STATISTICAL MODELS AND INFERENCE FOR DYNAMIC NETWORKS

BY

DANIEL K. SEWELL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Yuguo Chen, Chair
Associate Professor Feng Liang
Professor John Marden
Professor Annie Qu

Abstract

Dyadic data are ubiquitous and arise in the fields of biology, epidemiology, sociology, and many more. Such dyadic data are often best understood within the framework of networks. Network data can vary in many ways. For example, one might have binary or weighted networks, directed or undirected networks, and static or longitudinal networks. This last type of network, also called a dynamic network, is the focus of this work, with the goal of developing important tools and methodology for the analysis of dynamic networks.

A general framework is developed for modeling dynamic networks via a latent space approach. Using a latent space approach to model such networks allows the researcher to model both the local and global structure of the network, inherently accounts for transitivity, and yields rich and meaningful visualization which can easily be interpreted for qualitative inference on the network. A Markov chain Monte Carlo (MCMC) estimation method within a Bayesian setting is presented. Several useful tools for the researcher arise from this estimation method. First, a method of predicting future relations, or edges, is given. Second, missing data can easily be incorporated into the model, obtaining a posterior probability of each missing edge. Third, a novel concept called nodal influence is introduced which describes how one actor can influence the edges of another actor. Detection of such nodal influence is given via computationally efficient posterior estimation. This model is shown to outperform the existing method, as well as being able to handle richer and more complex data than the existing method. The MCMC algorithm is made scalable by utilizing a log likelihood approximation proposed in the literature, slightly adapted to allow for missing data.

Many of the dynamic networks that arise inherently have weighted edges. The latent space model is extended to handle a variety of types of weighted edges which arise. In particular, the model is extended to account for relational data that can be viewed as, conditioning on the latent actor positions, having come from an exponential family of distributions. An example is also given which demonstrates how, through data augmentation, a similar strategy can be employed when this is not the case. The log likelihood approximation method is then extended to make the MCMC

algorithms scalable for weighted networks.

Of particular interest is Newcomb's fraternity data, a network which captures the evolution and formation of a network beginning in its most nascent form and and ending at a stabilized form. The previous model is modified in two non-trivial ways; the first allows for the modeling of rank-order data, which does not fall into the broad categories of weighted network data given previously, and the second allows for the estimation of the evolution of the stability of the network. Next, it is shown how to use the uncertainties associated with the posterior estimation for subgroup detection and for determining the time at which these subgroups formed. Finally, the model parameters are used to find the association between individual stability and popularity.

A longitudinal mixture model is described which can be used to make hard or soft clustering assignments for p-dimensional real valued data. This model accounts for temporal dependence of both the clustering assignment and the object to be clustered. Additionally, the model allows for covariates which may aid in explaining the clustering assignments. The solutions for implementing the generalized EM algorithm are presented. Recursive relationships are derived which allow the computational cost to grow linearly with time rather than exponentially.

The latent space framework and the longitudinal clustering model are combined to perform community detection within dynamic network data, where the communities' characteristics are fixed but the membership of each community can evolve over time. This method can handle directed or undirected weighted dynamic network data. For community detection within directed or undirected binary networks, a novel model is given along with an efficient variational Bayes estimation algorithm. Both methods are shown to have better performance than using community detection methodology which does not borrow information across time.

Acknowledgments

I would like to thank my advisor, Dr. Yuguo Chen, for spending countless hours working with me and my numerous manuscript drafts, for acquiring funding support, and most of all for teaching me how to better conduct research and write research articles. I would also like to thank Dr. Annie Qu for supporting and working with me via the consulting office, through which I obtained the initial motivation for Chapter 5 of this manuscript. I would also like to thank Drs. Feng Liang and John Marden for being willing to serve on my graduate committee, giving up their time to thoughtfully consider my research.

I would finally like to express my gratitude towards my wife Marnie for being supremely patient and encouraging, all while enduring Spousal Income Frustration Syndrome, and also towards my parents and my grandmother who financially helped me through my first years of higher education.

Table of Contents

Chapter 1

Introduction

Longitudinal relational, or dyadic, data arise in a variety of fields, examples of which include sociology, biology, computer science, entomology and engineering. This type of data, called dynamic network data, consists of a set of actors and a sequence of sets of relations between the actors called "edges" corresponding to observations at discrete time points. Analyzing dynamic networks is key to seeing how friendships form or dissolve, how politicians form loyalties or break ranks with their parties, how co-authorship patterns develop and change over time, etc.

