 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 is worth pointing out that the standard ERGM is a probability mass function (PMF) with respect to the counting measure. Furthermore, letting $t(y) \equiv t$, the probability a graph is sampled by the ERGM is

$$p(t|\theta) = \frac{N(t)}{c(\theta)}\exp(\theta\cdot t) \quad t \in T, \quad T = \{s : \exists y \in G_N \, s.t. \, s = t(y)\}$$

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, Czabarka, and Toroczkai (2015) provide the following insight:

Result 2. Let $\tilde{N}(t(y))$ be a smoothed, continuous interpolator of $N(t(y))$. An ERGM is nondegenerate if and only if $\tilde{N}(t(y))$ is strictly log-concave.

Because of its 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 computationally impossible or at best extremely expensive. Under the Tapered ERGM, however, we have

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

We are now able to avoid computing $N(t)$ and we can guarantee the Tapered ERGM is nondegenerate so long as $\tilde{N}(t)\exp(-\tau\cdot(\mu-t)^2)$ is strictly log-concave. Neither Horvát, Czabarka, and Toroczkai (2015) nor Fellows and Handcock (2017) explicate a smoothing function $\tilde{N}(t)$, but we do so here. 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 $\mathbb{R}^d_{\geq 0}$. If we define T as the set of realizable sufficient statistics, one possible choice for $\tilde{N}(t)$ is

$$\tilde{N}(t) = \begin{cases} N(t), \text{if } t \in T \\ \sum_{s\in T} N(s)\exp(-||t-s||^2), \text{otherwise} \end{cases} \tag{3}$$

Note that in Fellows and Handcock (2017), the authors prove the nondegeneracy 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 in equation (2).

Theorem 3. Let chull(T) be the convex hull of the sample space of statistics, T. For any vector μ of mean parameters in chull(T), there exists a vector of tapering parameters $\tau \in \mathbb{R}^d_{\geq 0}$ such that the Tapered ERGM with tapering center μ is nondegenerate.

Proof. We will use $\tilde{N}(t)$ as defined in equation (3) for our smoothing function. It suffices to show that $\tilde{N}(t)\exp(-\tau\cdot(\mu - t)^2)$ 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 $h(t) = \tau\cdot(\mu - t)^2$. 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 $x^T\nabla^2 h x = \sum_i 2\tau_i x_i^2$. Thus, regardless of $\nabla^2\log(\tilde{N}(t))$, we can always choose τ large enough such that $x^T\nabla^2 r x < 0$. Thus, r is concave and the Tapered ERGM is nondegenerate by Results 1 and 2. □

2.2. Interpreting the Tapered ERGM Parameters

If the tapering parameters τ are zero, then the Tapered model is identical to the standard ERGM and an interpretation of the θ parameters is as conditional log-odds. However, nonzero τ has an effect on the interpretation of the parameters. To see this, let $P(Y_{ij} = 1|Y^c_{ij} = y^c_{ij}) \equiv P(Y^+_{ij})$ and $P(Y_{ij} = 0|Y^c_{ij} = y^c_{ij}) \equiv P(Y^-_{ij})$. Then, under the Tapered ERGM the log-odds of a tie conditional on Y^c_{ij} is

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

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. δ_{kij} is a measure of the deviation of

the network statistics from their 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 and the effect of the dyad change on the change statistics.

Note that when θ_k is the MLE, $\hat{\theta}_k$, $\mu_k = t_k(Y)$ and for any given dyad Y_{ij} it must be the case that $\mu_k = t_k(Y_{ij}^+)$ or $\mu_k = t_k(Y_{ij}^-)$. Hence, when θ_k is the MLE, the log-odds of a tie is $\sum_k \Delta t_k(Y_{ij}) \left[\hat{\theta}_k + \tau_k(2Y_{ij} - 1)\Delta t_k(Y_{ij})\right]$. The last expression suggests a measure of the bias in the Tapered ERGM parameter estimate $\hat{\theta}_k$ as an estimate of the conditional log-odds is the average over the dyads in the network of the penalty term: $-\tau_k \sum_{ij} (2Y_{ij} - 1)\Delta t_k(Y_{ij})$. This is easy to compute as the change statistics are available as a by-product of the computation of the maximum pseudo-likelihood estimator (MPLE) (van Duijn, Handcock, and Gile 2009), which is typically used as a starting value for the MCMC-MLE algorithm. We shall use this measure in the case-studies of Section 5 to empirically show that the bias in the Tapered ERGM parameter estimates are very small (on the order of 10^{-3} or smaller). The small magnitude of the bias, together with the fact that most statistics need not be tapered at all and incur no bias, points toward practically interpreting θ under the Tapered ERGM exactly as one would under the standard ERGM.

3. The Kurtosis and Bimodality

While Section 2 shows that we can prevent multimodality, we need a way to measure it. This brings us to a discussion of kurtosis. One of the hallmarks of near-degeneracy is bi/multimodality. When near-degeneracy strikes, often a large amount of mass is placed at or near the extremes (empty and complete graphs) of the graph space T with very little mass placed near the realistic graphs. Consider again the seven node graph model and suppose we observe a graph with sufficient statistics 10 edges and 10 triangles. Figure 3 shows two bimodal marginal distributions taken from an ERGM maximum likelihood fit. By construction, the MLE has mean parameters of 10 edges and 10 triangles, yet very little mass is put near those observed 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. We now consider measuring bimodality via kurtosis.

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. Our approach is to measure the bimodality using kurtosis. Specifically, the kurtosis of a random variable, X, is

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

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 nonnegligibly to the kurtosis since raising a number less than one 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; that is, 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.

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

From the above identity, Darlington argues the kurtosis can be interpreted 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, that is, one that is completely bimodal.

