**Title**
Reduced Degeneracy Statistics for Exponential-family Random Graph Models and Latent Space Network Models for Rating



**Author**
Carlen, Jane

**Publication Date**
2018

UNIVERSITY OF CALIFORNIA
Los Angeles

# Reduced Degeneracy Statistics for Exponential-family Random Graph Models and Latent Space Network Models for Rating

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Jane Carlen

2018

ABSTRACT OF THE DISSERTATION

Reduced Degeneracy Statistics for
Exponential-family Random Graph Models and
Latent Space Network Models for Rating

by

Jane Carlen
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Mark S. Handcock, Chair

With a rise in the amount of network data comes increased need for flexible and interpretable network models. Exponential-family random graph models (ERGM) are widely used to analyze small- to medium-sized networks, but suffer from model degeneracy which detracts from their application. In Part I of this dissertation we address this problem by developing novel statistics for ERGM. We focus on the modeling of transitivity in networks as it is a key feature of many real-world networks, but most attempts to account for it within ERGM have induced model degeneracy. The statistics we propose combine the strategies of the transformed statistics proposed by Horvát et al. (2015) and the regularized statistics proposed by Fellows (2012b). They include statistics to capture transitivity, clustering, and a new class of moment statistics to improve goodness of fit. We characterize our newly introduced statistics along with those of Horvát et al. (2015) and Fellows (2012a) using recent theoretical developments regarding ERGM degeneracy. We also compare them theoretically and in practice to the geometrically weighted statistics of Snijders et al. (2006) that are currently the most commonly used to model transitivity in ERGM.

In Part II of this dissertation we develop models to rate and rank items based on network data, and demonstrate many advantageous properties of these models. The impetus for this work came from research on ranking statistics journals by Varin et al. (2016). They

present a *quasi-Stigler* model that is a great improvement over the commonly used but statistically indefensible Impact Factor, especially in the quantification of ratings uncertainty. However, the quasi-Stigler model does not fully leverage the network structure of the data and underestimates uncertainty. In addition to applying latent space models to the network rating problem, we identify a fast computational method for fitting the models. We also develop a new latent network model that leverages the symmetric and asymmetric patterns in directed relational data. This model has many potential applications beyond item rating.

The dissertation of Jane Carlen is approved.

Jacob Foster

Alyson Fletcher

Qing Zhou

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2018

*To John, who waited patiently.*

*To my parents, who always supported me.*

*To my sister, a ficus.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# VITA

| | |
|---|---|
| 2002–2006 | B.A. (Mathematics) and B.A. (Political Science), Haverford College. |
| 2012–2017 | Teaching Assistant and Special Reader, Statistics Department, UCLA.<br>TA: Courses including Introduction to Probability and<br>Introduction to Design and Analysis of Experiments.<br>Reader: Courses including Matrix Algebra and Optimization,<br>Linear Models, and Social Statistics. |
| 2013, 2014<br>Summer | Statistical Software Intern, Fellows Statistics, UCLA.<br>Assistant developer of software to implement respondent driven sampling.<br>Funded by Center for Disease Control, Grant Number 3U2GPS001468-04S1<br>(Subaward Agreement Number 7438sc). |
| 2015, 2016<br>Summer | Academic Mentor, Research in Industrial Projects for Students, IPAM.<br>Mentored student teams in research for Arete Associates and Google.<br>Derived probabilistic bounds on frequency of DNA strings and<br>implemented recommendation algorithms for Yelp data. |
| 2014–2016 | Assistant Editor, Journal of Statistical Software.<br>Managed submissions and communications with authors and editors. |
| 2017 | Graduate Student Researcher, Statistics Department, UCLA.<br>NSF-supported project: Scalable Model-Based Inference for<br>Social Networks from Complex Sampling Designs. |
| 2016–present | Editor, Journal of Statistical Software.<br>Oversee reviews of papers on network modeling software. |
| 2017–present | Research Statistician, Economic Roundtable.<br>Analyze wage stagnation of union workers at Disneyland.<br>Design triage tool to allocate public resources to fight homelessness. |

# PUBLICATIONS

*"Discussion on the paper 'Statistical Modelling of Citation Exchange Between Statistics Journals' by Varin, Cattelan and Firth,"* with Mark S. Handcock. Non peer-reviewed, invited discussion. *Journal of the Royal Statistical Society A*, 179, 43-44.

*"A Latent Space Network Modelling Approach to Ratings with Applications to Journal and Film Rating,"* revised and resubmitted. *Journal of the Royal Statistical Society C.*

Part I

# Reduced Degeneracy Statistics for Exponential-family Random Graph Models

# CHAPTER 1

# Introduction

Network analysis is increasingly applied not only to social networks, but to any relational data that can be placed in a network format. Exponential-family random graph models (ERGM) are widely used to analyze small- to medium-sized networks of up to a few thousand nodes. Applications span numerous scientific fields, including:

- Epidemiology: to predict paths of disease spread and factors influencing spread (Goodreau et al., 2010) and estimate prevalence (Morris et al., 2009);
- Political Science: to analyze alliance formation (Cranmer et al., 2012a,b), political collaboration and planning (Gerber et al., 2013);
- Biology: to characterize the structure of the human brain (Simpson et al., 2011, 2012) and the architecture of biological networks (Saul and Filkov, 2007);
- Sociology: to track diffusion of ideas and formations of social structure in schools and professional settings (Goodreau et al., 2009; Wimmer and Lewis, 2010; Zappa and Mariani, 2011).

ERGM are well-suited to these applications because they are versatile, interpretable and easy to draw simulations from.

Although the study of statistical properties of social networks first emerged in the 1930s, ERGM were not proposed until the "second generation" of study beginning in the 1970s (Wasserman and Pattison, 1996). Holland and Leinhardt (1981) introduced ERGM for directed graphs, motivated by the study of social relationships. They advocated the model

class because of its ability to capture reciprocity in relationships and varying levels of popularity of actors, what they termed *differential attraction*.

A distinguishing feature of network data - what makes it both rich and challenging - is relational dependence (or *dyad dependence*). This concept expresses that the existence or strength of a relationship between two actors can affect other relationships in the network. The models of Holland and Leinhardt (1981) did not incorporate dyad dependence. Frank and Strauss (1986) introduced models in which relationships could be dependent if sharing a common actor.

One example of a dyad-dependent phenomena is transitivity: "A friend of my friend is a friend of mine." The tendency to form closed triangles in a social network is well-observed in real-world social networks, but problematic for modeling. It induces *model degeneracy*, and we describe how this appears in practice in Chapter 2. As a result, capturing dependent relationships in ERGM has demanded major compromises in model fit and/or interpretability (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson, 1992b; Wasserman and Pattison, 1996; van Duijn et al., 2009).

In Part I of this dissertation we develop novel statistics to increase flexibility and interpretability of ERGM while reducing degeneracy, paying particular attention to transitivity. In Sections 1.1 and 1.2 we introduce the model and several related models. In Section 1.3 we describe available goodness-of-fit measures for ERGM, which we will use to compare the efficacy of novel statistics. One of the difficulties in understanding and diagnosing model degeneracy is that several working definitions of degeneracy have appeared in the literature. We catalog and compare them in Chapter 2. In Chapter 3 we describe newly proposed statistics to model transitivity and evaluate their degeneracy using the criteria introduced in Chapter 2. In Chapter 4 we propose additional statistics to model clustering and higher order moments of the degree and shared partner distributions of a network. We apply the full collection of newly introduced statistics to model a real-world online social network in Chapter 5, illustrating the improvement over previously available models. We conclude with a summary of our work and discussion of ongoing challenges in Chapter 6.

## 1.1 Exponential-family Random Graph Models

ERGM assign the following probability to a network $y$ in $\mathscr{Y}_n$, the set of all graphs with $n$ nodes:

$$P(Y = y|\theta, X) = \frac{\exp(\theta^\top t(y, X))}{c(\theta, X)}, \text{ where} \tag{1.1}$$

- $\theta \in \mathbb{R}^q$ is a vector of parameters.
- $X$ is a set of covariates which may be nodal ($X_k$ a vector) or dyadic ($X_k$ a matrix).
- $t(y, X)$ is a vector of sufficient statistics.
- $c(\theta, X) = \Sigma_{y \in \mathscr{Y}_n} \exp(\theta^\top t(y, X))$ is the normalizing constant, usually intractable.

We refer to a node in the network as $y_i$ or $i$ and a dyad as $y_{ij}$, $1 < i, j < n$, with $y_{ii} = 0$ by convention. For directed networks $y_{ij}$ is distinct from $y_{ji}$, and for undirected networks $y_{ij} = y_{ji}$, calculating $t$ and $c$ accordingly. In this treatment we focus on networks that are undirected and binary: $y_{ij} \in \{0, 1\}$. Note, $\theta$ is sometimes replaced with $\eta(\theta)$. A linear $\eta$ function does not affect our interpretation, but non-linear $\eta$ extends the formulation to curved ERGM (Hunter and Handcock, 2006; Schweinberger, 2011).

The number of possible graphs configurations is $\mathcal{O}(2^n)$, which makes exact computation of the likelihood intractable for all but very small networks. This prevents exact computation of the maximum likelihood estimate (MLE) of $\theta$, where MLE $\equiv argmax_{\theta \in \Theta} P(Y = y_{obs}|\theta, X)$, for networks with dyad dependence. However, Snijders (2002) introduced a Markov chain Monte Carlo (MCMC) algorithm to find the MLE, drawing on earlier work by Geyer and Thompson (1992a). Subsequently, Hunter and Handcock (2006) proposed an alternate method of MCMC estimation of the MLE (MCMC-MLE). This method estimates the relative likelihood of a current and proposed parameter vector by sampling from the current parameter model. Sampling proceeds via Metropolis algorithm, in which an edge toggle is proposed and accepted with probability $\min(1, \pi)$, where:

$$\pi = \frac{P(Y_i j = 1 - y_{ij} | Y_{ij}^c = y_{ij}^c)}{P(Y_i j = y_{ij} | Y_{ij}^c = y_{ij}^c)} = \begin{cases} \theta^\top \delta_{ij} & \text{if } y_{ij} = 0, \\ -\theta^\top \delta_{ij} & \text{if } y_{ij} = 1. \end{cases}$$

$\delta_{ij}$ is referred to as the *change statistic* for $y_{ij}$ and is equal to $t(y_{ij} = 1, y_{ij}^c) - t(y_{ij} = 0, y_{ij}^c)$, the difference in the sufficient statistics when $y_{ij} = 1$ versus when $y_{ij} = 0$, with the rest of the network held constant. Once a sample is drawn, parameter updates are generated via Newton-Raphson algorithm. This framework for model estimation is used throughout our examples.

The change statistics, $\delta_{ij}$, are not only useful for model fitting, but also interpreting model output. The probability of an edge, conditioned on the rest of the network, can be put in terms of $\delta_{ij}$:

$$P(Y_{ij} = 1 | Y_{ij}^c) = \frac{P(Y_{ij} = 1, Y_{ij}^c)}{P(Y_{ij} = 1, Y_{ij}^c) + P(Y_{ij} = 0, Y_{ij}^c)} \frac{\exp(\theta^\top \delta_{ij})}{\exp(\theta^\top \delta_{ij}) + 1} \quad = \text{logit}^{-1}(\theta^\top \delta_{ij}) \quad (1.2)$$

### 1.1.1 Sufficient Statistics

An advantage of ERGM is the range of sufficient statistics or terms that can be used to capture network behavior. A small sample of these statistics and the dynamics they are intended to capture is described below. A more complete list of statistics and description can be found in (Handcock et al., 2008) and under `ergm-terms` within the **ergm** package (Handcock et al., 2015).

- *Edges* or *Density*: The count or proportion of edges in the network, respectively. These statistics measure overall activity levels or sociality in the network. In terms of edge count or density, most observed social networks are sparse. In general, density does not increase at pace with network size.

- *Degree/k-stars*: A degree term counts the number of nodes with a specific degree, $k$, and a $k$-degree node is also referred to as a $k$-star, as shown in Figure 1.1. We may be

interested in a specific degree statistic because it reflects a unique feature of the social structure or because it is an artifact of data collection methods.

- *Triangles* and *k-triangles (ESP)*: A triangle term counts the number of closed triangles in the network. The count is broken down into triad types for directed networks. A $k$-triangle refers to an edge that serves as the base for $k$ triangles, as shown in Figure 1.1. It is also referred to as an edge with $k$ shared partners or $ESP_k$. Count statistics of these formations are used to measure transitivity and, to some extent, clustering in the network. However, both counts are also correlated with network density.

- *Twopaths* and *k-twopaths (DSP, NSP)*: The twopath statistic counts the number of two-step paths from $y_i$ to $y_j, i \neq j$ over all $i < j$. If $y_{ij} = 0$, it is termed a non-edgewise $k$-shared partner ($NSP_k$). The count of dyadwise $k$-shared partners, $DSP_k$, includes all $NSP_k$ and $ESP_k$. $DSP$ and $NSP$ measure network connectivity and can serve as a control on the $ESP$ counts. They also measure clustering to some extent, and the distribution of these statistics is connected to network modularity.

- *Homophily* or *Mixing*: A homophily or "nodematch" statistic is a count of edges whose endpoints have the same level of a specific covariate. Mixing terms generalize this to count matches and non-matches. These terms measure preferential attachment between sub-populations in the graph. Homophily in a graph may be indicative of clustering, but a model with only edge and homophily terms is not guaranteed to replicate this structure.

An additional class of geometrically weighted terms was introduced by Snijders et al. (2006) and is discussed in Section 3.1.

### 1.1.2 Software

The proliferation of ERGM owes in part to the readily available software for implementation. Most notably, the `statnet` family of network packages for `R` includes the **ergm** package for fitting, analyzing, and visualizing ERGM. The statistics we develop are incorporated into

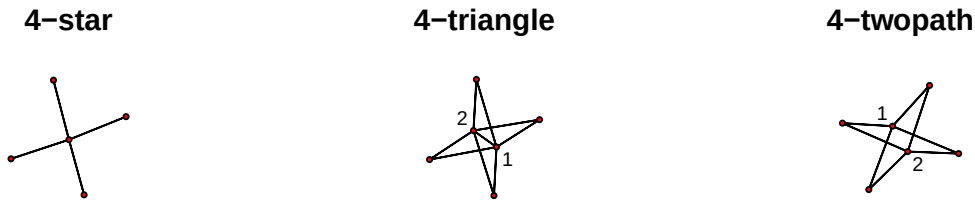


Figure 1.1: A k-star is a node with degree k. A k-triangle is k distinct triangles that share one common edge. A k-twopath is k distinct paths of length two joining one pair of nodes.

the **ernm** package, developed for exponential-family random network models, a superset of ERGM in which nodal covariates may be random (Fellows, 2012a). The functionality of the package mirrors **ergm**, but facilitates the future application of our work to ERNM as well as ERGM.

### 1.1.3 Related Models

There are several sub-classes and extensions of ERGM which have evolved to reduce model degeneracy or increase flexibility.

- *Dyad Independent ERGM*: These models assume that edges form independently in the network. As such, they can be fit by logistic regression with output variable $\delta_{ij}$. This is termed maximum pseudolikelihood estimation (MPLE), and is unbiased if all statistics in the model are dyad-independent. If not, it can generate a very biased estimate of the MLE (van Duijn et al., 2009). The appeal of this model class is that it is not degenerate and scales well to large networks. See for example Handcock (2003) and van Duijn et al. (2009), and for an application example see Traud et al. (2012).

- *Valued ERGM*: Valued ERGM extend the ERGM framework to allow for weighted ties. Though potentially very powerful for valued graphs, the formulation also expands

the size of the graph space potentially infinitely. To account for this, a reference distribution is placed on the dyads, such as a binomial or Poisson distribution, whose parameters must be estimated. For details see Krivitsky (2012) and Krivitsky and Butts (2015).

- *Hierarchical ERGM with Local Dependence*: Schweinberger and Handcock (2015) introduced these models due to the problem of "large and growing neighborhoods" in traditional ERGM. The model breaks an observed graph into local neighborhoods, and assumes dyad dependent relationships within neighborhoods, but independent relationships between neighborhoods. By keeping neighborhoods small, they open the door for using statistics that are asymptotically degenerate. This makes it even more important to refine our understanding of model degeneracy.

## 1.2 Alternative Network Models

We mention two other other popular network models which avoid some of the limitations of ERGM, and to which the solutions we propose can be compared or extended. We will refer to these models in a comparison of results later in the paper.

### 1.2.1 Stochastic Blockmodels

One of the most commonly used network models, especially to capture group structure, is the stochastic blockmodel (SBM). The guiding principle of blockmodels is *structural equivalence*, the idea that conditioned on block membership the within-block and between-block edge probabilities are independent and identical for each member of a block. Items in the same block form an equivalence class. More formally, for a random binary graph $Y$ under a basic SBM with $G$ blocks,

$$P(Y_{ij} = 1|g) = \eta(x_i, x_j) \tag{1.3}$$

where $g$ is an n-length vector of class assignments, and $\eta$ is a $G \times G$ matrix of block-to-

block edge probabilities (Nowicki and Snijders, 2001). If $Y$ is symmetric, i.e., the graph is undirected, then the lower triangle of $\eta$ is redundant. For early presentations see Lorrain and White (1971), Holland et al. (1983), Snijders and Nowicki (1997), and (Nowicki and Snijders, 2001). Recent advances have added flexibility to SBM, for example the mixed membership stochastic blockmodels of Airoldi et al. (2008), models with covariates implemented by INRA and Leger (2015), and the degree-corrected blockmodels of Karrer and Newman (2011). These extensions allow for flexibility in the equivalence of nodes, but the assumption of conditional edge independence holds. As a result, SBM do not fully capture dyad-dependent forces such as transitivity. This is evidenced by their typically underpredicting triangle counts in real-world networks (Fortunato and Hric, 2016).

### 1.2.2 Latent Space Models

Latent space network models assume that nodes occupy a position in latent space, usually in low dimension. The distance between nodes is inversely correlated to the probability of a tie between them. Additional characteristics may also influence edge probability, such as sender and receiver effects for each node, and cluster assignment. Dyads are independent conditional on positions and covariates, which ensures the models are tractable. We discuss latent space models at length in Part II, and we refer the reader to, for example, Hoff et al. (2002), Handcock et al. (2007), and Krivitsky et al. (2009b).

An advantage of latent space models is that visualizations from the model have a clear interpretation in terms of the model parameters. The dependence structure is also more complex than an SBM due to the flexibility of latent positions. The Euclidean norm is the most common distance function used in latent space models of undirected networks. Using this distance, the triangle inequality induces transitivity in the network, as the distances from node $i$ to node $k$ and $k$ to $j$ place an upper bound on the distance from $i$ to $j$ (Hoff, 2003). However, the level of transitivity is a function of the distance function, rather than being fit to the observed network. This limits the ability of the model to capture transitivity in real-world networks.

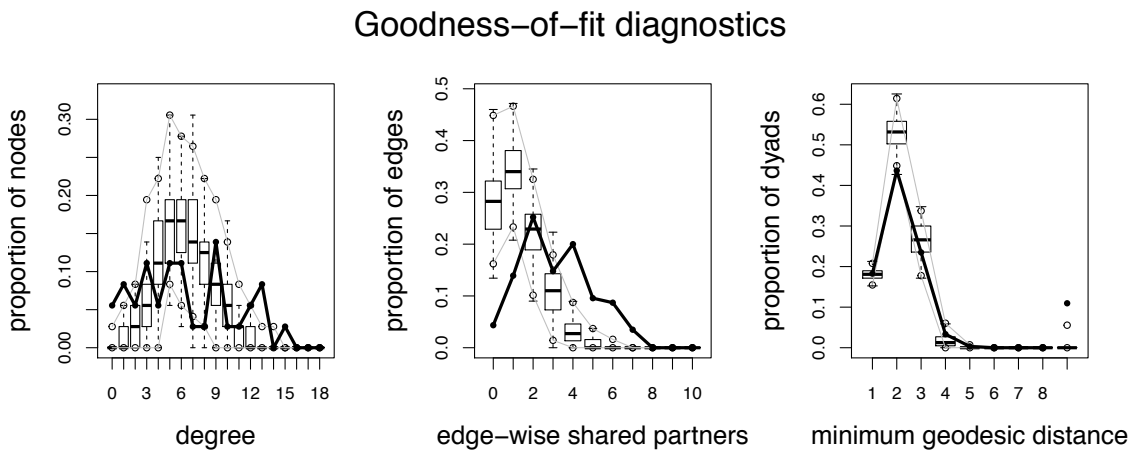


Figure 1.2: Sample graphical goodness-of-fit plots.

## 1.3 Goodness of Fit

In this section we introduce goodness-of-fit measures that we will use to compare models. Methods such as cross-validation are not directly applicable because the sample size for dyad-dependent models is usually one, i.e., we have a single network observation from which to base inference. As a result, simulation based methods are more prevalent.

### 1.3.1 Graphical Goodness of Fit

The most commonly used diagnostic for ERGM is graphical goodness of fit (GGOF). This is a simulation-based method available in **ergm** through the `gof.ergm` and `plot.gofobject` functions. Example output of `plot.gofobject` is presented in Figure 1.2.

Figure 1.2 shows boxplots for statistics over 100 simulations from the underlying ERGM. The statistics of the observed graph are shown as heavy black lines. The distributions shown are of degree counts, edgewise shared partner counts, and minimum geodesic distances. These statistics track sociality, clustering, transitivity, and centrality, which are often of interest to researchers (Hunter et al., 2008). Typically either none of these statistics are included in the model, or only a small subset are. The plots are not meant to capture the accuracy of the MCMC MLE, which if equal to the true MLE has expected statistics equal

to observed statistics (Barndorff-Nielsen, 1978), but the extent to which the global behavior in the network is captured by the parsimonious set of terms in the model. In the example above it is clear that the model is not capturing the full behavior of the network. The degree distribution is overly peaked and the ESP distribution is too right-skewed.

This graphical diagnostic provides no quantification of how poorly the model fits the data, which can make it difficult or cumbersome to compare the goodness of fit for several models. The `gof.ergm` function does provide simulation-based p-values for each individual statistic shown in the plots. However, these p-values are not adjusted for multiple comparisons between dependent statistics. In practice they do not provide much information beyond the corresponding plots.

### 1.3.2 AIC and BIC

The classic model selection tools, Akaike information criterion (AIC) and Bayesian information criterion (BIC), can be applied to ERGM. They are defined: $AIC \equiv 2(|\theta|) - 2\ln(L)$ and $BIC \equiv \ln(n)|\theta| - 2\ln(L)$, where $L$ is the maximum value of the likelihood function, $|\theta|$ is the number of estimated parameters, and $n$ is the number of independent observations. Despite their simple expressions, both are difficult to apply to ERGM.

Calculating $L$ directly is not possible due to the intractable normalizing constant, as mentioned in Section 1.1. However, we can express the log likelihood of the null model with $\theta = \mathbf{0}$ as the negative log of the number of possible graphs, $-\binom{n}{2}log(2)$, and it holds that

$$l(\theta) = [l(\theta) - l(\mathbf{0})] + \binom{n}{2}log(2). \tag{1.4}$$

Hunter and Handcock (2006) build on these facts to express the estimated log likelihood, $\hat{l}(\theta)$ as a function of two intermediate log likelihood ratios that are estimated by path sampling. See Equation 2.4 and Section 5 of Hunter and Handcock (2006) for, respectively, the formulation of $\hat{l}(\theta)$ and details of the path sampling algorithm. An additional challenge for BIC is that the seemingly straightforward $n$ value is not straightforward for dyad-dependent ERGM. (For dyad-independent models it is simply the number of dyads.) The true value of

$n$ is somewhere in the range $[1, \binom{n}{2}]$; **ergm** conservatively uses the maximum of the range to estimate BIC.

Neither AIC nor BIC have been shown to be particularly helpful in tuning and selecting ERGM. Their maximum likelihood criterion is not indicative of particular features of interest to a researcher, nor do they provide insight into how to improve a model. Accordingly, we do not apply them in our examples in later chapters. See Simpson et al. (2011) for an applied comparison of AIC and GGOF.

### 1.3.3 Spectral Goodness of Fit

Spectral goodness of fit (SGOF), introduced by Shore and Lubin (2015a), responds to the need for an interpretable, quantitative measure of goodness of fit. It is based on the graph Laplacian, which is defined for an undirected graph $Y$ as $L = D - A$, where $D$ is a diagonal matrix whose elements are the nodal degrees of $Y$, and $A$ is the adjacency matrix corresponding to $Y$. The eigenvalues of this matrix can be interpreted in terms of the connectivity of $Y$. The number of zero eigenvalues indicates the number of disconnected components in the network. The non-zero eigenvalues measure fine-grain levels of connectivity. They are indicative of the number of ties that must be cut to separate additional components (Shore and Lubin, 2015a).

The *spectral goodness of fit (SGOF)* is defined:

$$SGOF(model|Y_{obs}) = 1 - \frac{\overline{ESD}_{obs,fitted}}{\overline{ESD}_{obs,null}} \tag{1.5}$$

where $\overline{ESD}_{obs,fitted}$ is the mean Euclidean distance from the normalized spectrum of the observed Laplacian to that of graphs simulated by the fitted model. $\overline{ESD}_{obs,null}$ is the same quantity, but for simulations from the null model. The null model is determined by the user, but the most common choice is the edge- or density-only model.

A benefit of SGOF is that its output is on a consistent scale of $[-\infty, 1]$, which shrinks to $[0, 1]$ if the model is an improvement over the null. Like GGOF, SGOF also captures vari-

ability in model simulations. For example, the SGOF of the model underlying the GGOF example above is **0.046** (-0.182, 0.302), calculated using the `spectralGOF` package of Shore and Lubin (2015b). The interval indicates the 5th and 95th percentile of $ESD_{obs,fitted}$ with $\overline{ESD}_{obs,null}$ held fixed. The mean value indicates that only about five percent of graph structure beyond density is captured by the model.

Since SGOF, like GGOF, is simulation-based it can compare models regardless of functional form. Furthermore, SGOF is a holistic measure of multiple structural features. This can be an advantage, but also a drawback because it does not indicate where the model is failing. The tools for visualizing spectral error in the `spectralGOF` package can be of some use in this respect. SGOF succeeds where GGOF and AIC/BIC are weak because of its consistent and interpretable scale, but does not supply the detailed information of GGOF.

Because SGOF measures graph structure through the spectrum, it only captures statistics which are invariant to permutations of node labels. Homophily and mixing, for example, are not captured unless they influence the structure of the graph. In the above example, the model includes two significant homophily terms but neither seems to impact the structure strongly. Another weakness which we do not consider at length here is that SGOF can only be applied to undirected graphs, at least in its current form. The hurdles to extending SGOF to directed graphs are addressed by Shore and Lubin (2015a).

# CHAPTER 2

# ERGM Degeneracy

The concept of ERGM degeneracy encompasses several problematic behaviors. Some authors have attempted to theoretically distinguish its definition from its pathology, including Handcock (2003), Schweinberger (2011), and Horvát et al. (2015), and we will summarize this work below. In practice, degenerate models are often discovered when a model fails to fit. In **ergm** the simulated statistics from the MCMC sampling routine will signal empty or full models, or an oscillation between them. In the latter case, the variance of the simulated statistics would be extremely inflated, as parameter estimates jump between disparate values. This is not a failure of the MCMC routine, but of the model itself. What makes this even more difficult in practice is that some models known to be degenerate, such as the edge-triangle model, will still fit successfully in some cases. We would like to identify degenerate cases *a priori*.

One early method for illustrating model degeneracy uses the fact a small change in the degenerate statistic can create a sizable jump in one or more expected statistics, as formalized by Handcock (2003). This *phase transition* behavior, borrowing a term from physics literature, can be shown graphically by fixing the parameter values of non-degenerate statistics, and simulating networks under a range of values for the parameter of the degenerate statistic. An example of this from Schweinberger (2011) is shown in Figure 2.1. The x axis represents the range of the degenerate parameter. The plot identifies a symptom of model degeneracy, but not the underlying cause.

Attempting to model transitivity or clustering is often the proximate cause of model degeneracy. A workaround is to employ a dyad independence assumption and MPLE as described in Section 1.1.3, but this introduces bias and sacrifices the relational dependence

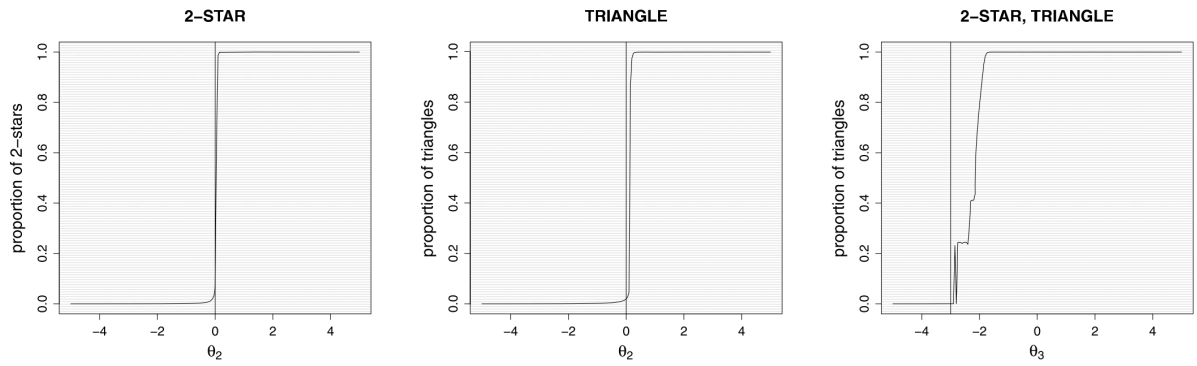


Figure 2.1: Proportion of two-stars (left panel) and triangles (middle and right panel) plotted against two-star parameter (left and middle) and triangle parameter with two-star parameter fixed at one (right). Reprinted from Schweinberger (2011).

that is the strength of network models. As Schweinberger and Handcock (2015) write, "most relational phenomena are dependent phenomena."

A more relaxed requirement than dyad independence is *Markov dependence* (Frank and Strauss, 1986), defined as follows:

- Dyads $y_{ij}$ and $y_{kl}$ are only *neighbors* if they share a node.
- If $y_{ij}$ and $y_{kl}$ are not neighbors, they are independent conditional on the rest of the graph.

The number of neighbors of a dyad in a Markov graph is therefore $2(n-2)$. Although Markov dependence seems only a shade stronger than dyad independence, it imparts a far-reaching dependence as $n$ grows (Schweinberger and Handcock, 2015).

Transitivity is intrinsically a Markov-dependent property, and practitioners attempting to model it currently face a trade-off between reality and degeneracy. Given this fundamental connection between transitivity and degeneracy, as well as the importance of transitivity in real-world networks, our primary goal in developing non-generate models is to develop a non-degenerate transitivity term. This is discussed in detail in Chapter 3.

## 2.1 Definitions of Degeneracy

Despite great attention to model degeneracy in the ERGM literature, there is no agreed upon definition of degeneracy. In addition, the term *degeneracy* is sometimes interchanged with two related concepts: model *instability* and *sensitivity*. Below we present two formal definitions of degeneracy for ERGM:

1. The MLE does not exist. This occurs if and only if $t(y_{obs})$ is on the *boundary* of the convex hull, $C$, of realizable graph statistics $t(Y) : Y \in \mathscr{Y}_n$ (Handcock, 2003). (Note we drop $X$ in the notation, but it is implied.)

2. $P^*(t(Y); \theta_{MLE})$, the smoothed probability of sufficient statistics $t(Y)$ under the MLE, is multimodal, with probability mass concentrated around two or more disjoint and well-separated (by $\mathcal{O}(1)$ distances) domains in $C$ (Horvát et al., 2015).

We also present two definitions of *near degeneracy*:

1. Let $\mu(\theta) = E_\theta(t(Y))$, the mean value statistics of a model.

   The model is *near degenerate* if $\mu(\theta)$ is close to the boundary of $C$, where

   "boundary" refers to the relative boundary of $C, rbd(C) = cl(C) \setminus rint(C)$.

   "Close" is described in several ways, including:

   a) $\mu(\lambda e) \to sup_{u \in rint(C)}(e^\top \mu)e$, where $\lambda \in \mathbb{R}$, and $e$ is a unit vector in $\in \mathbb{R}^q$.

   b) For every $d < sup_{u \in rint(C)}(e^\top \mu), P_{\lambda e}(e^\top t(Y) \leq d) \to 0$ as $\lambda \to \infty$ (Handcock, 2003).

2. $\theta$ defines a *near degenerate* model if for any $\epsilon \in (0, 1)$ :
   $P_\theta(\theta^\top t(Y) > (1 - \epsilon) max_{Y \in \mathscr{Y}_N}[\theta^\top t(y)]) \to 1$ as $N \to \infty$ (Schweinberger, 2011).
   $N$ indicates degrees of freedom for the graph, which is $\binom{n}{2}$ for an undirected graph with $n$ nodes. We assume without loss of generality that $min_{Y \in \mathscr{Y}_N}[\theta^\top t(y)] = 0$.

The definitions of degeneracy are distinct. In particular, the first states that the MLE does not exist while the second assumes that it exists, and does not depend on an asymptotic

condition. In the former respect, it is closer to the *near degeneracy* of Handcock (2003) and Schweinberger (2011).

On the other hand, the two definitions of near degeneracy are similar. Both imply that, under some limiting condition, all of the weight of the distribution is placed on a small, "uninteresting" subset of $\mathscr{Y}_n$, usually empty or full graphs (Handcock, 2003; Schweinberger, 2011). Under the first definition, the limiting condition applies to all parameter vectors associated with a choice of sufficient statistics. The limiting behavior occurs as the parameter vector extends in any direction. There is an implication that the sufficient statistics, if multiple, are not on the same scale. In contrast, the second definition applies to a certain choice of $\theta$ though it may be easy to show for large ranges of $\theta$. The limiting behavior is as the size of the network grows, with parameter values held fixed.

Finally, we present two definitions of *instability* and one of *excessive sensitivity*:

1. A discrete exponential family distribution is *unstable* if for any $C > 0$ there exists constant $N_C > 0$ such that $max_{Y \in \mathscr{Y}_n}[\theta^\top t(y)] > CN \; \forall \; N > N_C$ (Schweinberger, 2011). As a reminder, N indicates degrees of freedom for the graph, which is $\binom{n}{2}$ for an undirected graph with $n$ nodes. This condition implies that a sufficient statistic grows at a higher order than edge count for sufficiently large networks.

2. A small change in $\theta$ causes a large change in the probability structure of the model (Handcock, 2003). This is distinct from the stability of Schweinberger (2011), but related to his near degeneracy and to phase transitions.

*Excessive sensitivity*: Let *sensitivity* refer to nearest-neighbor log odds, $\theta^\top \delta_{ij}$, where $\delta_{ij}$ is the change statistic for $y_{ij}$. A model is excessively sensitive if sensitivity is unbounded as $N \to \infty$ (Schweinberger, 2011). This requires only that the maximum sensitivity over all dyads is asymptotically unbounded.

## 2.2 Theorems of Degeneracy

The following theorems of Schweinberger (2011) connect his three concepts of near degeneracy, instability, and sensitivity. (Note: The statement of theorems is condensed for clarity. See Schweinberger (2011) Theorems 1, 2, 4 and Corollary 2 for complete formulations.)

*Theorem 1*: If a discrete exponential family distribution $P_\theta, \theta \in \Theta$ is unstable it is excessively sensitive.

*Theorem 2*: If a discrete exponential family distribution $P_\theta, \theta \in \Theta$ is unstable it is near degenerate.

Again, these definitions and theorems apply to specific choices of $\theta$, but can sometimes be shown for large swaths of $\Theta$, e.g., all non-zero $\theta$. Not only do these theorems connect several ideas in the degeneracy literature, they help to identify degenerate models in practice. Stability and sensitivity are easier to observe than the relative position of $t(y_{obs})$ in $C$. For example, the triangle count for the full graph of $n$ nodes is $\binom{n}{3} \approx n^3$ or $\mathcal{O}(N^{1.5})$. Any model with a non-zero triangle parameter is therefore unstable. By Theorem 1 it is also excessively sensitive. We can also establish excessive sensitivity directly by noting that removing one edge from the full graph removes $n-2$ triangles, so $\delta_{ij}$ is unbounded as $N \to \infty$.

The definition of degeneracy of Horvát et al. (2015) implies a different strategy for identifying degenerate models:

*Theorem 3*: $P_\theta$ is non-degenerate iff the density of states $\mathscr{N}(t(Y))$ is strictly log-concave, where $\mathscr{N}(t(Y))$ is number of graphs with sufficient statistics $t(Y)$ (Horvát et al., 2015).

Theorem 3 readily establishes the degeneracy of the edge-triangle model - degeneracy in the sense of Horvát et al. (2015). The density of states for the edge-triangle model cannot be

log-concave because the domain of sufficient statistics is not convex. The expected triangle count conditioned on the number of edges, $e$, is $\mathcal{O}(e^3)$. This suggests taking a cube root of the triangle count as a sufficient statistic to convexify the domain of sufficient statistics when edge count is included in the model (Horvát et al., 2015). A comparison of a large sample of the domain before and after taking the cube root transformation is shown in Figure 2.2 for $n = 16$, suggesting that it has been made nearly convex. Note, however, that Horvát et al. (2015) do not prove the log-concavity of the density of states after their proposed transformation. In addition, there is no general rule to generate convex transformations.

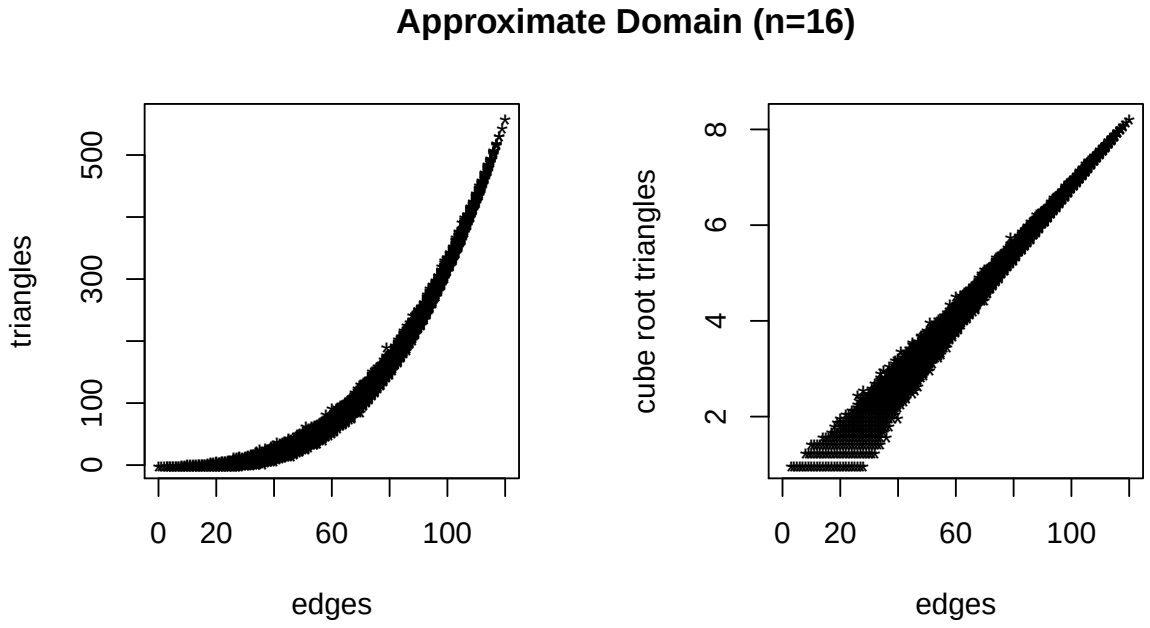


Figure 2.2: After a cube root transformation of triangle count, the domain of sufficient statistics appears nearly convex. The plot shows unique statistics from random samples of 500 networks at each possible edge count.