There exist numerous methods of modeling network data within a statistical framework. Some of these models are intended for static networks but have generative processes which can be thought of as dynamic, in the sense of building up the graph over a series of time points. Examples of this notion can be found in the rewiring of "small-world" networks (Watts and Strogatz, 1998), the subsequent addition of edges in an Erdös-Rényi random graph model (Durrett, 2007), or the addition of nodes and edges in a duplication-attachment model (Kumar et al., 2000). Other methods were developed for static networks and were then extended for the dynamic case. One of the most well known methods of analyzing static networks is the exponential random graph model (ERGM) developed by Frank and Strauss (1986), and much attention is still being given to this class of models (see, e.g., Robins et al., 2007). This was extended to analyzing networks observed over discrete time intervals by Hanneke et al. (2010) in the introduction of the temporal ERGM, or TERGM. Using continuous time Markov processes, Snijders (1996) began a series of works corresponding to what is known as stochastic actor-oriented models. These two approaches focus on the use of common network structures or user-defined objective functions. The last commonly used approach to modeling networks that will be mentioned is the latent space model. Latent space approaches aim to embed network information into some latent space (usually a low dimensional Euclidean space). Benefits of using such an approach is that both local and global structures are modeled, transitivity is inherently incorporated in the model, meaningful visualizations are obtained, and the output is easily interpreted, lending itself to much qualitative inference. While the bulk of the literature on

latent space models is concerned with static networks, in this dissertation this approach will be used to model longitudinal network data.

The ideas behind latent space models have long been in use. For example, Nakao and Romney (1993) used multidimensional scaling to visualize and analyze the latent positions of the nodes in Newcomb's fraternity data (Newcomb, 1956). Two formal latent space models were introduced for static networks by Hoff et al. (2002), one of which placed the latent node positions within a Euclidean space, the other placed the latent locations on a unit hypersphere while giving each node an activity level. This latter model was intended to allow for a lack of reciprocity in directed networks. Estimation was performed using Markov chain Monte Carlo (MCMC), hence giving the full posterior of parameters and latent positions. Handcock et al. (2007) expanded the Euclidean model of Hoff et al. (2002) by allowing the latent space positions to follow a mixture of normals, hence allowing clustering to occur simultaneously with embedding in a Euclidean space. Krivitsky et al. (2009) expanded on this work by allowing asymetrical edge probabilities. Schweinberger and Snijders (2003) used a similar approach as Hoff et al. (2002) but used an ultrametric space rather than a Euclidean or hypersphere space to perform model-based clustering. Further work was done in Hoff (2005), where the author extended previous notions of ANOVA models of networks by including as interaction effects the hypersphere latent positions from Hoff et al. (2002).

A limited number of works has considered the temporal aspect of networks while implementing a latent space approach. Robinson and Priebe (2012) presented a method of discovering change points in network behavior via using a k-dimensional simplex latent space. Foulds et al. (2011) developed a non-parametric infinite feature model, where the features are latent. The work most related to the model that will be proposed in Chapter 2 is that of Sarkar and Moore (2005), which extended the Euclidean latent space model of Hoff et al. (2002) to (undirected) dynamic networks. They developed a generalized multidimensional scaling (GMDS) to find the initial latent node positions across discrete time points. The authors then furthered this by using a conjugate-gradient method of optimizing an objective function,. While this is a speedy algorithm and hence can be used for larger data sets, the estimation is based on a filtering-like algorithm (hence not all the data is used for estimating the latent positions) which in the end leads to estimates that are not statistically meaningful. The models in Chapters 2 and 3 extend this to broader contexts while better modeling the network structures.