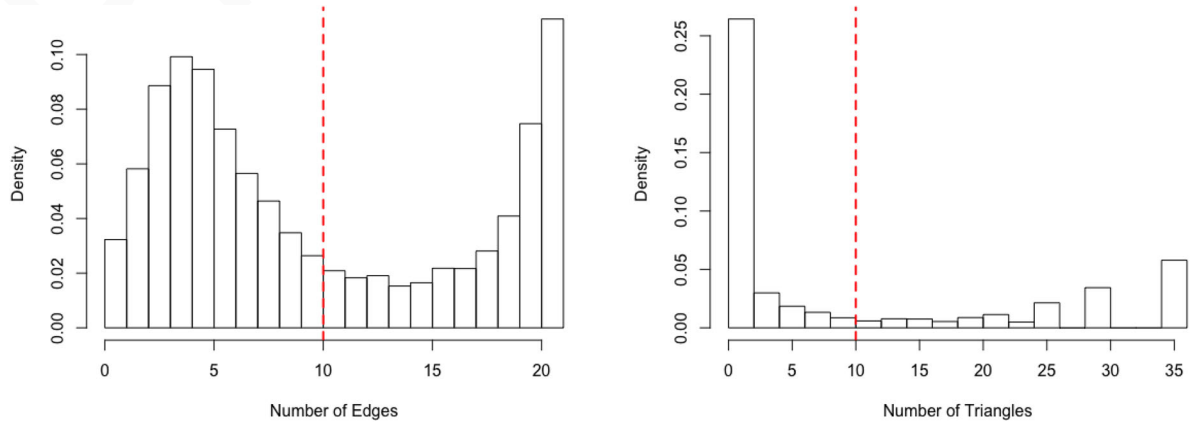

Figure 3. The marginal distributions of edges (left) and triangles (right) sampled from a near-degenerate ERGM. Much of the mass falls toward the empty and complete graphs with very little near the mean parameters (dashed line). A restriction to such polarized behavior is unrealistic for most social processes.

It would appear then that a lower kurtosis would indicate bimodality, where several benchmarks could be used (Kurt[X] = 3 for the Gaussian distribution and 9/5 for the uniform distribution). However, others (Hildebrand 1971; Westfall 2014) were quick to demonstrate counterexamples where bimodal distributions still had kurtosis values close to that of the Gaussian distribution, 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 large 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 near-degeneracy occur when significant probability mass is placed at the extremal configurations, that is, the empty and complete graphs (Horvát, Czabarka, and Toroczkai 2015; Handcock 2003). It is not possible to obtain a bimodal distribution with a high ($\approx$ 3) 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 near-degeneracy.

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 a Bernoulli with probability one-half.

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

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

$$\beta = \frac{\gamma_1^2 + 1}{\gamma_2} \tag{4}$$

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.

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. Tapering Methodology

In this section we address two main concerns when using the Tapered ERGM: (i) Will the level of tapering effect the numerical value of the parameter estimates; and (ii) What level of tapering should we use (and on which terms)? Our illustration in Section 2 suggests that the answer to (i) is most likely "no." Figure 5 shows that estimates of θ are remarkably stable across a wide range of tapering levels. In other words, the numerical value of the parameter estimates appear to be insensitive to the degree of tapering, as determined by τ. In addressing the first question, we can show that as τ goes to zero, the Tapered ERGM is identically the ERGM.

Theorem 4. Let $P_\theta(Y)$ denote the standard ERGM and $P_{\theta,\tau}(Y)$ denote the Tapered ERGM. Then as $\tau \to 0, D_{KL}(P_{\theta,\tau}||P_\theta) \to 0$, where $D_{KL}()$ is the Kullback-Leibler divergence of $P_{\theta,\tau}$ from P_θ.

Proof. Let $P_\theta(Y)$ be the standard ERGM and $P_{\theta,\tau}(Y)$ the Tapered ERGM. That is,

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

and

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

The Kullback–Leibler Divergence from P_θ to $P_{\theta,\tau}(Y)$ is

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

where $\sigma_k^2 = \mathbb{E}_{\theta,\tau}[(\mu_k - t_k(y))^2] = \text{var}_{\theta,\tau}[t_k(y)]$.
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, $D_{KL}(P_{\theta,\tau}||P_\theta) \to 0$ as $\tau \to 0$. □

This result has two important implications. First, it ensures the Tapered ERGM does not behave markedly different from the standard ERGM across certain thresholds of τ since the convergence to the ERGM distribution is smooth as τ goes to zero. Second, and more importantly, the equivalency of the distributions as τ approaches zero implies the parameter estimates of the two distributions also become equivalent (assuming the ERGM is minimal (Barndorff-Nielsen 1978, Corollary 8.1)).

The answer to question (ii) is more nuanced. While Theorem 4 indicates that the effect is negligible for sufficiently small τ, it does not ensure it is in real-world usage. Indeed, we should aspire to taper as few terms, and as little on each term, as possible. The argument for this is as follows. We saw in Theorem 4 above that the smaller τ is, the closer the Tapered ERGM is to the ERGM. Of course, in a nondegenerate scenario we would not need any tapering at all, but we most often cannot know a priori if the ERGM will be near-degenerate. So we should apply the minimum amount of tapering necessary in order to define a model with realistic behavior. This can be done in the following manner, with greater explanation of each step to follow.