# CHAPTER 3

# Network Transitivity Terms

In this chapter we examine the degeneracy of four transitivity terms besides the triangle count, including the cube root of the triangle count proposed by Horvát et al. (2015). The first is a geometrically weighted term that has frequently been used instead of the triangle count. The second is a regularized term which originated in the **ernm** package, but was never formally analyzed or applied (Fellows, 2012a). The last is a novel statistic we developed which combines the strategies of Horvát et al. (2015) and Fellows (2012b).

The concept of transitivity is colloquially understood as "a friend of my friend is a friend of mine," and associated with the tendency for triangles to close. Yet each of the terms we describe in this chapter implies a unique definition of transitivity. For example, the regularized transitivity term measures triangle density while controlling for edge density. We compare edge + transitivity models of a small network for each of the four terms discussed, applying goodness-of-fit diagnostics discussed in Section 1.3.

## 3.1 Geometrically Weighted Shared Partners

Because of the degeneracy of the triangle and two-star statistics, Snijders et al. (2006) developed new statistics to capture transitivity in ERGM. These terms were reformulated by Hunter (2007) as the geometrically weighted dyad shared partner (GWDSP), and geometrically weighted edgewise shared partner (GWESP) statistics:

$$GWDSP(y;\alpha) = e^{\alpha}\sum_{k=1}^{n-2}\{1-(1-e^{-\alpha})^{k}\}DSP_k(y). \tag{3.1}$$

$$GWESP(y;\alpha) = e^{\alpha}\sum_{k=1}^{n-2}\{1-(1-e^{-\alpha})^k\}ESP_k(y). \tag{3.2}$$

$ESP_k(y)$ is the number of edges with exactly $k$ shared partners, and $DSP_k(y)$ is the number of dyads with exactly $k$ shared partners. $\alpha$ is a decay parameter. (This parameterization is used by **ergm** and **ernm**.) Larger values of the decay parameter make for slower decay (Hunter, 2007). Although the decay parameters can be estimated as in the curved exponential family models of Hunter and Handcock (2006), here we assume they are fixed. The above formulations correspond to the `gwdsp` and `gwesp` functions respectively in **ergm**. GWESP captures, in a constrained way, the distribution of $k$-triangles. GWDSP can be thought of as a control to GWESP, by accounting for the number of $k$-twopaths, though in practice the correlation between the terms means they are not often included in the same model. In applications in later chapters we focus on the GWESP term because it is more closely associated with transitivity.

The change statistics of these terms are better behaved than those of triangle or two-star statistics. As long as $0 < e^{-\alpha} < 1$, which can be assumed in practice, the change statistics are positive, but tend to flatten as nodal degree increases, and become negligible where $k$ is large. An interpretation of this in context is that if a tie does not exist between two nodes despite many shared partners, adding more shared partners does not significantly increase the chance of the tie (Snijders et al., 2006). This mitigates against the "avalanche" effect "in which the MCMC routine, once a few new edges are created in the graph, is quickly forced to add edge after edge until the complete graph is reached" (Hunter and Handcock, 2006).

Many authors have used the geometrically weighted terms to successfully fit small and medium sized networks. For example, Goodreau (2007) used them to fit data from the National Longitudinal Study of Adolescent Health (AddHealth), including a 1681-node friendship network. Using a model with a `gwesp` term, Goodreau (2007) captured the full degree and edgewise shared partner distribution well, and all but the tail of the geodesic distance distribution, which is underpredicted. However, the data in that case had low density and triangle count – only 1236 edges (0.00087 density) and 157 triangles.

### 3.1.1 Constraints of Geometrically Weighted Terms

By geometrically weighting counts, these statistics impose a constraint on the natural parameter space. Instead of having a parameter for each possible order of $k$-triangle or $k$-twopath, each geometrically weighted term adds only one parameter to the model. This makes the existence of an MLE much more likely (Hunter and Handcock, 2006). However, it implies:

$$\theta_{S_{k+1}} - \theta_{S_k} = (1 - e^{-\alpha})^k$$

where $S_k$ stands in for $DSP_k$ or $ESP_k$. Under the usual condition that $\alpha > 0$, the right side is positive and the implied parameter on $S_k$ increases with $k$, but the increase is negligible after a certain $k$, usually less than 15. In **ergm** the sum is taken for $k$ up 30 by default. While $\alpha$ allows flexibility in setting the rate of decay, we can imagine networks where the constraints of GWESP and GWESP do not reflect the observed data, especially for denser networks.

### 3.1.2 Stability and Sensitivity

It was established by Schweinberger (2011) that the ERGM of edge + GWESP and edge + GWDSP are unstable if the geometrically weighted coefficient is not equal to zero and $e^{-\alpha} > 2$, i.e., $\alpha < -log(2)$. This may not be a hindrance to their application because such $\alpha$ values are out of the commonly used range. However, either model is excessively sensitive. The maximal degree for a node in $Y_n$ is $n-2$. Closing an isolated $(n-2)$-twopath to make an $(n-2)$-triangle results in a $2(n-2)$ change in GWESP or GWDSP, as all non-base edges in the $(n-2)$-triangle are now the base for one triangle. Meanwhile, the edge statistic only changes by one. The closing of the $k$-triangle itself has no effect on GWDSP, and its effect on GWESP approaches a constant if $e^{-\alpha} \leq 2$, or may itself cause instability otherwise.

### 3.1.3 Density of States

We consider through small examples the log-concavity of the density of states for the edge + GWESP and edge + GWDSP models. We first sketch the domain for a graph of 16 by

plotting unique sufficient statistics from 500 randomly drawn networks at each possible edge count, with the decay parameter $\alpha = 0.5$. This does not give a complete picture of the domain, but does indicate its denser regions. The approximate domain of edge + GWESP statistics (Figure 3.1, top left) is closer to convex than for the edge + triangle model (Figure 2.2, left). The edge + GWDSP domain (Figure 3.1, right) is non-convex.

To view the exact domains and log density of states for these statistics we must consider an even smaller example of n = 7. Figure 3.2 plots the log density of states for two-statistic sets containing an edge count and transitivity statistic. The middle row of Figure 3.2 shows that the domain of GWESP(.5) is roughly convex. A smoothed version of the log density of states may be concave; though un-smoothed, it is locally jagged. The plot for GWDSP(.5) (middle row, right) is not log-concave and the domain is not convex.

## 3.2 Cube Root of Triangles

### 3.2.1 Stability and Sensitivity

Next we evaluate the cube root of triangles statistic proposed by Horvát et al. (2015). We refer to this term in software as `triangles3`. The edge + triangles$^{1/3}$ model is stable, as the maximal edge count for a graph of $n$ nodes and $N$ dyads is N, and the maximal cube root of triangle count is $\mathcal{O}(n)$. As a result, $\theta^\top t(y)$ can be bound by $NC$.

The model is nonetheless excessively sensitive, as the maximum of $\delta_{triangles^{1/3}}$ is unbounded as $n \to \infty$. Consider a graph with only an isolated $(n-2)$-twopath. If we turn the $(n-2)$-twopath into an $(n-2)$-triangle by adding a single edge then the cube root of triangles increases $\mathcal{O}(n^{\frac{1}{3}})$. However, such a network is rare, and as the density of triangles increases the maximum change statistic decreases.

### 3.2.2 Density of States

As in Section 3.1, we consider via examples the log-concavity of the density of states for the edge + triangles$^{1/3}$ model. Figure 3.1 (right) plots unique sufficient statistics from 500

randomly drawn networks at each possible edge count for $n = 16$. Figure 3.2 (top right) shows the complete log-density of states for $n = 7$. The domain of edge + triangles$^{1/3}$ statistics appears roughly convex and $log(\mathcal{N}(t(y)))$ appears approximately concave.

These plots raise a potential problem with this statistic, which is that it captures density as well as transitivity. The domain of sufficient statistics narrows for high-density networks, where edge and cube root of triangle statistics are strongly correlated.

## 3.3 Regularized Transitivity

The final transitivity terms we consider originated in work by Fellows (2012a). He proposed a regularized triangle count over all nodes, though we adapt this statistic to count over edges. We refer to the statistic as *regularized transitivity*, or *transitivity* when it is clear that we are referring to this statistic rather than the general concept. In software we refer to it as `transitivity`. It is a count over all edges of the square root of observed triangles incident on that edge minus the expected square root of triangles incident on the edge.

$$transitivity(y) = \sum_{i,\ j>i} \sqrt{ESP(i,j)} - E(\sqrt{ESP(i,j)} \mid deg(i), deg(j), n) \tag{3.3}$$

where $ESP(i,j)$ is the number of triangles incident on edge $y_{ij}$ and $deg(i)$ is the degree of node $y_i$.

The expected count of triangles on an edge given the degrees of its endpoints is hypergeometric, where $deg(i) - 1$ is the number of successes and $deg(j) - 1$ is the number of draws, or vice versa. The population size is $n - 2$. The expectation of this distribution is

$$E(ESP(i,j)|deg(i), deg(j), n) = \frac{(deg(i) - 1)(deg(j) - 1)}{n - 2}.$$

By Jensen's inequality, the square root of the expectation is an upper bound on the expected square root. Because there is no closed form expression for the expectation of the hypergeometric square root, the software calculates it directly.

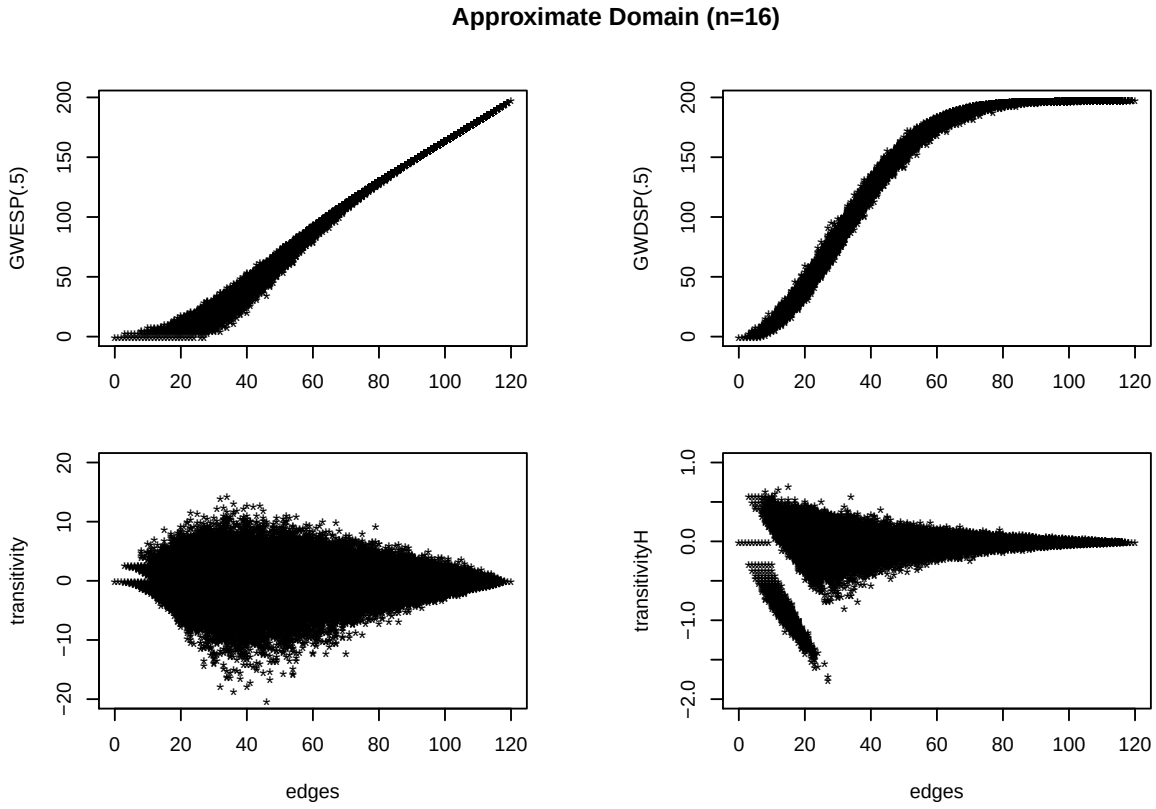


Figure 3.1: Samples of sufficient statistics for n = 16, generated from 500 samples from each edge count.

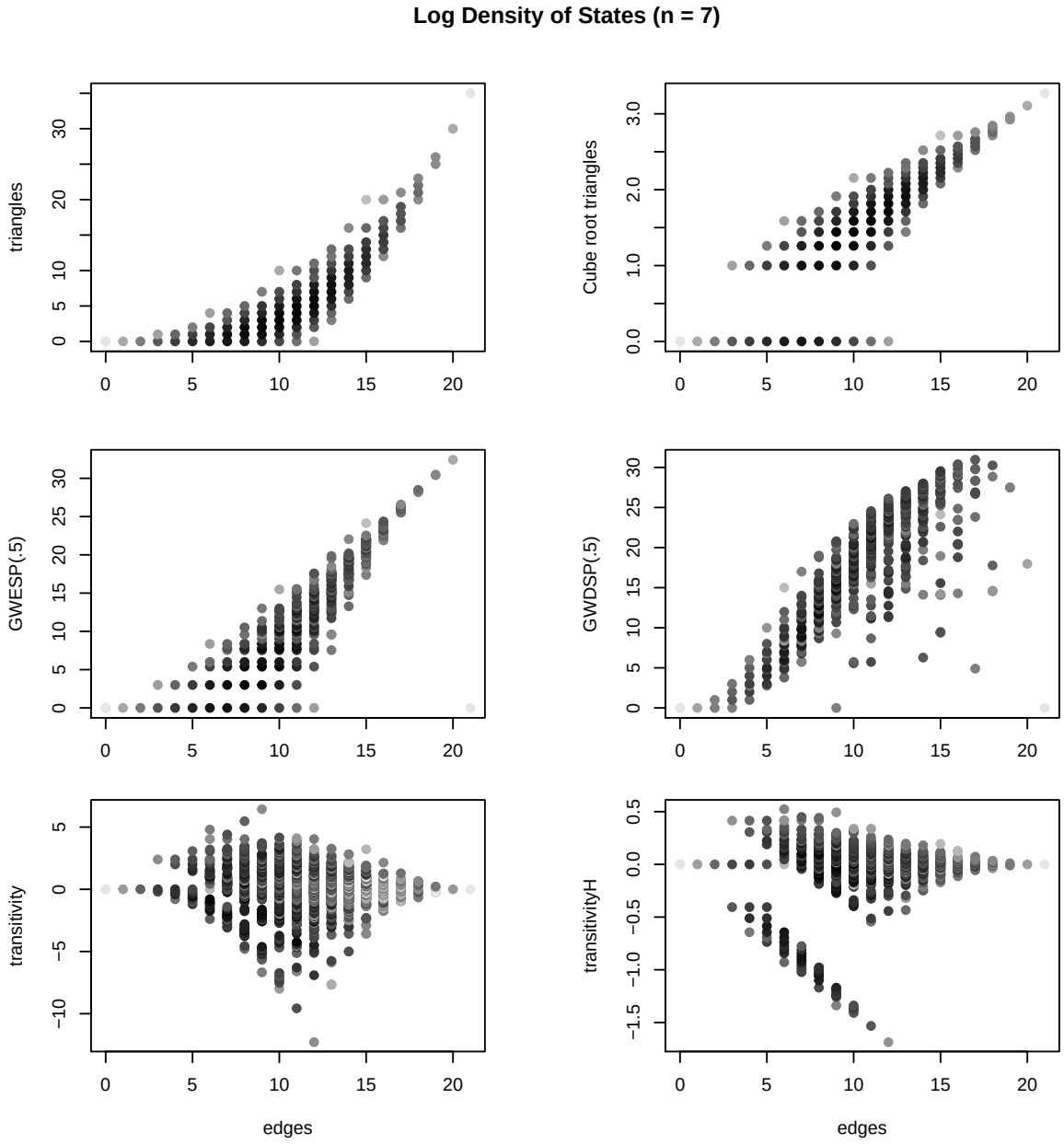


Figure 3.2: Log density of sufficient statistics for n = 7. Darker shades indicate higher density.

Fellows (2012b) provides heuristic arguments for the form of this statistic. Subtracting the expectation balances it against the "avalanche" effect referenced above (Hunter and Handcock, 2006). When network density is low, the expectation of triangle counts is very low and the observed triangle count dominates the statistic. Taking the square root prevents the term from behaving like the triangle count statistic. However, even with the square root adjustment this can be problematic, as the expectation fades for large, low-density graphs. In our **ernm** implementation the user may set an alternate root, most likely the cube root.

### 3.3.1 Stability and Sensitivity

The regularized transitivity statistic is unstable. For $\theta_{transitivity} < 0$, consider the left network of Figure 3.3 with nodes divided into two groups and a complete bipartite network between them. It has $\lfloor \frac{n^2}{4} \rfloor$ edges and no triangles. The expected square root of triangles per edge, i.e., the subtracted term in Equation 3.4, is greater than $\frac{1}{2}\sqrt{\frac{n}{4}}$. The transitivity statistic is approximately $-\frac{n^2}{4}\frac{1}{2}\sqrt{\frac{n}{4}}$. This is $\mathcal{O}(N^{1.25})$ and the product of the statistic with $\theta_{transitivity} < 0$ is positive.

For $\theta_{transitivity} > 0$, consider the right network of Figure 3.3 with complete cliques within each half of the nodes. The number of edges in this network is approximately $\frac{n^2}{4}$, the square root of number of triangles on each edge is approximately $\sqrt{\frac{n}{2}}$, and the expected number of triangles per edge is approximately $\frac{n}{4}$. By Jensen's inequality, the square root of the expectation is greater than or equal to the expectation of the square root, implying that transitivity is at least $\mathcal{O}(n^{2.5}) = \mathcal{O}(N^{1.25})$. The fact that we are summing over $\mathcal{O}(N)$ edges drives the value in this case. This suggests taking the root over the whole network triangle count and expected triangle count, rather than for each edge, to stabilize the statistic.

The regularized transitivity statistic is also excessively sensitive. Consider a graph of an isolated $(n-2)$-twopath, as in Section 3.1.2. If we add one edge to make a $(n-2)$-twopath a $(n-2)$-triangle, the regularized transitivity jumps from approximately $-2(n-2)$, with one unrealized triangle per edge, to 0. After the edge is added, observed triangles match expectation. The maximum change statistic is asymptotically unbounded.

Figure 3.3: (Left) Maximal edge, minimal triangle network. (Right) Two fully connected cliques.

### 3.3.2 Density of States

The sampled domain of edge + transitivity is roughly convex for the 16-node graphs shown in Figure 3.1 (bottom, left). It stands out for having little correlation between edge count and transitivity. However, the log density of states for $n = 7$, Figure 3.2 (bottom, left), does not appear concave, and the exact domain appears less convex.

## 3.4 Alternate Form of Regularized Transitivity

The lack of correlation between edge count and regularized transitivity is a desirable feature for modeling. However, the instability and excessive sensitivity of regularized transitivity imply near degeneracy. In light of this, we suggest an alternate form of the regularized transitivity statistic. We take the cube root rather than a square root, and apply it to the sum of triangle counts and expected triangle counts, rather than to each edge individually. Because the form of this modified statistic owes to Horvát et al. (2015), we refer to this statistic below and in software as *transitivityH*.

$$transitivityH(y) = \sqrt[3]{\sum_{i,\ j>i} ESP(i,j)} - \sqrt[3]{\sum_{i,\ j>i} E(ESP(i,j)|deg(i), deg(j), n)} \qquad (3.4)$$

### 3.4.1 Stability and Sensitivity

We established in Section 3.2.1 that the cube root of triangles is $\mathcal{O}(n)$ for the complete network. The sum of expected triangle counts also ranges from zero to $\mathcal{O}(n)$, indicating that transitivityH is stable. However, it is excessively sensitive. For example, using the same case as for regularized transitivity, if we add one edge to a graph with only an $(n-2)$-twopath to form a $(n-2)$-triangle, transitivityH goes from approximately $-(2(n-2))^{1/3}$, with one unrealized triangle per edge, to 0.

### 3.4.2 Density of States

The sampled domain of edge + transitivityH is not convex for the 16-node graphs shown in Figure 3.1 (bottom, right). The log density of states for $n = 7$, Figure 3.2 (bottom, right), appears closer to concave than regularized transitivity in the region not including zero-triangle graphs. The mass appears more concentrated in a central area, and because there is almost no correlation between edge count and transitivityH, the problem induced by a non-convex domain of averaging realized statistics over separated dense regions is less likely to occur.

## 3.5 Comparison of Phase Transitions

We conclude the comparison of transitivity terms with an examination of phase transition behavior. Phase transitions are symptomatic of model degeneracy, and observed when a slight change in a parameter value significantly alters the probability structure of the model, as described in Section 2. In Figure 3.4 we plot expected statistics for five types of edge + transitivity model, one with each of the transitivity terms discussed. The underlying graph size is 36, and the edge coefficient is held fixed at $-3$, while the transitivity parameter varies. The expected statistics are estimated from 1000 simulations at each parameter value.

The edge + triangle model (top, left) shows an obvious phase transition as the triangle parameter crosses 0.7. The cube root of triangles (top, right) and GWESP(.5) (middle, left)

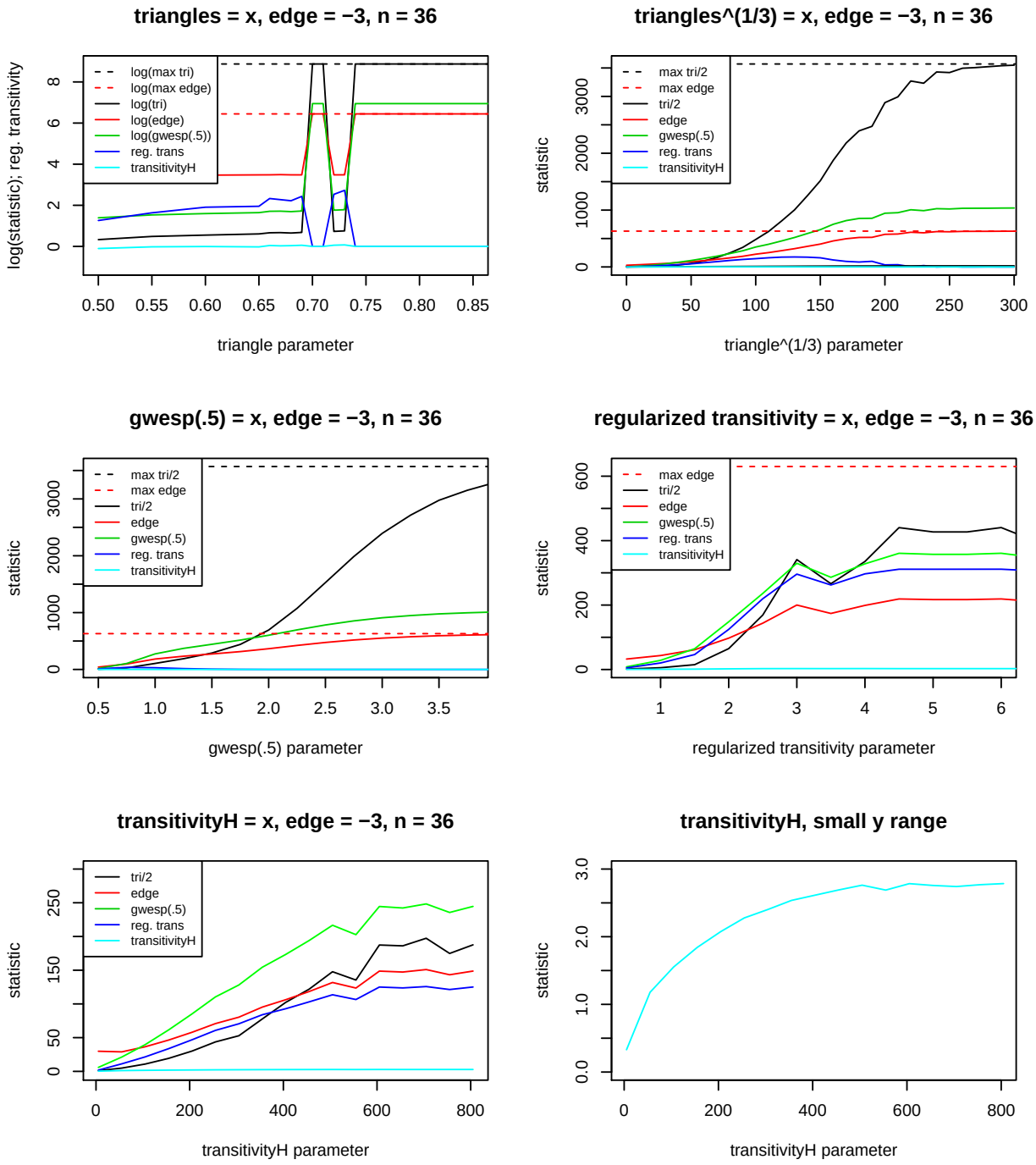


Figure 3.4: Trends in statistics as transitivity parameters increase for edge + transitivity term models. For the triangle model (top left) most statistics are presented on a log scale as noted in the legend. In other plots triangle count is halved to improve visiblity of trends.

models do not evidence a transition, though they do have an inflection point within a region of steeper increase. This is more exaggerated for larger graphs. The regularized statistics are the only ones that achieve a maximum well before the complete network and decrease to zero as network density approaches one.

The behavior of the models including regularized transitivity (middle, right) and transitivityH (bottom) are similar in that the maximum edge and triangle counts are not achieved. After a region of increase, higher values of these parameters do not impact expected statistics. The maximum achievable difference between triangle and edge count has already been reached. The bottom right panel makes visible the trend for transitivityH in the bottom left panel. Unlike other terms, the increase in transitivityH as the corresponding parameter increases is concave, and the concomitant increase in edges and triangles is almost linear. This is not suggestive of phase transition behavior.

## 3.6 Example: Model Comparison

Several of the transitivity statistics discussed above are stable, but all are excessively sensitive, and they vary with respect to phase transition behavior. To examine their behavior in practice we model a small network, the Lazega Lawyer Network. The 36 nodes in the network represent 36 law partners working in one of three offices (Lazega, 2001). An edge is said to exist between two partners if and only if both indicate that they collaborate with the other (Hunter and Handcock, 2006). There are 115 edges (undirected) and several nodal covariates including practice type (litigation or corporate), office location (Boston, Hartford or Providence), and seniority. The network is plotted twice in Figure 3.5, with nodes colored by practice type (left) and office location (right). Edges are drawn with a solid line if endpoints match with respect to the visualized covariate; otherwise the line is dashed. The placement of points is determined by a Fruchterman-Reingold force-directed algorithm (Fruchterman and Reingold, 1991). The algorithm balances repulsive forces that separate nodes with spring-like attractive forces on connected nodes. This spreads the nodes relatively evenly throughout the plotting area and imparts visual clarity.

Table 3.1: Selected parameter estimates and expected triangle counts from four models of the Lazega Lawyer Network. Models are listed by their transitivity statistic.

| | edges | trans | practice | nodeMatch practice | E(tri) |
|---|---|---|---|---|---|
| GWESP(.5) | -5.97 | 1.03 | 0.28 | 0.68 | 126.29 |
| triangles$^{1/3}$ | -4.40 | 30.95 | 0.25 | 0.72 | 121.01 |
| reg. transitivity | -3.89 | 0.85 | 0.41 | 0.60 | 102.62 |
| transitivitH | -3.52 | 47.53 | 0.39 | 0.56 | 100.52 |

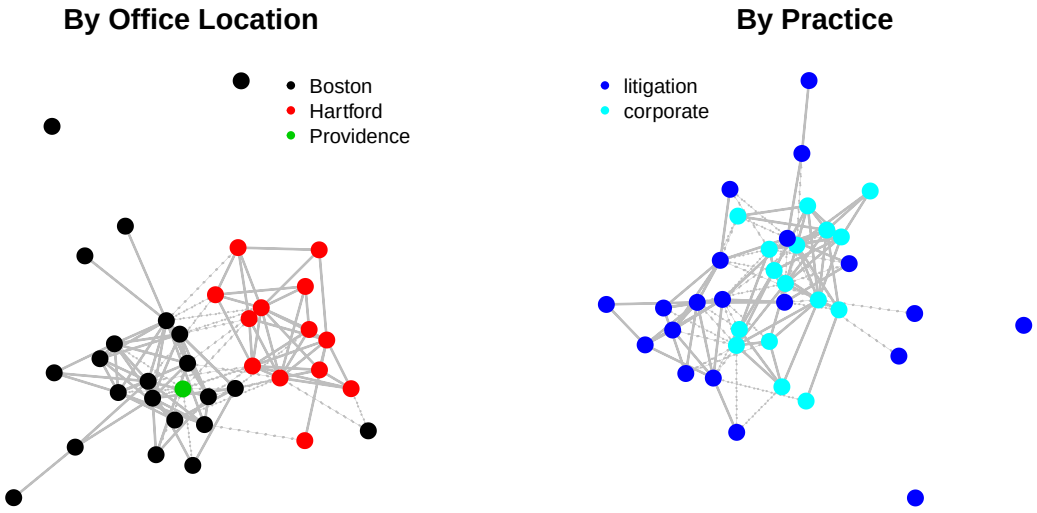


Figure 3.5: Visualiation of Lazega Lawyer network by office location and practice area.

Based on Hunter (2007), we fit models for this network that include one of the transitivity terms discussed above, along with statistics for edge count, homophily on practice, gender and office, and main effects for practice and seniority. We also add a term to count the number of isolates, an independent social process. A model with the triangle count statistic is not included because it is degenerate and fails to fit. All others do fit the data successfully, showing that excessive sensitivity is not a hindrance to modeling for this small example.

Table 3.1 compares selected results of the models. The row name indicates the transitivity statistic in the model, with all other terms unchanged. The first four columns show

the coefficients for edges, the transitivity term, the main effect of practice, and homophily on practice. The final column shows the expected number of triangles for each model as calculated through simulation. Not surprisingly, the model with the cube root of triangle statistic best captures the observed triangle count of 120, but the others are not too far off. The regularized transitivity model, row three, most drastically underestimates triangles. The choice of transitivity term affects values of other parameters and therefore interpretation of behavior in the network. For example, the edge coefficient is lower for the cube root of triangles model than the transitivityH model. This is due to the higher correlation between triangles and edges in the former model.

To visualize the differences between the models, Figure 3.6 shows a simulation from each model with nodes colored by office location. From a single simulation it seems that none of the models fully capture the two-cluster structure of the original data, though we must not place too much stock in single simulations given the variability of simulations under any model. A more reliable presentation of this lack of fit is shown in the plots of simulated degree distributions in the next section.

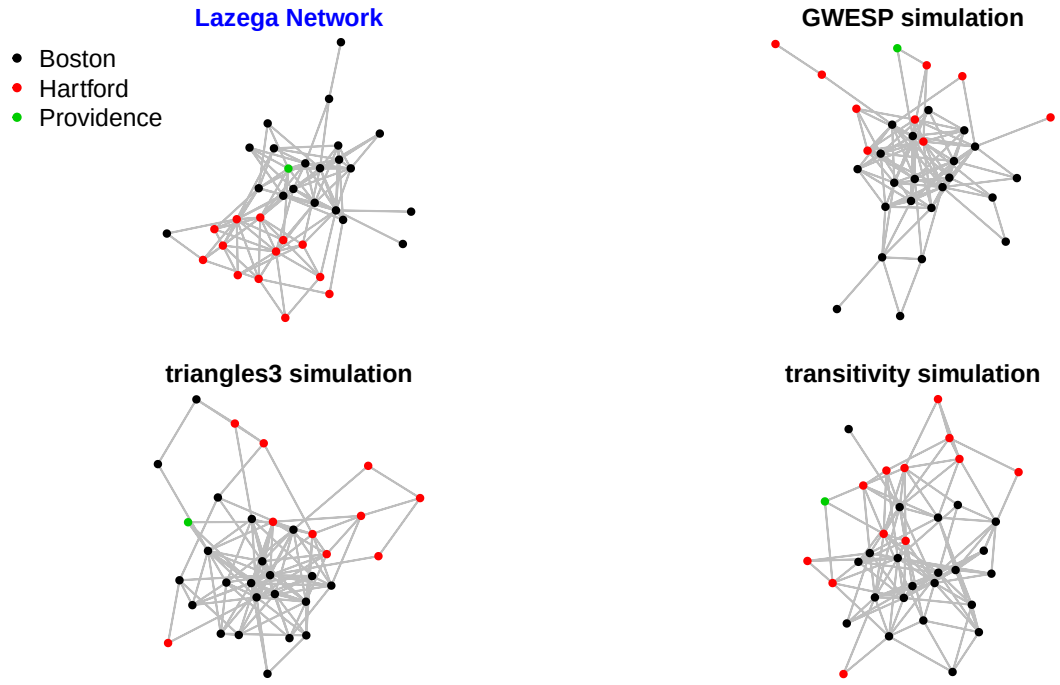


Figure 3.6: The observed Lazega network compared to a simulation from each model. Isolates are not shown.

### 3.6.1 Goodness of Fit

Graphical goodness-of-fit plots for degree and edgewise shared partner distributions are presented in Figure 3.7. In order to compare models only simulated means, not distributions, are shown. Geodesic distance distributions are close to the observed network for all models and are not shown. In Figure 3.7, the GWESP(.5) model seems to best capture the degree distribution, while the GWESP(.5) and regularized transitivity models best capture the ESP distribution. On the other hand, in spectral goodness of fit (see Table 3.2), the cube root of triangles model outperforms all others, though it leaves room for improvement.

Table 3.2: Comparison of spectral goodness of fit.

| | SGOF | SGOF5 | SGOF95 |
|---|---|---|---|
| GWESP(.5) | 0.45 | 0.09 | 0.71 |
| triangles$^{1/3}$ | 0.56 | 0.27 | 0.76 |
| reg. transitivity | 0.35 | 0.03 | 0.66 |
| transitivityH | 0.37 | 0.04 | 0.68 |

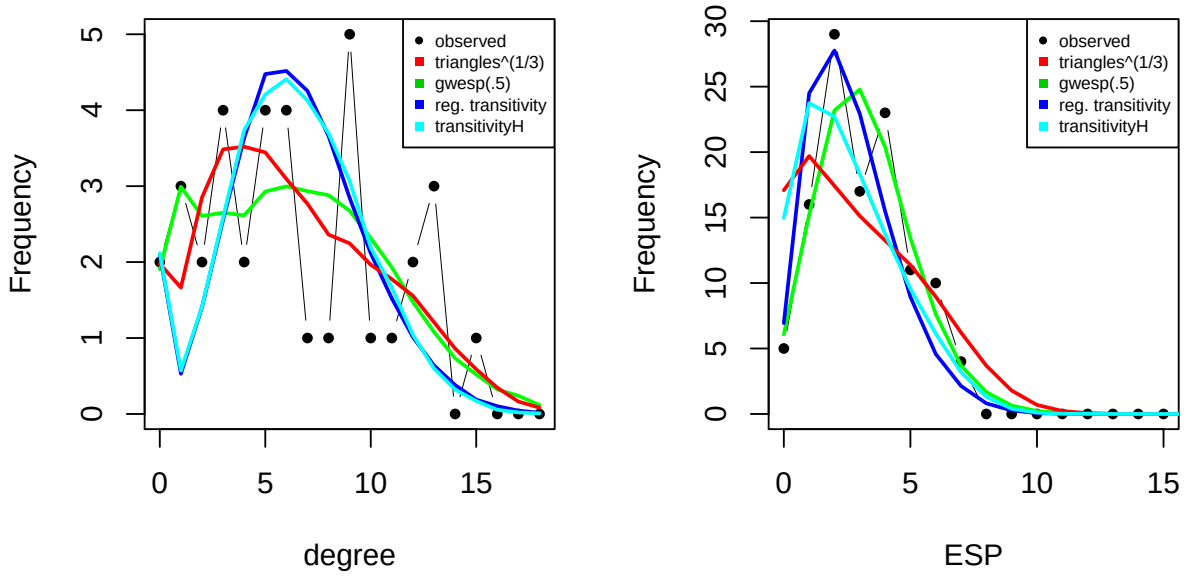


Figure 3.7: Observed degree and edgewise shared partner distributions compared to averages over 1000 simulations from each model of the Lazega Lawyer Network.

This small example shows that the transitivity terms discussed or introduced in this section are less degenerate than the triangle count statistic. However, the example also highlights the insufficiency of any of these models to fully capture the clustering and, relatedly, degree distribution of the observed network. In addition, the cube root of triangles statistic performs best in terms of spectral goodness of fit, but does not capture the edgewise shared partner distribution of the observed network. In the next chapter we propose stable statistics to make up for these shortcomings.

# CHAPTER 4

# Clustering and Moment Terms

Clustering, like transitivity, is a common feature of real-world networks. To some extent this results from homophily and can be modeled accordingly in an ERGM. However, in many environments the tendency to form groups goes beyond shared characteristics, and is a structurally generative force in its own right. The degree distribution of a graph is revealing of clustering and group structure. Most real-world networks have right-skewed degree distributions with a long tail of high-degree "hubs," and many low-degree nodes (Newman, 2010). Formally, it is common for degree distributions to approximately follow a power law, such that $deg_k = ck^{-\alpha}$, where $c$ is a constant and $\alpha$ is a decay parameter. However, under an edge-only ERGM, degree distribution is binomial or, in the limit of large graphs, Poisson (Newman, 2010).