In Chapter 2, I present a latent space model for dynamic networks with directed binary edges, which will be the foundation for more complex models. This model can represent the network

structure, both local and global, and the evolution of the network. The estimation is done in a Bayesian framework which allows for prediction, missing data, and detection of nodal influence. Chapter 3 shows how to extend this model for weighted dynamic network data. Specifically, the model is extended for the case where the edges belong to an exponential family of distributions; also is shown the case where a similar strategy can be employed through data augmentation. Chapter 4 takes a close look at Newcomb's fraternity data. This is a longitudinal network data set which uniquely captures the evolution of a network as it transitions from an unformed state (all actors were unacquainted) to a stabilized form. With this data I employ novel strategies for determining the time at which the network stabilizes, the time at which subgroups form, and the association between individual stability and popularity. Chapter 5 provides a method of clustering longitudinal p-dimensional real valued data, providing numerically efficient ways of computing the likelihood and an estimation method employing the generalized EM algorithm. Chapter 6 provides two methods of community detection in dynamic networks. The first method combines the models of Chapters 2, 3 and 5 for a model that allows the detection of communities in directed or undirected weighted or binary dynamic networks. The second method is a novel model and estimation algorithm for community detection within binary directed or undirected dynamic networks. In both cases, as in Chapter 5, the structure of each community is assumed fixed but the community memberships are allowed to vary with time.

Chapter 2

Latent Space Model for Binary Dynamic Networks

Network analysis, and in particular dynamic network analysis, is a ubiquitous area of study, used by scientists in many distinct fields (Vivar and Banks, 2011). Often studied are dynamic social networks, which come in a wide variety of forms (see the Special Issues on Network Dynamics in *Social Networks*, January 2010 and July 2012). Dynamic networks are also analyzed in epidemiological contexts (Bansal et al., 2010), in analyzing terrorist networks (Carley, 2006), and much more.

In this chapter, we propose a model which embeds dynamic directed, or undirected, network data into a latent Euclidean space, allowing each node to have a temporal trajectory in this latent space. Estimation of the model parameters and latent nodal positions occur within a Bayesian framework using MCMC. By using our approach, the user can observe much more easily how the network evolves over time, gain insight into global and local structures, handle missing data, make future predictions, and can detect the attracting influence one actor has on another actor's friendships (a concept we call nodal influence, which will be discussed later). To improve the speed of the MCMC algorithm for large networks, we describe an approximation method which reduces the computational cost.

The remainder of the chapter is organized as follows: Section 2.1 describes the proposed model for dynamic networks. Section 2.2 outlines the Bayesian estimation of the model parameters and latent nodal positions, as well as addressing the issue of scalability. Section 2.3 details how to handle missing data. Section 2.4 describes how to obtain network predictions. Section 2.5 gives a method for detecting and visualizing nodal influence. Section 2.6 shows simulation results. Section 2.7 presents the results from analyzing data collected from a Dutch classroom as well as from analyzing cosponsorship data collected on members of the U.S. House of Representatives. Sections 2.8 and 2.9 gives the proofs of the lemmas given earlier in the chapter.

2.1 Dynamic Latent Space Model

We assume that data come in the form of $(\mathcal{N}, \{\mathcal{E}_t : t \in \mathcal{T}\})$, where $\mathcal{N}$ is the set of all nodes, and $\mathcal{E}_t$ is the set of edges at time t. For simplicity let $\mathcal{T} = \{1, 2, \ldots, T\}$. For the majority of the chapter it will also be assumed that $\mathcal{E}_t$ consists of directed edges. The general idea of the latent space approach is that this time series of graphs can be represented as a state space model, with a latent state variable representing the nodes as positions in a low dimensional Euclidean space. The closer two nodes are in this latent Euclidean space, the more likely they are to form an edge. This low dimensional space can be thought of as a characteristic space where the distance between nodes represents how similar they are (Hoff et al., 2002), or as a social space where the distance between two nodes corresponds to the strength of the relationship between the two.

The notation to be used throughout the rest of the chapter is as follows: $n = |\mathcal{N}|$ is the number of nodes. For a latent space $\Re^p$, $\mathbf{X}_{it}$ is the p dimensional vector of the i^{th} node's latent position at time t, and $\mathcal{X}_t$ is the $n \times p$ matrix whose i^{th} row is $\mathbf{X}_{it}$. $Y_t = \{y_{ijt}\}$ is the adjacency matrix of the observed network at time t, and $y_{ijt} = 1$ if there is an edge from node i to node j at time t and 0 otherwise.