Algorithm 1. Setting the Tapering Parameter

1. Choose only the dyad-dependent terms to taper.
2. If there are K terms to taper, set a large value of τ_k in order to heavily taper each of the $k = 1, \ldots, K$ terms.
3. If the MCMC estimation for θ converges, proceed to the next step. If the MCMC does not converge, go back to step 1 and taper all terms.
4. Relax the amount of tapering by decreasing each τ_k until the estimate of 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 are explosive when near-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 4 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 (Section 2.2). We know empirically that $\hat{\theta}$ 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 2 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 $\hat{\theta}(\tau)_{\text{MLE}}$, we know that $\mu = t_{obs}(y)$. Thus, the computation is less sensitive to the value of τ when we are in the vicinity of the observed graph where $t(y) \approx t_{obs}(y)$. The heavy tapering ensures $\hat{\theta}(\tau)_{\text{MLE}}$ exists and can be estimated accurately during MCMC estimation. Once we have an estimate of $\hat{\theta}(\tau)_{\text{MLE}}$, we can restart our MCMC routine at that value for smaller values of τ. Convergence of the iterated MCMC should still be quick since our initial estimate of $\hat{\theta}(\tau)_{\text{MLE}}$ is likely very close to $\hat{\theta}(\tau)_{\text{MLE}}$ and the model is far from degenerate. Usually it is enough to taper only the dependent terms, since in doing so the independent terms (like edge count, e.g.,) 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 near-degeneracy may be occurring), or until the bimodality coefficient $\beta \geq 0.4$, where β will make use of the bias-corrected kurtosis (Blackburn 2021). 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 $\beta = 0.33$, and the uniform distribution has $\beta = 0.55$. 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 $\mu_k = \text{E}[t_k(y)]$. 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 $r = 0.5$ 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 of this section shows that it is theoretically possible to fit networks using the Tapered ERGM, and the algorithm above shows that it is also practical.

4.1. Penalized Likelihood via the Kurtosis

There is yet another way to use the kurtosis to assist in setting the tapering parameters τ. Instead of relying on the guesswork of Algorithm 1, if we set a target kurtosis value we can simply maximize the likelihood $p_{\theta,\tau}(Y)$ subject to a penalty on how far the kurtosis deviates from the target value. Note that in this framework there is no need to work with the bimodality coefficient and we can instead use the bias-corrected sample kurtosis, K_C, directly (Blackburn 2021).

We can always increase τ to make the kurtosis of the Tapered ERGM closer to a target kurtosis, say K_T. However, in doing so the values of τ will necessarily increase until $K_C = K_T$ on average. In order to avoid over-tapering, we must also set a penalty on the magnitude of τ. That is, we estimate τ as

$$\hat{\tau} = \arg\max_{\tau} \left[l(\theta, \tau, ; \mathbf{y}) - \tau - \gamma \left(\frac{K_C - K_T}{K_\sigma} \right)^2 \right]$$

where $l(\theta, \tau\,; \mathbf{y})$ is the log-likelihood of Equation (2). Hence, we actually seek to optimize a doubly penalized likelihood; we penalize kurtosis values too far from K_T while simultaneously penalizing values of τ that are too large. The value in this approach is that it does not require the user to manually adjust values of r in order to find the optimal level of tapering. Instead, a default value of $r = 2$ is used to initialize the optimization, and then the penalized likelihood is optimized with user-specified values for K_T, K_σ, and γ. Sensible default values are $K_T = 3$, the kurtosis of the Gaussian distribution; and $K_\sigma = 0.6$, half the distance from 3 to 1.8, the kurtosis of the Uniform distribution. We have found that setting $\tau_k = \frac{1}{r^2\mu_k}$ and optimizing over the scalar $r > 0$ is quite effective. The choice of penalty coefficient γ is somewhat arbitrary, though we recommend $\gamma = 1/2$. It is worthwhile to note the search for τ is in the region of minimal tapering and the standard ERGM is often chosen.

4.2. Likelihood-based Inference

We can sample from the model (2) using the same type of MCMC procedure as for the standard ERGM (Handcock et al. 2003). These draws can be used to create an MCMC estimate of the log-likelihood (Geyer and Thompson 1992). This, or the penalized likelihood of Section 4.1, can be estimated. For the latter, MCMC estimates of the bias-corrected kurtosis, K_C, can be computed from the same sample so the additional computation compared to the standard ERGM is small.

Krivitsky (2012) review likelihood-based inference for ERGM in finite, super and infinite population scenarios, including the asymptotic normality of the MLE. The standard errors of the MLE are often approximated based on the Hessian of the log-likelihood. The next result gives expressions for the Hessian, providing a minor correction to Equation (4) in Fellows and Handcock (2017).

Theorem 5. At the MLE, the Hessian of the log-likelihood is

$$\frac{\partial l(\theta,\tau\,;\mathbf{y})}{\partial\theta_i\partial\theta_j}\Big|_{\tau,\hat{\theta}_{\text{mle}}} = -\frac{\partial\mu_i(\theta,\tau)}{\partial\theta_j} - 2\sum_k \tau_k \frac{\partial\mu_k(\theta,\tau)}{\partial\theta_i}\frac{\partial\mu_k(\theta,\tau)}{\partial\theta_j}$$

where

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

and c^i is the vector with r^{th} element $c^i_r = \text{cov}(t_r(Y), t_i(Y))$.

Proof.

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

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

$$\begin{aligned}
&\frac{\partial\mu_r(\theta,\tau)}{\partial\theta_i} - \sum_k 2\tau_k\frac{\partial\mu_k(\theta,\tau)}{\partial\theta_i}\text{cov}(t_r(Y), t_k(Y))\\
&\quad = \text{cov}(t_r(Y), t_i(Y))
\end{aligned}$$

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 $B_{rk} = 2\tau_k\,\text{cov}(t_r(Y), t_k(Y))$ and vector c^i with $c^i_r = \text{cov}(t_r(Y), t_i(Y))$. Thus, the correct expression is

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

□

The applications in this article use Hessian-based standard errors, although it is also possible to compute standard errors using a parametric bootstrap around the MLE model fit.

5. Case-Studies of Social Networks

In this section we fit the Tapered ERGM to two real-world networks, each time noting the tapering methodology and the advantages of the model.

5.1. Friendship Structure Among Adolescents

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. While this is a simulated network, the simulation is based on real-data and the simulation is to preserve the privacy of the adolescents.

Additionally, we note there are 677 triangles in the network. We would like to know if these three-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 near-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 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 4 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 1), which in this model are the triangles, isolates, and the edges with zero shared partners (esp(0)) terms. The heavier

Table 1. ERGM fit versus Tapered ERGM fit on Faux Desert High Network.