Controlling the number of edges in an ERGM via an edge count statistic also controls the mean of the degree distribution. Let $deg_k$ be the number of nodes with degree $k$ and $e$ the total number of edges.

$$E(deg) = \frac{\sum_k (deg_k) * k}{n} = \frac{2e}{n}. \tag{4.1}$$

In addition, the counts of two-stars and edges together dictate the variance of the degree distribution. Let $star(Y, 2)$ return the number of two-stars in graph $Y$.

$$
\begin{aligned}
star(Y,2) &= \sum_{k=1}^{n-1} \binom{k}{2} * deg_k \\
&= \sum_{k=1}^{n-1} \frac{k(k-1)deg_k}{2} = \sum_{k=1}^{n-1} \frac{k^2 deg_k}{2} - \frac{k deg_k}{2} \\
&= \frac{E(deg^2)n}{2} - e. \qquad (4.2)
\end{aligned}
$$

The variance, $E(deg^2) - E(deg)^2$ is determined. The correspondence between two-stars and the variance of edges was raised by Horvát et al. (2015) (see their Figure 3). They show that a high two-star parameter can force high variance in the edge count, resulting in a bimodal distribution in an ERGM.

In this section we first introduce two statistics to model two-stars, which in turn control the variance of the degree distribution. Subsequently, we introduce stable statistics to model higher moments of the degree distribution and edgewise shared partner distribution.

## 4.1 Square Root of Two-stars

Although the edge-triangle model has received greatest attention as a degenerate ERGM, the edge-two-star model is also simple but degenerate, as shown in Figure 2.1. According to the criteria above, the edge-two-star model is unstable and excessively sensitive. The number of two-stars in a full network is $\mathcal{O}(N^{1.5})$, and therefore unstable. The maximum change in two-stars of a network of size $n$ is $\mathcal{O}(N^{.5})$, and therefore unbounded. Attempts to model even small networks via MCMC-MLE usually fail. The expected number of two-stars in a graph with $e$ edges is $\mathcal{O}(e^2)$. Therefore, following Horvát et al. (2015) we implement the square root of two-stars statistic to supplant this term.

### 4.1.1 Stability, Sensitivity and Density of States

The square root of two-stars is stable. The maximum number of two-stars in a graph of size $n$ is $\mathcal{O}(n^3)$. Therefore, the square root of the number of two-stars is at most $O(n^{1.5}) = O(N^{0.75})$.

Unlike the cube root of triangles, the square root of two-stars is not excessively sensitive. The largest change in two-star count is achieved by adding an edge between a graph of two isolated $(n-2)$-stars. The resulting change in the square root of two-star statistic is $\sqrt{(n-1)(n-2)}-\sqrt{(n-3)(n-2)}$. It is simple to show that this is a positive but decreasing function of $n$.

Figure 3.2 (top) plots the log density of states for $n = 7$ and a sample of the domain of sufficient statistics for $n = 16$, based on 500 draws at each each count. The square root transformation does seem to approximately convexify the domain, and the log density appears concave. However, the correlation between edge count and this statistic is high.

## 4.2 Expected Two-stars

Due to the strong correlation between edge count and the square root of two-stars, we propose a regularized statistic for two-stars. It is the square root of the count of two-stars minus the expectation of that count given the number of edges. We refer to it below and in software as *estar*:

$$estar(y) = \sqrt{\sum_i star(i,2)} - \sqrt{\sum_i E(star(i,2)|e)}, \tag{4.3}$$

where $star(i,2)$ is the number of two-stars incident on a given node, equal to $\binom{deg(i)}{2}$. The expected sum of two-stars conditioned on density is

$$\sum_i E(star(i,2)|e) = E[\sum_i star(i,2)|e] \tag{4.4}$$

$$= \binom{n}{3}\frac{3e(e-1)}{N(N-1)} = \frac{2e(e-1)(n-2)}{n(n-1)-2}. \tag{4.5}$$

$N$ as before is the number of dyads (degrees of freedom) in the graph. Simplifying,

$$estar(y) = \sqrt{\sum_i \binom{deg(i)}{2}} - \sqrt{\frac{2e(e-1)(n-2)}{n(n-1)-2}}. \tag{4.6}$$

The expectation part is approximately $\frac{2e^2}{n}$, so its square root is linear in the number of edges. In our implementation of `estar` in **ernm** we allow a multiplier on the expectation part so

that `estar(0)` is equal to the square root of two-stars statistic.

### 4.2.1 Stability, Sensitivity and Density of States

We consider the stability and sensitivity of the estar statistic. We showed previously that the observed part is $\mathcal{O}(N^{0.75})$. The same is true of the expectation part. As noted previously, the square root of the expectation is linear in the number of edges, so it must allow an $NC$ bound implying that the statistic is stable. We also noted previously that the observed part is not excessively sensitive, i.e., the maximum change statistic is bounded. The change statistic for the expectation part is

$$\sqrt{\frac{2(n-2)}{n(n-1)-2}}(\sqrt{e(e-1)}-\sqrt{e(e+1)}),$$

which is in turn bounded by $\sqrt{2}$, implying that the statistic is not excessively sensitive. Based on Figure 4.1 (right), the domain appears approximately convex and the log density of states appears concave.

## 4.3 Moment Statistics

We illustrated in the opening to this chapter that the two-star and edge count statistics capture the variance of the degree distribution. In this section, we introduce statistics to capture higher moments of this distribution and the edgewise shared partner distribution. To reduce correlation with edge, two-star, and triangle counts, we model the centralized moments of these distributions. We focus on the second-, third-, and fourth-order central moments which can be interpreted in terms of variance, skew, and kurtosis ("tailedness"). We estimate these values from the network. Let

$$\begin{aligned} degm_p(Y) &= (E[deg(i) - E(deg)]^p)^{1/p}. \\ &= (\frac{1}{n}\sum_i [deg(i) - \frac{1}{n}\sum_i deg(i)]^p)^{1/p}. \end{aligned} \tag{4.7}$$

This is the formula for the p$th$-order central moment with the addition of taking its p$th$ root. We do this to put the term on the scale of degrees and stabilize it. Because of the direct

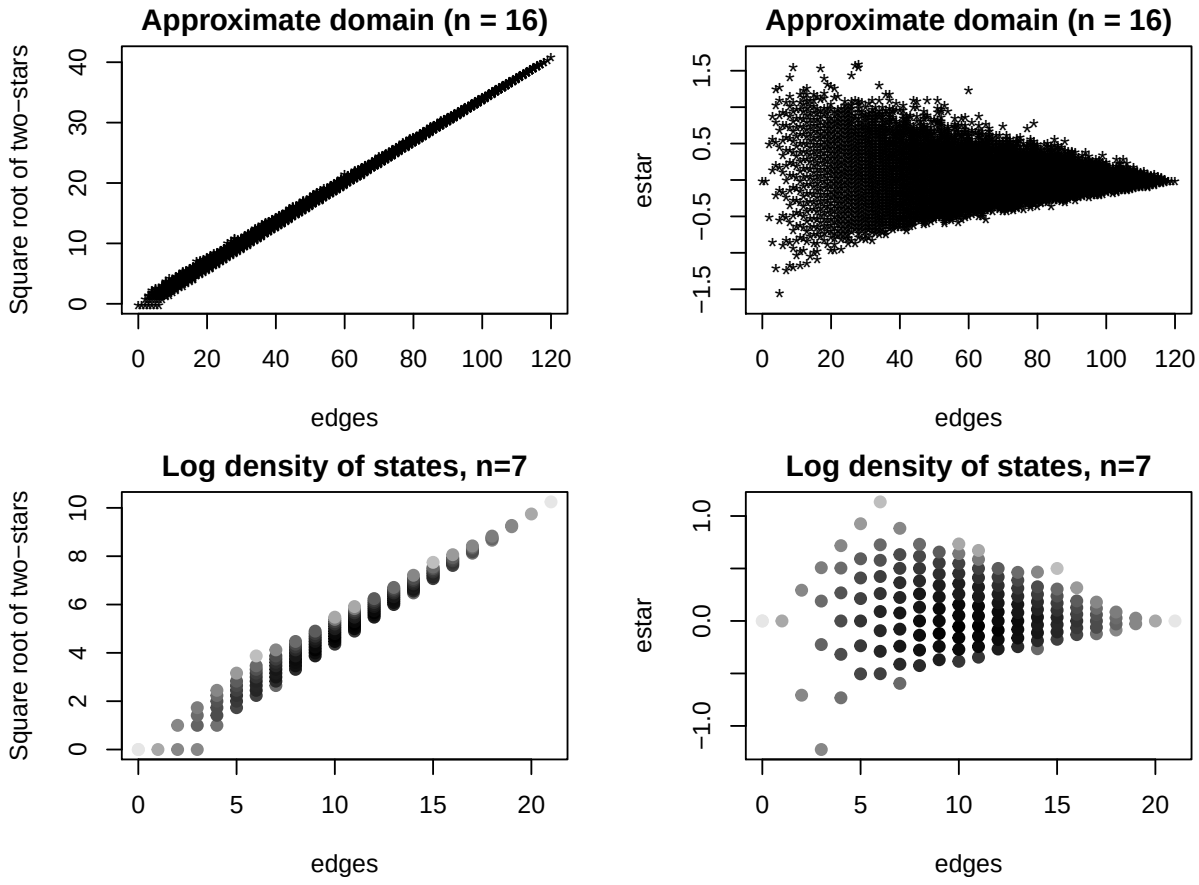


Figure 4.1: (Top) Sampled domain of sufficient statistics, based on 500 draws for each edge count. (Bottom) Log density of states for n = 7.

relationship between two-stars and the second-order moment of the degree distribution, we do not use $\text{degm}_2$ with estar or the square root of two-stars statistic in an ERGM.

The analogous statistic for the edgewise-shared partner distribution is as follows.

$$\begin{aligned} ESPm_p(Y) &= (E[ESP(i,j) - E(ESP)]^p)^{1/p} \\ &= (\frac{1}{e}\sum_{i,\ j>i}[ESP(i,j) - \frac{1}{e}\sum_k k(ESP_k)]^p)^{1/p}, \end{aligned} \tag{4.8}$$

Again, we are primarily interested in $p$ of $2, 3$ or 4. The first moment of the distribution of edgewise shared partners is captured by the triangle and edge count statistics.

$$E(ESP) = \frac{1}{e}\sum_k k(ESP_k) = \frac{1}{e}3tri(Y) \tag{4.9}$$

We implemented these statistics in **ernm** as `degm` and `espm` for $p$ of $2, 3$ and 4. To reduce computation time, the implementation calculates change statistics using expansions of the central moments in terms of non-central moments, for example

$$\begin{aligned} E[x - E(x)]^2 &= E(x^2) - E(x)^2 \\ E[x - E(x)]^3 &= E(x^3) - 3E(x^2)E(x) + 2E(x)^3, \end{aligned} \tag{4.10}$$

and so forth, following Pascal's triangle.

### 4.3.1 Stability and Sensitivity

The degm and espm terms are bounded by $n$ and therefore stable. Degm is not excessively sensitive, as adding or removing one edge changes the total degree count by two regardless of network size. However, espm is excessively sensitive as adding one edge can add $\mathcal{O}(n)$ triangles to a network that previously had zero.

We do not plot the log density of states for the moment statistics because models with only these terms and edges are generally not of interest. In practice, at least one additional clustering or transitivity term would be included.

Table 4.1: Summary of statistics as unstable or excessively sensitive.

| | Unstable | Excessively sensitive |
|---|---|---|
| triangles | yes | yes |
| triangles$^{1/3}$ | no | yes |
| GWESP($e^{-\alpha} \leq 2$) | no | yes |
| GWDSP($e^{-\alpha} \leq 2$) | no | yes |
| reg. transitivity | yes | yes |
| transitivityH | no | yes |
| two-star | yes | yes |
| two-star$^{1/2}$ | no | no |
| estar | no | no |
| degm | no | no |
| ESPm | no | yes |

## 4.4 Summary of Statistics

To facilitate comparison, we summarize the stability and sensitivity of the statistics we have discussed in the following table. The dark gray rows of Table 4.1 are statistics that we have introduced. The light gray rows are previously introduced statistics, which we implemented in the **ernm** package. In the case of regularized transitivity we altered and expanded the previous implementation.

All of the statistics that we have introduced are stable. However, as Schweinberger (2011) points out, stability is not a guarantee against near degeneracy. In fact, for large enough parameter values, near degeneracy is achievable for all ERGM (Barndorff-Nielsen, 1978; Schweinberger, 2011). In the next section we model a significantly larger and more dense network than the Lazega Lawyer Network. This provides further insight into the relative utility of the stable statistics. We also describe how correlation between statistics affects modeling in the context of our larger example.

# CHAPTER 5

# Example: Facebook Network

## 5.1 ERGM with Reduced Degeneracy Terms

In this section we model an online social network from the social media service Facebook. The network we consider is from a data set published by Traud et al. (2011), and expanded upon by Traud et al. (2012). In total it consists of complete snapshots of Facebook networks within 100 American colleges and universities from a day in September, 2005. Each network consists of all within-school ties, where nodes are mostly students, but also include some alumni, faculty, and staff. Edges represent undirected "friendships" within Facebook's online medium.

Traud et al. (2012) studied the largest connected components of four networks for each institution: the full network, the student-only network, the female-only network and the male-only network. They focused on two qualities of these networks: 1) the strength of homophily by gender, class year, major, high school, and residence (e.g., dorm), and 2) the similarity between communities based on features (e.g., residence) and ones detected via algorithm, to determine which features had an organizing role in the network. To study homophily, both Traud et al. (2011) and Traud et al. (2012) used the *assortativity coefficient* of Newman (2003), defined

$$r = \frac{\sum_r e_{rr} - \sum_r e_{r\cdot} e_{\cdot r}}{1 - \sum_r e_{r\cdot} e_{\cdot r}} \tag{5.1}$$

where $e$ is a matrix such that $e_{rs}$ is equal to the proportion of edges that connect a node of type $r$ to a node of type $s$; $e_{r\cdot}$ is the $rth$ row sum of $e$, the proportion of edges connected to a node of type $r$; and $e_{\cdot r}$ is the $rth$ column sum, equal to the $rth$ row sum if the network is

undirected.

Additionally, Traud et al. (2012) fit a logistic regression model and an ERGM to each of the smallest 16 schools. The predictors used were number of edges and number of node-matching edges by residence, class year, major, and high school. Rather than include a term for gender, Traud et al. (2012) modeled the gender-specific networks. According to the authors, the ERGM retained the statistics of the logistic regression model and added a triangle count statistic. In fact, both of the models considered are logistic regression models or, equivalently, ERGM fit by maximum pseudolikelihood estimation (MPLE). The response variable is equal to the change statistics of each dyad given the model terms. Although it is not explicitly stated, we infer that Traud et al. (2012) employed MPLE to fit their ERGM with triangles. We confirm this by recreating the paper results using MPLE, but failing to fit any model, even on a small college network, using MCMC-MLE. For equivalence of MPLE to logistic regression see van Duijn et al. (2009).

Although MPLE is not prone to the estimation degeneracy of MCMC-MLE, it may be an extremely biased estimator of the MLE. The inadequacy of MPLE for dyad-dependent ERGM is addressed by Geyer and Thompson (1992b) and van Duijn et al. (2009), among others. MPLE may mask the degeneracy of the underlying model. For example, when we simulate a sample school network using MPLE with a triangle term as described above, the average edge count is several times that of the observed network. On the other hand, without the triangle term the simulations contain less than half the observed number of triangles.

The model-based approach of Traud et al. (2012) is insufficient to capture Facebook networks. Yet, a model-based approach in general is valuable for accurately capturing the effects of homophily on social structure, controlling for multiple forces, and estimating uncertainty. Using the statistics that we have discussed and introduced in Chapters 3 and 4 we are able to produce more accurate models of a particular Facebook school network.

We focus on one of the 16 smallest schools in the network, Haverford College (the author's alma mater). We also subset the network by graduation year, creating five smaller networks of the largest connected component for each of the graduation years from 2005 to 2009. We

present results for the network of students with graduation year 2005, which includes a mix of students and alumni. This is the smallest of the year-specific networks, with 174 nodes and 2653 edges, as opposed to 1446 nodes and 59589 for the entire college network.

We impute missing values in covariates of interest. To do this we use a simple, iterative imputation method. Nodes with missing covariate information are assumed to have the covariate value of the majority of their neighbors, and are imputed in decreasing order of the percentage of non-missing values among their neighbors. The covariates in the Haverford network to which we apply this imputation are amenable to it because their percentages of missingness are low, while network density is high. All models we consider contain baseline terms for edge count (`edges`) and node-matching variables on gender and major (`nodeMatch(gender_complete)`, `nodeMatch(major_complete)`). Because the majority of the 2005 network is alumni, and because most high schools listed are unique, we do not include node-matching variables for residence or high school.

In addition to the baseline terms, we experimented with many permutations of statistics for transitivity, clustering, and degree distribution. We were unable to fit a model using the regularized transitivity statistic, which is not surprising given its instability. Another statistic that proved problematic was the square-root of two-star statistic. As expected, the correlation between this term and edge count disrupted estimation. However, the regularized square root of two-star statistic, `estar`, did not suffer this problem, and is included in model three of Figure 5.1.

We were also unable to fit some models containing transitivityH. We believe that this is due to the relationship between transitivityH and degree assortativity. The expected number of triangles for each edge includes a product of the degrees of its endpoints. The sum of these expectations is a measure of degree assortativity. Decreasing transitivityH while holding edge count constant is achieved by either breaking up triangles or increasing degree assortativity. For networks with high degree assortativity, the expected number of triangles is correlated with $\sum_i deg(i)^2$, which, along with edge count, determines the variance of the degree distribution. High degree assortativity and high variance in degree distribution bring out the instability of the transitivityH statistic, as many dense clusters are formed. We

found that by adding a term to an edge + transitivityH model which controls the variance of the degree distribution, either `degm(2)` or `estar`, the model stabilizes.

On the other hand, models containing the cube root of triangles statistic (`triangles3`), GWESP (`gwesp`), and ESPm (`espm`) terms produced successful models. We found these models preferable to those with transitivityH due to the straightforward interpretation of the former terms and the aforementioned difficulty with the latter.

Figure 5.1 compares simulated edgewise shared partner distributions based on 1000 simulations from six competing models, numbered one through six. Each model includes the baseline edge and node-matching terms mentioned above, and the additional terms listed in each panel title. Comparing models one and two we see that the addition of `triangles3` drastically improves the fit to the observed edgewise shared partner distribution. Model one is insufficient in this regard, though until recently it would have been close to the best available ERGM.

The other model in Figure 5.1 that successfully captures the edgewise shared partner distribution is the baseline model plus cube root of triangles with second through fourth moments of the ESP distribution, model six. We see the improvement in modeling ESP distribution as the third and fourth moment terms are added, from model four to model six.

A subtle difference between the two best-performing models is the slightly higher variance in the tail of ESP distribution for model two. We highlight other differences in these models by comparing statistics of interest under 1000 simulations from each model. The results are presented in Tables 5.1 and 5.2. For both models, simulated statistics of terms included in the model are near observed values, as expected. Though not shown, the underlying `nodeMatch` parameters are also very similar in the models. For model two, all other simulated statistics are not near observed values. On the other hand, model six captures GWESP(1.5) fairly well, despite the term not being included in the model. However, neither model captures the regularized two-star statistic or second moment statistic of the degree distribution. This is also shown in Figure 5.2, which depicts simulated degree distributions from both of these models. Yet, when we attempt to add either of the statistics capturing the variance of the

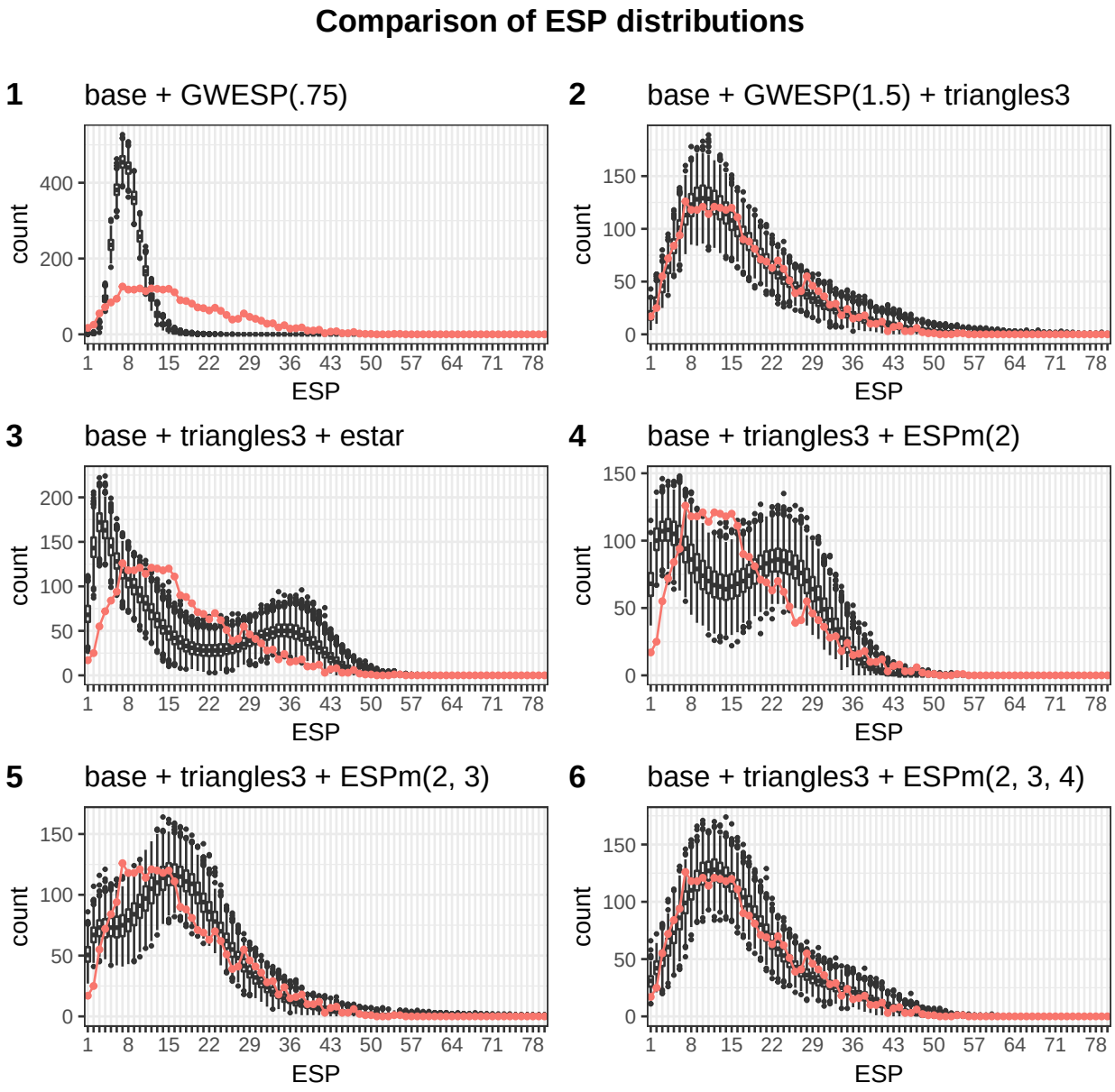


Figure 5.1: Comparison of simulated edgewise shared partner distributions. Each model contains baseline edge and nodematch terms for gender and major, and additional terms listed in each plot title.

Table 5.1: Simulated statistics under model two containing GWESP(1.5). Gray rows are terms in the model.

| | obs | mean | min | max | sd sim | MC p-val |
|---|---|---|---|---|---|---|
| edges | 2653.00 | 2632.71 | 2318.00 | 2901.00 | 111.03 | 0.84 |
| triangles3 | 23.93 | 23.94 | 22.95 | 25.22 | 0.32 | 0.98 |
| estar | 43.37 | 56.52 | 46.87 | 65.49 | 3.42 | 0.00 |
| nodeMatch(gender_complete) | 1511.00 | 1498.05 | 1299.00 | 1677.00 | 74.89 | 0.90 |
| nodeMatch(major_complete) | 377.00 | 367.35 | 326.00 | 415.00 | 14.45 | 0.53 |
| degm(2) | 18.14 | 20.64 | 19.18 | 21.84 | 0.37 | 0.00 |
| espm(2) | 9.55 | 10.15 | 8.74 | 11.62 | 0.45 | 0.18 |
| espm(3) | 8.87 | 10.36 | 8.39 | 12.23 | 0.60 | 0.01 |
| espm(4) | 12.82 | 14.46 | 12.10 | 16.70 | 0.72 | 0.02 |
| gwesp(1.5) | 10656.01 | 10570.42 | 9173.86 | 11857.53 | 487.88 | 0.85 |
| transitivityH | 3.60 | 2.84 | 2.34 | 3.42 | 0.22 | 0.00 |

Table 5.2: Simulated statistics under model six containing ESPm terms. Gray rows are terms in the model.

| | obs | mean | min | max | sd sim | MC p-val |
|---|---|---|---|---|---|---|
| edges | 2653.00 | 2648.86 | 2513.00 | 2824.00 | 55.20 | 0.94 |
| triangles3 | 23.93 | 23.90 | 22.93 | 25.18 | 0.32 | 0.92 |
| estar | 43.37 | 53.77 | 46.52 | 61.01 | 2.34 | 0.00 |
| nodeMatch(gender_complete) | 1511.00 | 1527.51 | 1377.00 | 1678.00 | 47.81 | 0.75 |
| nodeMatch(major_complete) | 377.00 | 374.09 | 334.00 | 434.00 | 16.39 | 0.88 |
| degm(2) | 18.14 | 20.19 | 18.89 | 21.46 | 0.41 | 0.00 |
| espm(2) | 9.55 | 9.54 | 8.57 | 10.53 | 0.35 | 0.97 |
| espm(3) | 8.87 | 8.88 | 7.79 | 10.16 | 0.38 | 0.99 |
| espm(4) | 12.82 | 12.80 | 11.41 | 14.33 | 0.47 | 0.98 |
| gwesp(1.5) | 10656.01 | 10598.81 | 9885.56 | 11653.42 | 270.00 | 0.81 |
| transitivityH | 3.60 | 2.85 | 2.47 | 3.18 | 0.12 | 0.00 |

degree distribution to these models the estimation process fails. The reason for this failure is a subject of ongoing investigation. It may be a result of both the model and the MCMC procedure.

Another apparent problem with both model two and six is the high correlation between certain model terms, illustrated in Tables 5.3 and 5.4. In model two, the correlation between edges and GWESP(1.5) is near one. In model six, the correlation between the third and fourth moment terms of the ESP distribution is almost as high. Such high levels of correlation bring instability to the estimates of the corresponding parameters.

Finally, we compare spectral goodness of fit for these two models. Model six slightly outperforms model two, and both mean SGOF values are fairly high despite degree distribution not being fully captured. Both models are a vast improvement over model one. Despite

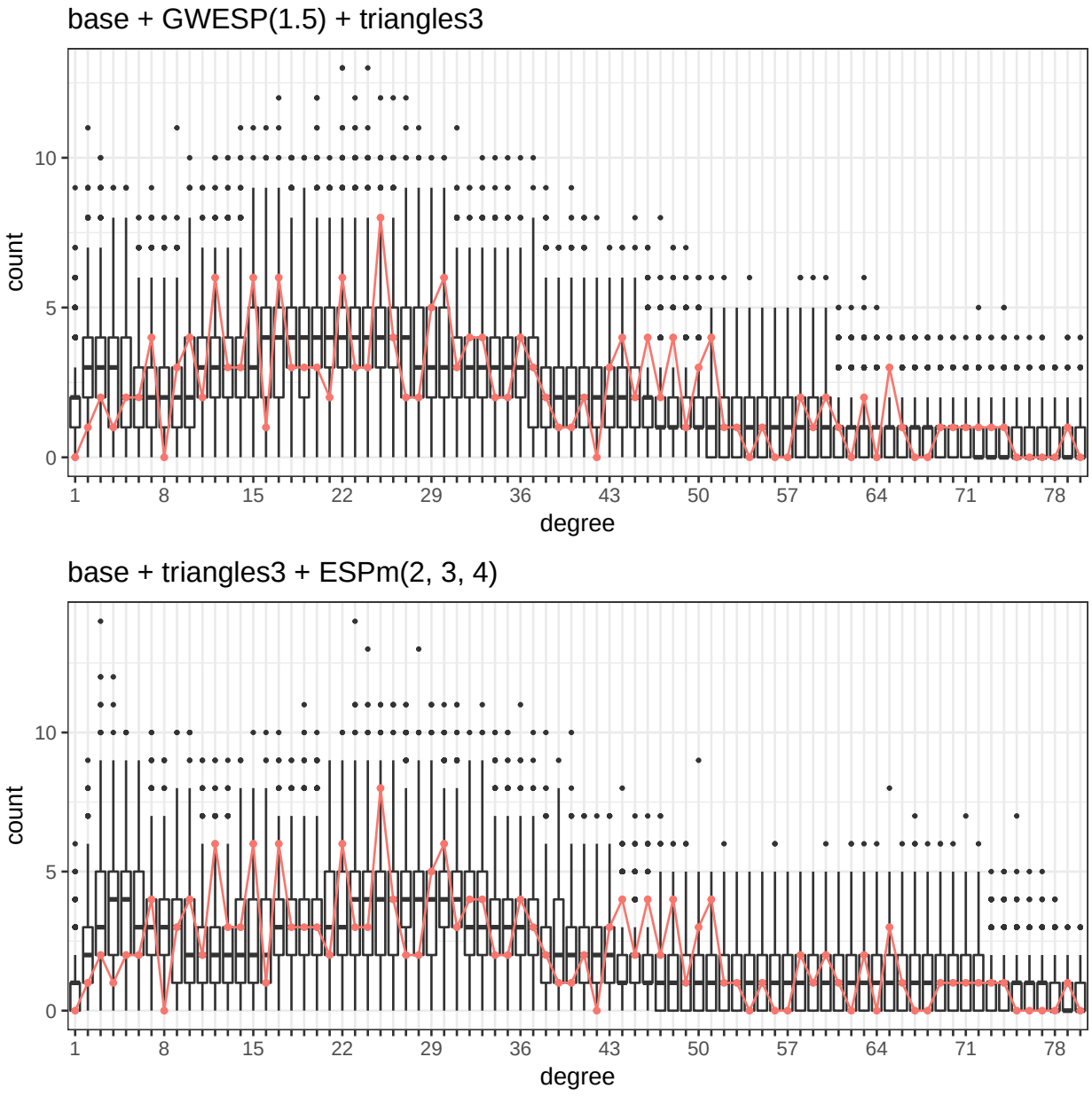


Figure 5.2: Comparison of simulated degree distributions for models two and six.

Table 5.3: Correlation of simulated statistics under model two. Gray rows are terms in the model. Correlation between edges and GWESP is extremely high.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. edges | 1.00 | 0.74 | -0.85 | 0.88 | -0.06 | -0.42 | -0.49 | 0.08 | -0.01 | 0.99 | -0.91 |
| 2. triangles3 | 0.74 | 1.00 | -0.32 | 0.62 | 0.15 | 0.23 | 0.11 | 0.29 | 0.27 | 0.77 | -0.42 |
| 3. estar | -0.85 | -0.32 | 1.00 | -0.78 | 0.15 | 0.83 | 0.80 | 0.24 | 0.34 | -0.82 | 0.92 |
| 4. nodeMatch(gender_complete) | 0.88 | 0.62 | -0.78 | 1.00 | -0.13 | -0.42 | -0.44 | 0.04 | -0.04 | 0.86 | -0.81 |
| 5. nodeMatch(major_complete) | -0.06 | 0.15 | 0.15 | -0.13 | 1.00 | 0.19 | 0.23 | 0.00 | 0.02 | -0.05 | 0.18 |
| 6. degm(2) | -0.42 | 0.23 | 0.83 | -0.42 | 0.19 | 1.00 | 0.86 | 0.52 | 0.59 | -0.36 | 0.62 |
| 7. espm(2) | -0.49 | 0.11 | 0.80 | -0.44 | 0.23 | 0.86 | 1.00 | 0.58 | 0.64 | -0.48 | 0.69 |
| 8. espm(3) | 0.08 | 0.29 | 0.24 | 0.04 | 0.00 | 0.52 | 0.58 | 1.00 | 0.98 | 0.10 | -0.02 |
| 9. espm(4) | -0.01 | 0.27 | 0.34 | -0.04 | 0.02 | 0.59 | 0.64 | 0.98 | 1.00 | 0.01 | 0.09 |
| 10. gwesp(1.5) | 0.99 | 0.77 | -0.82 | 0.86 | -0.05 | -0.36 | -0.48 | 0.10 | 0.01 | 1.00 | -0.89 |
| 11. transitivityH | -0.91 | -0.42 | 0.92 | -0.81 | 0.18 | 0.62 | 0.69 | -0.02 | 0.09 | -0.89 | 1.00 |

Table 5.4: Correlation of simulated statistics under model six. Gry rows are terms in the model. Correlation between ESPm terms is extremely high.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. edges | 1.00 | 0.60 | -0.42 | 0.48 | 0.13 | -0.00 | -0.19 | -0.12 | -0.11 | 0.95 | -0.57 |
| 2. triangles3 | 0.60 | 1.00 | 0.44 | 0.47 | 0.14 | 0.77 | 0.42 | 0.36 | 0.49 | 0.76 | 0.29 |
| 3. estar | -0.42 | 0.44 | 1.00 | -0.03 | -0.02 | 0.91 | 0.62 | 0.53 | 0.65 | -0.17 | 0.87 |
| 4. nodeMatch(gender_complete) | 0.48 | 0.47 | -0.03 | 1.00 | -0.01 | 0.19 | -0.04 | -0.01 | 0.03 | 0.53 | -0.08 |
| 5. nodeMatch(major_complete) | 0.13 | 0.14 | -0.02 | -0.01 | 1.00 | 0.04 | 0.07 | 0.03 | 0.07 | 0.12 | 0.01 |
| 6. degm(2) | -0.00 | 0.77 | 0.91 | 0.19 | 0.04 | 1.00 | 0.60 | 0.52 | 0.66 | 0.25 | 0.69 |
| 7. espm(2) | -0.19 | 0.42 | 0.62 | -0.04 | 0.07 | 0.60 | 1.00 | 0.88 | 0.92 | -0.16 | 0.65 |
| 8. espm(3) | -0.12 | 0.36 | 0.53 | -0.01 | 0.03 | 0.52 | 0.88 | 1.00 | 0.96 | -0.09 | 0.47 |
| 9. espm(4) | -0.11 | 0.49 | 0.65 | 0.03 | 0.07 | 0.66 | 0.92 | 0.96 | 1.00 | -0.05 | 0.61 |
| 10. gwesp(1.5) | 0.95 | 0.76 | -0.17 | 0.53 | 0.12 | 0.25 | -0.16 | -0.09 | -0.05 | 1.00 | -0.36 |
| 11. transitivityH | -0.57 | 0.29 | 0.87 | -0.08 | 0.01 | 0.69 | 0.65 | 0.47 | 0.61 | -0.36 | 1.00 |

Table 5.5: Comparison of spectral goodness of fit.

| | SGOF | SGOF5 | SGOF95 |
|---|---|---|---|
| Model 1 | 0.17 | 0.13 | 0.24 |
| Model 2 | 0.74 | 0.65 | 0.84 |
| Model 6 | 0.80 | 0.70 | 0.84 |

remaining challenges, these models vastly outperform those with only dyad independent and geometrically weighted terms.

## 5.2 Degree-corrected Stochastic Blockmodel

In this section we compare our results to results from a different modeling approach. We fit a degree-corrected stochastic blockmodel to the Haverford 2005 network. Under the degree-corrected blockmodel introduced by Karrer and Newman (2011), the expected value of $y_{ij}$

depends on the degrees of nodes $i$ and $j$, their class memberships denoted $g(i)$ and $g(j)$, and a function which captures the overall traffic between $g(i)$ and $g(j)$.

$$y_{ij,\ i<j}|\theta,\eta,g \sim Poisson(\theta_i\theta_j\eta(g(i),g(j)),$$
$$\theta_i = \frac{deg(i)}{deg(g(i))},\ 1 \leq i \leq n \tag{5.2}$$

where $\theta_i$ is a degree-correction factor for each node and $deg(g(i))$ is the sum of degrees in group $g(i)$. $\eta(g(i),g(j))$ is estimated as the number of edges between class $g(i)$ and class $g(j)$. The model assumes Poisson distributions of edge weights, and therefore valued edges. However, Karrer and Newman (2011) note that this does not significantly compromise the fit when applied to a binary network, and the error is $\mathcal{O}(1/n)$. Despite this drawback, there are major benefits to implementing this over an uncorrected SBM. The formulation preserves the expected degrees between blocks and expected degree sequence of the network. It is better-suited to real-world networks with high degree heterogeneity than the uncorrected blockmodel. As Karrer and Newman (2011) show, using an uncorrected blockmodel for community detection tends to group nodes by activity level, whereas the degree-corrected classes correspond to meaningful roles.

We implement the degree-corrected SBM by adapting the C++ code by Karrer (2010) for R with **Rcpp** (Eddelbuettel and Francois, 2011). Estimated edgewise shared partner and degree distributions based on a nine-block model and 1000 simulated networks are shown in Figure 5.3. The choice of number of blocks is large enough to ensure that communities are detected, but small enough to ensure good representation in each. Trials with three to twenty blocks show that the ESP distribution is not sensitive to this choice. While the simulated degree distribution is a good match to the observed, as guaranteed by the model, the ESP distribution is much more skewed than the observed distribution. The simulated average triangle count is 5655, which is less than half of the observed count of 13702. A small amount of this misfit is due the use of Poisson distribution for edge values. Because edges with values greater than one are reduced to binary edges, the simulated networks have about 400 fewer edges than the observed one, roughly fifteen percent less. This marginally drives

down the triangle count, but does not explain the large gap. Rather, the assumption of edge independence conditional on the degree and block structure under-accounts for transitivity in the network. Our best-performing ERGM capture transitivity far better, though the degree-corrected SBM is preferable for capturing the degree sequence.