The latent node positions are modeled by a Markov process with the initial distribution

$$\pi(\mathcal{X}_1|\boldsymbol{\psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{i1}|\mathbf{0}, \tau^2 I_p), \tag{2.1}$$

and transition equation

$$\pi(\mathcal{X}_t|\mathcal{X}_{t-1}, \boldsymbol{\psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_p) \tag{2.2}$$

for $t = 2, 3, \ldots, T$, where I_p is the $p \times p$ identity matrix, $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the normal probability density function with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$, and $\boldsymbol{\psi}$ is a vector of parameters which will be defined shortly.

The observed networks at different time points are conditionally independent given the latent positions. This dependence structure is illustrated in Figure 2.1. Further, it is assumed that for any two (distinct) pairs (i, j) and (i', j'), y_{ijt} and $y_{i'j't}$ are independent conditioning on $(\mathcal{X}_t, \boldsymbol{\psi})$. In formulating the observation equation of our model, we desire two main properties: first, the probability of an edge from actor i to actor j at time t should increase as the distance between their latent positions decreases; second, the probability of an edge should depend on both who is sending and who is receiving the link, and we should further be able to determine the importance of each

in edge formation; i.e., whether the identity of the sender or the identity of the receiver is more important in edge formation. To this end, we use the formulation

$$\mathbb{P}(Y_t|\mathcal{X}_t,\boldsymbol{\psi}) = \prod_{i\neq j}\mathbb{P}(y_{ijt}=1|\mathcal{X}_t,\boldsymbol{\psi})^{y_{ijt}}\cdot\mathbb{P}(y_{ijt}=0|\mathcal{X}_t,\boldsymbol{\psi})^{1-y_{ijt}} = \prod_{i\neq j}\frac{\exp(y_{ijt}\eta_{ijt})}{1+\exp(\eta_{ijt})}, \tag{2.3}$$

where

$$\eta_{ijt} \triangleq \log\left(\frac{\mathbb{P}(y_{ijt}=1|\mathcal{X}_t,\boldsymbol{\psi})}{\mathbb{P}(y_{ijt}=0|\mathcal{X}_t,\boldsymbol{\psi})}\right) = \beta_{IN}\left(1-\frac{d_{ijt}}{r_j}\right)+\beta_{OUT}\left(1-\frac{d_{ijt}}{r_i}\right), \tag{2.4}$$

and $d_{ijt} = \|\mathbf{X}_{it}-\mathbf{X}_{jt}\|$ and $\boldsymbol{\psi} = (\tau^2,\sigma^2,\beta_{IN},\beta_{OUT},r_{1:n})$ are the model parameters. Here $r_{1:n} = (r_1,r_2,\ldots,r_n)$; similar notation will be used throughout the rest of the chapter. β_{IN} and β_{OUT} are global parameters which reflect the importance of popularity and social activity respectively. The r_i's are positive node specific parameters that represent each node's social reach and is reflective of the tendency to form and receive edges. Within the latent space, there is also the geometrical interpretation of r_i forming a radius around the i^{th} node, as we will see later. For model identifiability, the r_i's are constrained so that $\sum_{i=1}^n r_i = 1$. This parameterization emulates both the distance and projection models of Hoff et al. (2002) for static networks, given as $\eta_{ij} = \beta(1-d_{ij})$ and $\eta_{ij} = \beta + \mathbf{X}_i'\mathbf{X}_j/\|\mathbf{X}_j\|$ respectively, by utilizing the visually appealing and intuitive Euclidean space for the latent positions while incorporating the individual actors' "sociability," or social reach, while also accounting for both activity and popularity.

Krivitsky et al. (2009) built onto Hoff et al.'s model by including additive random individual effects. Here our parameterization links the actors' individual effects to the latent space, in the sense that the social reach dampens or augments the effect of the distance between the two actors, rather than having the individual actor effects be constant additive effects; thus these two parameterizations are in fact different, rather than being subsets of each other. In some sense their model is more flexible in that an actor has both an indegree effect and an outdegree effect. Our model can be trivially extended to account for this by simply allowing $r_{1:n}$ to be replaced by two sets of parameters $r_{1:n}^{(IN)}$ and $r_{1:n}^{(OUT)}$. We applied this more complex model on the two real data sets presented in Section 2.7 with no improvement in model fit. Hence our focus remains on the simpler model, given in (2.4).