Term	ERGM	Tapered ERGM
edges	−3.48 (0.10)	−3.49 (0.10)
triangles	−0.008 (0.038)	−0.002 (0.054)
isolates	1.16 (0.47)	1.20 (0.63)
esp(0)	−1.35 (0.13)	−1.35 (0.15)
match.grade.7	2.22 (0.23)	2.19 (0.24)
match.grade.8	2.07 (0.17)	2.05 (0.17)
match.grade.9	1.99 (0.16)	1.98 (0.16)
match.grade.10	1.57 (0.11)	1.57 (0.11)
match.grade.11	1.78 (0.15)	1.77 (0.15)
match.grade.12	1.28 (0.28)	1.28 (0.28)

NOTE: In the Tapered ERGM, the optimal tapering scaling factor of $r = 2.484$ was found with automatic tapering via the kurtosis-penalized likelihood method of Section 4.1, where tapering was done on the dyad-dependent terms.

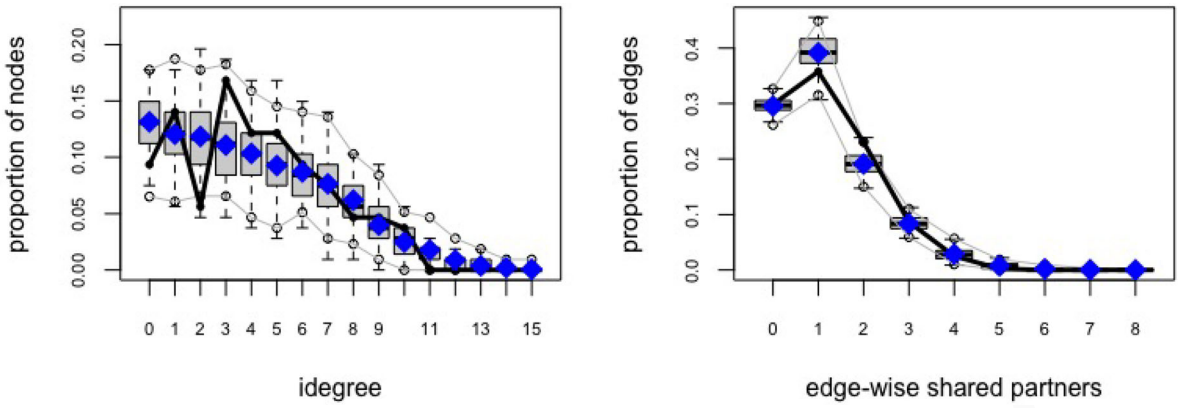

Figure 4. The indegree distribution and edgewise shared partners distribution from 100 networks simulated from the Tapered ERGM MLE compared to the observed network statistics (thick black line), where the Tapered ERGM was fit to the Faux Desert High Network.

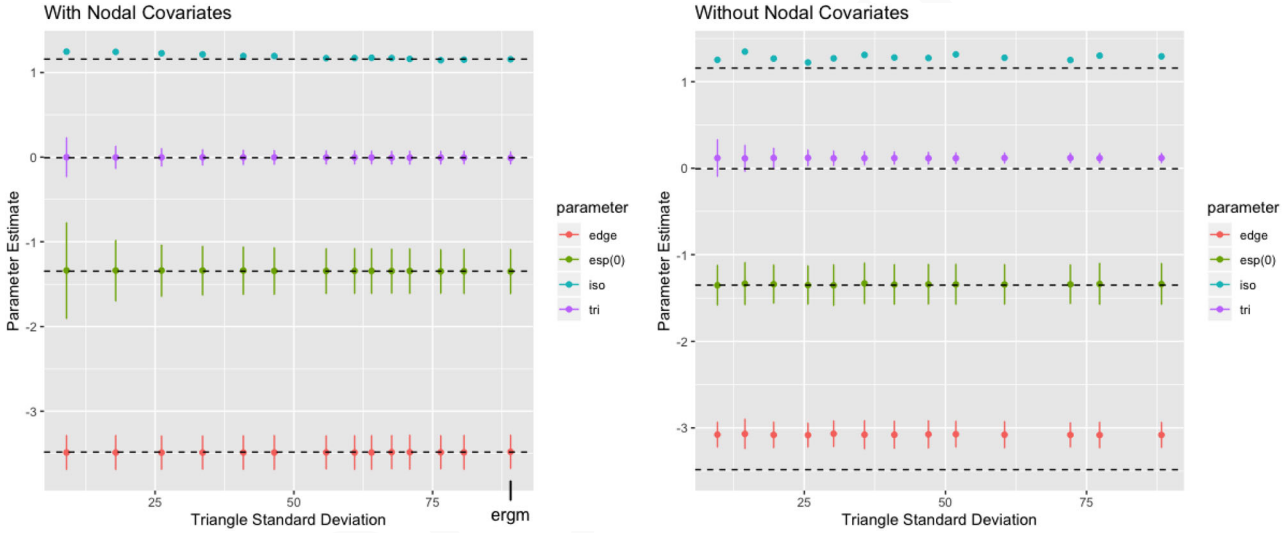

Figure 5. Similarity of parameter estimates across levels of tapering in the Faux Desert High Network. *LEFT*: Tapered models in which the nodal attribute 'grade' is included. The points on the far right of the plot are the estimates from the standard (untapered) ERGM, and the dashed line is set at those 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.

the tapering, the smaller the standard deviation of the counts of each term. The left panel of Figure 5 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 horizontal-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 near-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 5 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 near-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 2 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 near-degeneracy often causes the triangle count to explode as the MCMC progresses), 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.