## 5.3 Latent Space Model

We conclude by applying a latent space model to the Haverford 2005 network. We delve into much greater detail about this type of network model in Part II, especially Sections 7.3 and 7.4. The discussion there pertains to valued, directed networks. Here we employ a Euclidean latent space model for binary, undirected data (Hoff et al., 2002; Hoff, 2003; Krivitsky et al., 2009a).

$$P(Y|Z, X, \beta, \gamma) = \prod_{i<j} P(y_{ij}|z_i, z_j, x_{ij\cdot}, \beta, \gamma_i, \gamma_j), \tag{5.3}$$

where $z_i$ is the position of node $i$ in $d$ dimensions and $\gamma_i$ is the random "sociality" effect of node $i$. The edge covariates, $X_{ijk}$, include an intercept and node-matching terms for gender and major, as above. Edge values are modeled by logistic regression,

$$P(y_{ij,\ i<j} = 1|z_i, z_j, x_{ij\cdot}, \beta, \gamma_i, \gamma_j) = logit^{-1}(\beta' x_{ij\cdot} + \gamma_i + \gamma_j - \|z_i - z_j\|). \tag{5.4}$$

Edge probabilities are conditionally independent. This independence assumption eliminates the model degeneracy problem. However, as shown in Figure 5.4, neither the edgewise shared partner or degree distribution is well captured by this model. The results presented are for a four-dimensional model, but results were similarly poor for our trials of up to ten dimensions. Unlike the degree-corrected SMB, this model overestimates transitivity. The high density of the network pulls nodes together, and the inherent transitivity of the distance-based model leads to inflated clustering and degree counts.

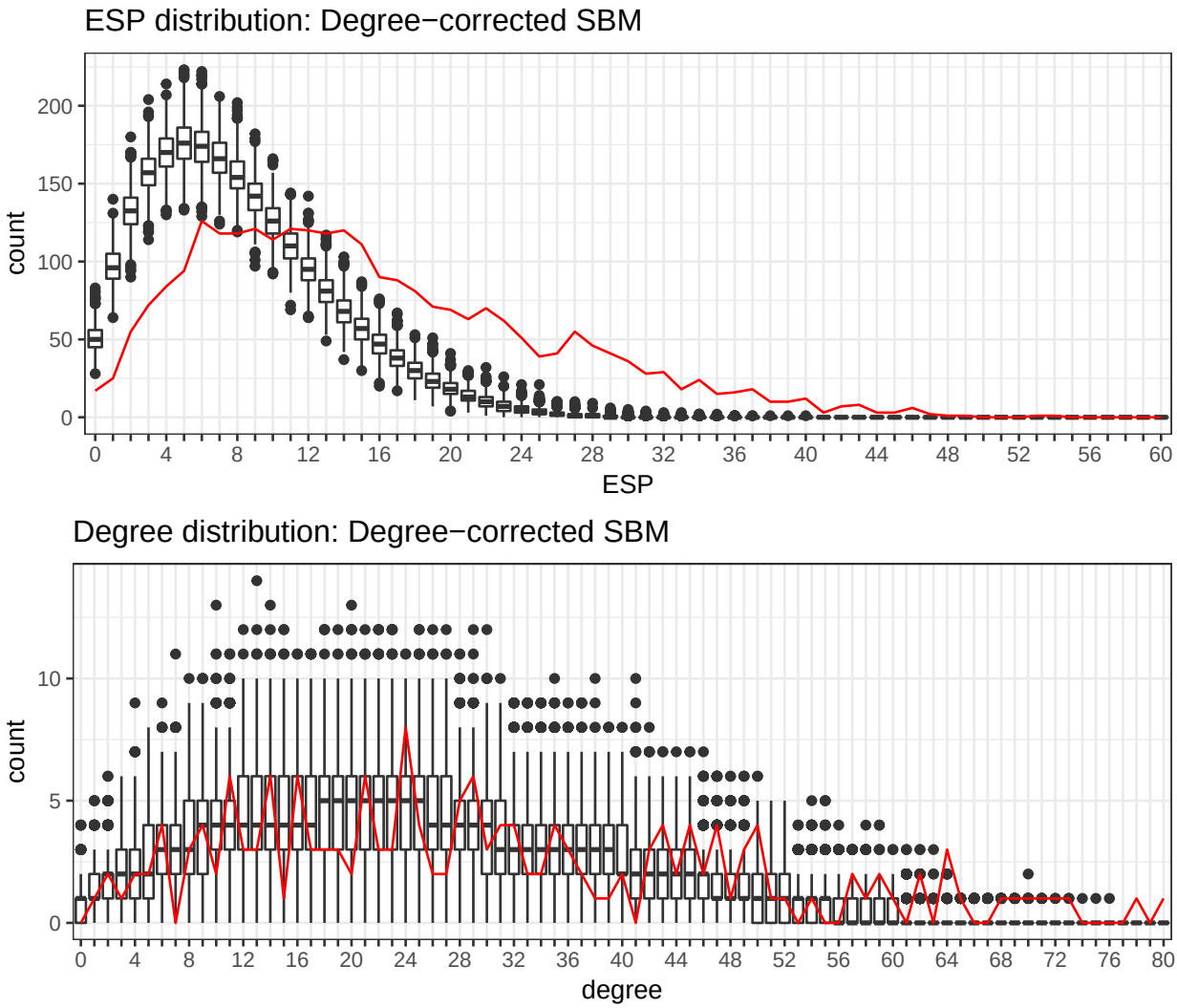


Figure 5.3: Simulated edgewise shared partner and degree distributions of the Haverford 2005 network from the degree-corrected stochastic block model.

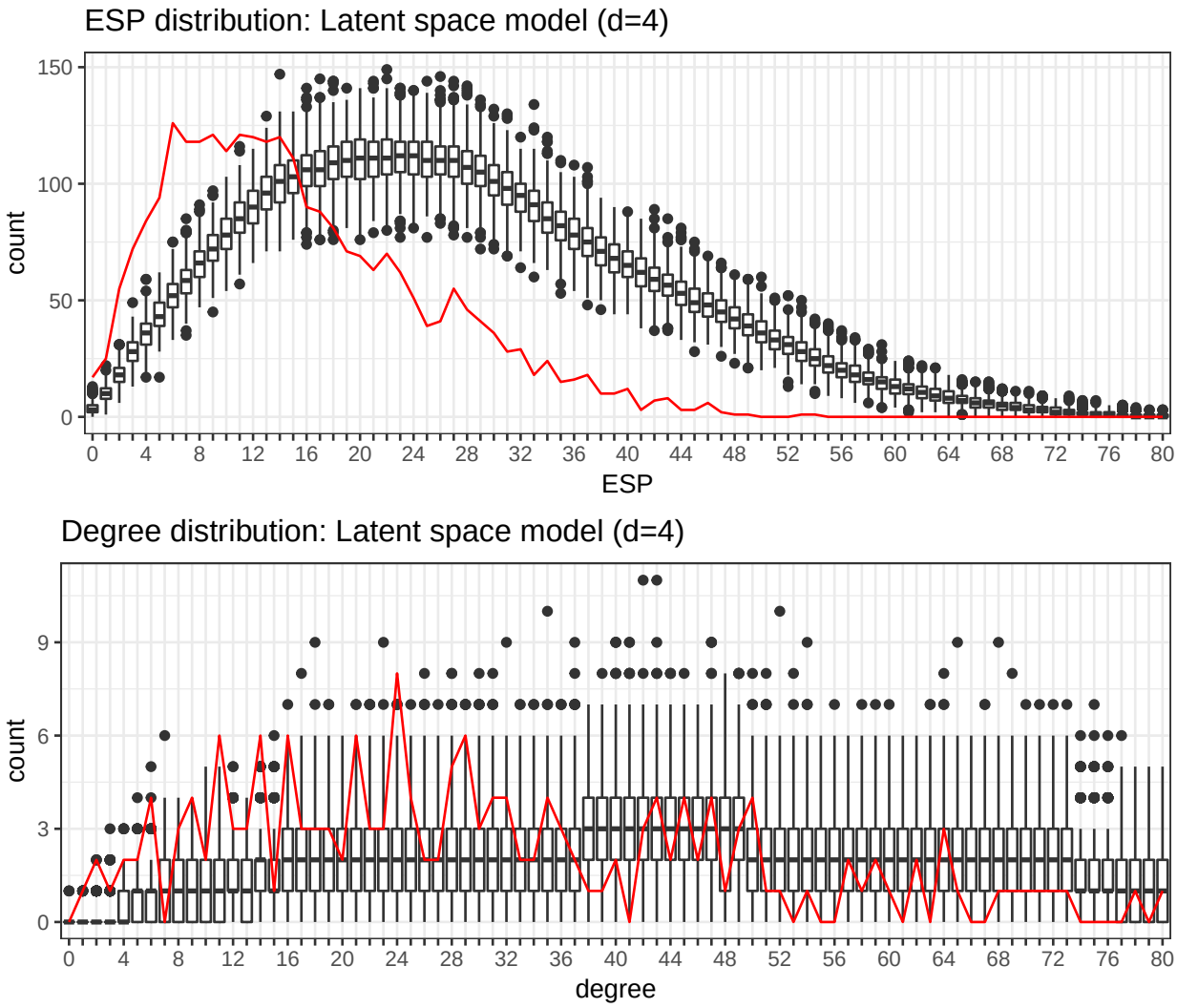


Figure 5.4: Simulated edgewise shared partner and degree distributions of the Haverford 2005 network from the latent space model.

# CHAPTER 6

# Conclusion

## 6.1 Summary of Contributions

Below we summarize the main contributions of Part I to the field of network modeling. Details of these contributions and subjects of future work are subsequently discussed. We have:

- Cataloged and compared existing definitions of model degeneracy and related concepts;
- Proposed a novel transitivity statistic, clustering (two-star) statistic, and class of central moment statistics. We illustrated that the latter improves ERGM goodness of fit without displaying degeneracy in theory or estimation;
- Evaluated the degeneracy of all statistics discussed using theorems that connect model degeneracy to observable characteristics;
- Analyzed recently proposed statistics in models of a real-world online social network, identifying an interpretable set of statistics to model transitivity. These statistics had not yet been incorporated into models outside of introductory examples;
- Compared results from newly developed ERGM to non-degenerate models, establishing that these ERGM best capture transitivity.

## 6.2 Discussion

In Part I we introduced and analyzed statistics that drastically improve the ability of ERGM to capture transitivity in real-world networks while avoiding model degeneracy. We analyzed

a range of statistics to identify those which do not exhibit degeneracy *a priori*. Both the stability criterion of Schweinberger (2011) and the convexity criterion of Horvát et al. (2015) proved useful in this regard. These criteria lead to a more general guideline for developing non-degenerate statistics. The rate of increase in the expected or maximal value of a statistic should not be greater than the rate of increase in edges, both as density increases in a network and as network size increases.

The cube root of triangles statistic proposed by Horvát et al. (2015) proved to be the most useful of those we considered. The performance of the GWESP statistic which has been in use for over a decade is greatly augmented by joint use with the the cube root of triangles term. In Chapter 4 we introduced a new class of stable statistics. The second through fourth central moment terms of the ESP distribution also work well in conjunction with the cube root of triangles statistic, which controls the first moment of the ESP distribution. Taken together, the cube root of triangles and ESP moment terms form an interpretable set of statistics to model transitivity in ERGM, with far less degeneracy than previously applied statistics. Although the performance of the model containing these terms was only slightly better than the model including GWESP when applied to a sample Facebook network, the interpretability of the former model is much greater. It is clear how to add or subtract terms from this model to better capture the ESP distribution.

We also found that the *estar* statistic and higher moment terms of the degree distribution successfully model the degree distribution of a network. However, we face difficulty in modeling the degree and edgewise shared partner distributions simultaneously. One reason for model sensitivity is the high correlation between certain statistics. To address both of these challenges we may consider more complex MCMC updating procedures to more efficiently explore the space of graphs.

It would benefit network practitioners to implement these reduced-degeneracy statistics within the **ergm** package. This would also make recent advances in MCMC-MLE available, in particular the contrastive divergence technique of Krivitsky (2017), which improves initial values for MCMC estimation. Implementing the statistics in **ergm** would also facilitate comparison with MPLE for our reduced-degeneracy models. In some cases the bias of

MPLE may be reduced to a tolerable level, especially when weighed against the drawbacks of cumbersome MCMC estimation.

Expanding these statistics to directed networks is another subject of future work. The extensions of degree-based moments terms are straightforward, as in- and out-degree distributions may be substituted for undirected degree distributions. The extension of triangle-based terms to directed networks is more complicated as we must consider the range of triad configurations as classified by Davis and Leinhardt (1972).

The value of the statistics presented here goes beyond network modeling. They are also useful for generating random networks with specific structural properties. The generation of random networks with comprehensible structure that embody real-world properties has high value for network researchers (Newman et al., 2001; Watts and Strogatz, 1998; Barabási and Albert, 1999; Newman, 2002). Through simulation studies, the models that we have developed here can further our understanding of the properties of transitive networks.

Part II

# Latent Space Network Models for Rating

In Part II we introduce a method for rating items based on network data or pairwise comparison data, employing the latent space network models developed by Hoff et al. (2002), Hoff (2003) and Krivitsky et al. (2009a). The method estimates positions of items in latent space, along with individualized sender and receiver coefficients that capture powers of transmission for each item. Ratings are derived from the difference in sender and receiver coefficients, and are presented with a measure of uncertainty. We later extend the underlying model to incorporate elements of the additive and multiplicative effects models of Hoff (2015) and Minhasa et al. (2016). We illustrate that by decomposing the non-parametric part of our model into symmetric and asymmetric elements we improve the overall fit while retaining the benefits of interpretation and visualization supplied by latent positions. The methods we introduce are ideal for rating items with nebulous similarities which exert a range of influences on the strength of ties. We illustrate that latent space rating is a more appropriate measure of influence or importance than widely used methods such as PageRank, which better capture centrality or popularity.

Furthermore, we show that quasi-Newton estimation produces results on par with commonly used MCMC methods in a fraction of the time. This extends the practical use of latent space rating methods and of the class of latent network models more generally. Our method is implemented in R, using the packages **latentnet** (Krivitsky and Handcock, 2017) and **visNetwork** (Almende and Thieurmel, 2016), and supplementary code.

Chapter 7 provides background and describes the latent space rating model and estimation methods. Chapters 8 and 9 describe two applications of the model, first to ranking statistics journals using citation data and second to rating films and identifying their genre. Chapter 8 also presents methods for dimension selection for the latent space rating model. Chapter 10 introduces the mixed latent model and demonstrates its value in the context of the journal and film applications. Chapter 11 concludes with a discussion of benefits, limitations, and future work.

# CHAPTER 7

# The Latent Space Network Model for Rating

## 7.1 Background to Network-Based Rating

The motivating problem for the latent space rating models we introduce here is the challenge of ranking academic journals based on citations between them. Given many longstanding criticisms of the most widely used rating method, the Impact Factor, Varin et al. (2016) took up this problem. The model they present is a vast improvement over the Impact Factor for several reasons that we discuss in Section 8.1. However, it does not leverage the full structure of the network data in either the modeling framework or presentation of results. Before introducing our model, we present a brief survey of related work.

Considerable attention has been given to the problem of ranking or sorting items based on pairwise comparisons. What we refer to as ranking is elsewhere referred to as exact ranking, and what we refer to as rating, the continuous scoring of items, is elsewhere referred to as approximate ranking. The most common parametric model for rating, or approximate ranking, is the Bradley Terry model (Bradley and Terry, 1952). In its most general form, each element $i$ to be rated is associated with a score, $p_i$, such that for a pair of items, $i, j$,

$$P(i > j) = \frac{p_i}{p_i + p_j}. \tag{7.1}$$

In network terminology $i > j$ determines that a directed edge between nodes $i$ and $j$ points from $i$ to $j$ and not the other way around. In more common parlance we associate this with $i$ defeating $j$ or being preferred to $j$. A common form of this model assumes that $p_i = e^{\mu_i}$, such that

$$logit(P(i > j)) = log\left(\frac{P(i > j)}{1 - P(i > j)}\right) = log\left(\frac{P(i > j)}{P(j > i)}\right) = \mu_i - \mu_j. \tag{7.2}$$

The model presented by Varin et al. (2016) that we discuss in Section 8.2 is a type of Bradley-Terry model adapted for citation data.

Though the Bradley-Terry model is often a touchstone, research on ratings from pairwise comparisons spans parametric models, more general classes of models, and model-independent algorithms. For example, Shah et al. (2016) study error rates of parametric ordinal models, of which the Bradley-Terry model is a special case. Shah et al. (2015) examine an even broader class of models in which the latent probability matrix underlying pairwise comparison data need only satisfy strong stochastic transitivity. Braverman and Mossel (2007) describe a permutation-based model where items have an unobserved order and any item has greater than a one-half chance of "beating" an item lower in the order. This is reflected as a positive entry in an observed signed network. Model-independent approaches include the win-counting algorithm analyzed by Shah and Wainwright (2015) and the random-walk algorithm of Negahban et al. (2012).

Within this body of work, much attention is paid to finding optimal ratings assuming incomplete pairwise data, i.e., a partially observed network, and, conversely, the number of pairs or optimal set of pairs needed to achieve certain bounds on rating error. This problem has many common applications, such as ranking sport and game competitors who have only competed against a subset of others, and forming product recommendations given limited information about consumer preferences between products. Examples of recent work in this vein are Chatterjee (2015), Chen et al. (2016) and Mao et al. (2017).

There are many specific settings for which rating systems based on pairwise comparison data have been developed. For example, competitive chess players are ranked by the Elo system, which accounts for an individual's win-loss record and the strength of their opponents (Elo, 1978). Several variations of the Elo system exist, and it has additionally been applied to other gaming and sport competitions (see Park and Yook, 2014; Silver and Fischer-Baum, 2015). The ranking of sports teams and competitors has received much attention, and we

refer the reader to, for example, the surveys of sports ranking and rating methods by Barrow et al. (2013) and Stefani (1997). One example of a network-based ranking system developed for US college football is that of Park and Newman (2005), which we mention because of its similarity to informetric rating systems discussed below. The method calculates a *win* score for each team as the exponentially decreasing sum of its wins, opponents' wins, opponents' opponents' wins, etc. A *loss* score is analogously calculated and the final rank of the team is its win score minus its loss score. This system can be viewed as an extension of Katz centrality.

Citation data differs in several key respects from data generated by contests, which motivates most rating methods. First, the outcome of a competition is presumed to be the result of a disparity in quality between the competitors, noise, and potentially covariates such as home-field advantage in sports. In contrast, for one journal to cite another is only somewhat indicative of their relative quality. It more strongly relates to similarity in topic matter between the two journals.

Furthermore, schedules of competitions are almost always externally determined, whether by tournament organizers or test designers. They are also usually incomplete, as it would not be feasible to carry out all possible matchups. On the other hand, a journal citation matrix contains complete pairwise comparison data, in the sense that a lack of citations between two journals indicates a lack of communication between them, rather than missing information. (We could still encounter truly missing data due to data corruption or censorship. This is not the focus of our work, and our considered applications draw on complete data.)

The latent space rating models we propose are designed for scenarios like that of journal ranking, in which pairwise comparison data is not only indicative of the relative quality of the underlying items, but of similarities between them; and the "schedule" of comparisons is endogenous. Our challenge is how to decompose the data into factors that compose the rating and those that capture item similarity. We also seek a model that elucidates the structure of the network and drivers of the ratings, especially given the high level of uncertainty that we find in estimated rankings.

We turn our attention now to previous work on the specific task of ranking journals and the related task of ranking web pages. Both have drawn major interest over the last several decades. Networks of links between web pages are more similar to journal citation networks in that links represent the flow of information and shared subject matter between items, rather than outcomes of competitions. In addition, the data is similarly complete in the sense that lack of connection between objects indicates lack of communication, rather than missing information. Perhaps the best-known ranking algorithm for network data in general is the PageRank algorithm (Page et al., 1999), which formed the foundation of the Google search engine. In brief, it ranks web pages by the eigenvector of the dominant eigenvalue of a Markov transition matrix which describes traffic flow among web pages. The rating corresponds roughly to the equilibrium amount of time an internet user would spend on a specific web page. It is worth noting that the development of PageRank was influenced by earlier work in citation analysis. Pinski and Narin (1976) proposed a similar eigenvalue-based method for scoring journals, with an application to ranking physics journals. Because PageRank is generalizable, fast, and has a guaranteed solution, it has been applied to many settings, including biology, chemistry, ecology, neuroscience, physics and sports, as well as author and journal ranking (Gleich, 2014; Maslov and Redner, 2008).

The main advantage of a method like PageRank over raw count-based metrics (including Impact Factor, described below) is that references from highly rated pages are more highly valued. As Page et al. (1999) put it in an early paper, "we give the following intuitive description of PageRank: a page has high rank if the sum of the ranks of its backlinks is high." This is crucial in ranking web pages where most linking pages are not of any interest to a user. However, it is not as important when ranking fairly homogeneous catalogs of items, and in that context can lead to overemphasizing popularity, as we will illustrate in Chapter 8. A related algorithm to PageRank is Hyperlink-Induced Topic Search (HITS), also called hubs and authorities Kleinberg (1999). It is similarly iterative, and explicitly aimed at extracting the most valuable web pages from a massively heterogeneous pool. Accordingly, it assumes a certain structure of hubs and authorities which is not appropriate for journal ranking.

Eigenfactor is a method similar to PageRank, but tailored specifically to rank journals (Bergstrom, 2007). The main differences are 1) the data is normalized, i.e., we model the percentage of citations journals send to each other instead of raw counts, and 2) rather than affording every journal a uniform minimum weight this amount is scaled to the number of articles published by each journal. These changes reflect the greater uniformity and much smaller scale of the journal ranking problem as compared to web page ranking. In tandem, the Article Influence score also proposed by Bergstrom (2007) is a measure of the "average influence" of an article in a given journal. It is proportional to the Eigenfactor score divided by the number of articles published by the journal. In Chapter 8 we compare our journal ranking results to rankings by these methods.

There is a close relationship between the network-based rating methods described above and established measures of network centrality. In particular, PageRank and Eigenfactor are related to eigenvector and Katz centrality; the Park and Newman method is related to Katz centrality; and the Impact Factor, described in Section 8.1, is related to degree centrality. In contrast, the latent space method introduced below makes allowance for the fact that influence and centrality are not synonymous. An item may not be central, but may nonetheless have the ability to influence disparate items, and this is reflected in its rating.

## 7.2 Applications of Latent Space Network Models

Latent space network models have been used for various applications, but not specifically for rating, as far as we know. For example, Hoff and Ward (2004) used a latent space model to visualize the structure of relationships between political actors in Central Asia. Gormley and Murphy (2007) developed a latent space model for rank data and used it to co-locate voters and candidates in an Irish parliamentary election. Sewell and Chen (2015) employed such a model to dynamic network data to study network stability and the relationship between popularity and stability. They subsequently extended their model to fit dynamic clusters (Sewell and Chen, 2016).

Although latent space network models have not been used to rank authors or journals,

as we do in Section 8, they have been applied previously to analyze citation networks. For example, Sarkar and Moore (2005) developed a dynamic latent space model that can track the relationships between authors and their level of influence over time, which they illustrated on NIPS co-authorship data. Latent class network models have also been used to discover communities in citation networks, for example in Leicht et al. (2007), but the addition of latent positions adds capability to identify externally mislabeled nodes, as we demonstrate in Section 9.

## 7.3 The Latent Space Network Model for Rating

The latent space network rating method we introduce incorporates the following features uniquely: 1) accounting for similarity between nodes; 2) providing measures of uncertainty in estimates; 3) meaningfully and easily visualizing results; 4) distinction between influence and centrality; and 5) simple implementation in R (R Development Core Team, 2016). The method is applicable to directed networks, including those derived from pairwise comparison data. We focus on the case where the edges are valued and can be reasonably modeled as Poisson-distributed.

We denote a network of $n$ nodes by its adjacency matrix $Y = \{y_{ij}\}, 1 \leq i, j \leq n$, where $y_{ij}$ denotes the value of the edge from node $i$ to node $j$. Self-edges are disallowed. A dyad in the network consists of two directed edges, $Y_{ij}$ and $Y_{ji}$. The latent space models introduced by Hoff et al. (2002) assume that nodes in a network have implicit positions in "social space". Given these $d$-dimensional positions, $Z$, as well as possible covariates, $X$, and corresponding parameters $\beta$, the probability of an edge is independent of all other edges. Thus, the probability of a graph $Y$ is the product over its edges

$$P(Y|X, \beta, Z) = \prod_{i \neq j} P(y_{ij}|x_{ij}, \beta, z_i, z_j).$$

Hoff (2003) recast the parameters with unobserved random effects,

$$P(Y|X,\beta,\gamma) = \prod_{i \neq j} P(y_{ij}|x_{ij},\beta,\gamma_{ij}), \tag{7.3}$$

where here we model $\gamma$ as in the "distance model" of Hoff (2003):

$$\gamma_{ij} = a_i + b_j + \epsilon_{ij},$$

$$\epsilon_{ij} = f(z_i, z_j) = -\|z_i - z_j\|. \tag{7.4}$$

We consider $a_i$ and $b_j$ to be node-specific sender and receiver effects. $\|z_i - z_j\|$ is the Euclidean distance between nodal positions $z_i$ and $z_j$. (Hoff et al. (2002) also considered an asymmetric projection model which we do not employ.) Although the positions can be in high-dimensional space, we usually consider one to three dimensions for reasons of interpretability, visualization, or parsimony.

Adapting machinery of the generalized linear model (GLM), let

$$\begin{aligned} E(y_{ij}) &= g^{-1}(\eta_{ij}) \\ \eta_{ij} &= \beta' x_{ij} + a_i + b_j - \|z_i - z_j\|. \end{aligned} \tag{7.5}$$

In the context of our applications, we assume $y_{ij}$ is Poisson distributed and let $g$ be the standard log link function. In some cases other distributions may be more appropriate, but here we use the Poisson because we are dealing with count data. This is in contrast to the binomial distributions of edge weights in the quasi-Stigler model described in Section 8.2. Unlike the quasi-Stigler model the estimates are not conditioned on the total weight of each dyad $(y_{ij} + y_{ji})$.

We note some features of the model:

- Increasing distance between $z_i$ and $z_j$ implies decreasing expectation of $y_{ij}$. One way to view this is as controlling for similarity between nodes. Nodes with salient similarities

are likely to have fitted positions relatively close together. Some of the magnitude of their connection is attributable to their similarly, and the rest to their individual sender ("push") and receiver ("pull") effects.

- The effect of positions on expected edge weights is symmetric, affecting both edges in a dyad equally. To condition on total weight, as in the quasi-Stigler method discussed below, greatly diminishes the value of estimated positions.

- The *rating* or *score* of node $i$ is its receiver minus sender coefficient,

$$rating_i = b_i - a_i. \tag{7.6}$$

  Its *rank* is derived from its order among the ratings.

## 7.4 Parameter Estimation

With the introduction of latent space network models, Hoff et al. (2002) developed a Markov chain Monte Carlo (MCMC) estimation algorithm. Hoff (2003) added the capacity to fit random effects, and Krivitsky et al. (2009a) extended the model further and refined the underlying algorithm. In this section we present an overview of the MCMC estimation described by those authors, adapted for the latent space rating model and embellished with details from the **latentnet** implementation (Krivitsky and Handcock, 2008, 2017). We conclude the section by discussing previous use of quasi-Newton estimation for latent space models and trade-offs between this strategy and MCMC.

First, we describe the Bayesian framework and initialization method. This applies to both MCMC estimation and direct optimization by a quasi-Newton method. Our aim is to return a sample from the posterior parameter distribution and desired point estimates, such as a maximum likelihood estimate, posterior mean and mode. We adapt Equation 7.3 for the case when the random effects are as described in (7.4), and we have no covariates, i.e., $\beta$ is reduced to an edge intercept.

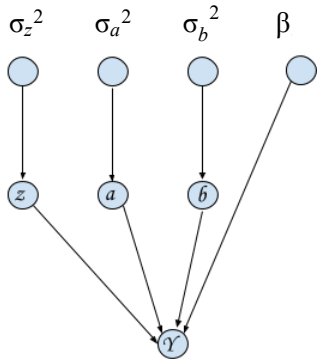


Figure 7.1: Dependence structure of the latent space sender-receiver model.

$$P(\theta|Y) \propto P(Y|\theta)P(\theta) = \prod_{i \neq j} P(y_{ij}|\beta, z_i, z_j, a_i, b_j)P(\theta). \tag{7.7}$$

We posit independent normal distributions for the components of $\gamma_i$, $i = 1, ...., n$ (Hoff, 2003).

$$a_i \sim N(0, \sigma_a^2))$$

$$b_i \sim N(0, \sigma_b^2))$$

$$z_i \sim MVN_d(0, I_d * \sigma_z^2))$$

We expand Equation 7.7:

$$P(Y|\theta)P(\theta) = \prod_{i \neq j} P(y_{ij}|\beta, z_i, z_j, a_i, b_j)P(a|\sigma_a^2)P(b|\sigma_b^2)P(z|\sigma_z^2)P(\beta)P(\sigma_a^2)P(\sigma_b^2)P(\sigma_z^2) \tag{7.8}$$

Equation 7.8 reflects the dependence structure of the parameters as displayed in Figure 7.1. As stated above we assume $y_{ij}|\beta, z, a, b$ is Poisson distributed with mean parameter given by Equation 7.5, where $g^{-1} = exp$.

We place priors on the elements of $\beta$, $\sigma_z^2$, $\sigma_a^2$ and $\sigma_b^2$.

$$\beta \sim N(\mathbf{0}, \sigma_\beta^2)$$

$$\sigma_a^2 \sim \text{Scale-inv-}\chi^2(v_a, s_a^2))$$

$$\sigma_b^2 \sim \text{Scale-inv-}\chi^2(v_b, s_b^2))$$

$$\sigma_z^2 \sim \text{Scale-inv-}\chi^2(v_z, s_z^2))$$

The estimation algorithm of **latentnet** sets default values for the hyperparameters to generate diffuse distributions on the prior parameters. The default value for $\sigma_\beta^2$ is 9 to allow a wide range of $\beta$ values. Low degrees of freedom ($v_a = v_b = 3$) reflect uncertainty in the values of $\sigma_a^2$ and $\sigma_b^2$, while the default scale parameters ($s_a^2 = s_b^2 = 1$) curtail them to a wide but reasonable range. The default values for $v_z$ and $s_z^2$ are $\sqrt{n}$ and $\frac{1}{8}\sqrt[d/2]{n}$. These values reflect that larger networks tend to take up more space, but as observed network size increases the influence of prior variance should decline. For discussion of the choice of hyperparameters see Krivitsky et al. (2009a) and Krivitsky and Handcock (2008). These values are fixed throughout the estimation process.

We must supply initial parameter values. Below we list the default initializations implemented in **latentnet**. They are functions of the observed network. While they may speed convergence in some cases, in our applications in low dimension we found that random initialization on a reasonable scale performs as well or better.

$z^{(0)}$: The positions are initialized through either multidimensional scaling (MDS) or normal draws. In the former, the geodesic distances for all dyads are computed from the binary adjacency matrix $Y_b$. Disconnected pairs are given distances of $n$. The initial value $z^{(0)}$ is then computed by multidimensional scaling, returning optimal $d$-dimensional coordinates whose Euclidean distances best approximate the geodesic distances between nodes. In the latter, $z^{(0)}$ is generated via independent draws from a normal distribution. Hoff et al. (2002) noted that the choice of initialization does not impact their results. We found that the scale of the initial positions was more important than the values, with preference for smaller starting values.

$$a_i^{(0)} = logit\left(\frac{Y_{b_i\cdot} + 1}{n - 1 + 2}\right) - \frac{1}{n}\sum_{j=1}^{n} logit\left(\frac{Y_{b_i\cdot} + 1}{n - 1 + 2}\right)$$

The initialization of sender parameters shown above is derived by considering initial nodal degrees as binomially distributed with $n-1$ trials, observed success probability $\frac{Y_{b_{i\cdot}}}{n-1}$, and a uniform prior on success probability. As a reminder, $Y_b$ denotes the binary adjacency matrix and $Y_{b_{i\cdot}}$ denotes the $i$th row sum of $Y_b$. The initialization of receiver coefficients is analogous.

$$b_i^{(0)} = logit\left(\frac{Y_{b\cdot i}+1}{n-1+2}\right) - \frac{1}{n}\sum_{i=1}^{n} logit\left(\frac{Y_{b\cdot i}+1}{n-1+2}\right)$$

$$\beta^{(0)} = logit\left(\frac{1}{n(n-1)}\sum_{i,i\neq j}\mathbb{1}(y_{ij} > \overline{y_{ij}})\right) + \frac{1}{\binom{n}{2}}\sum_{i,i<j}\|z_i^{(0)} - z_j^{(0)}\|$$

The initial intercept is composed of a "weight intercept," a valued-network analog of graph density, plus a "distance intercept," the average initial pairwise distance.

The initial values of $\sigma_z^{2(0)}, \sigma_a^{2(0)}$, and $\sigma_b^{2(0)}$ are the variances of $z^{(0)}, a^{(0)}$, and $b^{(0)}$, respectively.

*Parameter Updates*:

Before starting an MCMC chain, **latentnet** employs an intermediate optimization step. It uses the bounded quasi-Newton optimization routine of Byrd et al. (1995), as implemented in the `optim` function of the R base package **stats**. This returns starting values for MCMC with higher posterior probability than the initial values described above. In our applications the results are competitive with the final MCMC output, and we compare them in Section 8.5.3.

After intermediate optimization, the MCMC chain runs through a suitably long burn-in period. If the automated burn-in length is insufficient, proper length can be determined by carrying out MCMC diagnostics on the results, such as those in the `mcmc.diagnostics` function of **latentnet**. Once starting values for MCMC sampling are determined, parameters are updated as follows:

- $\sigma_a^2, \sigma_b^2, \sigma_z^2$:

The variances of the sender, receiver, and position parameters may be sampled directly from their posterior distributions because they were assigned conjugate priors.

$$\sigma_a^2|a \sim \text{Scale-inv-}\chi^2(v_a + n, \frac{v_a s_a^2 + \sum_{i=1}^{n} a_i^2}{v_a + n}))$$

$$\sigma_b^2|b \sim \text{Scale-inv-}\chi^2(v_b + n, \frac{v_b s_b^2 + \sum_{i=1}^{n} b_i^2}{v_b + n}))$$

$$\sigma_z^2|z \sim \text{Scale-inv-}\chi^2(v_z + n * d, \frac{v_z s_z^2 + \sum_{i=1}^{n*d} z_{i,d}^2}{v_z + n * d}))$$

- Actor-specific parameters, $a, b, z$:

  The sender, receiver and position parameters cannot be sampled directly and may be strongly correlated. They are updated for each actor in random order by Metropolis-Hastings block updates.

  1. Propose $z_i^*, a_i^*, b_i^*$ from symmetric proposal distributions.

  $$z_i^* \sim MVN_d(z_i, \tau_z^2 I_d)$$

  $$a_i^* \sim N(a_i, \tau_a^2)$$

  $$b_i^* \sim N(b_i, \tau_b^2)$$

  2. Accept as a block with probability $min(1, \frac{P(Y|z^*,a^*,b^*,\beta)P(z^*)P(a^*)P(b^*)}{P(Y|z,a,b,\beta)P(z)P(a)P(b)})$.

- $\beta$, shift of random effects, position scale:

  To speed convergence, a simultaneous shift in $\beta$ and the random effects and a rescaling of the positions is proposed. The magnitude of the shifts and multiplier is proposed by:

  $$(h_\beta, h_a, h_b, h_z) \sim MVN_4(0, \tau_{\beta,a,b,z})$$

  $$\beta^* = \beta + h_\beta$$

  $$a^* = a + h_a$$

$$b^* = b + h_b$$

$$z^* = exp(h_z)z$$

$$\sigma_z^{2*} = exp(2h_z)\sigma_z^2$$

The move is block-accepted or rejected. For discussion of the acceptance probability see Section 3.2 of Krivitsky et al. (2009a).

- *Proposal Variances* $\tau_z^2, \tau_a^2, \tau_b^2, \tau_{\beta,a,b,z}$:

  The variances of the proposal distributions are set adaptively during the burn-in period to stay near a fixed acceptance rate, with a default target rate of 0.234 (Krivitsky and Handcock (2008), following Neal and Roberts (2006)).

*MCMC Post-processing*:

The likelihood depends on positions only through their pairwise distances, and is invariant to rotations, reflections and translations of the positions. We are interested in the posterior variance in positions that comes from changing distances between points rather than distance-preserving transformations. One way to address this and stabilize our estimates is to store not the sampled positions, but a transformed set that has minimal squared distance to a set of reference positions. This is the Procrustean transformation used by Hoff et al. (2002). They let $z^*_{store} = argmin_{Tz^*} tr(z_{ref} - Tz^*)^\top (z_{ref} - Tz^*)$, where $T$ is the set of distance-preserving transformations.

There may be strong correlation or near non-identifiability between a node's actor-specific parameters, especially if it is poorly connected. To reduce instability in the estimate Shortreed et al. (2006) considered the parameter estimate that is optimal in the sense of minimizing Bayes risk with a Kullback-Leibler (KL) loss function. Their "MKL" estimate minimizes the posterior expectation of the KL divergence from its predictive distribution of networks to the posterior predictive distribution of networks. As such, the MKL estimate only pertains

to the parameters on which networks are immediately dependent.

$$\theta_{MKL} = argmin_{\widetilde{Z}\widetilde{a},\widetilde{b},\widetilde{\beta}} \left[ E_{Z,a,b,\beta|Y_{obs}} \left[ \sum_{Y} log(\frac{P(Y|Z,a,b,\beta)}{P(Y|\widetilde{Z}\widetilde{a},\widetilde{b},\widetilde{\beta})}) P(Y|Z,a,b,\beta) \right] \right] \tag{7.9}$$

The MKL positions are more stable than standard point estimates because they average over all networks. They require the posterior sample to calculate, so, unlike Hoff et al. (2002), the sampled positions are transformed with MKL positions as reference *after* the sampling is complete.

### 7.4.1 Quasi-Newton Estimation

Typically, quasi-Newton estimation has been used as an initialization step of MCMC, as in Hoff et al. (2002) and (Krivitsky and Handcock, 2017). Handcock et al. (2007) employed it for the first stage of a two-stage maximum likelihood method to estimate a latent space cluster model. They found that the two-stage MLE gives a good match to the MCMC fit in terms of cluster membership and relative positions, but the clusters are more spread out. However, in that case, it was only applied to the simplified model without clusters, so it could not capture dependence between positions and cluster assignments. In addition, the networks being modeled were binary, so edges provided less granular information than in valued networks.

Methods proposed to speed up latent space mode fitting, such as the variational Bayesian method of Salter-Townshend and Murphy (2009, 2013) or the case-control approximate likelihood of Raftery et al. (2012), treat Bayesian MCMC as the benchmark for estimation speed and complexity. These methods compromise the true likelihood function and introduce bias on behalf of speed, justified by impracticality of MCMC estimation. In examples of Salter-Townshend and Murphy (2009), positions fit by the variational Bayesian algorithm also show greater within-cluster variance than MCMC estimates. In our trials on networks of up to several hundred nodes, results from quasi-Newton estimation closely approximate those from MCMC, as we discuss in Sections 8.5.3 and 9.2.