To better understand the interpretation of the β coefficients and the radii $r_{1:n}$, we consider the following two cases which one would reasonably expect to occur in practice: (1) both $\beta_{IN} > 0$ and $\beta_{OUT} > 0$ (which we would expect to happen most often), and (2) either $\beta_{IN} < 0$ and $\beta_{OUT} > |\beta_{IN}|$ or $\beta_{OUT} < 0$ and $\beta_{IN} > |\beta_{OUT}|$.

In case (1), the interpretation of the radii that comes naturally from (2.4) is that r_i marks the

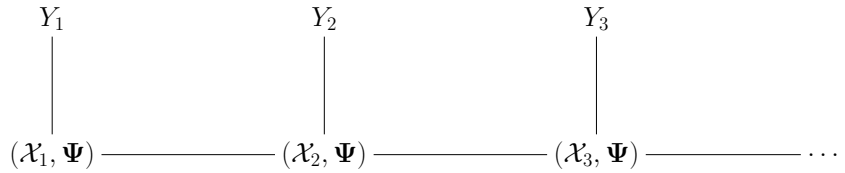


Figure 2.1: Illustration of the dependence structure for the latent space model. Y_t is the observed graph, $\mathcal{X}_t$ is the unobserved latent node positions, and $\boldsymbol{\psi}$ is the vector of model parameters.

radius within the latent social space of the i^{th} node's social reach. This is evident in that if the distance between two nodes are within each other's radii, i.e., $d_{ijt} < \min(r_i, r_j)$, then the probability of an edge is greater than 1/2; if they are outside each other's radii, i.e., $d_{ijt} > \max(r_i, r_j)$, then the probability of an edge is less than 1/2; and if the distance between the two nodes equals both radii, i.e., $d_{ijt} = r_i = r_j$, then the probability of an edge equals 1/2. These scenarios are illustrated in Figure 2.2. Thus in case (1), a larger radius implies an increasing propensity to send and receive ties. This fact is illustrated in Figure 2.3a, where the probability $\mathbb{P}(y_{ijt} = 1|\mathcal{X}_t, \boldsymbol{\psi})$ is shown in a contour plot, allowing r_i and r_j to vary, with distance $d_{ijt} = 0.01$, $\beta_{IN} = 2$ and $\beta_{OUT} = 1/2$. Now if $\beta_{IN} > \beta_{OUT}$ ($\beta_{OUT} > \beta_{IN}$) then we can conclude that the probability of an edge from node i to node j (from node j to node i) is determined more by the radius of j than by the radius of i. This is also illustrated in Figure 2.3a, where it is apparent that the probability of an edge from i to j increases much faster when we fix a value of r_i and allow r_j to increase than vice versa. Thus in case (1), if $\beta_{IN} > \beta_{OUT}$ then the edges of the network are determined more by the popularity of the actors than by their activity, i.e., the identity of the receiver of the edge is more important than the identity of the sender, and if $\beta_{OUT} > \beta_{IN}$ the edges of the network are determined more by the activity of the actors than by their popularity, i.e., the identity of the sender is more important than the identity of the receiver.

In case (2), one of the β coefficients is negative; without loss of generality assume $\beta_{OUT} < 0$ and $\beta_{IN} > |\beta_{OUT}|$. Considering the probability of an edge from node i to node j, note that as r_j increases the probability of an edge increases; thus in case (2), a larger radius implies an increasing propensity to receive ties. This is illustrated in Figure 2.3b, where $\mathbb{P}(y_{ijt} = 1|\mathcal{X}_t, \boldsymbol{\psi})$ is shown in a contour plot, allowing r_i and r_j to vary, with distance $d_{ijt} = 0.01$, $\beta_{IN} = 2$ and $\beta_{OUT} = -1/2$. Similar to case (1), the probability of an edge from node i to node j is determined more by r_j than r_i. This can be seen in Figure 2.3b where it is apparent that when we fix r_i and allow r_j to increase, $\mathbb{P}(y_{ijt} = 1|\mathcal{X}_t, \boldsymbol{\psi})$ increases at a more rapid rate than the rate at which $\mathbb{P}(y_{ijt} = 1|\mathcal{X}_t, \boldsymbol{\psi})$ decreases