5.2. Ethnic Heterogeneity in the Activity and Structure of a London Street Gang

The data for this network were 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 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 dataset 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 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 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 an ERGM 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

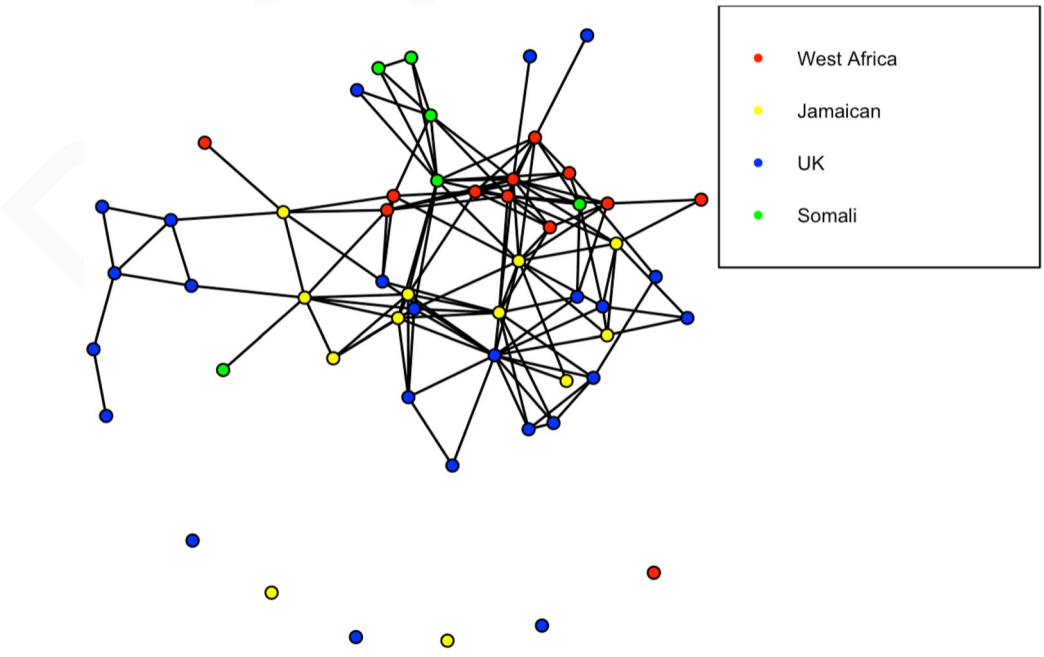

Figure 6. The London Gang Network. A tie exists between two gang members if they have committed at least one crime together. All gang members are Black but the gang is comprised of four distinct ethnicities, categorized by the authors as their countries of origin.

on ethnicity to measure homophily, would provide a conclusion to their hypothesis. Unfortunately, the authors note that counting triangles elicits near-degeneracy and cannot include such terms. As a workaround to measuring the effects of triad closure, they include a geometrically weighted edgewise shared partner (GWESP) term (Snijders et al. 2006) 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 Tapered ERGMs, 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 near-degeneracy during maximum likelihood estimation of parameters. With Tapered ERGMs 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 homogeneous 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 matching term but remove the Somali triangle term, for the purpose of model stability. Table 2 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 automatic tapering via the kurtosis-penalized likelihood method of Section 4.1. In both cases, a mild value of $r = 2$ was used as a starting value, and each time the tapering was allowed to decrease further until the maximum of the penalized likelihood was found ($r = 2.466$ and $r = 2.486$ for Model 1 and 2, respectively).

Unsurprisingly, the Somali matching term is highly significant (as would be a Somali triangle term had it been included instead of the Somali matching 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, and the borderline significant Jamaican matching term in Model 1). 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 borderline 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 7 and 8 show that both Model 1 and 2 provide superior fits to the data than that of the ERGM of Grund and Densley (2015), especially with regard to the edgewise shared partner distribution, further showing the importance of 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. Furthermore, including additional terms in the model, for example, edgewise shared partner terms, would improve the overall fit but were intentionally left out as to give a fair comparison to the ERGM fit by Grund and Densley (2015), who were only concerned with triad closure and ethnic homophily. This example clearly demonstrates the vital need for Tapered ERGMs, since without the ability to fit fundamental terms like the triangle it is very possible to make incorrect inferences.

Table 2. Summary of Tapered ERGMs fit on London Gang Network.

Term	Model 1	Model 2	τ	bias
edges	−3.23 (0.18)***	−3.34 (0.17)***	0.001	−0.0001
triangles	0.68 (0.10)***	0.71 (0.09)***	0.001	−0.0012
triangles(West Africa)	0.11 (0.38)	0.12 (0.37)	0.011	−0.0023
triangles(Jamaican)	0.17 (0.61)	0.41 (0.54)	0.027	0.0000
triangles(UK)	0.56 (0.38)	0.61 (0.42)	0.021	−0.0015
match(West Africa)	0.96 (0.60)	0.95 (0.56)	0.008	−0.0005
match(Jamaican)	1.35 (0.66)*	0.94 (0.55)	0.012	0.0006
match(UK)	0.27 (0.40)	0.31 (0.42)	0.007	−0.0004
match(Somali)	2.17 (0.59)***	2.33 (0.50)***	0.027	0.0004
isolates		0.98 (0.67)	0.027	−0.0027

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

5.3. Supplemental Case-Study: Going to Extremes with the 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, with over 60 million users across the globe (Last.fm 2020). This dataset was collected by Toivonen et al. (2009) and used in their comparative study of social network models. The network is very large, consisting of 8,003 nodes and 16,824 undirected ties. This network contains only mutual friendship structure and does not include any nodal covariates or information on musical preference.