The main argument against a quasi-Newton estimation method is that it is not guaranteed to converge to a global optimum since the likelihood function is not convex. However, there

are several reasons why it is often still successful in practice. First, in applications well suited to latent space network rating the search space for the positions is relatively small. This is especially true when positions are in low dimension. Second, although MCMC estimation is theoretically guaranteed to converge to the true distribution, it may face prohibitively slow mixing time and fail to converge to the global optimum. We see a minor example of this in Section 8.5.3. Third, the speed of the quasi-Newton method means we can consider many initial values to increase our chance of finding the global optimum. If necessary, we can still use MCMC estimation to validate the quasi-Newton results, but with a much smaller burn-in than would otherwise be required.

Although quasi-Newton estimation does not return a sample from the posterior parameter distribution like MCMC estimation does, we can still approximate the uncertainty in ratings using a Poisson GLM. This is described in Section 8.5.3. While there may be strong dependence between positions and sender or receiver parameters, the dependence between the positions and the ratings (receiver minus sender) is much less. This is confirmed by the similarity we find between MCMC estimates of ratings uncertainty and those by the Poisson GLM.

The particular quasi-Newton algorithm that we employ is the L-BFGS-B algorithm developed by Byrd et al. (1995). For a complete description of the algorithm see Byrd et al. (1995) and Byrd et al. (1994), but we note some key points. This method is well-suited to our rating problem, which is a non-linear optimization with simple bounds on some parameters - the variance parameters must be greater than zero. We also have well-defined gradients and second-order partial derivatives as long as positions are unique, which is generally the case and can be easily dealt with if not. In addition, L-BFGS-B is a limited memory algorithm that does not calculate the full Hessian matrix, which has $\mathcal{O}(n^2)$ terms. Rather, it makes use of the BFGS quasi-Newton method (see for example Nocedal and Wright (2006)) with compact representations of limited memory matrices developed by Byrd et al. (1994), such that storage and updates to the approximate Hessian are $\mathcal{O}(n)$.

At each iteration of the algorithm, the aim is to minimize a quadratic form of the objective function. In our case we minimize the negative of the posterior probability function, which

for convenience refer to below as $f(x_k)$, where $x_k$ is our estimate of the optimum at the $k^{th}$ iteration.

$$m_k(x) = f(x_k) + g_k^\top (x - x_k)^\top B_k (x - x_k), \tag{7.10}$$

where $g_k$ is the gradient of $f$ at $x_k$ and $B_k$ is the limited memory approximation to the Hessian at iteration $k$. The approximation $B_k$ is formed from the updates to $x_k$ and $g_k$ over the previous $m$ iterations, where $m$ is a small integer. (In the R **optim** implementation of this algorithm the default value of $m$ is five.) These updates provide information about the curvature of the function near the current value. They are concatenated into a matrix of size $n \times 2m$, and $B_k$ is represented using this term in a product of size $n \times 2m, 2m \times 2m, 2m \times 2m$. Using this product expression, nearly all calculations involving $B_k$ can be done in $\mathcal{O}(m^2)$ operations. Even so, the dependence structure of the model results in some costly computations. In particular the gradient is calculated at each iteration and this takes $\mathcal{O}(n^2)$ operations (see Appendix 7.6 and 10.5 for gradient calculations). However, the number of iterations needed is a tiny fraction of the number needed for MCMC estimation, which involves likelihood calculations costing $\mathcal{O}(n^2)$ at each iteration.

## 7.5 Comparison to the Gravity Model

Before moving on to applications of the latent space rating model, we consider its shared attributes with the gravity model of social science, which is typically used to model exchange between nodes in trade and transportation networks. Combining formulations of the gravity model in Sarzynska et al. (2016) and Ward et al. (2013), we describe a generalized gravity model as,

$$E(Y_{ij}) = G m_i^{\beta_a} m_j^{\beta_b} f(d_{ij}), \tag{7.11}$$

where f is the *deterrence function* describing the damping effect of distance, $d_{ij}$, between nodes on expected edge weights. In the equation for gravitational force, $m_i$ is an object

mass, $G$ is the gravitational constant, and f is the reciprocal of squared Euclidean distance between $i$ and $j$. Other common $f$ functions use different powers of Euclidean distance or the natural exponential function of distance. Under the latter formulation with $f = e^{-d_{ij}}$, the log of the expected edge weight is

$$log(E(Y_{ij})) = log(G) + \beta_a log(m_i) + \beta_b log(m_j) - d_{ij}. \tag{7.12}$$

If we consider $G$ and $m$ to be unknown values rather than empirically determined then this lies somewhere between a symmetric sociality model (see Krivitsky et al. (2009a)), with equivalence if $\beta_a = \beta_b$, and our asymmetric rating model,

$$log(E(Y_{ij})) = \beta + \beta_a * m_i + \beta_b * m_j - d_{ij}. \tag{7.13}$$

Under this model expected edge weights are not symmetric, but the difference is constrained by

$$log(E(Y_{ij})) - log(E(Y_{ji})) = (\beta_a - \beta_b)(m_i - m_j). \tag{7.14}$$

In other words, we may consider our latent space rating model of Equation 7.5 to be an extension of the gravity model for directed networks. Rather than having a single empirical mass for each node we estimate a sender and receiver mass for each node. We inherently account for a family of exponential deterrence functions, $c_1 e^{-c_2 d_{ij}}$, by this model as $c_1$ is absorbed by $\beta$ and $c_2$ is implicit in the optimal scale of the fit positions. Note that if the deterrence function $f$ is instead a power of Euclidean distance then the log of expected edge weight contains a log-distance term. This is likely to be a better model in the context of trade and transportation networks given the large distances between nodes. However, in our applications the positions are relatively close together so we prefer the non-logged distance (exponential deterrence function) which avoids complications near zero and facilitates interpretability.

## 7.6 Appendix: Calculations for Quasi-Newton Algorithm

**1. Log posterior probability of the Euclidean distance model, up to a constant:**

Calculations assume no self-edges.

Let $\lambda_{ij} = exp(\beta + a_i + b_j - \|z_i - z_j\|)$.

$$
\begin{aligned}
log(P(\theta|Y)) = & \sum_i \sum_{j \neq i} y_{ij}(log(\lambda_{ij})) - \lambda_{ij} - log(y_{ij}!) + \\
& log(\frac{1}{\sqrt{2\pi\sigma_\beta^2}} exp(\frac{-\beta^2}{2\sigma_\beta^2})) + \sum_i log(\frac{1}{\sqrt{2\pi\sigma_a^2}} exp(\frac{-a_i^2}{2\sigma_a^2})) + \sum_i log(\frac{1}{\sqrt{2\pi\sigma_b^2}} exp(\frac{-b_i^2}{2\sigma_b^2})) + \\
& \sum_{i,d} log(\frac{1}{\sqrt{2\pi\sigma_z^2}} exp(\frac{-z_{i,d}^2}{2\sigma_z^2})) + log(\frac{s_a^{2\,\frac{v_a}{2}}}{\Gamma(\frac{v_a}{2})}) - \frac{v_a s_a^2}{2\sigma_a^2} - log(\sigma_a^2)(1 + \frac{v_a}{2}) + \\
& log(\frac{s_b^{2\,\frac{v_b}{2}}}{\Gamma(\frac{v_b}{2})}) - \frac{v_b s_b^2}{2\sigma_b^2} - log(\sigma_b^2)(1 + \frac{v_b}{2}) + log(\frac{s_z^{2\,\frac{v_z}{2}}}{\Gamma(\frac{v_z}{2})}) - \frac{v_z s_z^2}{2\sigma_z^2} - log(\sigma_z^2)(1 + \frac{v_z}{2})
\end{aligned}
$$

**2. Gradient calculations:**

$$
\begin{aligned}
\frac{\partial \mathbb{LP}}{\partial a_i} &= -\frac{a_i}{\sigma_a^2} + \sum_{j \neq i} y_{ij} - \lambda_{ij} \\
\frac{\partial \mathbb{LP}}{\partial b_i} &= -\frac{b_i}{\sigma_b^2} + \sum_{j \neq i} y_{ji} - \lambda_{ji} \\
\frac{\partial \mathbb{LP}}{\partial \beta} &= -\frac{\beta}{\sigma_\beta^2} + \sum_i \sum_{j \neq i} y_{ij} - \lambda_{ij} \\
\frac{\partial \mathbb{LP}}{\partial z_{i,d}} &= -\frac{z_{i,d}}{\sigma_z^2} + \sum_{j \neq i} \frac{(z_{i,d} - z_{j,d})}{\|z_i - z_j\|} \Big[ - y_{ij} - y_{ji} + \lambda_{ij} + \lambda_{ji} \Big] \\
\frac{\partial \mathbb{LP}}{\partial \sigma_a^2} &= -n(2\pi\sigma_a^2)^{-1}\pi + \sum_i \frac{a_i^2}{2}(\sigma_a^2)^{-2} + \frac{v_a s_a^2 (\sigma_a^2)^{-2}}{2} - (1 + \frac{v_a}{2})(\sigma_a^2)^{-1} \\
&= (\sigma_a^2)^{-2}(\frac{v_a s_a^2 + \sum_i a_i^2}{2}) - (\sigma_a^2)^{-1}(\frac{n + v_a + 2}{2}) \\
\frac{\partial \mathbb{LP}}{\partial \sigma_a^2} &= (\sigma_b^2)^{-2}(\frac{v_b s_b^2 + \sum_i b_i^2}{2}) - (\sigma_b^2)^{-1}(\frac{n + v_b + 2}{2}) \\
\frac{\partial \mathbb{LP}}{\partial \sigma_z^2} &= (\sigma_z^2)^{-2}(\frac{v_z s_z^2 + \sum_{i,d} z_{i,d}^2}{2}) - (\sigma_b^2)^{-1}(\frac{n * D + v_b + 2}{2})
\end{aligned}
$$

# CHAPTER 8

# Ranking Statistics Journals from Citation Data

In this chapter we examine rankings of 47 statistics and probability journals. The data set we consider was gathered from Journal Citation Reports and analyzed by Varin et al. (2016). It consists of a $47 \times 47$ matrix of directed citation counts, encompassing within-network citations from 2001 to 2010. We compare latent space model rankings to several competing methods in Section 8.3, visualize our results in Section 8.4, and evaluate competing models and estimation methods in Section 8.5.

## 8.1 Impact Factor

Impact Factor is the most commonly referenced journal rating measure, despite widespread criticism (Seglen, 1997; Amin and Mabe, 2003; Marx and Bornmann, 2013). Impact Factor measures how frequently articles from a specific journal are cited. An Impact Factor of 1.0 means that articles published by that journal in the last two years have been cited once on average (JCR, 2012). Journal Citation Reports also publish modified versions of Impact Factor that exclude journal self-citations or alter the size of the time window to one or five years. These modifications address two problems with Impact Factor, but it has additional pitfalls as a proxy measure of journal quality.

First, Impact Factor does not normalize for article length or out-citations. Whether planned or not, there is documented reciprocity in citations between journals. (In our data the correlation in one-way citations is 0.57.) Second, it does not account for differences in citation patterns between fields, such as mathematics papers tending to have relatively few citations while bioscience papers have many (Leydesdorff et al., 2013). Third, the distribution

of citations counts by article is very long-tailed, with a few articles receiving many citations and most receiving only a few. As Colquhoun wrote in the Discussion on the paper by Varin et al. (2016), "It has been obvious for a long time that it is statistically illiterate to characterize very skew distributions by their mean. And it is statistically illiterate to present point estimates with no indication of their uncertainty" (Colquhoun et al., 2016).

## 8.2 The Quasi-Stigler Model

Varin et al. (2016) introduced the *quasi-Stigler* model to address the criticisms above. The second criticism is not accounted for by the model, but by restricting the data to only 47 out of 110 journals of statistics and probability. The quasi-Stigler model requires that journals are fairly homogeneous and have a relatively high level of citation exchange.

The model measures each journal's "'propensity to export intellectual influence"' (Varin et al., quoting Stephen Stigler). It is a type of Bradley-Terry model, as described in Equation 7.1, adapted to valued citation data. The rank of journal $i$ is determined by its *export score*, $\mu_i$, under the assumption that citations counts, $C_{ij}$, are quasi-binomially distributed as follows:

$$\begin{aligned} E(C_{ij}) &= t_{ij}\pi_{ij} \\ \pi_{ij} &= logit^{-1}(\mu_j - \mu_i) \\ &= \frac{exp(\mu_j - \mu_i)}{1 + exp(\mu_j - \mu_i)} \\ var(C_{ij}) &= \phi t_{ij}\pi_{ij}(1 - \pi_{ij}) \end{aligned} \tag{8.1}$$

where $t_{ij}$ is the observed total number of citations between journals $i$ and $j$, i.e., $t_{ij} = c_{ij} + c_{ji}$. The notation here is slightly different than in Varin et al. (2016) due to transposition of the citation matrix $\{c_{ij}\}$. We note as Varin et al. that the export scores could be obtained as estimates from a quasi-binomial GLM with logit link.

Uncertainty in the export scores is conveyed through the *quasi-variance*, $qvar_i$, of each $\mu_i$. The quasi-variances are estimated to minimize the difference between the true pairwise

variances, $var(\hat{\mu}_i - \hat{\mu}_j)$, and a quasi-variance approximation, $qvar_i + qvar_j$. Although they could convey exact variances, the authors preferred quasi-variances and quasi-standard errors (QSE) because they can be succinctly displayed alongside export scores.

To connect the quasi-Stigler model to the latent space model, consider the quasi-symmetry formulation of the model, $E(C_{ij}) = t_{ij}exp(a_i + b_j)$, where the export score is expanded as $\mu_i = b_i - a_i$ (Varin et al., 2016). As in the latent space model, $a_i$ and $b_i$ are sender and receiver coefficients. If we constrain $E(C_{ij}) + E(C_{ji}) = t_{ij}$ in this formulation it is equivalent to the original formulation. The quasi-symmetry formulation makes clear how quasi-Stigler ratings, like latent space ratings, control for article length and number of articles per journal. The sender effect of each journal is a controlled measure of its tendency to send citations, which is in turn closely related to number and length of articles published by the journal. Subtracting the sender effect from the receiver effect in the rating is therefore a proxy control for the amount published by the journal.

## 8.3 Comparison of Journal Rankings

We compare the rankings from our latent space model to others discussed. Unless otherwise stated, our results are based on two-dimensional MKL parameter estimates from MCMC estimation. These had the highest posterior probability and graph probability of any two-dimensional estimates considered. In Section 8.5 we provide a foundation for the choice of model dimension and estimation method.

Table 8.1 compares journal rankings from the latent space model, quasi-Stigler model, PageRank, Eigenfactor, Article Influence, and Impact Factor. (See Appendix 8.6 for a table of journal names and abbreviations.) Comparisons to other versions of the Impact Factor can be found in Table 4 of Varin et al. for a slightly different data set. Table 8.1 presents ranks rather than ratings to facilitate comparisons across the methods. According to Varin et al. (2016), there is "diffuse opinion" among statisticians that the most prestigious statistics journals are, in alphabetical order, *Annals of Statistics* (AoS), *Biometrika* (Bka), the *Journal of the American Statistical Association* (JASA) and the *Journal of the Royal*

Table 8.1: Comparison of journal rankings. The "big four" have gray background.

| | Latent Space | quasi Stigler | PageRank | Eigenfactor | Article Inf | Impact Fac |
|---|---|---|---|---|---|---|
| AmS | 12 | 12 | 26 | 25 | 22 | 23 |
| AISM | 15 | 15 | 25 | 30 | 30 | 24 |
| AoS | 3 | 2 | 2 | 3 | 3 | 2 |
| ANZS | 21 | 22 | 38 | 35 | 32 | 36 |
| Bern | 9 | 7 | 16 | 12 | 11 | 22 |
| BioJ | 35 | 33 | 19 | 23 | 27 | 16 |
| Bcs | 5 | 5 | 5 | 6 | 12 | 13 |
| Bka | 2 | 3 | 4 | 7 | 5 | 11 |
| Biost | 10 | 9 | 11 | 9 | 6 | 3 |
| CJS | 14 | 16 | 18 | 26 | 18 | 33 |
| CSSC | 46 | 45 | 35 | 31 | 45 | 45 |
| CSTM | 41 | 41 | 20 | 18 | 46 | 44 |
| CmpSt | 39 | 38 | 42 | 36 | 37 | 40 |
| CSDA | 37 | 36 | 8 | 4 | 28 | 18 |
| EES | 32 | 35 | 44 | 37 | 23 | 14 |
| Envr | 24 | 27 | 29 | 32 | 33 | 29 |
| ISR | 17 | 14 | 37 | 38 | 34 | 27 |
| JABES | 26 | 28 | 40 | 39 | 31 | 30 |
| JASA | 4 | 4 | 1 | 1 | 4 | 8 |
| JAS | 47 | 47 | 41 | 40 | 47 | 47 |
| JBS | 43 | 43 | 36 | 27 | 35 | 19 |
| JCGS | 7 | 10 | 12 | 14 | 13 | 17 |
| JMA | 34 | 34 | 10 | 10 | 24 | 21 |
| JNS | 38 | 37 | 28 | 41 | 38 | 42 |
| JRSS-A | 8 | 6 | 24 | 15 | 8 | 5 |
| JRSS-B | 1 | 1 | 3 | 5 | 1 | 1 |
| JRSS-C | 22 | 20 | 22 | 28 | 20 | 35 |
| JSCS | 44 | 44 | 30 | 33 | 42 | 41 |
| JSPI | 31 | 31 | 7 | 8 | 36 | 32 |
| JSS | 42 | 42 | 33 | 16 | 10 | 4 |
| JTSA | 13 | 13 | 34 | 29 | 25 | 34 |
| LDA | 20 | 19 | 27 | 42 | 26 | 26 |
| Mtka | 29 | 29 | 32 | 43 | 39 | 38 |
| SJS | 6 | 8 | 15 | 19 | 15 | 28 |
| StataJ | 23 | 24 | 47 | 20 | 7 | 9 |
| StCmp | 25 | 23 | 17 | 21 | 9 | 10 |
| Stats | 36 | 39 | 39 | 44 | 40 | 39 |
| StMed | 18 | 21 | 6 | 2 | 17 | 7 |
| SMMR | 33 | 32 | 31 | 24 | 14 | 12 |
| StMod | 30 | 30 | 43 | 46 | 29 | 31 |
| StNee | 27 | 26 | 45 | 47 | 41 | 46 |
| StPap | 45 | 46 | 46 | 45 | 43 | 37 |
| SPL | 28 | 25 | 14 | 11 | 44 | 43 |
| StSci | 19 | 18 | 13 | 13 | 2 | 6 |
| StSin | 16 | 17 | 9 | 17 | 21 | 25 |
| Tech | 11 | 11 | 21 | 22 | 16 | 15 |
| Test | 40 | 40 | 23 | 34 | 19 | 20 |

Table 8.2: Correlation of journal rating methods

| | Latent Space | quasi Stigler | PageRank | Eigenfactor | Article Inf | Impact Fac |
|---|---|---|---|---|---|---|
| Latent Space | 1.00 | 0.99 | 0.65 | 0.51 | 0.74 | 0.57 |
| quasi Stigler | 0.99 | 1.00 | 0.66 | 0.52 | 0.75 | 0.59 |
| PageRank | 0.65 | 0.66 | 1.00 | 0.91 | 0.70 | 0.55 |
| Eigenfactor | 0.51 | 0.52 | 0.91 | 1.00 | 0.59 | 0.58 |
| Article Inf | 0.74 | 0.75 | 0.70 | 0.59 | 1.00 | 0.87 |
| Impact Fac | 0.57 | 0.59 | 0.55 | 0.58 | 0.87 | 1.00 |

*Statistical Society*, Series B (JRSS-B). (These journals have gray background in Table 8.1.) Accordingly, they argue that a good rating method will put them near the top.

Although PageRank ranks the "big four" journals highest, it places *Statistics in Medicine* (StMed), *Journal of Statistical Planning and Inference* (JSPI), and *Computational Statistics and Data Analysis* (CSDA) in positions six through eight, much higher than most other methods. These journals have the three highest out-citation counts in our data, and are among the most prolific citers of the top four journals. However, their ratios of in- to out-citations rank 20th, 30th, and 35th. Eigenfactor behaves similarly to PageRank, with some differences reflecting its use of normalized data. (Eigenfactor is strongly correlated with PageRank, 0.91, as shown in Table 8.2.) We conclude that PageRank and Eigenfactor are better measures of centrality or activity level than importance, influence or prestige.

The Impact Factor rankings show the detriment of averaging per article and not controlling for out-citations or a citation's field of origin. For example, *Environmental and Ecological Statistics* (EES) ranks 14th and the *Journal of Statistical Software* (JSS) ranks 4th, though they have the second- and ninth-lowest in-citation counts in our data. On the other hand, the "big four" journals Bka and JASA are ranked 11th and 8th. The Impact Factor ratings are most strongly correlated with Article Influence (0.87, see Table 8.2), which is also normalized by articles per journal and calculated using all citations, not just ones from the 47 journals in our data set. The two methods share some anomalous rankings, such as *Statistical Science* placing 2nd and 6th respectively. We argue that the high rank for this journal is not reflective of its importance within the field. As a review journal it is more likely to disseminate than publish cutting-edge research. It owes its high rank to connections to top journals (confirmed by PageRank and Eigenfactor ranking it 13th), ci-

tations from outside statistics, and a relatively low number of articles published. Its ratio of in- to out-citations ranks 24th in the network. The visualizations provided by the latent space model help to further explain the position of *Statistical Science*, which will be revisited below.

The latent space and quasi-Stigler rankings are very similar (correlation 0.99, see Table 8.2) and seem to provide the best measures of influence or importance. They rank the "big four" journals in the top four positions, the *Journal of Statistical Software* 42nd, and *Statistical Science* 19th and 18th respectively. On the other hand, the methodological journal *Scandinavian Journal of Statistics* (SJS) is highly ranked by both, at 6th and 8th respectively. SJS states its mission as "reporting significant and innovative original contributions to statistical methodology, both theory and applications" (SJS, 2017). Varin et al. (2016) use data from the UK Research Assessment Exercise (RAE), a periodic evaluation of UK university departments, as an external check on the quasi-Stigler rankings. They find some evidence that the quasi-Stigler model provides stronger correlation to RAE assessment of research quality than other methods. Details of that comparison, and its many caveats, are found in Section 6 of Varin et al. (2016).

### 8.3.1 Comparison of Latent Space and Quasi-Stigler Model Output

As the quasi-Stigler and latent space models emerge as the best suited to our ranking priorities, we compare their results most closely. Figure 8.1 (left) plots the compared ranks, with lighter labels for larger differences in rank. Three is the highest observed difference. However, the underlying differences in scores are very small, as shown in the right panel of Figure 8.1. Figure 8.2 plots the posterior distributions of latent space scores. It confirms that the small differences in rank shown in Figure 8.1 are not significant given the uncertainty in the estimated scores, highlighting the importance of capturing model uncertainty.

Uncertainty in ratings is very similar between the two models, with quasi-Stigler standard deviations being 0.04 smaller on average, and at most 0.069 smaller. Under both models, *Stata Journal* (StataJ) has the largest standard error, which is due to the fact that it has by

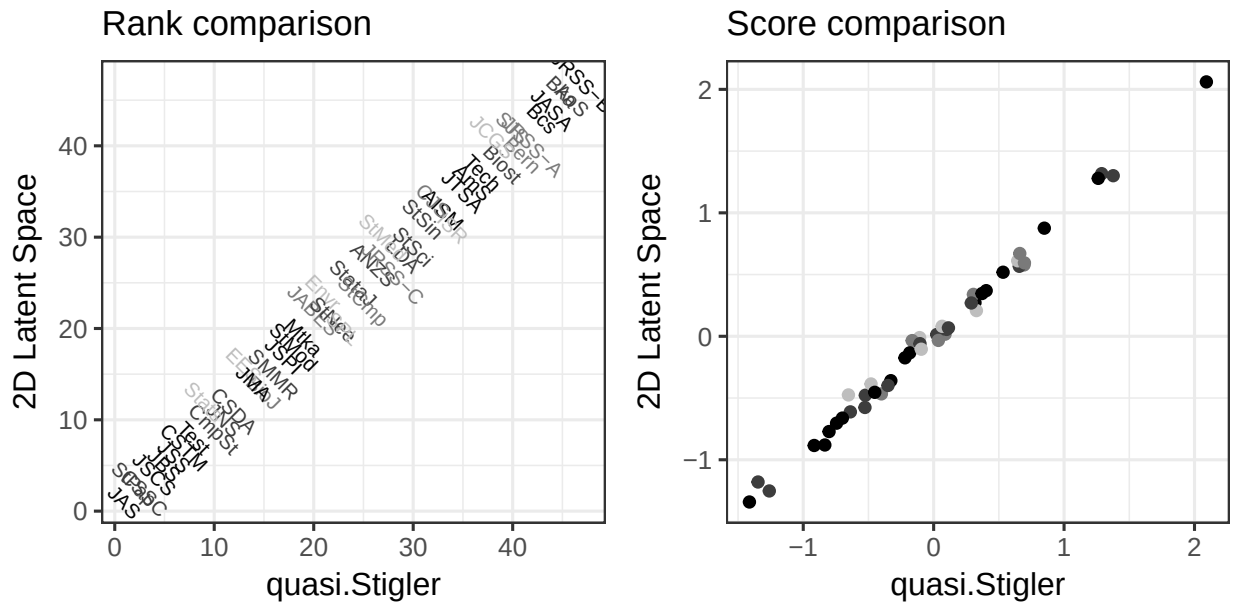


Figure 8.1: (Left) Latent space vs. quasi-Stigler rankings. Better-ranked journals are at the top right, corresponding to higher numbers. (Right) Comparison of scores rather than rankings. Lighter labels means larger differences. Maximum observed rank difference is 3.

far the fewest in- and out-citations. As expected, standard error of scores and citation counts are inversely correlated. Figure 8.3 shows point estimates from each model inside intervals of 1.96 times the standard error in either direction. On the left, the variances of the latent space scores are calculated from a sample of 5000 draws from the posterior distribution of parameters stored during MCMC estimation. On the right, the longer intervals are calculated from standard errors extracted from the scaled covariance matrix of the model. The interval for *American Statistician* (AmS) is missing because its coefficients were fixed at zero for identifiability. The interior intervals are 1.96 quasi-standard errors (QSE) in each direction. These "comparison intervals" are analogous to those in Figure 4 of Varin et al. (2016). We see that they are smaller and more variable than the true standard error intervals. The justification by Varin et al. to present uncertainty through quasi-standard errors is that they can be listed alongside estimates in a table, and allow a familiar Pythagorean estimate of the standard error of a difference of two export scores. However, in a centipede plot as shown, the true standard errors are just as compact and easy to compare, and reveal that

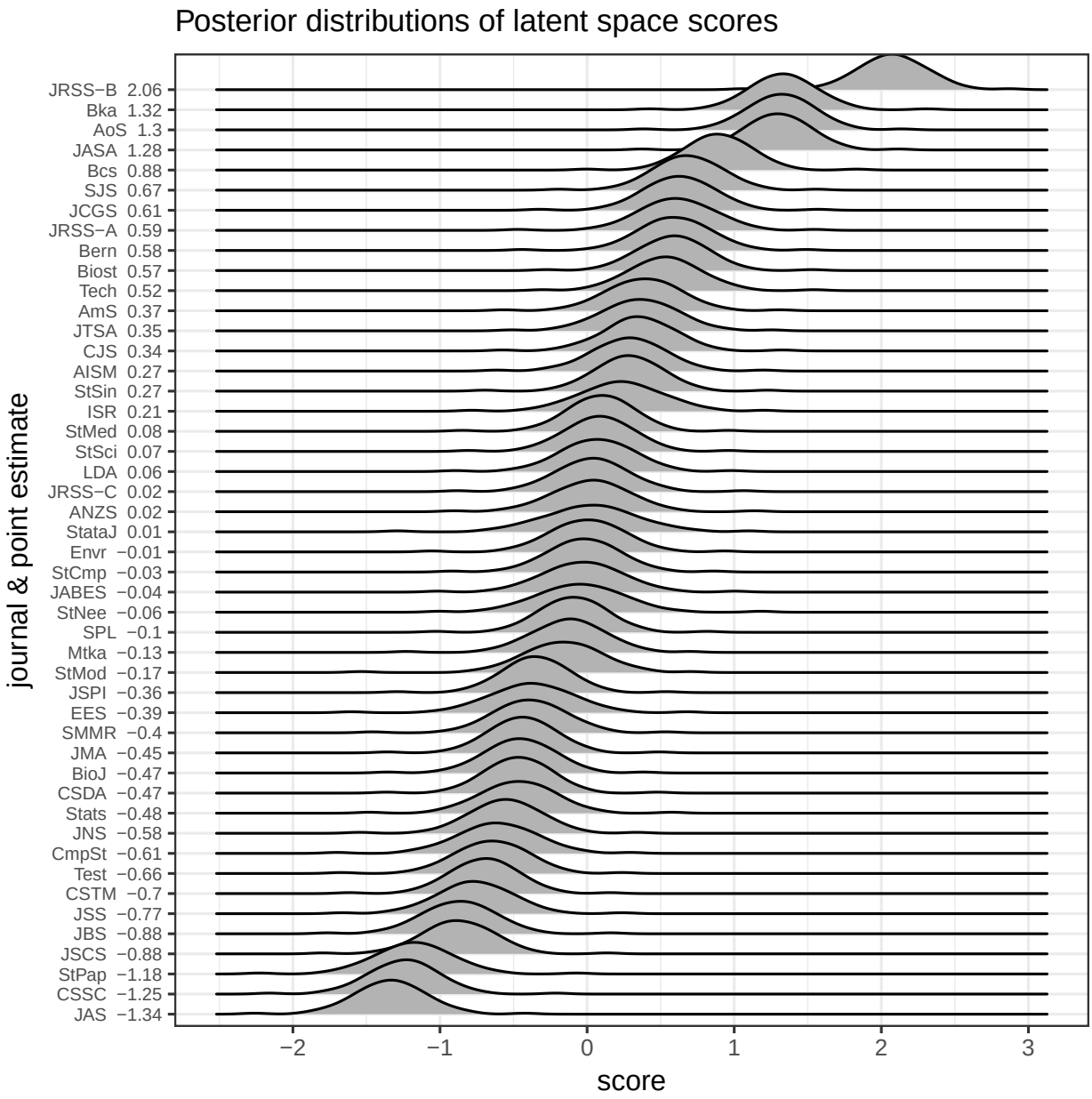


Figure 8.2: Posterior distributions of latent space score. Most small differences in ratings are not signficiant.

the uncertainty in scores is roughly equivalent between the two models.

The near identical rankings from these two models belie the fact that the quasi-Stigler model is conditioned on dyad totals. In contrast, the positions in the latent space model help to explain those totals. The distances resulting from the positions exert the only dyad-

specific symmetric effect in the model. (We fit a binomial two-dimensional latent space model conditioned on dyad totals and found the estimated positions to be very close to zero.) A major advantage of the latent space model is that the positions can be visualized and help us to better understand the ratings.

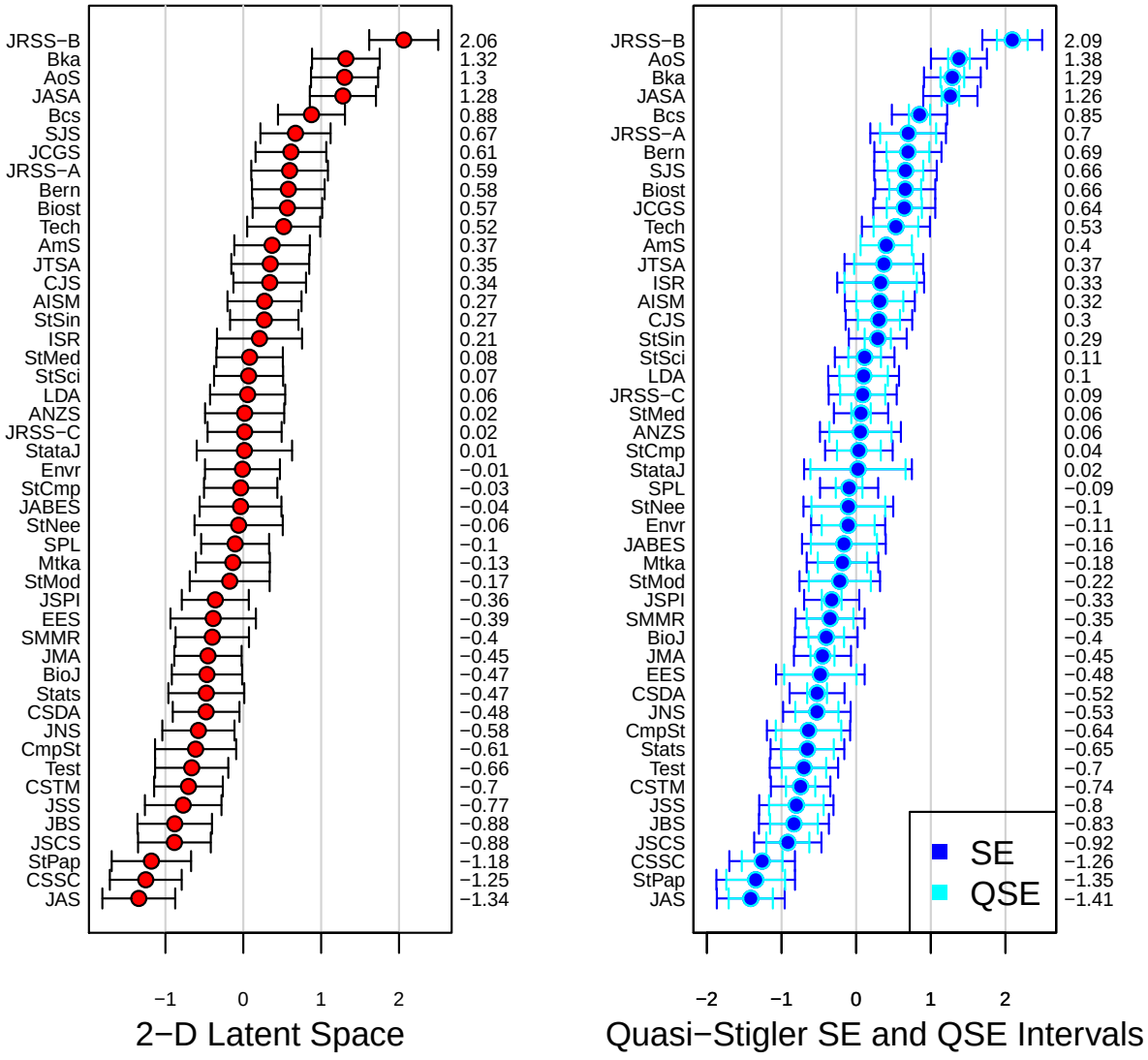


Figure 8.3: Visualizing uncertainty in latent space scores (left) and quasi-Stigler scores (right). The error bars are +/- 1.96*SE in each direction, and inner intervals on the right are "comparison intervals" equal to 1.96∗QSE in each direction. Due to quasi-Stigler model constraints, AMS only has an estimated QSE.

## 8.4 Visualization of Latent Space Journal Rankings

Figure 8.4 (top) shows estimated MKL positions. The size of the nodes is scaled to the estimated scores. The coloring and cluster labels for the nodes in both panels are based on the hierarchical clustering of Varin et al. (2016). Although there is no clustering term in our latent space model, the estimated positions are consistent with these clusters. (Some labels are difficult to read, but this is addressed by the dynamic plot referenced below.) The bottom panel of Figure 8.4 shows a sample of 1000 posterior draws of positions, visualizing the uncertainty. Although individual journal positions are variable, the clusters occupy discernible areas. However, their irregular shapes caution against applying the latent space clustering model of Handcock et al. (2007), in which positions given clusters have spherical Gaussian distributions. The plot shows how they fit together and provides more information than discrete labels. For example the *Journal of Biopharmaceutical Statistics* is most deeply embedded in the applied/health cluster, while *Statistical Science* is on the border. In fact, it should be classified with the review journals it lies near. (We discuss the use of latent space visualization to identify mislabeled classes in Section 9.)

The citation network is too dense to display network edges in a static plot. However, using the **visNetwork** package we render a dynamic plot of the citation network to explore its connections. A version of the plot is included as an html file in the supplementary material (citation_net.html). Because of the density of the network, edges in the dynamic visualization are only shown if they account for at least three percent of a journal's out-citations, and their width when highlighted is scaled to that percentage. Rankings based on the latent space model are reported next to the journal titles in the drop-down menu on the left and in the hover text. Coloring of highlighted edges is determined by the cluster of the originating node, the same clusters as in Figure 8.4.