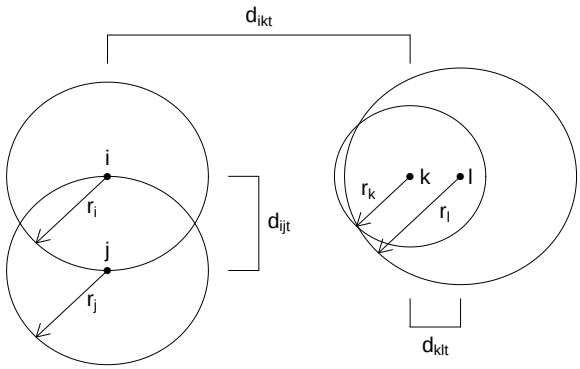
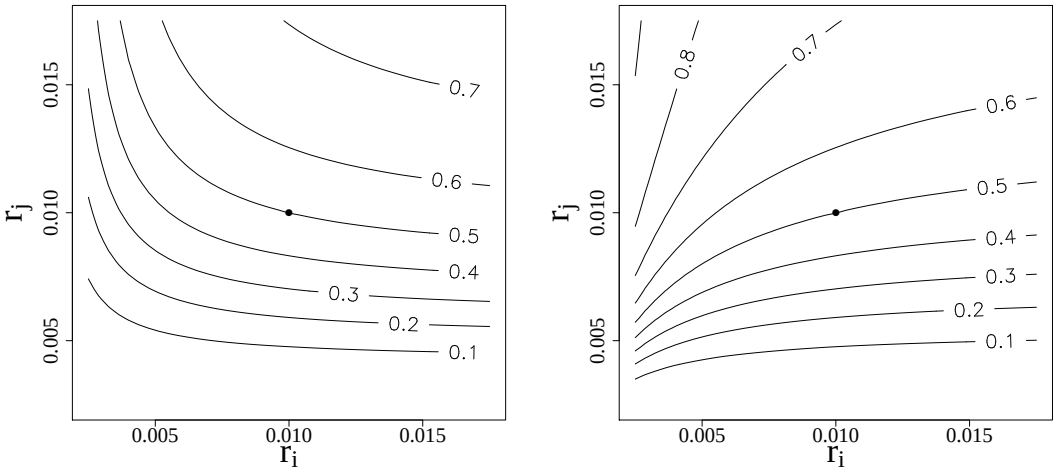


Figure 2.2: Illustration of how to interpret social reach parameters $r_{1:n}$ in the case that $\beta_{IN}, \beta_{OUT} > 0$; the probability of an edge from i to k is less than $1/2$, from k to l is greater than $1/2$, and from i to j is equal to $1/2$.

when we fix r_j and allow r_i to increase. Thus in case (2) the edges of the network are determined more by the popularity of the actors than by their activity. Similar conclusions can be made if we assume $\beta_{IN} < 0$ and $\beta_{OUT} > |\beta_{IN}|$.

Any other case than these two would contradict the intuition of the model that a shorter distance between two nodes is guaranteed to lead to an increased probability of an edge; additionally the possibility of groups of nodes (communities) would no longer exist.



(a) Case (1)

(b) Case (2)

Figure 2.3: Contour plots of $\mathbb{P}(y_{ijt} = 1 | \mathcal{X}_t, \boldsymbol{\psi})$; $\beta_{IN} = 2$, $d_{ijt} = 0.01$, and $\beta_{OUT} = 1/2$ in (a), and $\beta_{OUT} = -1/2$ in (b). The point $r_i = r_j = d_{ijt}$ is marked with a dot.

2.2 Estimation

We adopt a Bayesian approach, and hence we wish to make inferences based on $\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}|Y_{1:T})$, where $\boldsymbol{\psi} = (\tau^2, \sigma^2, \beta_{IN}, \beta_{OUT}, r_{1:n})$. We implement a Metropolis-Hastings (MH) within Gibbs MCMC scheme as suggested by Geweke and Tanizaki (2001) to sample from the posterior, hence giving point estimates and uncertainties. We set the priors on the parameters as follows: assume that $\beta_{IN} \sim N(\nu_{IN}, \xi_{IN})$, $\beta_{OUT} \sim N(\nu_{OUT}, \xi_{OUT})$, $\sigma^2 \sim IG(\theta_\sigma, \phi_\sigma)$, $\tau^2 \sim IG(\theta_\tau, \phi_\tau)$ and $(r_1, r_2, \ldots, r_n) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, where IG is the inverse gamma distribution. The inverse gamma priors were chosen to be conjugate, and the Dirichlet prior is a natural selection for such constrained parameters.