Whereas a standard ERGM may struggle with a network this size, the Tapered ERGM is able to fit the data with relative ease. Nonetheless, fitting a network of this magnitude is not without some difficulties and nuances that are worth mentioning. The large size of the network coupled with the lack of exogenous information requires an extreme amount of tapering in order to achieve a fit. With such heavy tapering, estimation of the standard errors can become strained and results should be

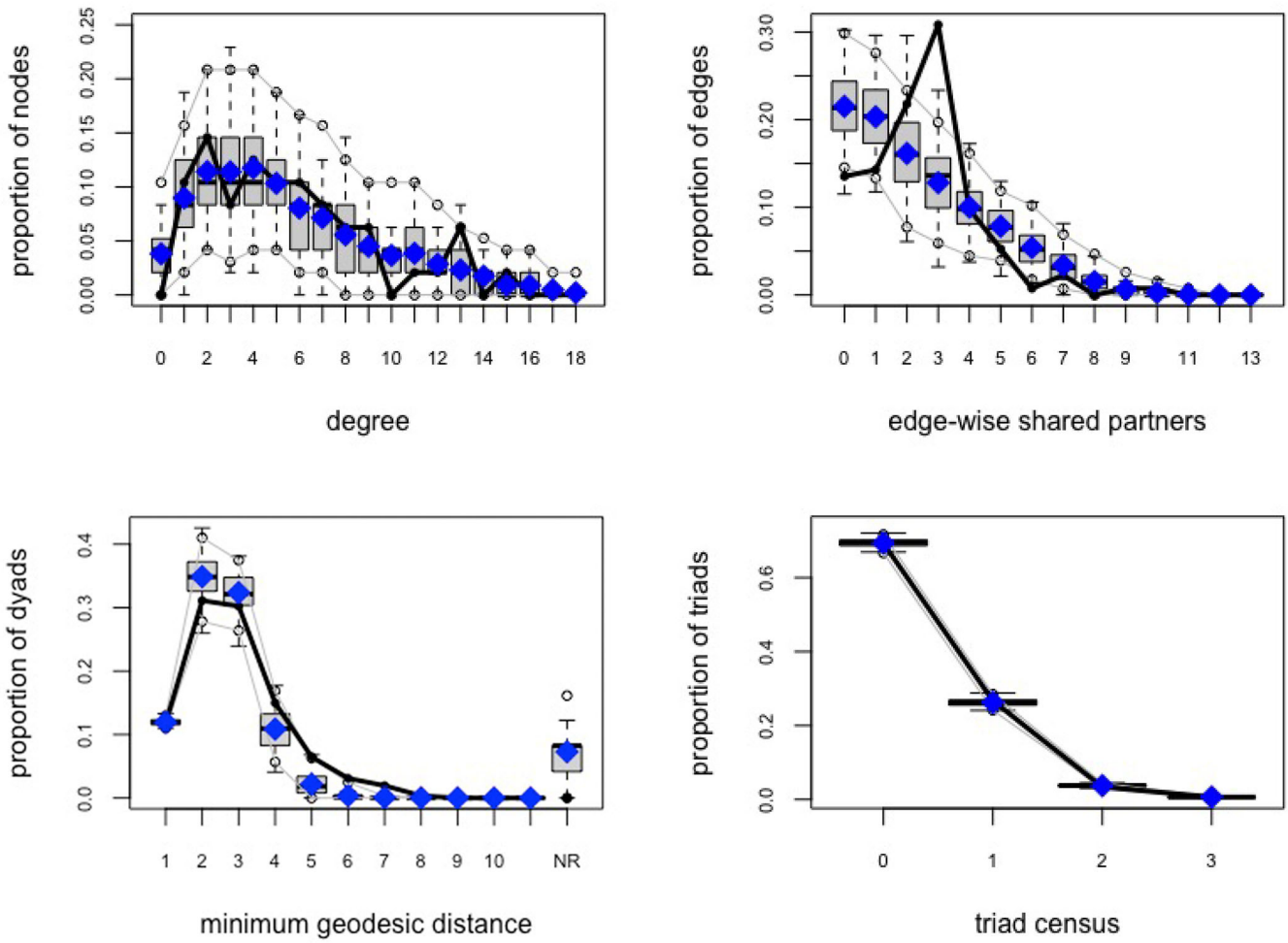

Figure 7. Goodness of fit diagnostic plots for the Tapered ERGM fit on the largest connected component of the London gang network (Model 1 in Table 2).

interpreted carefully. This is an rare example of an extreme case, but in general tapering values are often small and the effect on the standard errors is limited. We refer the reader to the Appendix to "Practical Network Modeling via Tapered Exponential-family Random Graph Models" published in the Journal of Computational and Graphical Statistics for further details.

6. Discussion

For too long, practical modeling via ERGMs has been hindered by concerns about near-degeneracy. Near-degeneracy constrains the space of ERGMs in that many intuitive terms, like the triangle, most often cannot be used within the ERGM as they induce near-degeneracy. The Tapered ERGM of Fellows and Handcock (2017) frees the ERGM, ironically, by constraining it; that is, by placing variance constraints on select statistics the Tapered ERGM can incorporate any term with a guarantee of nondegeneracy. Knowing what level of tapering to use was left as an open question that was unanswered until now. The data provide no insight as to how much tapering is necessary, so we developed two methods here for determining the proper amount of tapering.

In this article we have expounded upon the idea of tapering and have shown the Tapered ERGM to be highly effective in modeling networks. The concept of the kurtosis and why it is appropriate in the context of ERGMs is at the core of how to apply Tapered ERGMs. Employing a novel bias-corrected measure of the kurtosis, we can use a benchmark bimodality coefficient threshold of 0.4 to know if we have tapered enough. This is an integral part of Algorithm 1 which lays out exactly how, what, and when to taper the terms of the Tapered ERGM.

Alternatively, we may also use the kurtosis within a penalized likelihood setting to inform how much tapering is necessary, as outlined in Section 4.1. Theorems 3, 4, and 5 prove that the Tapered ERGM lies on a firm theoretical foundation in addition to its practicality. With all of the benefits and fewer of the downsides of ERGMs, the Tapered ERGM can be used as a replacement for ERGM as the default modeling framework for network analysis. One appealing feature of the Tapered ERGM is that it includes ERGM as a special nested model. Hence, it allows a standard ERGM model to be selected if supported by the data and a better model to be used in cases where the standard ERGM is not appropriate. Tapered ERGMs are also naturally appropriate for curved exponential families. Curved exponential families are complex because the structure is determined by the curved parameterization. However, our penalization still naturally applies and should be well behaved if the curved parameterization itself is.