Now we revisit the rankings of journals discussed in Section 8.3. The "big four" journals all draw citations widely from the network but tend to cite journals fairly nearby. In contrast, StMed, JSPI and CSDA, which are ranked highly by PageRank and Eigenfactor but not the latent space model, give and receive citations from a wide range of journals. JSS, which is

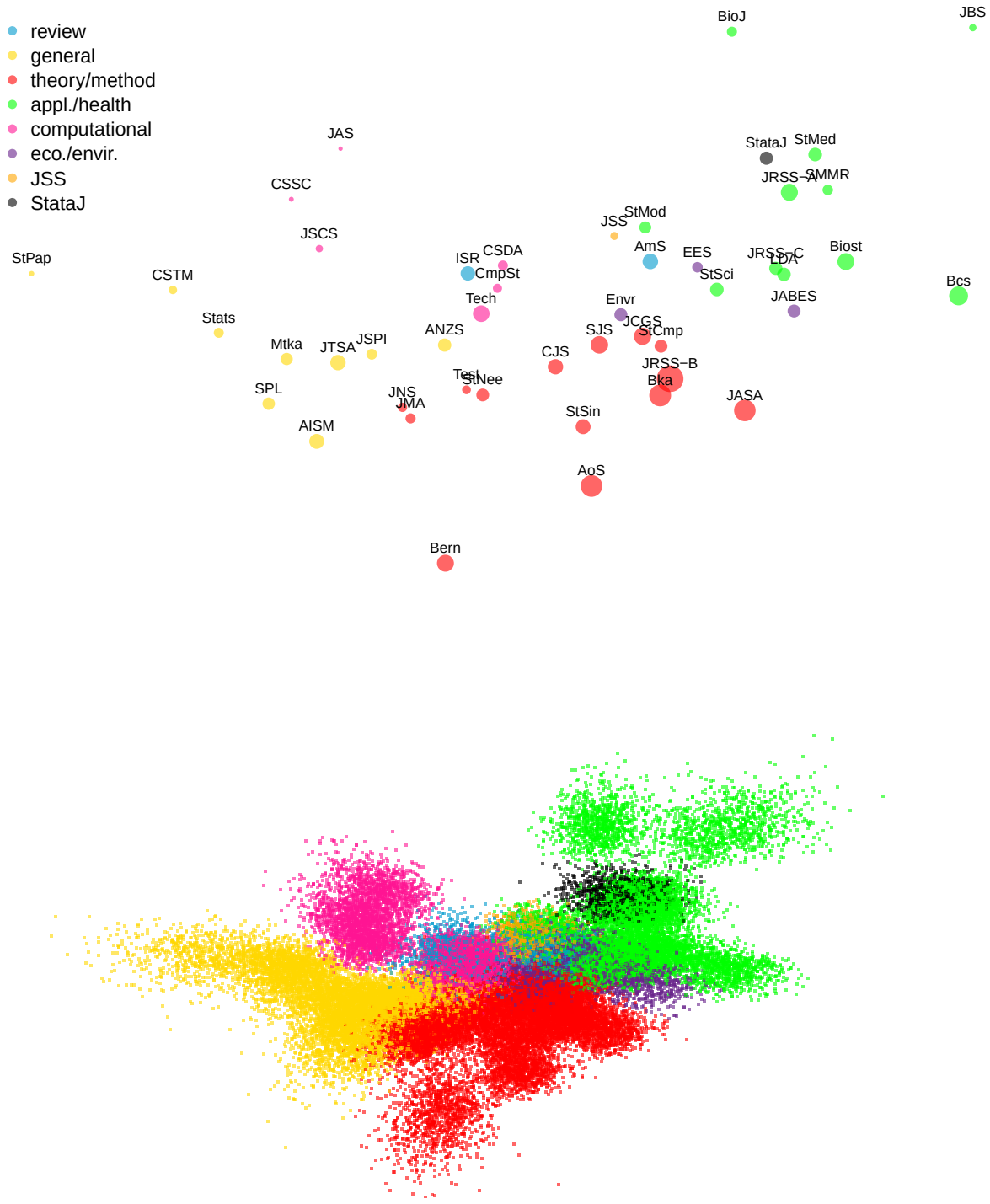


Figure 8.4: Estimated journal positions from the two-dimensional latent space model. Top: Point estimates with node size scaled to receiver minus sender coefficient. Bottom: Sample of positions from the model. Colouring is due to the hierarchical clustering of Varin et al.

ranked highly by Impact Factor and Article Influence but 42nd by the latent space model, only accounts for significant out-citations from StataJ. *Statistical Science* does draw citations from two top-four journals, but not at the rate of the highest-ranked journals. Its in-citations have roughly the same span as its out-citations, but generally come from theoretical or computational journals and go to theoretical or applied ones. The visualization helps us to see how research flows through the network.

## 8.5 Model Evaluation

Having illustrated the value of the latent space rating model, in this section we compare competing latent space models and estimation methods. We discuss the choice of model dimension, model fit to the observed network, trade-offs between estimation methods, and sensitivity to hyperpriors.

### 8.5.1 Latent Dimension

To choose the latent space dimension of the model we consider adaptations of the Bayesian information criterion (BIC), an estimate of the integrated likelihood of the data. The classic Bayesian information criterion (BIC) of a model is an approximation to $-2log(P(Y))$, assuming that the data under the model follows an exponential family distribution.

$$BIC = -2log(P(Y|\hat{\theta})) + d \cdot log(s), \tag{8.2}$$

where $\hat{\theta}$ is the maximum-likelihood parameter estimate, $d$ is the dimension of the parameter space, and $s$ is the effective sample size. In the case of the latent space rating model, this is complicated by several factors. The MCMC-generated maximum-likelihood estimate may not be globally optimal, the parameter dimension is constrained by prior distributions, and different effective sample sizes are used in estimating various parameters. As an alternative to BIC we consider the BIC for model selection (BICM) of Raftery et al. (2007).

$$BICM = -2\hat{\ell}_{max} + \sum_{k=1}^{K} log(n_k) + (\hat{d} - K)log(n_K), \tag{8.3}$$

where $K$ is the number of fixed effects in the model. We let $\hat{\ell}_{max} = log(P(Y|\hat{\theta}))$, where $\hat{\theta}$ is our highest-likelihood MCMC estimate. We consider the estimate of $\hat{d}$ derived by Raftery et al. (2007), which draws on the asymptotic distribution

$$\ell_{max} - \ell_t \sim Gamma(\alpha, 1), \tag{8.4}$$

where $\ell_t$ is the log likelihood of a draw from the posterior distribution of $\theta$, and $\alpha = d/2$. The variance of this gamma distribution is $\alpha$, which we estimate by $var(\ell_t)$, and thereby estimate

$$\hat{d} = 2var(\ell_t). \tag{8.5}$$

We use the MCMC sample for this estimate, so it is important that it has been thinned enough that the posterior log likelihoods are independent. (We verify this using the `mcmc.diagnostics` function of **latentnet**.) In addition, we note that the assumed scale parameter implies another estimate of $\alpha$ if we have a pre-existing estimate of $l_{max}$, namely $\alpha = E[l_{max} - l_t]$. This recovers the estimate of parameter dimension introduced by Spiegelhalter et al. (2002),

$$p_D = -2 * \hat{E}[l_t - l_{max}] = \hat{D}_{avg}(Y, \theta) - D(Y, \hat{\theta}), \tag{8.6}$$

where deviance $D(Y, \theta) = -2log(P(Y|\theta))$ and $\hat{D}_{avg}(Y)$ is calculated by averaging over the MCMC sample. Spiegelhalter et al. (2002) use the posterior mean as the point estimate $\hat{\theta}$, but note that others can be justified. Discrepancy between $\hat{d}$ and $p_D$ reveals the extent to which the asymptotic gamma assumption does not fit the data. This may imply that the scale parameter is not exactly one, though it should be close. Alternately, it may indicate error in our estimate of $l_{max}$ or $var(l_t)$, but if the magnitude of the discrepancy error is small, as we find in this example, then we can proceed with model selection.

For the effective sample size, we are primarily concerned with the number of data points used to estimate positions, since here our models only differ in the dimension of positions. Although Raftery et al. (2007) based the effective sample size for positions in a binary network on the number of realized edges, in a valued network zero-weight edges are informative, so we use $2(n-1)$. This poses a slight problem because the sender and receiver random effects would logically halve that effective sample size to $n-1$, since they only depend on out- and in-edges respectively. However, the over-penalization that results is consistent across dimensional models, so we can ignore it when comparing models.

Table 8.3: Comparison of BIC(M), estimated parameter dimension and ratings correlation for models in zero to four dimensions. Correlations listed are to the two-dimensional ratings.

| | BIC | $BICM_{\hat{d}}$ | $BICM_{p_D}$ | $\hat{d}$ | $p_D$ | Rating Cor. |
|---|---|---|---|---|---|---|
| 0 | 15806 | 15474 | 15486 | 89 | 92 | 0.9652 |
| 1 | 13214 | 12775 | 12731 | 146 | 137 | 0.9857 |
| 2 | 10597 | 9991 | 9915 | 189 | 173 | 1 |
| 3 | 10492 | 9723 | 9628 | 233 | 212 | 0.9991 |
| 4 | 10619 | 9695 | 9443 | 278 | 223 | 0.9984 |

Table 8.3 shows, from left to right, a traditional BIC estimate with parameter dimension equal to the number of parameters and sample size equal to $n*(n-1)$; BICM using $\hat{d}$; BICM using $p_D$; the $\hat{d}$ and $p_D$ estimates of parameter dimension; and the correlation in ratings from each model to the two-dimensional model. The two-dimensional model is a major improvement over the zero- and one-dimensional models by all measures. The three- and four-dimensional models offer small improvement over the two-dimensional model. However, the correlation in ratings between the two and higher-dimensional models is extremely high ($> 0.998$). These factors combined with the increased difficulty of visualizing three- and four-dimensional positions leads us to select the two-dimensional model. Although the error of BICM estimates is not of primary interest here, bootstrap estimates of the standard error in $BICM$ due to MCMC sampling variation show that it is small relative to differences between models.

### 8.5.2 Model Fit

The latent space and quasi-Stigler models produce roughly equivalent journal rankings, but how well do they model the observed network? We consider elements of the fit of our two-dimensional latent space model in comparison to the quasi-Stigler model.

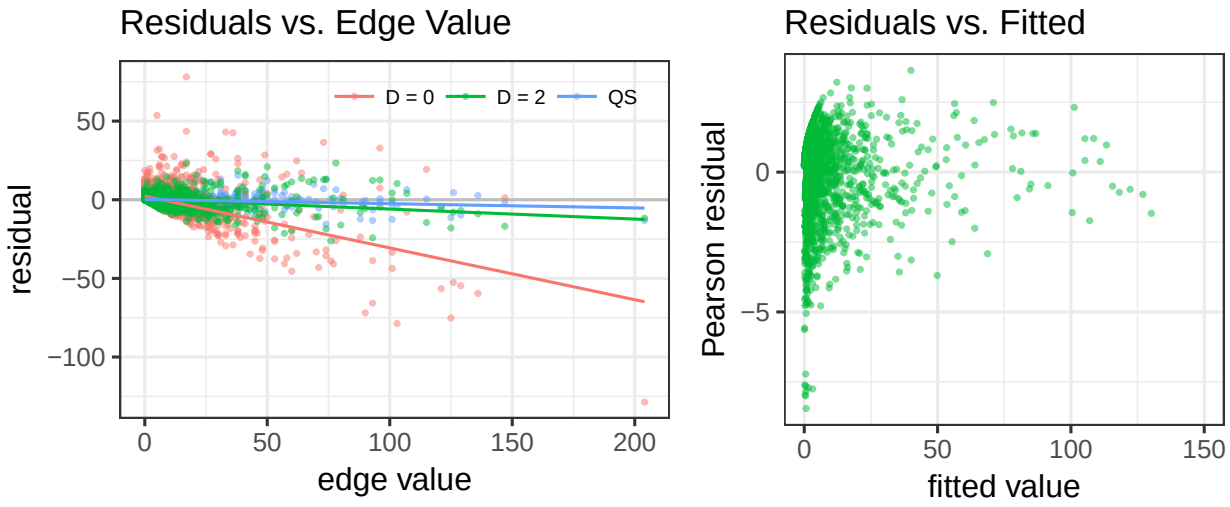


Figure 8.5: Comparison of 2162 residuals for latent space models in zero and two dimensions and the quasi-Stigler model. (Left) Residuals ordered by value of the corresponding edge. A linear model is fit to each set. (Right) Pearson residuals are plotted against fitted values. The highest fitted value is left out to enhance detail.

Figure 8.5 (left) displays model residuals ordered by the value of the underlying edge, i.e. number of citations, for a Poisson sender-receiver model with no latent positions (D = 0), the two-dimensional latent space model (D = 2), and the quasi-Stigler model. We include the zero-dimensional model to help distinguish the impact of modeling edge weights as Poisson from the effect of adding positions. The quasi-Stigler model has the smallest residuals overall ($SD = 2.2$), which we expect since the model is constrained by dyad totals. The two-dimensional latent space model is a bit worse ($SD = 3.7$), but of course much better than the no-position model ($SD = 8.8$). The no-position model is systematically biased, in that edges with lower citation counts are more likely to be overestimated, and heavier edges more likely to be underestimated. This pattern is evident in the other two models, but

much less severe. Examining high-value edges underestimated by the two-dimensional model reveals that they are outliers relative to the citation patterns of the underlying journals. This may be a result of the long-tailed distribution of citation counts per article, but we do not have access to the itemized counts to investigate. There may be additional covariates that we could add to the model to account for these seeming outliers. We also note that by far the highest-valued edge in the network is from *Statistics in Medicine* to *Biometrika*. These journals are related to biological sciences, a field with much higher average citations counts than statistics (Leydesdorff et al., 2013). This hints at the importance of a model extension to account for differing activity patterns within journal fields. Although the data for related fields is not currently available for publication, we revisit relevant model extensions in the discussion.

The right-hand plot of Figure 8.5 displays Pearson residuals against fitted values for the two-dimensional model. There is heightened variation where fitted values are close to zero and residuals must be negative. Although these mostly correspond to small absolute differences, the standardized scale reveals some lack of fit. To account for this in future implementations we may add a small constant to the citation data matrix to stabilize the model for low counts. This is reasonable, since we presume that all statistics journals in the network have at least a small chance of exchanging citations. Outside the near-zero area, the model spread appears roughly constant given the amount of data across fitted values, and the model does not appear overdispersed.

Although dyad-total constraints in the quasi-Stigler model lower the residuals, they also restrict the space of simulated networks. To further compare the model fits we conduct a posterior predictive check on a feature of interest, network reciprocity.

We employ a version of the reciprocity statistic defined by Newman (2010), which we adapt for weighted networks,

$$reciprocity(Y) = \frac{\sum_{i \neq j} y_{ij} y_{ji}}{\sum_{i \neq j} (\frac{y_{ij} + y_{ji}}{2})^2}. \tag{8.7}$$

The value of the statistic ranges from zero to one, achieving the maximum if the network

is symmetric. The denominator controls for the influence of variation in the dyad totals. We also consider an unnormalized version with the denominator removed. Figure 8.6 shows distributions of reciprocity statistics, each based on 5000 simulated networks. They are simulated from an MKL point estimate for the two-dimensional latent space model; the posterior parameter distribution of the two-dimensional latent space model; and the point estimate for the quasi-Stigler model. To simulate networks from the quasi-Stigler model we use a beta-binomial approximation with mean and variance matching the overdispersed binomial. The left panel shows reciprocity while the right panel shows unnormalized values.

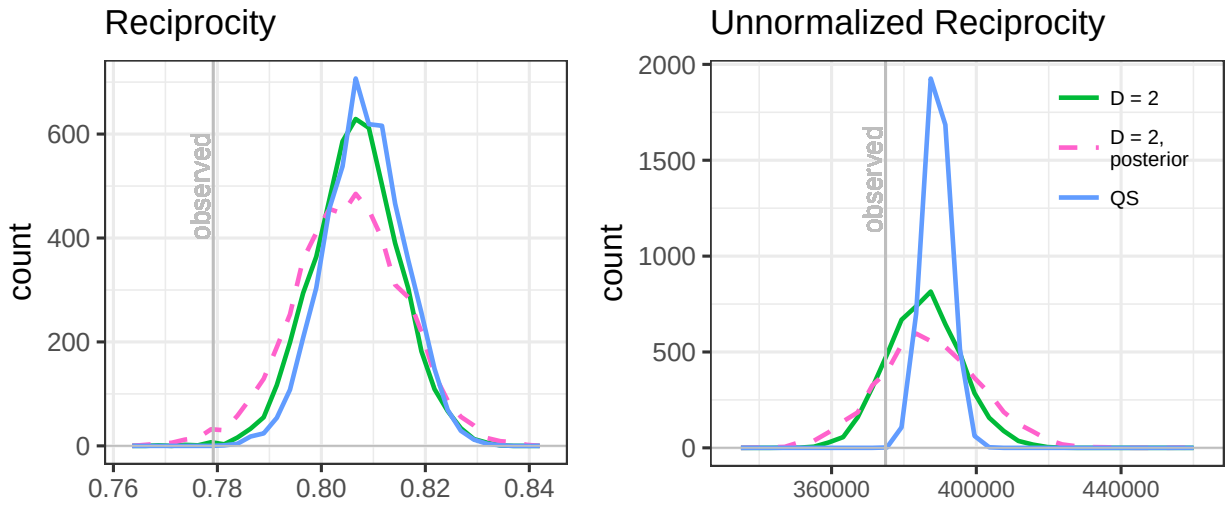


Figure 8.6: Comparison of simulated reciprocity distributions based on 5000 simulations

The average reciprocity estimates are similar across models and biased to the right. This is likely due to the underestimation of high-valued outliers. However, the variance is larger for the two-dimensional models, especially for unnormalized reciprocity. As a result, only the two-dimensional models include the observed values in their simulated ranges. For reciprocity, the percents of simulated values to the left of the observed value are 0.012 and 0.001 for the posterior distribution and point estimate respectively; for unnormalized reciprocity they are 0.19 and 0.15. The quasi-Stigler model is overly restrictive in that simulations from the model do not include the observed network with respect to this statistic.

### 8.5.3 Comparison of Estimation Methods

We now turn to the comparison of estimation methods for the latent space model. We illustrate that quasi-Newton optimization returns ratings nearly identical to MCMC results in much less time.

Table 8.4: Comparison of estimation methods. The $\ell$ function is the log of the probability.

| | Time (min) | $l(Y\|\hat{\theta})$ | $l(\hat{\theta})$ | $l(\hat{\theta}\|Y)$ | Rating Cor. |
|---|---|---|---|---|---|
| QN (L-BFGS-B) | $< 1$ | -4585 | -269 | -4854 | 0.9985 |
| MCMC (MKL) | 15 | -4575 | -274 | -4849 | 0.9973 |
| MCMC.init.r (MKL) | 15 | -4561 | -273 | -4834 | 1.0000 |

Table 8.4 compares results for the two-dimensional latent space rating model by three different estimations techniques: 1) The limited-memory bounded quasi-Newton method as implemented by `optim` (QN L-BFGS-B). 2) MCMC estimation as implemented by **latentnet**; and 3) MCMC estimation as in method two, but with random initialization of sender and receiver coefficients and MDS initial positions scaled so that the maximum coordinate size is one. (We henceforth refer to this as the "random" initialization method.) Note that this initialization is also employed for the quasi-Newton estimate. For MCMC chains we allowed a burn-in period of 500000, and collected a sample of size 5000 by storing every $500th$ draw. We checked for convergence and appropriate MCMC interval and burn-in values using the `mcmc.diagnostics` function of **latentnet**. The MCMC method of row two seems to converge based on diagnostic plots, but given the higher-probably estimate resulting from a different initialization method, we see that it has in fact converged to a local maximum.

Timing of computations was in R 3.3.2 with a MacBookPro, 2.6 GHz Intel Core i5 processor with 8 GB 1600 MHz DDR3 memory. Although not employed here, **latentnet** does allow for parallel processing of multiple MCMC chains. The time estimates should be considered ballpark for each method, as the number of iterations for the quasi-Newton methods and control parameters for the MCMC could be augmented to achieve some speed-up. However, the process of optimizing these would mostly likely negate the time benefits.

We present MKL estimates from MCMC methods because they have higher posterior and graph likelihood than other point estimates, such as the posterior mode. To calculate

$\ell(\theta)$ for these estimates, where $\ell$ is the log of the probability function, we use the random effect variances estimated by the MCMC posterior mode. Results in previous sections have all drawn on the MKL estimates from MCMC estimation with random initialization (row three), because they have even higher probably than other MKL estimates. The MCMC methods are clearly much slower than the quasi-Newton methods due to the number of iterations required, although they have the advantage of returning a posterior sample. The much faster estimates from the quasi-Newton methods are very highly correlated with these results (column five), and the log likelihood is only a tiny bit worse.

To more formally examine the accuracy of quasi-Newton estimation we conduct a simulation study. We generate 100 simulated citation networks from the best MKL point estimate and fit each by QN L-BFGS-B. We then calculate the differences in ratings (sender minus receiver coefficient) and total actor parameters (sender plus receiver coefficient plus intercept) between the data generating model and the estimate from each simulation. Although the zero-mean prior reduces the non-identifiability of the random effects, with diffuse priors it is not entirely eliminated in practice. To account for this when calculating differences in ratings between fits, we center the sender and receiver estimates to mean zero. The results are presented in Figure 8.7. The top panel shows that the differences in ratings are all concentrated near zero, indicating that QN L-BFGS-B does reliably recreate the "true" ratings in this case. Slightly larger variances occur for journals such as StataJ which have fewer non-zero edges. The bottom panel shows the differences in total actor parameters. Deviations from zero in this panel indicate changes in positions relative to the data-generating model. Journals estimated to be closer on aggregate to other journals compensate with smaller random effects, and appear below the x-axis. Three journals lie significantly below the x-axis. We can see in the network visualization that these three are located in the periphery of the network space. Thus, there is a linear range in which their position is nearly identified with the total of their sender and receiver coefficients. This is not a problem with the estimation method but stems from underlying uncertainty in the model. However, it does not affect our estimated rankings, as we see in the top panel, nor the overall configuration of the network. That said, we must recognize that, in general, estimated positions of peripheral nodes may

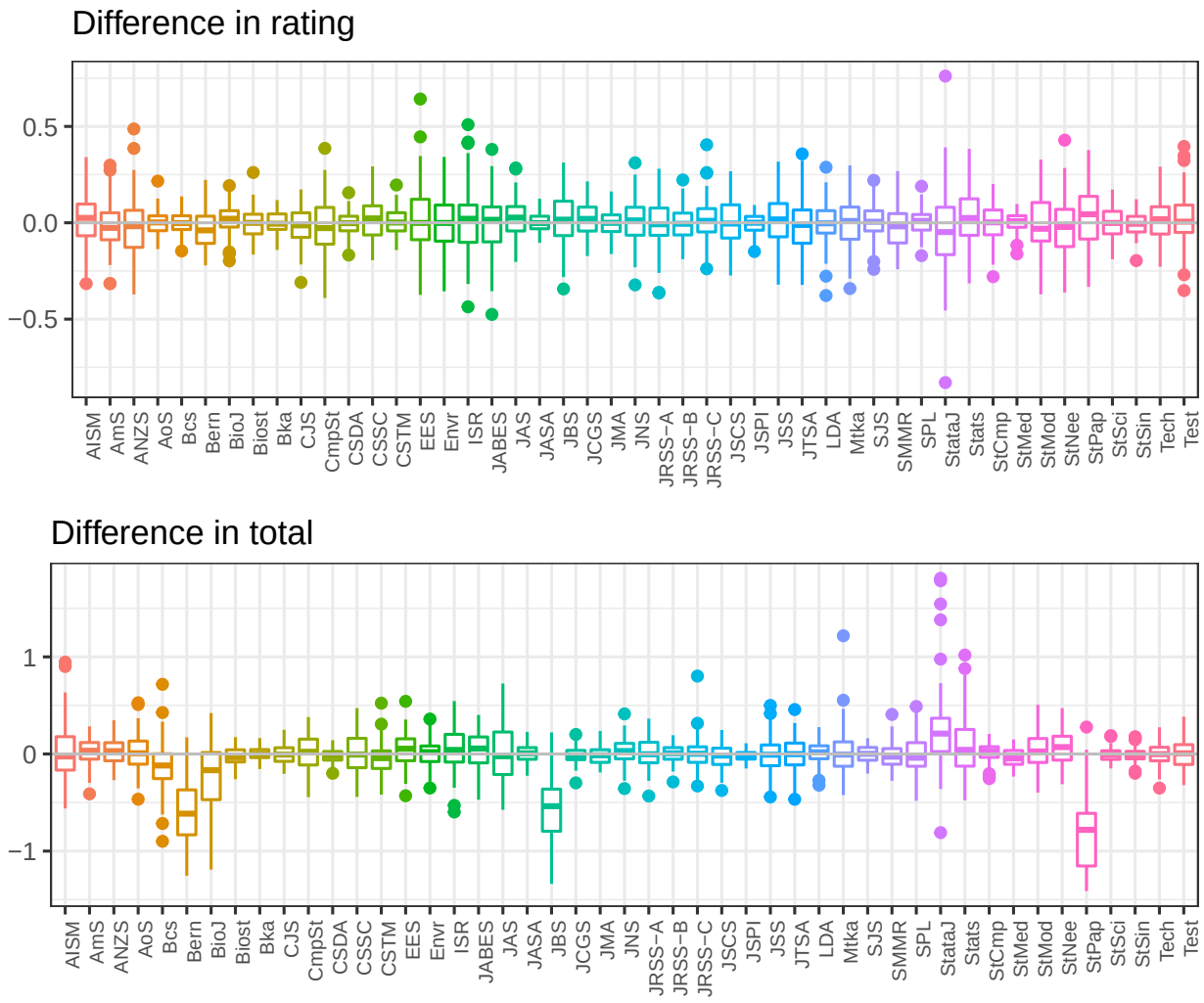


Figure 8.7: Distribution of error in parameter estimates as compared to the data-generating model, based on 100 simulations. The top panel shows error in ratings (receiver - sender) and the bottom panel shows error in total actor parameters (sender + receiver + intercept).

be much more variable than central ones.

A drawback to quasi-Newton estimation is that it does not generate a sample of the posterior distribution which which to estimate MKL positions or the variance of ratings. However, we can still approximate the ratings uncertainty. One way to do this is to fit a Poisson GLM to the data matrix with estimated positions as an offset, i.e the distance term in Equation 7.5 is treated as a constant. We implement such a model and find similar estimates of ratings uncertainty to those derived from the MCMC sample, although we do lose one journal comparison due to a model-fitting constraint. The estimated standard deviations range from 0.22 to 0.28, compared to the MCMC range of 0.22 to 0.29, with differences having mean 0.002 and maximum 0.013. Not only does this provide an estimate of rating uncertainty, it also confirms that there is minimal dependence between positional parameters and ratings estimates, as was alluded to by Figure 8.7.

As with any Bayesian model we must consider the sensitivity of estimates to the assumptions contained in the prior structure and hyperprior values. The near-identical ratings between MLE and MKL estimates suggest that the diffuse priors do not impact the ratings. Still, we fit a model with adjustments to the hyperprior degrees of freedom and variance which narrow the prior variance distributions to more realistic ranges based on previous results. The effect on the results is trivial, both in the ratings and their standard errors. In some cases it may be valuable to limit the spread of position estimates with stronger priors, but unless they restrict beyond likely observable ranges the impact on ratings should be minimal.

## 8.6 Appendix: Journal Names and Abbreviations

| | Journal Name | Abbreviation |
|---|---|---|
| 1 | American Statistician | AmS |
| 2 | Annals of the Institute of Statistical Mathematics | AISM |
| 3 | Annals of Statistics | AoS |
| 4 | Australian and New Zealand Journal of Statistics | ANZS |
| 5 | Bernoulli | Bern |
| 6 | Biometrical Journal | BioJ |
| 7 | Biometrics | Bcs |
| 8 | Biometrika | Bka |
| 9 | Biostatistics | Biost |
| 10 | Canadian Journal of Statistics | CJS |
| 11 | Communications in Statistics—Simulation and Computation | CSSC |
| 12 | Communications in Statistics—Theory and Methods | CSTM |
| 13 | Computational Statistics | CmpSt |
| 14 | Computational Statistics and Data Analysis | CSDA |
| 15 | Environmental and Ecological Statistics | EES |
| 16 | Environmetrics | Envr |
| 17 | International Statistical Review | ISR |
| 18 | Journal of Agricultural, Biological and Environmental Statistics | JABES |
| 19 | Journal of the American Statistical Association | JASA |
| 20 | Journal of Applied Statistics | JAS |
| 21 | Journal of Biopharmaceutical Statistics | JBS |
| 22 | Journal of Computational and Graphical Statistics | JCGS |
| 23 | Journal of Multivariate Analysis | JMA |
| 24 | Journal of Nonparametric Statistics | JNS |
| 25 | Journal of the Royal Statistical Society, Series A | JRSS-A |
| 26 | Journal of the Royal Statistical Society, Series B | JRSS-B |
| 27 | Journal of the Royal Statistical Society, Series C | JRSS-C |
| 28 | Journal of Statistical Computation and Simulation | JSCS |
| 29 | Journal of Statistical Planning and Inference | JSPI |
| 30 | Journal of Statistical Software | JSS |
| 31 | Journal of Time Series Analysis | JTSA |
| 32 | Lifetime Data Analysis | LDA |
| 33 | Metrika | Mtka |
| 34 | Scandinavian Journal of Statistics | SJS |
| 35 | Stata Journal | StataJ |
| 36 | Statistical Methods in Medical Research | SMMR |
| 37 | Statistical Modelling | StMod |
| 38 | Statistica Neerlandica | StNee |
| 39 | Statistical Papers | StPap |
| 40 | Statistical Science | StSci |
| 41 | Statistica Sinica | StSin |
| 42 | Statistics | Stats |
| 43 | Statistics and Computing | StCmp |
| 44 | Statistics in Medicine | StMed |
| 45 | Statistics and Probability Letters | SPL |
| 46 | Technometrics | Tech |
| 47 | Test | Test |

# CHAPTER 9

# Movie Rating and Genre Identification

In this section we apply latent space rating to a set of films with viewer-supplied star ratings. Film rating systems based on viewer ratings are biased by the fact that viewers do not choose movies to rate randomly, and they may skew high or low in their ratings. In addition, average ratings obscure the information provided by the volume of ratings for different films. The latent space rating method efficiently addresses those challenges. The concomitant visualizations aid in genre detection and enhance our ability to explore and compare related films. Through a dynamic network plot we increase the amount of data we can present. Movie ratings are extremely subjective, so without knowing detailed individual preferences the ability to explore a network of films may be more useful than finding "correct" ratings.

## 9.1 Movie Data

Our data came from the MovieLens data set (Harper and Konstan, 2015), collected by the GroupLens Research Project at the University of Minnesota. It consisted of about one million ratings of 3,952 movies from approximately 6,000 users who joined MovieLens in 2000. The data were released in 2003. Ratings were on a one to five integer (star) scale with five being the best. Each included user supplied at least 20 ratings.

To convert the data into a network format we aggregated differences in individual users' ratings to form a $3952 \times 3952$ ratings-difference matrix. To be explicit, entry $i, j$ in the matrix represented the sum of positive values of $rating(j) - rating(i)$ over users who rated both movies. (If $rating(i) > rating(j)$ the difference was added to entry $j, i$.) For example, if a user prefers movie $j$ to $i$ by one point, then one is added to entry $i, j$: $i$ "sends" a point

to $j$. The corresponding network is positive-valued and directed.

To illustrate certain points without too much computational burden we restricted ourselves to a subset of the MovieLens data, retaining only movies assigned genre "Action", "Crime", "Western" or some combination therein. (Only two of the possible combinations are present in the data.) We removed a small number of isolates before modeling. The resulting network has 128 nodes and 11,393 edges.

## 9.2 Latent Space Model

By modeling the network of differences in ratings we capture the tendency for a movie to be frequently and consistently rated above movies that draw an overlapping audience. We apply the latent space Poisson model in two dimensions to facilitate visualization, which is of greater interest in this application than ratings precision. Results are presented in Table 9.1. The methods presented in the first three rows are the same as in Table 8.4. The MCMC sampling parameters are also the same as above. We tried increasing them for this larger network, but found it not to be necessary. The MKL estimates are again the best of the estimates produced by the MCMC estimation. For comparison, we include in rows four and five the MLE and posterior mode estimated from the MCMC sample. Although the fit from MCMC estimation is not sensitive to initialization method in this example, the time is somewhat affected. In addition, the quasi-Newton method achieves better results in less time, about two minutes instead of ten, when using random initialization.

As above we consider adjustments to the hyperpriors that narrow the range of prior variance distributions and find the impact on ratings to be trivial (row 6). However, the hyperprior adjustments do benefit the visualization by limiting the amount that very low-connectivity nodes drift from the rest. These results are subsequently used for visualization and clustering analysis. For consistency, the calculations in columns three and four of Table 9.1 all assume the more diffuse priors.

There is a strong correlation of 0.95 (or 0.976 when weighted by the log of co-review counts) between a film's average star rating and its ratings from the latent space model.

Table 9.1: Comparison of estimation methods. The $\ell$ function is the log of the probability.

| | Time (min) | $l(Y\|\hat{\theta})$ | $l(\hat{\theta})$ | $l(\hat{\theta}\|Y)$ | Rating Cor. |
|---|---|---|---|---|---|
| QN (L-BFGS-B) | 2 | -32688 | -793 | -33481 | 0.9996 |
| MCMC (MKL) | 137 | -32687 | -793 | -33480 | 0.9999 |
| MCMC.init.r (MKL) | 104 | -32688 | -796 | -33484 | 1.0000 |
| MCMC.init.r (MLE) | - | -32857 | -808 | -33665 | 0.9951 |
| MCMC.init.r (P.mode) | - | -32865 | -792 | -33657 | 0.9971 |
| MCMC.h (MKL) | 117 | -32685 | -796 | -33481 | 0.9998 |

Figure 9.1 shows this correlation, plotting the latent space scores against the average star ratings and labeling points by review counts for each film. Although not pictured, the standard errors in the ratings are fairly consistent. They range from 0.13 to 0.26 (95th percentile) with a median of 0.14. The high-end outliers have very few co-review counts, and only points with few co-reviews deviate strongly from the overall linear trend. We take a closer look at a couple of the films that deviate from the trend but have more than 20 reviews, highlighted in green. The corresponding films are, from left to right, *Shaft in Africa (1973)* and *Assassination (1987)*. These films have a potential "cult" following based on their lead actors, Richard Roundtree (as Shaft) and Charles Bronson respectively. That may boost some of the user reviews, but when comparative reviews are considered in the latent space model the derived rating is lower.

It is somewhat surprising to find such a strong correlation between the two ratings methods pictured. This may be due in part to homogeneity in the MovieLens reviewer pool, which reduces the sources of bias discussed above. MovieLens was organized by a university and all reviewers in our data joined in 2000. The top four occupations of the reviewers are (in decreasing order): college/grad student, other/not specified, executive/managerial and academic/educator. This is certainly not representative of the population at large. The impact of our difference-based network model as compared to average star ratings may be more evident when reviewers are more heterogeneous.

We next consider the additional insight gained through latent space model positions and visualization. An interactive plot of the model output is included as an html file in the supplementary material (movie_net.html). The nodes are colored by genre with size scaled

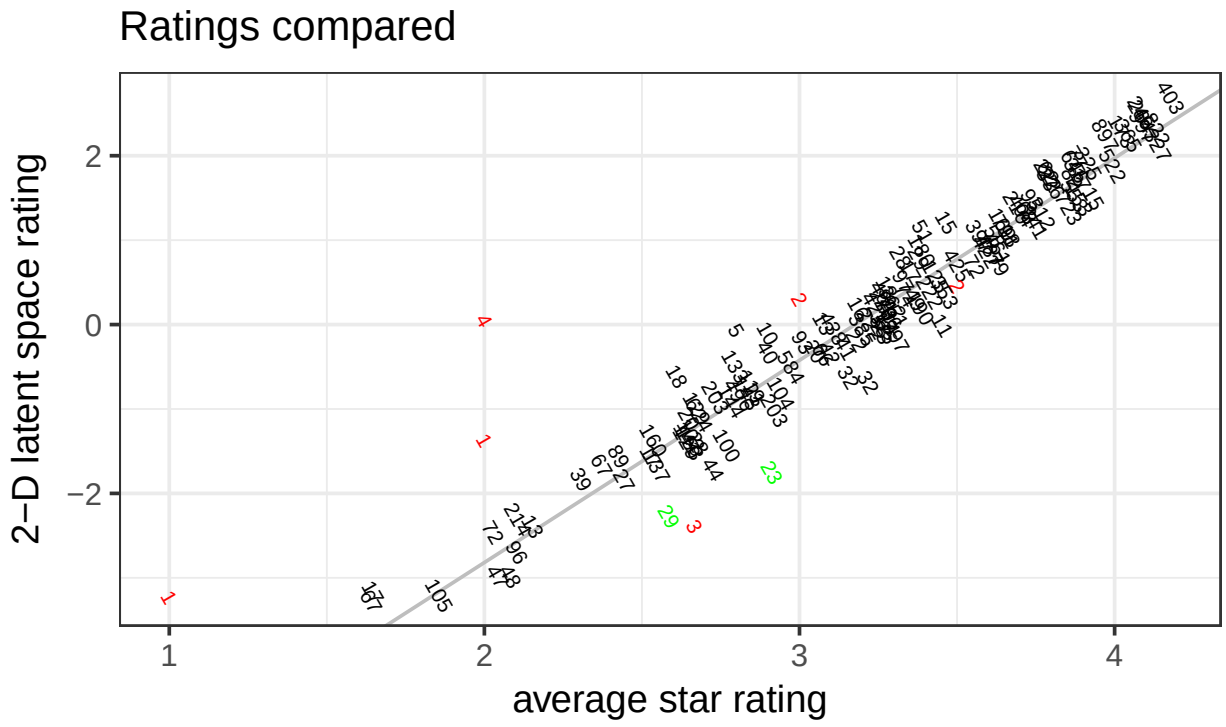


Figure 9.1: Latent space model scores vs. average star ratings. The plotting characters are the review counts for each film. The red points are strong outliers to the linear trend, have very few reviews, and are excluded in calculating the best fit line. The green points are discussed in the paper.

to latent space model rating. For comparison, the average star ratings are listed next to the movie titles in the drop-down menu and when hovering over a node. For clear visualization, the nodes are limited to displaying their (at most) seven strongest out-edges. In-degree is not limited in the plot.

## 9.3 Genre Detection

The visualizations shows that movies clearly cluster by genre even though genre was not a term in the model. Movies with hybrid genres are placed roughly between their two component genres. However, there are a few films that reside outside of their genre cluster. In those cases the plot can highlight incomplete or incorrect classification. For example, the film *Coogan's Bluff (1968)* is categorized as a crime film by MovieLens, but its latent position is among actions and westerns. The Internet Movie Database (IMDB) entry for this film describes it as "An Arizona deputy goes to New York City to escort a fugitive back into custody," and the lead role is played by action/crime/western star Clint Eastwood (coo, 2016). This film has heavy action and western influences which its latent position reveals. Another example of misclassification is the two "action" films *The Kid (1921)* and *Minnie and Moskowitz (1971)*, positioned between crimes and westerns. Their genres listed on IMDB are comedy/drama/family, and comedy/drama/romance, respectively. Neither is well-classified as an action film, and their positions close together and outside the action cluster reflect this, and also recognize their similarity.

The continuous positions from the latent space model are more precise identifiers of movie type than discrete cluster labels. Like the journal citation network visualized in Section 8.4, the clusters are discernible but irregularly shaped and blend into each other. The positions aid in identifying sub-genres, which may be valuable for recommendation systems. For example, the westerns that tail into the crime cluster, *Unforgiven (1992)*, *Tombstone(1993)*, and *Dead Man (1995)* reveal a "modern western" sub-genre. These films are much newer than most westerns in the data, with median year 1968.

If no genre labels were given the fit positions provide rich input for a clustering algorithm.