2.2.1 Initialization

The number of MCMC iterations required to reach convergence can be greatly reduced by appropriate initial values of the latent positions $\mathcal{X}_{1:T}$ and the parameters τ^2, σ^2, β_{IN}, β_{OUT}, and $r_{1:n}$.

We want the radii $r_{1:n}$, or social reaches, to be strongly associated with the in and outdegree of the nodes. Therefore we have the initial estimates (denoted by superscript $^{(1)}$) of each radius to be

$$r_i^{(1)} = \frac{\sum_{t=1}^{T} \sum_{j \neq i} (y_{ijt} + y_{jit})/2}{\sum_{t'=1}^{T} \sum_{j' \neq i'} y_{i'j't'}}. \tag{2.5}$$

A nice method of obtaining initial latent positions $\mathcal{X}_{1:T}^{(1)}$ is GMDS, given by Sarkar and Moore (2005). This GMDS method extends the classical multidimensional scaling (CMDS) by using CMDS to obtain $\mathcal{X}_1^{(1)}$, and then for $t = 2, 3, \ldots, T$ minimizing an objective function that balances the CMDS estimates for $\mathcal{X}_t^{(1)}$ with the previous estimates of $\mathcal{X}_{t-1}^{(1)}$. To perform GMDS, T distance matrices $\{d_{ijt}\}$ are required; Sarkar and Moore suggested constructing these distance matrices by setting d_{ijt} to be the length of the shortest path from i to j at time t. We too do this, though we rescale by $1/n$ to keep the distances on the same scale as the radii, as (2.4) seems to suggest is necessary. From the initial estimates of the radii and the latent positions, we find initial estimates of β_{IN} and β_{OUT} by maximizing the likelihood numerically.

An intuitive initial value for τ^2 can be found using the initial latent positions $\mathcal{X}_1^{(1)}$ in the following way:

$$\frac{1}{np} \sum_{i=1}^{n} \|\mathbf{X}_{i1}^{(1)}\|^2. \tag{2.6}$$

One could determine the initial value for σ^2 by using the initial positions $\mathcal{X}_{1:T}^{(1)}$ in much the same

way as was done for τ^2. Note that such an initial estimate would be heavily influenced by the user specified tuning parameter in the GMDS algorithm, and hence the user may be misled into thinking that the data are supporting a particular initial value when in fact the user has already predetermined it. In our simulation study and real data analyses we chose a large initial value for σ^2 in order to allow the latent positions $\mathbf{X}_{it}$ to move more during the beginning iterations of the MCMC algorithm.

There typically is no intuition as to what values to set the hyperparameters of the prior distributions, and so we suggest using some of these initial values of the parameters to aid in this. To make the prior of τ^2 flat one can set the shape and scale parameters of the inverse gamma prior to be respectively $\theta_\tau = 2 + \delta$ and $\phi_\tau = (1+\delta) \cdot \mathbb{E}(\tau^2)$ for some small $\delta > 0$, where $\mathbb{E}(\tau^2)$ is set to be the initial estimate in (2.6). The initial estimates of β_{IN} and β_{OUT} are natural selections for ν_{IN} and ν_{OUT}. The prior can then be made flat by simply choosing ξ_{IN} and ξ_{OUT} to be very large values. The hyperparameters for $r_{1:n}$ may all be set equal to 1 in order to obtain a flat and uninformative prior, as well as to simplify the computations involved in the MCMC algorithm. There are no automatic decisions for determining the hyperparameters for σ^2. However, as we will see in Section 2.6, the MCMC algorithm is not sensitive to this selection.

2.2.2 Posterior Sampling

To sample via Metropolis-Hastings within Gibbs algorithm, we draw from the full conditional distributions iteratively. These conditional distributions are either known in closed form or up to a normalizing constant and are given below.