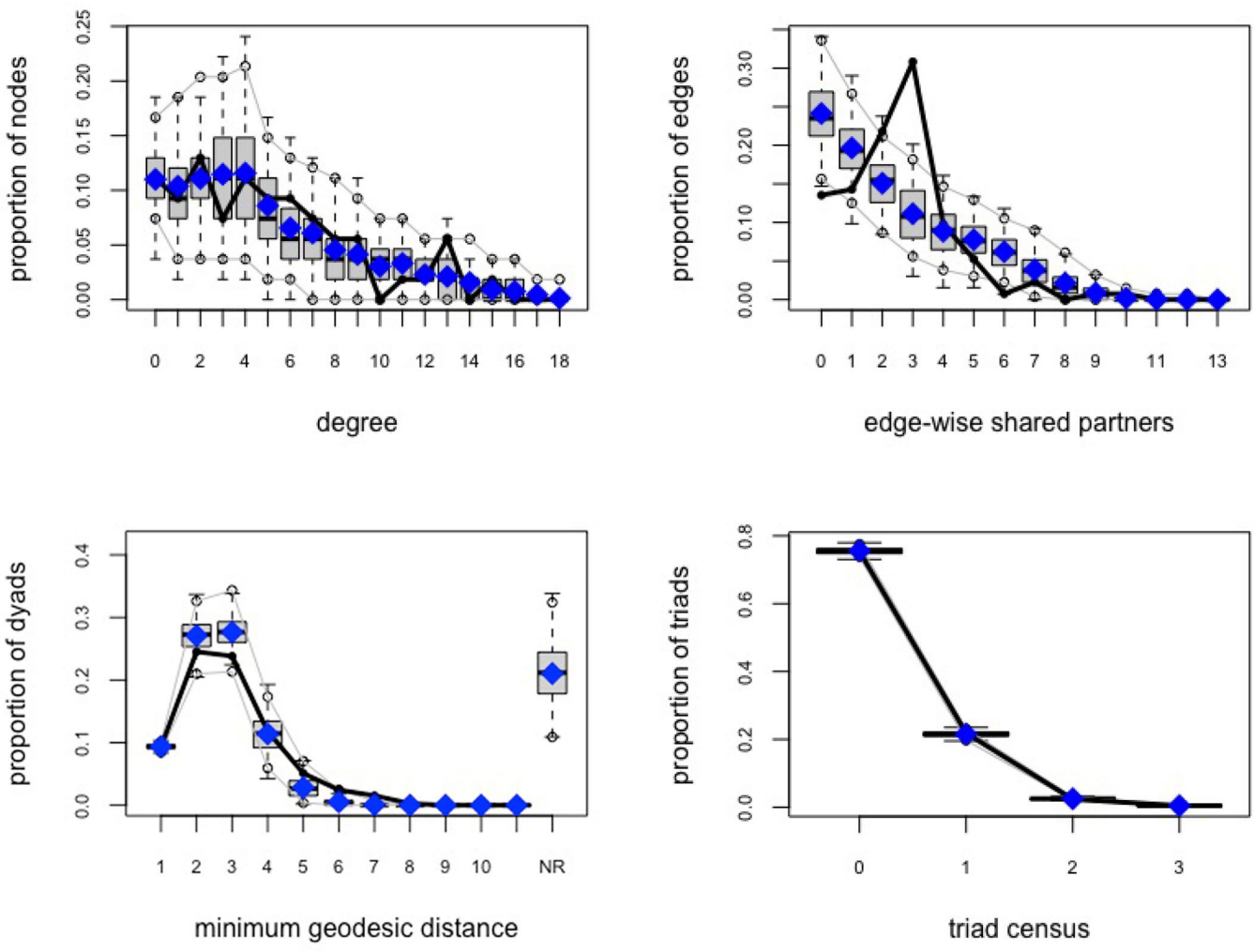

Figure 8. Goodness of fit diagnostic plots for the Tapered ERGM fit on the London gang network (Model 2 in Table 2).

The networks analyzed here provide several insights. The analysis of the friendship network of adolescents 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 scientific analysis. In situations where the data supports severe tapering, one can choose between accepting the tapering and assessing how realistic the terms represent the underlying social processes. The London street gang network demonstrated how important Tapered ERG models really are. Without them, substantively desirable terms like the triangle cannot be fit and incorrect inferences may occur. The analysis of the London street gang network resulted in conclusions nearly polar opposite of those reached by the authors of the original analysis done using standard ERGMs as they were unable to use triangle terms.

It is important to recognize the behavioral modification in modeling, that is, induced by concerns about near-degeneracy. Most practitioners model dependency by including sufficient statistics in the model from a very narrow palate (e.g., GWESP from Snijders et al. 2006). An alternative approach is taken by Wilson et al. (2017) who consider raising network statistics to a positive power less than one. This sub-linear curving produces statistics that are less degenerate as the power deceases, but comes at the cost of making such statistics difficult to interpret. Statistics in these cases are usually chosen not because they make the most sense, but because they provide a computational fit. The Tapered ERGM allows practitioners to fit their model of choice.

Finally, the additional computational burden of Tapered ERGMs is modest. They can be fit using the same MCMC machinery as standard ERGMs. No new terms need to be coded. An open-source R package implementing the methods developed in this article, `ergm.tapered`, (Handcock, Krivitsky, and Fellows 2021; Krivitsky et al. 2003-2020; R Core Team 2020), was used to do the simulation studies and analyze the case-studies. It is publicly available.

Acknowledgments

We are grateful for support from the National Science Foundation BIG-DATA: Applications program, grant NSF IIS-1546259, and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, population research infrastructure grants P2C-HD041041 and P2C-HD041022 and training grant T32-HD007545. We would also like to thank the reviewers for their insightful comments which lead to many improvements to the manuscript.
The authors report there are no competing interests to declare.