To illustrate, we apply k-means clustering with three clusters, which we expect to correspond roughly to the action, crime and western films (Hartigan and Wong 1979, R Development Core Team 2016). We use pair-counting to measure accuracy, with the caveat that not all pre-assigned genres are correct, as discussed above. Films originally labeled with two genres are considered a match if assigned to either of those genres. Table 9.2 compares the k-means output to pre-assigned genre labels. Of the 128 films in the network, 105 are assigned to matching labels. The most common change in classification is from action to crime. The comparison is visualized in Figure 9.2. Most films that switch labels are positioned near the boundary of the three classes, indicating that a single label is probably insufficient. To put these results in perspective, we implemented two popular community detection methods to the film matrix, a Louvain modularity maximization method and Infomap, as implemented in **igraph** (Csardi and Nepusz, 2006). Both are unsupervised and require undirected networks, so we input the symmetric matrix of dyad totals. The Louvain method returned three communities, one that corresponded to westerns, one to both action and crime films, and a smaller third one to action films in the center of the plot. There were 38 total misclassifications. Infomap returned only one community, highlighting the lack of distinct divisions between genres that latent positions are able to capture.

| | Action | Crime | Western |
|---|---|---|---|
| Action | 45 | 14 | 1 |
| Action\|Crime | 6 | 7 | 0 |
| Action\|Western | 0 | 0 | 2 |
| Crime | 3 | 17 | 1 |
| Western | 1 | 3 | 28 |

Table 9.2: Classification of films by pre-assigned genres (row) vs. k-means cluster (column)

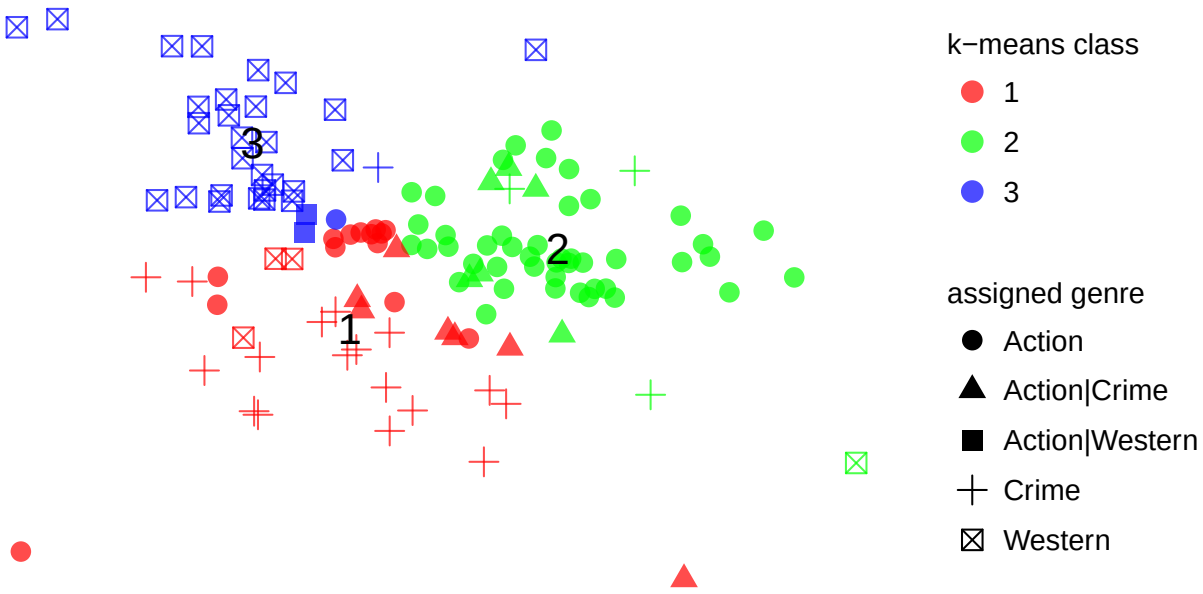


Figure 9.2: Film network colored by k-means class with shape determined by pre-assigned genre. Centers of the k-means classes are labeled 1-3.

# CHAPTER 10

# Mixed Latent Model

In this chapter we describe an extension to the latent space rating model which generalizes the sender and receiver coefficients to a set of additive and multiplicative effects. While the Euclidean distances of the latent space model capture multidimensional undirected effects, this extension allows for multidimensional directed effects as well. We show through journal and film applications how this improves model performance relative to competing models with the same number of parameters.

## 10.1 Additive and Multiplicative Effects Model

Minhasa et al. (2016) present a latent-factor model for networks in which the latent factors are multidimensional, directed and multiplicative. The model also contains additive row and column effects, analogous to sender and receiver coefficients, and as such we consider it an alternative model to the latent space rating model with Euclidean distances. The additive portion of the model derives from the social relations model, an ANOVA-type decomposition of data into random row and column effects and Gaussian noise. The model of Minhasa et al. (2016), referred to as AME for its additive and multiplicative effects, is formed as follows:

$$y_{ij} = \beta^\top X_{ij\cdot} + e_{ij} \tag{10.1}$$

$$\begin{aligned} e_{ij} &= a_i + b_j + \epsilon_{ij} + \alpha(u_i, v_j), \text{ where} \\ \alpha(u_i, v_j) &= u_i^\top D v_j = \sum_{r \in R^*} d_r u_{ir} v_{jr}. \end{aligned} \tag{10.2}$$

The variance structure of the effects, $e$, is given by

$$\{(a_1, b_1), ..., (a_n, b_n)\} \sim_{iid} N(0, \Sigma_{ab}) \tag{10.3}$$

$$\{(\epsilon_{ij}, \epsilon_{ji}) : i \neq j\} \sim_{iid} N(0, \Sigma_\epsilon), \text{ where} \tag{10.4}$$

$$\Sigma_{ab} = \begin{pmatrix} \sigma_a^2 & \sigma_{ab}^2 \\ \sigma_{ab}^2 & \sigma_b^2 \end{pmatrix}, \quad \Sigma_\epsilon = \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{10.5}$$

Letting $\beta$ consist of only an intercept, as in our latent space rating model, gives

$$y_{ij} = \beta + a_i + b_j + u_i^\top D v_j + \epsilon_{ij}. \tag{10.6}$$

Hoff (2009) interprets this model through the lens of matrix decomposition. The observed data matrix $Y$ is a sum of covariate effects, row and column effects, noise and any leftover systematic patterns, which we denote as the $n \times n$ matrix $M$. It is a well-established result that the best rank-$r$ approximation of $M$ in terms of error by the Frobenius norm is formed from the singular value decomposition of M (Eckart and Young, 1936),

$$M = UDV^\top,$$

where $U$ and $V$ are $n \times n$ matrices with orthonormal columns, and $D$ is an $n \times n$ diagonal matrix whose entries are the square roots of the eigenvalues of $M^\top M$. The best rank-$r$ approximation of $M$ is given by forming an $r \times r$ diagonal matrix from the $r$ largest elements of $D$, and the corresponding columns of $U$ and $V$.

$$\tilde{M} = U_{n \times r} D_{r \times r} V_{n \times r}^\top.$$

In the AME model and subsequently in this chapter, references to $U, V$ and $D$ refer to these truncated matrices. Accordingly, the $UDV^\top$ term in the AME model is a low-rank approximation of systematic patterns not accounted for by covariates, row and column effects, and noise.

We summarize the main differences between this model and our latent space rating model. First, the AME model uses multiplicative, directed latent factors rather than symmetric, Euclidean latent distances. Second, the AME model minimizes normal error in the transformed space of predictors, rather than Poisson-distributed error in the domain of observed data. This leads to differing covariance structures. Under the AME model, there is a correlation term between the directed edges in a dyad, while under the latent space rating model dyad dependence is captured by Euclidean distances. The AME model also includes a covariance term between sender and receiver effects, while under the latent space model with Bayesian inference they are drawn from independent prior distributions, though they are dependent in the posterior. In future work we may consider a joint prior distribution of $(a_i, b_i)$ for the latent space rating model.

It remains to describe how ratings can be calculated under the AME model. A natural measure in line with our latent space rating is to subtract estimated out-flow from in-flow for each node. We thus consider the rating

$$rating_i = b_i - a_i + \overline{UV_{\cdot i}^\top} - \overline{UV_{i\cdot}^\top}, \tag{10.7}$$

which is node $i$'s receiver minus sender effect plus the $i$th column mean of $UV^\top$ minus its $i$th row mean. Means are used rather than sums to place the additive and multiplicative effects on the same scale. Under the AME framework, this rating correlates highly with column means minus row means of the observed data matrix. For example, when using a rank-two decomposition, the correlation is 0.9985 for our citation data and 0.9999 for our film data. This is a testament to how well this model with low-rank decomposition captures the data. However, it also shows that this implementation of the AME model does not provide much insight into ratings beyond the raw data.

The Hoff et al. (2017) R package **amen** is available to implement the AME model as described above. However, to make this model more appropriate for our count data we first adapt it to our Poisson GLM framework. For simplicity and correspondence with our latent space rating model we retain independent priors for sender and receiver effects. This

implementation is described below in the context of a novel model extension that captures advantages of both the Euclidean latent space model and the AME model.

## 10.2 Mixed Latent Model

The singular value decomposition that underpins the AME model provides the best low-rank approximation of $M$, the matrix of unexplained systematic patterns in the data. However, by decomposing $M$ in this way we lose the notion of latent positions, and with that their utility for model interpretation and visualization. Relying on the asymmetric factors of $U$ and $V$ nullifies any symmetric effect in the model, but similarities between items are inherently symmetric. In addition, Although SVD provides the best low-rank matrix approximation, the use of asymmetric factors increases the number of parameters that must be estimated relative to a symmetric distance model. The number of parameters that must be estimated in the rank-$r$ AME model is $n(2r+2)+5$, while the $d$-dimensional Euclidean model requires $n(d+2)+4$.

To address these shortcomings, we consider another decomposition of $M$ which takes into account its symmetric and asymmetric patterns. $M$, like any asymmetric matrix, can be re-written as the sum of a symmetric matrix and a skew-symmetric matrix as follows:

$$M = \frac{(M + M^\top)}{2} + \frac{(M - M^\top)}{2}.$$

This splits $M$ into a symmetric "total" matrix, and an asymmetric "difference" matrix. The former captures total strength of ties between two nodes and the latter captures directional difference, after accounting for additive sender, receiver, and covariate effects. Employing the GLM framework of the latent space ranking model this implies

$$E(Y) = exp(\beta^\top X_{ij\cdot} + outer(A, B, +) + \frac{(M + M^\top)}{2} + \frac{(M - M^\top)}{2}),$$

excluding the diagonal of $Y$ because we do not allow self-edges.

This leads to two low-rank matrix decompositions, one for the symmetric matrix, $M_{sym} =$

$\frac{(M+M^\top)}{2}$, and one for the asymmetric matrix, $M_{asym} = \frac{(M-M^\top)}{2}$. The decomposition of the symmetric part is based on an eigenvalue decomposition, $Q\Lambda Q^\top$, where $Q$ is a matrix of orthogonal eigenvectors of $M_{sym}$, and $\Lambda$ is a diagonal matrix of eigenvalues of $M_{sym}$. For a low rank approximation to $M_{sym}$, let $Q_{n\times d}$ be the rank-$d << n$ matrix of the $d$ columns of $Q$ corresponding to the $d$ largest values of $\Lambda$, and let $\Lambda_{d\times d}$ be the diagonal matrix of those values. Let $Z = Q_{n\times d}(\Lambda^{1/2}_{d\times d})$. Then $\tilde{M} = ZZ^\top$ is a rank-$d$ approximation to $M_{sym}$. Implicitly, this restricts us to $d$ less than or equal to the number of positive eigenvalues of $M_{sym}$, a fact that we will revisit below.

Let $UDV^\top$ be the rank-$r$ singular value approximation of $M_{asym}$, as described above. Let $U$ and $V$ absorb $D$ by multiplying each by the square root of $D$. Singular values are non-negative so this does not pose a restriction on $r$. Accordingly,

$$E(Y) = exp(\beta^\top X_{ij.} + outer(A, B, +) + ZZ^\top + UV^\top). \tag{10.8}$$

A symmetric inner product term, $ZZ^\top$, was previously incorporated in the generalized bilinear mixed effect model of Hoff (2005), and the asymmetric latent projection distance model of Hoff et al. (2002) as $\frac{z_i z_j}{|z_j|}$. However, neither of those models was introduced in the context of matrix decomposition, and neither considered dual symmetric and asymmetric latent factors. Hoff (2015) and Minhasa et al. (2016) considered a symmetric eigenvalue decomposition of $M$ for an undirected network, but this was to supplant rather than augment the singular value decomposition.

Minhasa et al. (2016) noted that the symmetric factors for each node, the rows of $Q$, describe the stochastic equivalence of the nodes, while the signs of eigenvalues in $\Lambda$ indicate whether the corresponding dimension exhibits homophily or its opposite, heterophily. Our low-rank approximation $ZZ^\top$ is suboptimal if the $d$ largest eigenvalues are not the largest by magnitude. However, by restricting our low-rank representation of $M_{sym}$ to dimensions with positive eigenvalues, we implicitly model only homophily through symmetric factors. This corresponds to our aim of modeling and visualizing similarity through symmetric effects, in a context where dissimilarity decreases dyad totals.

We now connect Equation 10.8 directly to the Euclidean latent space model. An inner product relates to Euclidean distance as follows, and implies a relationship between symmetric decomposition and Euclidean distance.

$$z_i \cdot z_j = \frac{1}{2}(|z_i|^2 + |z_j|^2 - \|z_i - z_j\|^2), \text{ implying} \tag{10.9}$$

$$E(y_{ij}) = exp(\beta^\top X_{ij\cdot} + a_i + \frac{|z_i|^2}{2} + b_j + \frac{|z_j|^2}{2} - \|z_i - z_j\|^2 + u_i^\top v_j). \tag{10.10}$$

Without loss of generality we absorb the additive effects of the magnitudes of the positions into the additive sender and receiver effects. This has no impact on the derived component of the rating, $b_i - a_i$. Assuming again that $\beta$ is reduced to an edge intercept, we are left with the following expression,

$$E(y_{ij}) = exp(\beta + a_i + b_j - \|z_i - z_j\|^2 + u_i^\top v_j). \tag{10.11}$$

Note that this expression includes the *squared* Euclidean distance rather than the distance. Although squared distance is optimal in the sense of matrix decomposition, we prefer to employ the un-squared Euclidean distance because of its value in visualization and interpretation. In addition, the Euclidean distance, like any $p$-norm with $p$ greater than or equal to one, satisfies the triangle inequality, $\|z_i - z_j\| \leq \|z_i - z_k\| + \|z_k - z_j\|$, which induces transitivity in the network. As Hoff (2003) points out, if the distances from $i$ to $k$ and $k$ to $j$ are small, then the "friend-of-a-friend" distance cannot be too large. This places a lower bound on the strength of the tie from $i$ to $j$, given other effects. This property is also true of similarities between items. If $i$ is similar to $j$ and $j$ is similar to $k$, we expect a lower bound on the dissimilarity of $i$ and $j$.

Taking this into account, we present our final mixed latent model. Assuming the same Poisson GLM framework as Equation 7.5, and adding variance terms and hyperparameters for $U$ and $V$ that are analogous to those of $Z$, the model is described by

$$E(y_{ij}) = exp(\beta + a_i + b_j - \|z_i - z_j\| + u_i^\top v_j), \tag{10.12}$$

where $z_i$ has length $d$ and $u_i$ and $v_j$ have length $r$. The rating we derive from this model is the same as the one presented in Equation 10.7 for the AME model.

This model offers both the benefits of Euclidean nodal positions and the optimality of the singular value decomposition. In addition, increasing $d$ by one adds half the number of parameters to the model as increasing $r$ by one. As long as there are symmetric patterns in the data, this offers an improvement in model efficiency over only asymmetric decomposition. In the next section we compare models of various Euclidean dimension and latent factor rank for the journal and film data analyzed previously. An analog to the AME model for Poisson-distributed count data is obtained by setting $d$ to zero in the mixed latent model. We include such Poisson AME models in our comparisons. The Euclidean latent space rating model is also a special class of the mixed latent model, obtained by setting $r$ to zero.

## 10.3 Comparison of Latent Models for Journal Rating

We implement the mixed latent model of Equation 10.12 using the L-BFGS-B quasi-Newton method discussed in Section 7.4.1, having shown by simulation study that it is fast and accurate. The same default hyperprior values and random initialization techniques are used. The implementation of the extended model only requires expressions for the updated likelihood and gradient formulas (see Appendix 10.5). Because we embed the model in a Bayesian framework and do not restrict estimated latent factors to be orthogonal, we presume that they are not. This does not compromise the justification of the rank-$r$ decomposition, though it should be noted when interpreting latent dimensions. Post-hoc orthogonalization of the latent factors may be useful in some applications to facilitate interpretation.

Table 10.1 and Figure 10.1 summarize models of the journal citation network in varying low rank and dimension. In each case we use several initializations to optimize results. From Figure 10.1, the best fit model in terms of probability of the observed graph is the two-dimensional, rank-two model. The higher likelihood of this model compared to the rank-three Poisson AME model confirms our earlier assertion that a mixed model more efficiently captures a network with symmetric and asymmetric patterns. Next we replicate the analysis

of Section 8.5.2 for the citation network to explore how the mixed latent model has altered or improved our model fit, ratings, and interpretation of results.

Table 10.1: Comparison of models of varying dimension and rank. Gray rows correspond to models which exhibited possible collinearity. The final column shows improvement in log likelihood per additional parameter over the two-dimensional Euclidean model.

| | d | r | Var Params | $\|\theta\|$ | $l(\hat{\theta})$ | $l(Y\|\hat{\theta})$ | Inc/Param |
|---|---|---|---|---|---|---|---|
| Euclidean (d = 2) | 2 | 0 | 3 | 192.00 | -269.00 | -4585 | - |
| Euclidean (d = 3) | 3 | 0 | 3 | 239.00 | -328.00 | -4358 | 4.83 |
| Euclidean (d = 4) | 4 | 0 | 3 | 286.00 | -593.00 | -4242 | 3.65 |
| Mixed (d = 1, r = 1) | 1 | 1 | 5 | 241.00 | -305.00 | -4490 | 1.94 |
| Mixed (d = 2, r = 1) | 2 | 1 | 5 | 288.00 | -621.00 | -4162 | 4.41 |
| Mixed (d = 3, r = 1) | 3 | 1 | 5 | 335.00 | -694.00 | -4073 | 3.58 |
| Mixed (d = 1, r = 2) | 1 | 2 | 5 | 335.00 | -390.00 | -4111 | 3.31 |
| Mixed (d = 2, r = 2) | 2 | 2 | 5 | 382.00 | -798.00 | -4011 | 3.02 |
| Poisson AME (r = 2) | 0 | 2 | 4 | 287.00 | -615.00 | -4347 | 2.51 |
| Poisson AME (r = 3) | 0 | 3 | 4 | 381.00 | -840.00 | -4038 | 2.89 |

First, we investigate improvements in model fit using goodness-of-fit diagnostics in line with Section 8.5.2. We employ the dyad dependence and triad dependence statistics used by Hoff (2015) and Minhasa et al. (2016) to capture reciprocity and transitivity. This dyad dependence statistic is the standardized correlation between in- and out-edge values. It is similar to the reciprocity term of Equation 8.7, but we prefer to use correlation here because we are no longer comparing models with vastly different numbers of constraints.

$$dyad\ dependence(Y) = cor(Y, Y^\top) \tag{10.13}$$

The triad dependence statistic is a standardized measure of the weight of closed triangles.

$$triad\ dependence(Y) = \frac{tr((Y - \bar{Y})^3)}{n(n-1)(n-2)sd(Y)^3} \tag{10.14}$$

For both of these statistics, $Y$ is vectorized as necessary and diagonal entries (self-edges) are ignored. Figure 10.2 and Figure 10.3 compare the distributions of these terms over 1000 simulations from four models of interest. The models we compare are the two-dimensional Euclidean model analyzed previously; the two-dimensional, rank-one mixed model which

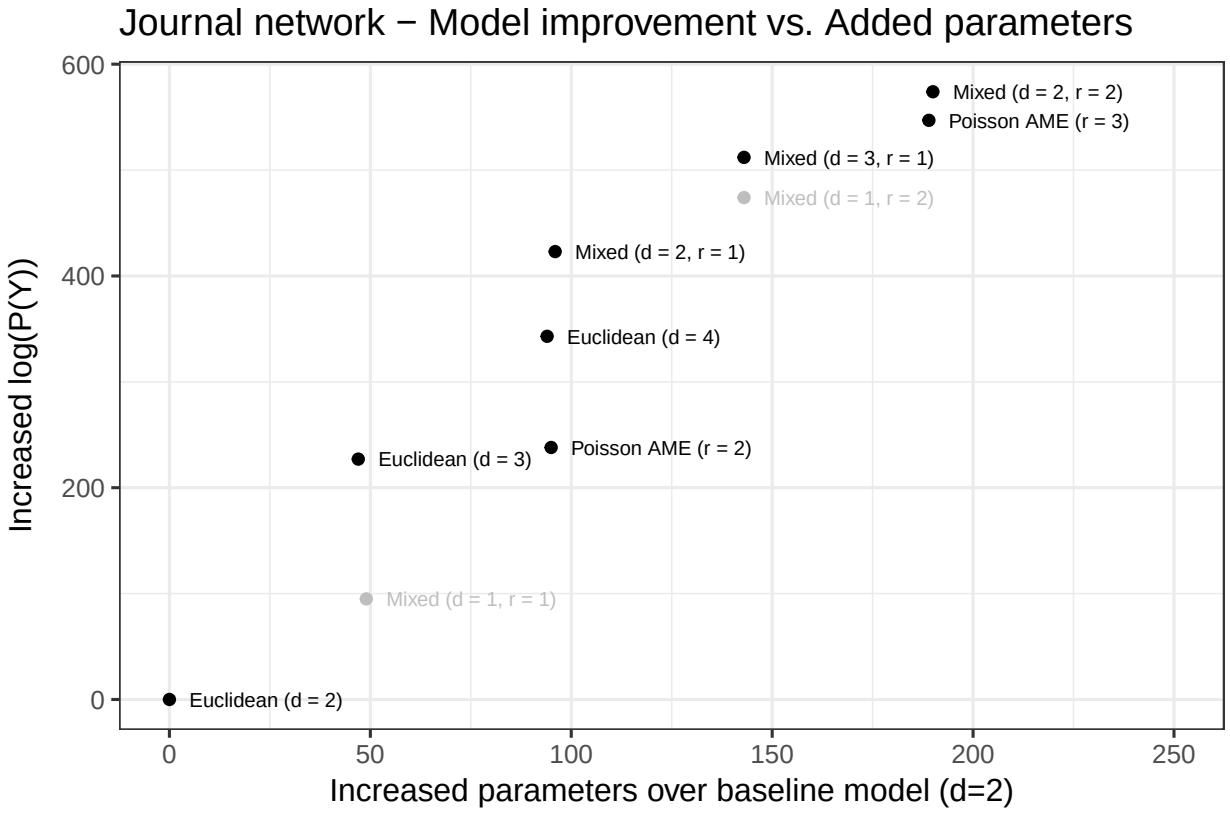


Figure 10.1: Improvement in log probability relative to the two-dimensional Euclidean model. Gray points indicate models with possible collinearity.

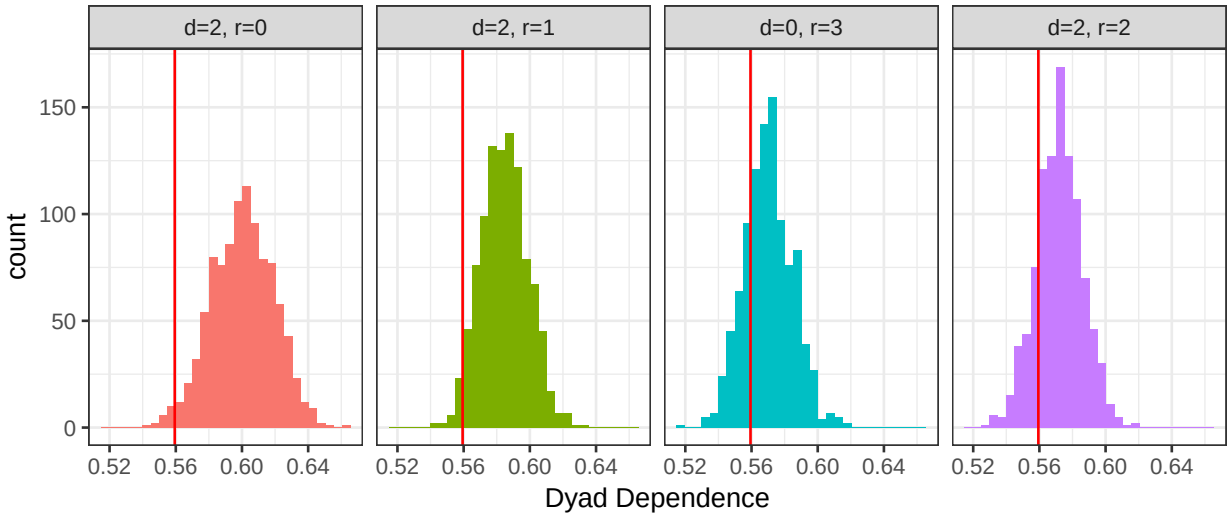


Figure 10.2: Comparison of dyad dependence statistics from 1000 simulated networks per model. The red vertical lines show the observed value and the corresponding lower tail probabilities are 0.015, 0.03, 0.225, and 0.18.

has the best per-parameter increase in likelihood of the mixed models; the two-dimensional, rank-two model which has the highest likelihood of all models considered; and the zero-dimensional, rank-three model which has the second highest likelihood.

Figure 10.2 shows that the two highest-likelihood models perform much better than the others in capturing dyad dependence, with the rank-three model slightly outperforming the two-dimensional, rank-two model. On the other hand, the two-dimensional, rank-two model far outperforms the others in capturing triad dependence, as shown in Figure 10.3. In summary, this mixed model offers clear improvement in goodness of fit relative to the Euclidean model, and slightly outperforms the Poisson AME model with approximately the same number of parameters.

Next we compare residuals of the four models. Figure 10.4 presents standardized residuals as in the right panel of Figure 8.5, which is also the left-most panel of Figure 10.4 to facilitate comparison. We see fewer large standardized residuals for large fitted values as we consider models from left to right in the figure, though the difference is not dramatic. Unfortunately, the overdispersion near fitted values of zero that we observed previously is still present for

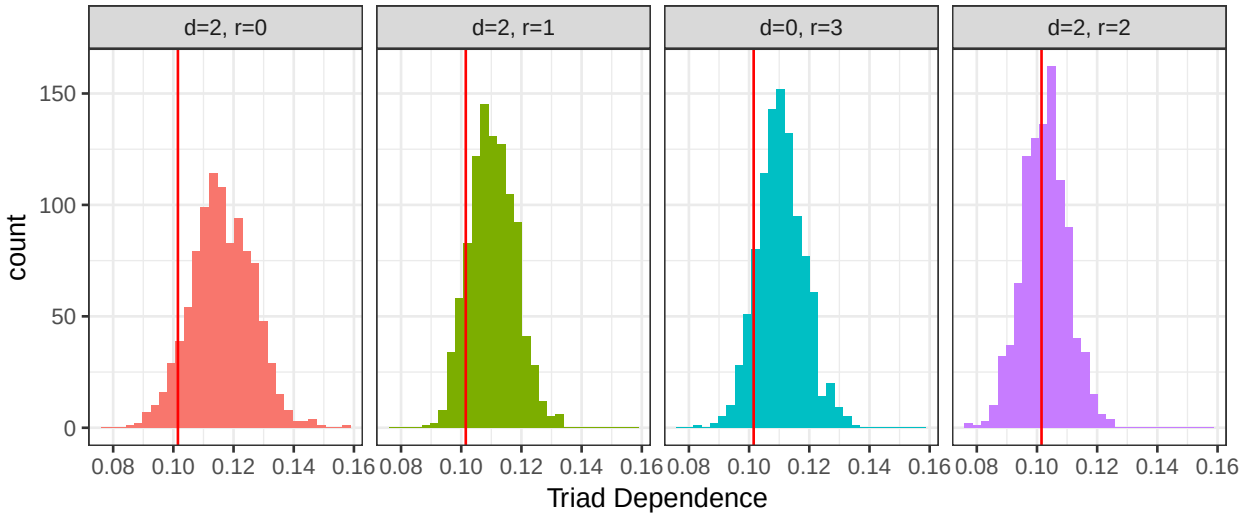


Figure 10.3: Comparison of triad dependence statistics from 1000 simulated networks per model. The red vertical lines show observed statistics and the corresponding lower tail probabilities are 0.079, 0.122, 0.121, and 0.435.

these higher parameter models. It makes sense the the behavior of journals that cite each other infrequently is less systematic and exhibits higher standardized error. In future work we may account for this explicitly, but presently the cost to model complexity outweighs this lack of fit, which does bear heavily on derived ratings or latent parameters.

Finally, we return to our motivating question and consider how the mixed model improves journal ratings. We focus on the two-dimensional, rank-two model because it has the highest likelihood and best captures a combination of dyad and triad dependence. It also facilitates visualization of results since latent positions and factors are in two dimensions. We cannot employ the dimension selection criteria of Section 8.5.1 because the L-BFGS-B method does not return a sample from the model posterior, but the goodness-of-fit diagnostics show that this model captures real patterns in the network and not just noise. We plot the latent positions from this mixed model next to those from the two-dimensional Euclidean model in Figure 10.5. An interactive plot of the mixed model positions is included in supplementary material (citation_net_22.html). Figure 10.5 plots previously analyzed ratings from the two-dimensional Euclidean model against the ratings derived from the two-dimensional, rank-two

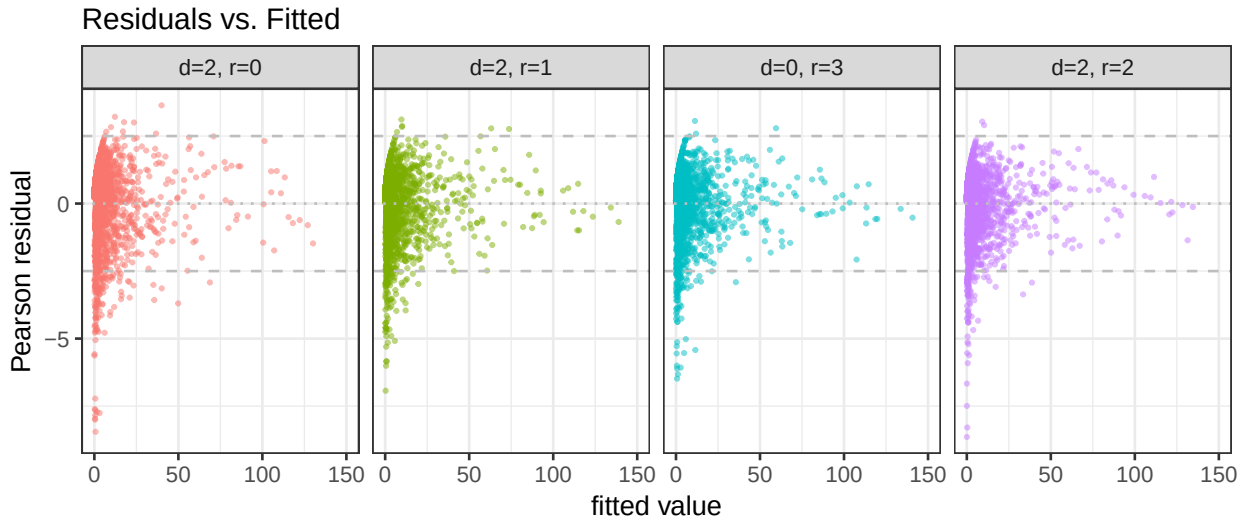


Figure 10.4: Comparison of the 2162 residuals for each model. Pearson residuals are plotted against fitted values. The highest fitted value is left out to enhance detail.

mixed model. They are similar overall, with a correlation of 0.9695.

We focus on the three journals with the largest changes in rating: StNee fares better under the mixed model, and StataJ and JBS fare worse. We generate a circle plot, Figure 10.7 to explore the directed latent factors. This is the native plotting method for AME models. The inner ring of the circle plot shows normalized $u$ (sender) vectors for each journal, with the size of the plotting character proportional to the magnitude of $u$. The outer ring does the same for $v$ (receiver) vectors. StNee has by far the highest magnitude of $u$ vector, suggesting that its out-citations are to a small group of journals. Its direction indicates that its behavior as a sender is similar to other theoretical and general journals. On the other hand, JBS has the largest $v$ vector, indicating that it only received citations from a narrow group of journals. In fact, over half of its citations come from two journals, BioJ and StMed. Similarly, 20 out of 34 citations to StataJ are from StMed, which is by far the highest in-percentage of any pair of journals. In the position plot of the mixed model, StataJ has moved to the periphery of the network.

The directed latent factors have enabled us to distinguish sub-field influence from global influence in the network. Under the two-dimensional Euclidean model the ranking of StNee

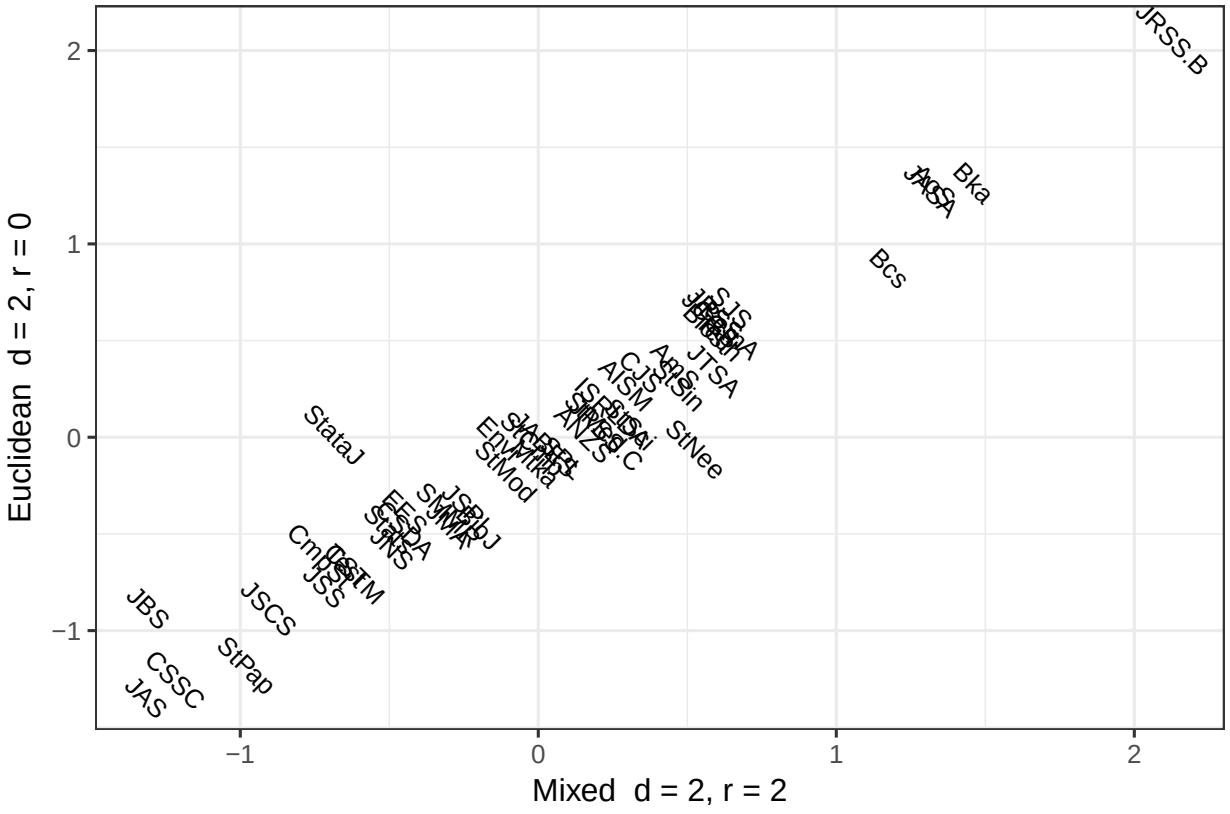


Figure 10.5: Comparison of journal rankings under the two-dimensional Euclidean model and the two-dimensional, rank-two mixed model.

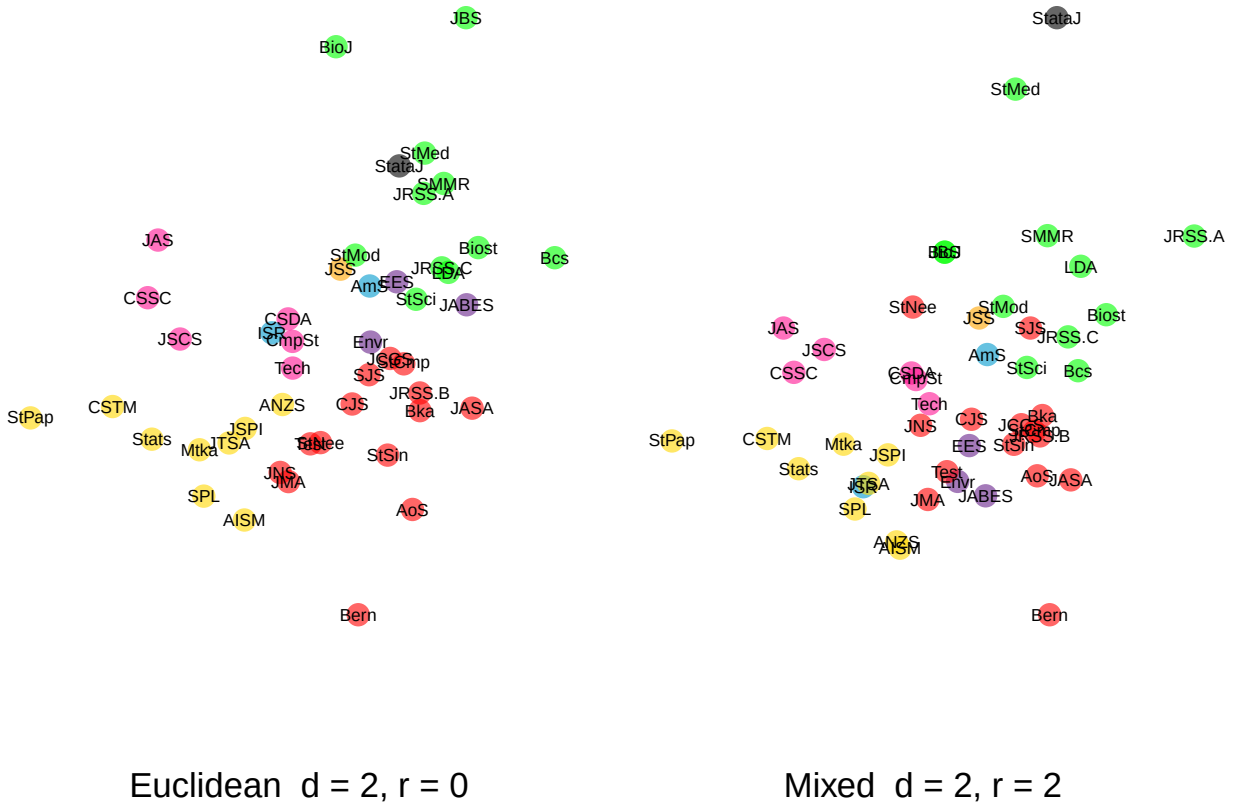


Figure 10.6: Comparison of journal positions under the two-dimensional Euclidean model (left) and the two-dimensional, rank-two mixed model (right). The coloring of the labels corresponds to the groups in Figure 8.4.

was overly reduced by its sender behavior, when in fact it only heavily cites a small group of journals in the network. On the other hand, StataJ and JBS were overly rewarded for receiving citations from a few journals. The quality of the ratings has improved as a result of the decomposition of latent effects.