Supplementary Materials

The folder *TaperedERGMCode* contains the R packages, R code, and data needed to reproduce the results presented in the manuscript. The README file contained within gives detailed instructions on how to install the `ergm` and `ergm.tapered` packages, as well as descriptions of each individual R file. Also contained within the folder are the exact fitted model objects, saved as RDS files, used in this manuscript.

ORCID

Bart Blackburn http://orcid.org/0000-0002-4954-0855

References

Balanda, K. P., and MacGillivray, H. (1988), "Kurtosis: A Critical Review," *The American Statistician*, 42, 111–119. [5]

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley. [2,6]

Blackburn, T. (2021), "Novel Approaches to Degeneracy in Network Models," Ph.D. thesis, UCLA, *https://escholarship.org/uc/item/5fp7403t*. [3,7]

Chissom, B. S. (1970), "Interpretation of the kurtosis statistic," *The American Statistician*, 24, 19–22. [5]

Darlington, R. B. (1970), "Is Kurtosis Really 'Peakedness?'," *The American Statistician*, 24, 19–22. [5]

DeCarlo, L. T. (1997), "On the Meaning and Use of Kurtosis," *Psychological Methods*, 2, 292. [5]

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. [6]

Fellows, I. E. (2012), "Exponential Family Random Network Models," PhD in Statistics, University of California, Los Angeles, Advisor: Mark S. Handcock. [1]

Fellows, I. E., and Handcock, M. S. (2017), "Removing Phase Transitions from Gibbs Measures," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, eds. A. Singh and J. Zhu, Proceedings of Machine Learning Research, 54, pp. 289–297. [2,3,4,7,8,12]

Geyer, C. J., and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood Calculations,", (with Discussion), *Journal of the Royal Statistical Society*, Series B, 54, 657–699. [8]

Goodreau, S. M. (2007), "Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network," *Social Networks*, 29, 231–248. [1]

Goodreau, S. M., Kitts, J., and Morris, M. (2009), "Birds of a Feather, or Friend of a Friend? Using Statistical Network Analysis to Investigate Adolescent Social Networks," *Demography*, 46, 103–125. [1]

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. [10]

——— (2015), "Ethnic Homophily and Triad closure: Mapping internal gang structure using exponential random graph models," *Journal of Contemporary Criminal Justice*, 31, 354–370. [10,11]

Handcock, M. S. (2003), "Assessing Degeneracy in Statistical Models of Social Networks," Working paper #39, Center for Statistics and the Social Sciences, University of Washington. [2,6]

Handcock, M. S., and Gile, K. J. (2010), "Modeling Social Networks from Sampled Data," *Annals of Applied Statistics*, 4, 5–25. [2]

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2003), *ergm: Fit, Simulate and Analyze Exponential-Family Models for Networks*, Statnet Project. [8]

——— (2008), "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Social Network Data," *Journal of Statistical Software*, 24, 1548–7660. [1]

Handcock, M. S., Krivitsky, P. N., and Fellows, I. (2021), *ergm.tapered: Tapered Exponential-Family Models for Networks*, Los Angeles, CA, R package version 1.1. *https://github.com/statnet/ergm.tapered*. [13]

Hildebrand, D. K. (1971), "Kurtosis Measures Bimodality?" *The American Statistician*, 25, 42–43. [6]

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs. With Comments by Ronald L. Breiger, Stephen E. Fienberg, Stanley S. Wasserman, Ove Frank and Shelby J. Haberman and a Reply by the Authors," *Journal of the American Statistical Association*, 76, 33–65. [2]

Horvát, S., Czabarka, É., and Toroczkai, Z. (2015), "Reducing Degeneracy in Maximum Entropy Models of Networks," *Physical Review Letters*, 114, 158701. [4,6]

Krivitsky, P. N. (2012), "Exponential-Family Random Graph Models for Valued Networks," *Electronic Journal of Statistics*, 6, 1100–1128. [1,8]

Krivitsky, P. N., Handcock, M. S., Hunter, D. R., Butts, C. T., Klumb, C., Goodreau, S. M., and Morris, M. (2003-2020), *statnet: Software Tools for the Statistical Modeling of Network Data*, Statnet Development Team. [13]

Last.fm. (2020), "About Us," *https://store.last.fm/pages/about-us*, Accessed March 10, 2020. [11]

Moors, J. J. A. (1986), "The Meaning of Kurtosis: Darlington Reexamined," *The American Statistician*, 40, 283–284. [5]

R Core Team. (2020), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [13]

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," *Journal of the American Medical Association*, 278, 823–832. [3,8]

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models," *Electronic Journal of Statistics*, 3, 446–484. [2]

Schweinberger, M. (2011), "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families," *Journal of the American Statistical Association*, 106, 1361–1370. [2]

Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. R. (2020), "Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios," *Statistical Science*, 35, 627–662. [2]

Shalizi, C. R., and Rinaldo, A. (2013), "Consistency Under Sampling of Exponential Random Graph Models," *The Annals of Statistics*, 41, 508–535. [2]

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153. [2,11,13]

Strauss, D. (1986), "On a General Class of Models for Interaction," *SIAM Review*, 28, 513–527. [1,2]

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. [11]

van Duijn, M. A. J., Handcock, M. S., and Gile, K. J. (2009), "A Framework for the Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models," *Social Networks*, 31, 52–62. [5]

Westfall, P. H. (2014), "Kurtosis as Peakedness, 1905–2014. RIP," *The American Statistician*, 68, 191–195. [5,6]

Wilson, J. D., Denny, M. J., Bhamidi, S., Cranmer, S. J., and Desmarais, B. A. (2017), "Stochastic Weighted Graphs: Flexible Model Specification and Simulation," *Social Networks*, 49, 37–47. [13]