## 10.4 Comparison of Latent Models for Movie Rating

In this section we apply the same set of low-dimension, low-rank models to the movie network as we applied to the citation network, excluding the two one-dimensional mixed models which seemed to suffer from collinearity. Before applying these models we remove three low-connectivity films from the network which would do not provide enough information to

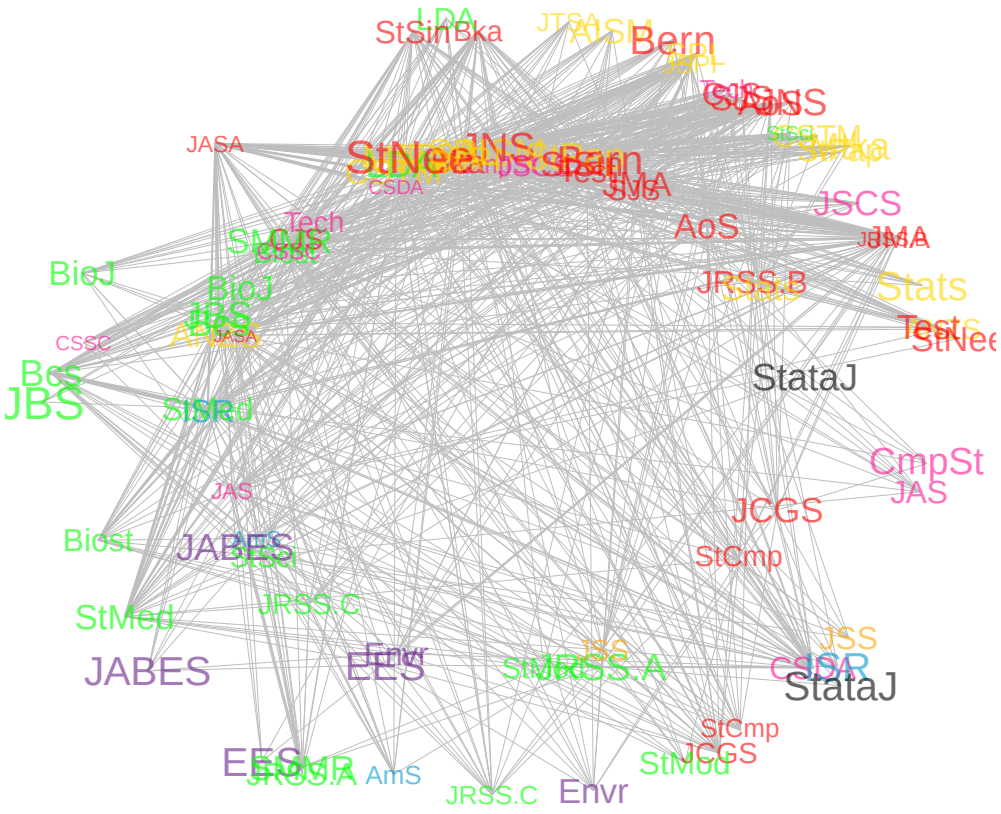


Figure 10.7: Circle plot of the rank-two latent factors of the two-dimensional, rank-two mixed model. The outside ring represents receiver factors (V) and the inside ring represents sender factors (U). The size of the labels are scaled to factor magnitudes. The coloring of the labels corresponds to the groups in Figure 8.4.

Table 10.2: Comparison of models of varying dimension and rank. The final column shows improvement in log likelihood per additional parameter over the two-dimensional Euclidean model.

| | d | r | Var Params | $\|\theta\|$ | $l(Y\|\hat{\theta})$ | $l(\hat{\theta})$ | Inc/Param |
|---|---|---|---|---|---|---|---|
| Euclidean (d = 2) | 2.00 | 0.00 | 3.00 | 504.00 | -32548 | -1036.00 | - |
| Euclidean (d = 3) | 3.00 | 0.00 | 3.00 | 629.00 | -31578 | -1294.00 | 7.76 |
| Euclidean (d = 4) | 4.00 | 0.00 | 3.00 | 754.00 | -31186 | -1448.00 | 5.45 |
| Mixed (d = 2, r = 1) | 2.00 | 1.00 | 5.00 | 756.00 | -30895 | -1596.00 | 6.56 |
| Mixed (d = 3, r = 1) | 3.00 | 1.00 | 5.00 | 881.00 | -30559 | -1763.00 | 5.28 |
| Mixed (d = 2, r = 2) | 2.00 | 2.00 | 5.00 | 1006.00 | -29782 | -2145.00 | 5.51 |
| Poisson AME (r = 2) | 0.00 | 2.00 | 4.00 | 755.00 | -31555 | -1985.00 | 3.96 |
| Poisson AME (r = 3) | 0.00 | 3.00 | 4.00 | 1005.00 | -30239 | -2492.00 | 4.61 |

estimate several directed effects.

Table 10.2 and Figure 10.8 summarize the results. The order of the models in terms of improved log likelihood relative to increased number of parameters is the same as for the citation network models. Again, the two-dimensional, rank-two mixed model performs the best in terms of likelihood of the observed network. The correlation between ratings from this model and those of the two-dimensional Euclidean model is 0.993, and a visual comparison of rankings does not show major fluctuations. An interactive plot of positions derived from this model is included in supplementary material (movie_net_22.html). For brevity, we do not recreate the detailed analysis of model fit that we carried out for models of the citation network. However, we do revisit the genre identification of films via latent positions.

Figure 10.9 shows an updated position plot for the two-dimensional, rank-two model with the results of a three-group k-means clustering algorithm indicated by the plotting color. Misclassifications have dropped from 23 to 9, as detailed by Table 10.3. This is not a result of removing three low-connectivity nodes, after which the two-dimensional model actually exhibited one more classification than before, or 24 total. Rather, comparing Figure 9.2 and Figure 10.9 we see that a central cluster of action films is better-separated from western films under the mixed model, which drives the gain in classification accuracy. Cross-referencing this with the interactive plot of film positions we see that this central chunk consisted mostly of popular films, and we infer that the directed latent factors of the mixed model helped to

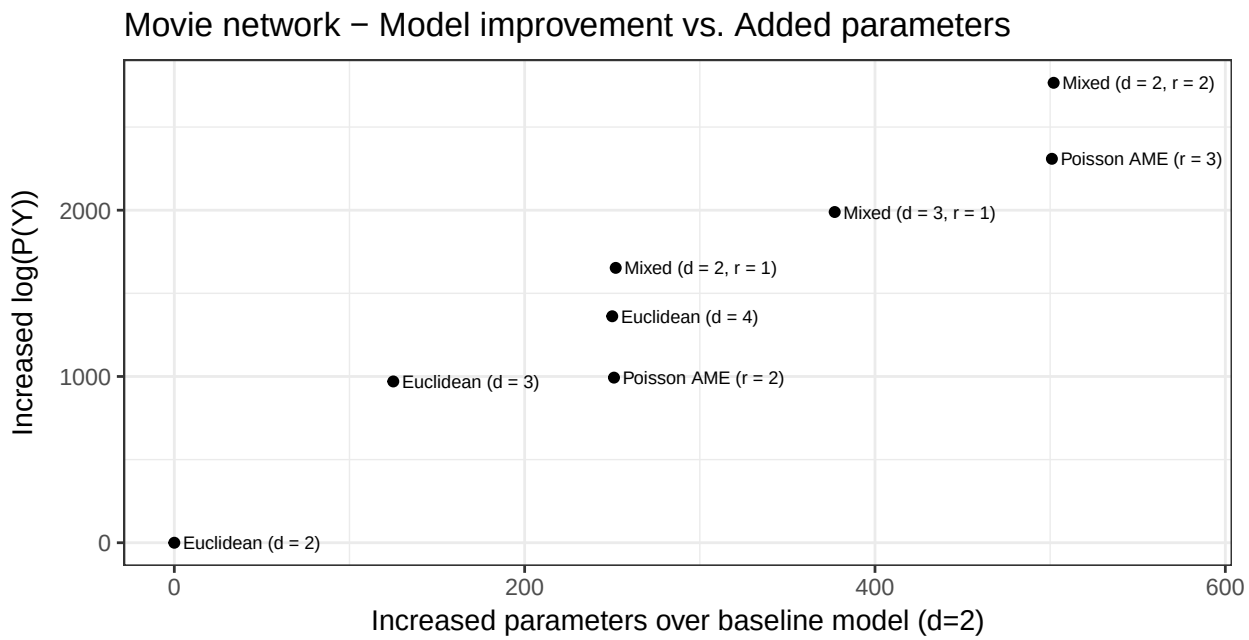


Figure 10.8: Improvement in log probability relative to the two-dimensional Euclidean model.

disambiguate their shared popularity from their similarity.

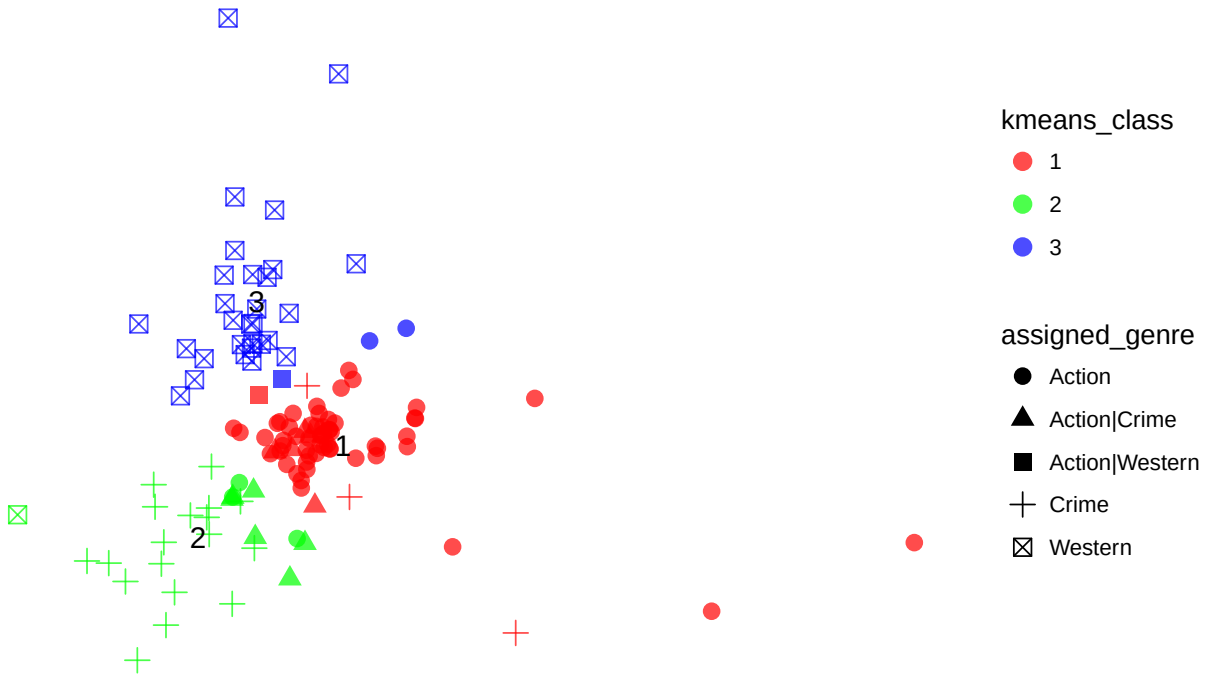


Figure 10.9: Film network colored by k-means class with shape determined by pre-assigned genre. Centers of the k-means classes are labeled 1-3.

Table 10.3: Classification of films by pre-assigned genres (row) vs. k-means cluster (column) for the two-dimensional, rank-two model. Misclassifications have dropped from 23 to 9.

| | Western | Crime | Action |
|---|---|---|---|
| Action | 12 | 4 | 44 |
| Action\|Crime | 0 | 5 | 7 |
| Action\|Western | 2 | 0 | 0 |
| Crime | 3 | 15 | 3 |
| Western | 28 | 1 | 1 |

## 10.5 Appendix: Calculations for Mixed Model Estimation

**1. Log posterior probability of the mixed model, up to a constant:**

Calculations assume no self-edges.

Let $\lambda_{ij} = exp(\beta + a_i + b_j + u_i^\top v_j - \|z_i - z_j\|)$.

$log(P(\theta|Y)) =$

$$\sum_i \sum_{j \neq i} y_{ij}(log(\lambda_{ij})) - \lambda_{ij} - log(y_{ij}!) +$$
$$log(\frac{1}{\sqrt{2\pi\sigma_\beta^2}} exp(\frac{-\beta^2}{2\sigma_\beta^2})) + \sum_i log(\frac{1}{\sqrt{2\pi\sigma_a^2}} exp(\frac{-a_i^2}{2\sigma_a^2})) + \sum_i log(\frac{1}{\sqrt{2\pi\sigma_b^2}} exp(\frac{-b_i^2}{2\sigma_b^2})) +$$
$$\sum_{i,d} log(\frac{1}{\sqrt{2\pi\sigma_z^2}} exp(\frac{-z_{i,d}^2}{2\sigma_z^2})) + \sum_{i,r} log(\frac{1}{\sqrt{2\pi\sigma_u^2}} exp(\frac{-u_{i,r}^2}{2\sigma_u^2})) + \sum_{i,r} log(\frac{1}{\sqrt{2\pi\sigma_v^2}} exp(\frac{-v_{i,r}^2}{2\sigma_v^2})) +$$
$$log(\frac{s_a^{2\,\frac{v_a}{2}}}{\Gamma(\frac{v_a}{2})}) - \frac{v_a s_a^2}{2\sigma_a^2} - log(\sigma_a^2)(1 + \frac{v_a}{2}) + log(\frac{s_b^{2\,\frac{v_b}{2}}}{\Gamma(\frac{v_b}{2})}) - \frac{v_b s_b^2}{2\sigma_b^2} - log(\sigma_b^2)(1 + \frac{v_b}{2}) +$$
$$log(\frac{s_z^{2\,\frac{v_z}{2}}}{\Gamma(\frac{v_z}{2})}) - \frac{v_z s_z^2}{2\sigma_z^2} - log(\sigma_z^2)(1 + \frac{v_z}{2}) + log(\frac{s_u^{2\,\frac{v_u}{2}}}{\Gamma(\frac{v_u}{2})}) - \frac{v_u s_u^2}{2\sigma_u^2} - log(\sigma_u^2)(1 + \frac{v_u}{2}) +$$
$$log(\frac{s_v^{2\,\frac{v_v}{2}}}{\Gamma(\frac{v_v}{2})}) - \frac{v_v s_v^2}{2\sigma_v^2} - log(\sigma_v^2)(1 + \frac{v_v}{2})$$

**2. Gradient calculations:**

$$\frac{\partial \mathbb{LP}}{\partial a_i} = -\frac{a_i}{\sigma_a^2} + \sum_{j \neq i} y_{ij} - \lambda_{ij}$$
$$\frac{\partial \mathbb{LP}}{\partial b_i} = -\frac{b_i}{\sigma_b^2} + \sum_{j \neq i} y_{ji} - \lambda_{ji}$$
$$\frac{\partial \mathbb{LP}}{\partial \beta} = -\frac{\beta}{\sigma_\beta^2} + \sum_i \sum_{j \neq i} y_{ij} - \lambda_{ij}$$
$$\frac{\partial \mathbb{LP}}{\partial z_{i,d}} = -\frac{z_{i,d}}{\sigma_z^2} + \sum_{j \neq i} \frac{(z_{i,d} - z_{j,d})}{\|z_i - z_j\|} \Big[ - y_{ij} - y_{ji} + \lambda_{ij} + \lambda_{ji} \Big]$$
$$\frac{\partial \mathbb{LP}}{\partial u_{i,r}} = -\frac{u_{i,r}}{\sigma_u^2} + \sum_{j \neq i} v_{j,r}(y_{ij} - \lambda_{ij})$$
$$\frac{\partial \mathbb{LP}}{\partial v_{i,r}} = -\frac{v_{i,r}}{\sigma_v^2} + \sum_{j \neq i} u_{j,r}(y_{ji} - \lambda_{ji})$$

# CHAPTER 11

# Conclusion

## 11.1 Summary of Contributions

We summarize below the main contributions of Part II. Details of these contributions and subjects of future work are subsequently discussed. We have:

- Developed a novel rating method for pairwise comparison data using latent space distance models with random effects;
- Implemented our latent space rating model in two applications, highlighting the value of latent space positions for interpretation and genre detection;
- Adapted dimension-selection criteria for the latent space rating model;
- Implemented quasi-Newton estimation and illustrated that it returns accurate parameter estimates in far less time than commonly used MCMC estimation;
- Introduced a novel network model, the mixed latent model, which separates symmetric and asymmetric latent effects. This model offers various improvements over the latent space rating model, and has many applications beyond item ratings.

## 11.2 Discussion

In Part II we introduced methods for rating objects based on relational data, using network models with latent effects. Unlike standard network-based ranking methods like PageRank, these methods are not variants of centrality measures. Instead, they compare the in- and

out-flow of a node while controlling for its position in space – a position that is determined by its network ties. They are thereby an appropriate measure of object importance, influence, or prestige within the network, rather than centrality or popularity. These methods are ideal for the case when elements to be rated have some underlying, but difficult to measure, similarities. In our applications these were the statistics journals with one or more sub-field (theoretical, computational, etc.) and films of one or more genre. The clusters in these networks were fuzzy and irregularly shaped, and not amenable to the latent position cluster models of Handcock et al. (2007). We attempted to apply latent position cluster models to our journal and movie networks, but the models failed to fit meaningful clusters.

Beyond our ratings application, the mixed latent model we have introduced has significant value as a new class of network model. Applications of latent space models such as the analysis of political actors of Hoff and Ward (2004) and of Irish parliamentary elections by Gormley and Murphy (2007) may benefit from the decomposition of symmetric and asymmetric effects provided by the mixed latent model. We plan to apply this model beyond the context of item rating. Another subject of future work is to compare our mixed latent model with one that replaces Euclidean distance with squared Euclidean distance. The latter is optimal in terms of matrix decomposition, but does not guarantee transitivity as un-squared distance does. We may also replace the distance term with the eigenvalue decomposition directly to allow for analysis of heterophilic symmetric factors. These models may pose additional estimation challenges due to collinear factors.

The latent space rating model implemented with Bayesian MCMC estimation provides estimates of uncertainty in rankings that allow us to determine the significance of ranking differences. The quasi-Stigler model returns similar estimates of uncertainty in ratings, but as a network-generating model it underestimates uncertainty because it is conditioned on dyad totals. The mixed latent model does not provide estimates of uncertainty as currently implemented with quasi-Newton estimation, but it would be straightforward to extend the Bayesian MCMC process to this model to generate a posterior sample. Alternatively, we could implement a parametric bootstrap approach to posterior sampling. Comparison of the speed and uncertainty estimates of these methods is a subject of future work.

Although latent space journal rankings are very similar to those of the quasi-Stigler model, the visualizations provided by the latent space model help us better understand what is driving each journal's rating. Furthermore, the visualizations allow for nuanced genre detection. Though our models do not include a term to capture genre, the latent space positions have proven to cluster in agreement with pre-assigned genres. In the case where labels are not provided, k-means clustering on the estimated positions can produce good results. The mixed latent model performs best in this respect, as symmetric and asymmetric effects are most precisely distinguished. In preliminary comparisons this technique has performed better than common clustering methods on the raw data, but further work is needed to establish these results. More work is also needed to establish rigorous standards for the performance of our models in recovering rankings, including in the case of missing data.

The patterns of activity in the networks we studied were fairly homogeneous. The journal set was restricted to a subset of a single field to be suitable for the quasi-Stigler model. Unfortunately, we are not currently able to obtain and present citation data for journals of related fields. The mixed latent model is well-suited to model a network with heterogeneous activity patterns, more so than the latent space rating model or quasi-Stigler model. The directed latent factors of the mixed model can identify nodes that act as conduits between clusters. In future applications we will explore this capacity.

Another type of heterogeneity that may limit the performance of our models is degree heterogeneity, in particular nodes with few connections. We saw in the movie example that films with very few reviews are most likely to deviate from an overall ratings correlation. One way to combat this is to make prior variance distributions less diffuse, which stabilizes the position estimates for low-connectivity nodes without sacrificing ratings estimates, in our experience. To further combat this we could incorporate a film's average star rating as a prior expectation for its receiver coefficient.

We have established that quasi-Newton estimation for the latent space rating model can perform as well as Bayesian MCMC estimation. The speed of quasi-Newton estimation enables us to fit larger networks than were practical with MCMC estimation. However,

we still face the problem of the cost of evaluating the likelihood and gradient functions growing $O(n^2)$. An area of future work is to integrate a local dependence structure into the model to speed up these calculations. For many large networks we expect that the parameters for each node depend primarily on a neighborhood around the node and a few high-activity nodes with wide reach. This would reduce the number of calculations needed for each update step, assuming we can develop a simple method for identifying the nodal neighborhoods. The number of edges to consider would grow roughly $O(n)$ in the case that the mean number of non-zero valued edges incident on a node is not an increasing function of graph size, which is a realistic assumption of many social networks. Raftery et al. (2012) took a related stratified sampling approach in their case-control approximate likelihood model, which they incorporated into MCMC estimation. Integrating a local dependence structure into L-BFGS-B quasi-Newton estimation would multiply the gains in speed.

Finally, we note that many of the arguments in favor of L-BFGS-B quasi-Newton estimation apply to other direct optimization methods. Comparing competing optimization methods and non-convex methods for these models is a subject of future work. Variational Bayesian inference methods have been developed for latent space network models and implemented in the **R** package **VBLPCM**, but only for binary networks (Salter-Townshend and Murphy, 2013). Given the speed of L-BFGS-B estimation and uncertainty in the level of bias introduced by variational techniques, we did not undertake to expand the variational methods to valued networks. Another line of work would be to implement that extension and compare the speed and results to the optimization methods we have employed.

Bibliography

Web of Knowledge, May 2012. URL http://ipscience-help.thomsonreuters.com/inCites2Live/indicatorsGroup/aboutHandbook/usingCitationIndicatorsWisely/jif.html.

Coogan's Bluff, October 2016. URL http://www.imdb.com/title/tt0062824/.

Scandinavian Journal of Statistics, 2017. URL http://www.wiley.com/WileyCDA/WileyTitle/productCd-SJOS.html.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed Membership Stochastic Blockmodels Edoardo. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Almende, B. and Thieurmel, B. *visNetwork: Network Visualization using 'vis.js' Library*. https://CRAN.R-project.org/package=visNetwork, 0.2.1 edition, 2016.

Amin, M. and Mabe, M. M. Impact factors: use and abuse. *Medecina (B Aires)*, 63(4): 347–54, 2003.

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

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

Barrow, D., Drayer, I., Elliott, P., and Osting, B. Ranking rankings: An empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2), 2013.

Bergstrom, C. T. Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68:314–316, 2007.

Bradley, R. A. and Terry, M. E. Rank Analysis of Incomplete Block Designs. *Biometrika*, 39(3):324–45, 1952.

Braverman, M. and Mossel, E. Noisy sorting without resampling. *ArXiv e-prints*, jul 2007. URL `http://adsabs.harvard.edu/abs/2007arXiv0707.1051B`.

Byrd, R. H., Nocedal, J., and Schnabel, R. B. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63:129–156, 1994.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Scientific Computing*, 16(5):1190–1208, 1995.

Chatterjee, S. Matrix Estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Chen, Y., Suh, C., and Goldsmith, A. J. Information Recovery from Pairwise Measurements. *IEEE Transaction on Information Theory*, 62(10):5881–5905, 2016.

Colquhoun, D., Aston, J. A. D., et al. Discussion on the paper by Varin, Cattelan and Firth. *Journal of the Royal Statistical Society A*, 179(1):33–63, 2016.

Cranmer, S. J., Desmarais, B. A., and Kirkland, J. H. Toward a Network Theory of Alliance Formation. *International Interactions*, 38(3):295–324, 2012a.

Cranmer, S. J., Desmarais, B. A., and Menninga, E. J. Complex Dependencies in the Alliance Network. *Conflict Management and Peace Science*, 29(3):279–313, 2012b.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

Davis, J. A. and Leinhardt, S. The Structure of Positive Interpersonal Relations in Small Groups. In Berger, J., editor, *Sociological Theories in Progress, Volume 2*, pages 218–251. Boston: Houghton Mifflin, 1972.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Eddelbuettel, D. and Francois, R. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

Elo, A. *The Rating of Chessplayers, Past and Present.* Arco, New York, 1978.

Fellows, I. Ernm Vignette, November 2012a.

Fellows, I. *Exponential-Family Random Network Models.* PhD thesis, University of California, Los Angeles, University of California, Los Angeles CA, 90095, 2012b.

Fortunato, S. and Hric, D. Community Detection in Networks: A User Guide. *Physics Reports*, 659:1–44, 2016.

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

Fruchterman, T. and Reingold, E. Graph Drawing by Force-directed Placement. *Software – Practice and Experience*, 21(11):1129–1164, 1991.

Gerber, E. R., Henry, A. D., and Lubell, M. Political Homophily and Collaboration in Regional Planning Networks. *American Journal of Political Science*, 57(3):598–610, 2013.

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

Geyer, C. J. and Thompson, E. A. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society, Series B*, 54(3):657–699, 1992b.

Gleich, D. F. PageRank Beyond the Web. *SIAM Review*, 57(3):321 – 363, 2014.

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

Goodreau, S. M., Kitts, J., and Morris, M. Birds of a Feather, or Friend of a Friend? Using Statistical Network Analysis to Investigate Adolescent Social Networks. *Demography*, 45: in press, 2009.

Goodreau, S. M., Cassels, S., Kasprzyk, D., Montano, D. E., Greek, A., and Morris, M. Concurrent Partnerships, Acute Infection and HIV Epidemic Dynamics Among Young Adults in Zimbabwe. *AIDS and Behavior*, 16(2):312–322, 2010.

Gormley, I. C. and Murphy, T. B. *A Latent Space Model for Rank Data*, pages 90–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-73133-7.

Handcock, M. S. Assessing Degeneracy in Statistical Models of Social Networks. Working paper #39, Center for Statistics and the Social Sciences, University of Washington, 2003. URL `http://www.csss.washington.edu/Papers`.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. Model Based Clustering for Social Networks. *Journal of the Royal Statistical Society, Series A*, 170(2):301–354, 2007.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. **ergm**: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3), 2008. URL `http://www.jstatsoft.org/v24/i03/`.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project, 3.5.1 edition, 2015.

Harper, M. F. and Konstan, J. A. The MovieLens Datasets: History and Context. In *ACM Transactions on Interacive Intelligent Systems (TiiS)*, volume 5, 4, page 19, DOI=http://dx.doi.org/10.1145/2827872, December 2015.

Hartigan, J. A. and Wong, M. A. A K-means clustering algorithm. *Applied Statistics*, 28: 100–108, 1979.

Hoff, P., Fosdick, B., Volfovsky, A., and He, Y. *amen: Additive and Multiplicative Effects Models for Networks and Relational Data*, 2017. URL `https://CRAN.R-project.org/package=amen`. R package version 1.3.

Hoff, P. D. Random effects models for network data. In Breiger, R., Carley, K., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis*, volume 126, pages 302–322.

Committee on Human Factors, Board on Behavioral, Cognitive, and Sensory Sciences, National Academy Press, Washington, DC., 2003.

Hoff, P. D. Bilinear Mixed-Effects Models for Dyadic Data. *Journal of the American Statistical Association*, 100:286–295, 2005.

Hoff, P. D. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15:261–272, 2009.

Hoff, P. D. Dyadic data anlysis with amen. Technical Report 638, Department of Statistics, Univeristy of Washington, arXiv:1506.08237, 2015.

Hoff, P. D. and Ward, M. D. Modeling Dependencies in International Relations Networks. *Political Analysis*, 12:160–175, 2004.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Holland, P. W. and Leinhardt, S. An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373), 1981.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic Blockmodels: First Steps. *Social Networks*, 5(2):109–137, 1983.

Horvát, S., Czabarka, E., and Toroczkai, Z. Reducing Degeneracy in Maximum Entropy Models of Networks. *Phys. Rev. Lett.*, 114:158701, Apr 2015. doi: 10.1103/PhysRevLett.114.158701. URL `http://link.aps.org/doi/10.1103/PhysRevLett.114.158701`.

Hunter, D. R., Goodreau, S. M., and Handock, M. S. Goodness of Fit of Social Network Models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

Hunter, D. R. Curved Exponential Family Models for Social Networks. *Social Networks*, 29 (2):216–230, 2007.

Hunter, D. R. and Handcock, M. S. Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics*, 15:565–583, 2006.

INRA and Leger, J.-B. *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*. INRA, 1.1.1 edition, 2015. URL `https://CRAN.R-project.org/package=blockmodels`. R package version 1.1.1.

Karrer, B. Information and code for the degree-corrected block model, November 2010. URL `http://www-personal.umich.edu/~mejn/dcsbm/KLOptimization.cpp`.

Karrer, B. and Newman, M. Stochastic Blockmodels and Community Structure in Networks. *Physical Review E*, 83(1), 2011.

Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

Krivitsky, P. N. Exponential-Family Random Graph Models for Valued Networks. *Electronic Journal of Statistics*, 6:1100–1128, 2012.

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

Krivitsky, P. N. and Butts, C. T. *Modeling Valued Networks with statnet*. The Statnet Development Team, May 2015. URL `http://statnet.csde.washington.edu/workshops/SUNBELT/previous/Valued/Valued.pdf`.

Krivitsky, P. N. and Handcock, M. S. Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24(5), 2008.

Krivitsky, P. N. and Handcock, M. S. *latentnet: Latent Position and Cluster Models for Statistical Networks*. The Statnet Project, http://www.statnet.org, 2.8.0 edition, 2017.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. Representing Degree Distributions, Clustering, and Homophily in Social Networks With Latent Cluster Random Effects Models. *Social Networks*, 27(5):417–428, 2009a.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. Representing Degree Distributions, Clustering, and Homophily in Social Networks With Latent Cluster Random Effects Models. *Social Networks*, 27(5):417–428, 2009b.

Lazega, E. *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press, 2001. URL `http://www.stats.ox.ac.uk/~snijders/siena/Lazega_lawyers_data.htm`.

Leicht, E., Clarkson, G., Shedden, K., and Newman, M. Large-scale structure of time evolving citation networks. *European Physics Journal B*, 59(1):75–83, 2007.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., and de Nooy, W. Field-Normalized Impact Factors (IFs): A Comparison of Rescaling and Fractionally Counted IFs. *Journal of the American Society for Information Science and Technology*, 64(11): 2299–2309, 2013.

Lorrain, F. and White, H. Structural Equivalence of Individuals in Social Networks block-structures with covariates. *Journal of Mathematical Sociology*, 1:49–80, 1971.

Mao, C., Weed, J., and Rigollet, P. Minimax Rates and Efficient Algorithms for Noisy Sorting. *ArXiv e-prints*, oct 2017. URL `http://adsabs.harvard.edu/abs/2017arXiv171010388M`.

Marx, W. and Bornmann, L. Journal Impact Factor: "the poor man's citation analysis" and alternative approaches. *European Science Editing*, 39(2):62–63, 2013.

Maslov, S. and Redner, S. Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks. *Journal of Neuroscience*, 28(44):11103–11105, 2008.

Minhasa, S., Hoff, P. D., and Ward, M. D. Inferential Approaches for Network Analyses: AMEN for Latent Factor Models. *ArXiv e-prints*, November 2016. URL `http://adsabs.harvard.edu/abs/2016arXiv161100460M`.

Morris, M., Kurth, A. E., Hamilton, D. T., Moody, J., and Wakefield, S. Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice. *American Journal of Public Health*, 99(6):1023–1031, 2009.

Neal, P. and Roberts, G. Optimal Scaling for Partially Updating MCMC Algorithms. *Annals of Applied Probability*, 16(2):475–515, 2006.

Negahban, S., Oh, S., and Shah, D. Iterative ranking from pair-wise comparisons. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2474–2482. NIPS, Curran Associates, Inc., 2012. URL `http://papers.nips.cc/paper/4701-iterative-ranking-from-pair-wise-comparisons.pdf`.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.

Newman, M. Assortative Mixing in Networks. *Physical Review Letters*, 89(20), 2002.

Newman, M. Mixing Patterns in Networks. *Physical Review E*, 2(026126), 2003.

Newman, M. *Networks: An Introduction*. Oxford University Press, 2010.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Spring Series in Operations Research and Financial Engineering. Springer New York, 2 edition, 2006.

Nowicki, K. and Snijders, T. A. B. Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.

Park, J. and Newman, M. A Network-Based Ranking System for US College Football. *Journal of Statistical Mechanics: Theory and Experiment*, P10014, 2005.

Park, J. and Yook, S.-H. Bayesian Inference of Natural Rankings in Incomplete Competition Networks. *Scientific Reports*, 4(6212):1–8, 2014.

Pinski, G. and Narin, F. Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. *Information Processing and Management*, 12:297–312, 1976.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. *Bayesian Statistics*, 8:1–45, 2007.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. Fast Inference for the Latent Space Network Model Using a Case-Control Approximate Likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919, 2012.

R Development Core Team. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `http://www.R-project.org/`.

Salter-Townshend, M. and Murphy, T. B. Variational Bayesian Inference for the Latent Position Cluster Model. Technical report, School of Mathematical Sciences, University College Dublin, Dublin, 2009.

Salter-Townshend, M. and Murphy, T. B. Variational Bayesian Inference for the Latent Position Cluster Model. *Computational Statistics and Data Analysis*, 57(1):661–671, 2013.

Sarkar, P. S. and Moore, A. W. Dynamic Social Network Analysis using Latent Space Models. *SIGKDD Explorations: Special Edition on Link Mining*, 7(2):31–40, 2005.

Sarzynska, M., Leicht, E. A., Chowell, G., and Porter, M. A. Null models for community detection in spatially embedded, temporal networks. *Journal of Complex Networks*, 4(3): 363–406, 2016.

Saul, Z. M. and Filkov, V. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(9):2604–2611, 2007.

Schweinberger, M. Instability, Sensitivity, and Degeneracy of Discrete Exponential Families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

Schweinberger, M. and Handcock, M. S. Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society B*, 77(3):647–676, 2015.

Seglen, P. O. Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314:498–502, 1997.

Sewell, D. K. and Chen, Y. Analysis of the formation of the structure of social networks by using latent space models for ranked dynamic networks Authors Analysis of the formation of the structure of social networks by using latent space models for ranked dynamic networks. *Journal of the Royal Statistical Society C*, 64(4):611–633, 2015.

Sewell, D. K. and Chen, Y. Latent Space Approaches to Community Detection in Dynamic Networks. *Bayesian Analysis*, In Press, 2016.

Shah, N. B. and Wainwright, M. J. Simple, Robust and Optimal Ranking from Pairwise Comparisons. *ArXiv e-prints*, dec 2015. URL `http://adsabs.harvard.edu/abs/2015arXiv151208949S`.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. J. Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues. *ArXiv e-prints*, oct 2015. URL `http://adsabs.harvard.edu/abs/2015arXiv151005610S`.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. J. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence Nihar. *Journal of Machine Learning Research*, 17:1–47, 2016.

Shore, J. and Lubin, B. Spectral Goodness of Fit for Network Models. *Social Network*, 43: 16 – 27, 2015a.

Shore, J. and Lubin, B. *spectralGOF: Spectral goodness of fit for network models*, 1.0 edition, January 2015b. URL `http://people.bu.edu/jccs/spectralGOF.html`.

Shortreed, S., Handcock, M. S., and Hoff, P. D. Positional Estimation within the Latent Space Model for Networks. *Methodology*, 2(1):24–33, 2006.

Silver, N. and Fischer-Baum, R. How We Calculate NBA Elo Ratings, May 2015. URL https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/.

Simpson, S. L., Hayasaka, S., and Laurienti, P. J. Exponential Random Graph Modeling for Complex Brain Networks. *PLoS ONE*, 6(5):e20039, 2011.

Simpson, S. L., Moussa, M. N., and Laurienti, P. J. An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks. *NeuroImage*, 60(2):1117–1126, 2012.

Snijders, T. A. B. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure*, 3(2), 2002.

Snijders, T. A. B. and Nowicki, K. Estimation and Prediction for Stochastic Block-Structures for Graphs with Latent Block Structure. *Journal of Classification*, 14:75–100, 1997.

Snijders, T. A., Pattison, P., Robins, G., and Handcock, M. S. New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 36:99–153, 2006.

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

Stefani, R. T. Survey of the major world sports rating systems. *Journal of Applied Statistics*, 24(6):635–646, 1997.

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

Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. A. Comparing Community Structure to Characteristics in Online Collegiate Social Networks. *SIAM Review*, 53(3): 526–543, 2011.

Traud, A. L., Mucha, P. J., and Porter, M. A. Social Structure of Facebook Networks. *Physica A*, 391:4165–4180, 2012.

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

Varin, C., Cattelan, M., and Firth, D. Statistical Modelling of Citation Exchange Between Statistics Journals. *Journal of the Royal Statistical Society A*, 179(1):1–33, 2016.

Ward, M. D., Ahlquist, J. S., and Rozenas, A. Gravity's Rainbow: A dynamic latent space model for the world trade network. *Network Science*, 1(1):95–118, 2013.

Wasserman, S. S. and Pattison, P. Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and $p^*$. *Psychometrika*, 61(3):401–425, 1996.

Watts, D. J. and Strogatz, S. H. Collective dynamics of small-world networks. *Nature*, 393: 440–442, 1998.

Wimmer, A. and Lewis, K. Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. *American Journal of Sociology*, 116(2):583–642, 2010.

Zappa, P. and Mariani, P. The interplay of social interaction, individual characteristics and external influence in diffusion of innovation processes: An empirical test in medical settings. *Procedia - Social and Behavioral Sciences*, 10:140–147, 2011.