Letting $p_{ijt} \triangleq \mathbb{P}(y_{ijt}|\mathcal{X}_t, \boldsymbol{\psi}) = \exp(y_{ijt}\eta_{ijt})/(1+\exp(\eta_{ijt}))$, the conditional distribution for $\mathbf{X}_{it}$ is

$$\pi(\mathbf{X}_{it}|Y_{1:T}, \boldsymbol{\psi}) \propto \begin{cases} \big(\prod_{j:j\neq i} p_{ijt}p_{jit}\big) \cdot N(\mathbf{X}_{it}|\mathbf{0}, \tau^2 I_p) \cdot N(\mathbf{X}_{i(t+1)}|\mathbf{X}_{it}, \sigma^2 I_p), & \text{if } t = 1 \\ \big(\prod_{j:j\neq i} p_{ijt}p_{jit}\big) \cdot N(\mathbf{X}_{i(t+1)}|\mathbf{X}_{it}, \sigma^2 I_p) \cdot N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_p), & \text{if } 1 < t < T \\ \big(\prod_{j:j\neq i} p_{ijt}p_{jit}\big) \cdot N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_p), & \text{if } t = T. \end{cases} \tag{2.7}$$

The conditional distributions for β_{IN} and β_{OUT} are

$$\pi(\beta_{IN}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{OUT}, r_{1:n}) \propto \left[\prod_{t=1}^{T}\prod_{i\neq j} p_{ijt}\right] \cdot N(\beta_{IN}|\nu_{IN}, \xi_{IN}), \tag{2.8}$$

$$\pi(\beta_{OUT}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{IN}, r_{1:n}) \propto \left[\prod_{t=1}^{T}\prod_{i\neq j} p_{ijt}\right] \cdot N(\beta_{OUT}|\nu_{OUT}, \xi_{OUT}). \tag{2.9}$$

Assuming the prior of $r_{1:n}$ is Dirichlet$(1, \ldots, 1)$, the conditional distribution for $r_{1:n}$ is

$$\pi(r_{1:n}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{IN}, \beta_{OUT}) \propto \prod_{t=1}^{T}\prod_{i\neq j} p_{ijt}. \tag{2.10}$$

The conditional distributions for τ^2 and σ^2 are

$$\tau^2|Y_{1:T}, \mathcal{X}_{1:T}, \sigma^2, \beta_{IN}, \beta_{OUT}, r_{1:n} \sim IG\Big(\theta_\tau + np/2,\ \phi_\tau + \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{X}_{i1}\|^2\Big), \tag{2.11}$$

$$\sigma^2|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \beta_{IN}, \beta_{OUT}, r_{1:n} \sim IG\Big(\theta_\sigma + np(T-1)/2,\ \phi_\sigma + \frac{1}{2}\sum_{t=2}^{T}\sum_{i=1}^{n}\|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|^2\Big). \tag{2.12}$$

The posterior sampling algorithm is then

0. Set the initial values of $(\mathcal{X}_{1:T}, \boldsymbol{\psi})$ (e.g., to those described in Section 3.1).
1. For $t = 1, \ldots, T$ and for $i = 1, \ldots, n$, draw $\mathbf{X}_{it}$ via MH using a normal random walk proposal.
2. Draw τ^2 from (2.11).
3. Draw σ^2 from (2.12).
4. Draw β_{IN} via MH using a normal random walk proposal.
5. Draw β_{OUT} via MH using a normal random walk proposal.
6. Draw $r_{1:n}$ via MH using a Dirichlet proposal.
Repeat steps 1-6.

Due to the constraint on the radii ($\sum_{i=1}^{n} r_i = 1$), it is necessary to, within the MH step, accept or reject all n values simultaneously; hence it is important to keep the movements small, i.e., the variance of the proposal small. We also desire to keep the means at the current values. Therefore the proposal used to draw the new values $r^*_{1:n}$ is another Dirichlet distribution with parameters $(\kappa r_1, \kappa r_2, \ldots, \kappa r_n)$, where the r_i's are the current values and κ is some large constant. Thus the acceptance rate for $r^*_{1:n}$ is

$$\min\left\{1, \frac{\pi(r^*_{1:n}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{IN}, \beta_{OUT})}{\pi(r_{1:n}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{IN}, \beta_{OUT})} \cdot \frac{Dir(r_{1:n}|\kappa r^*_{1:n})}{Dir(r^*_{1:n}|\kappa r_{1:n})}\right\},$$

where $Dir(r_{1:n}|\alpha_{1:n})$ is the Dirichlet probability density function with parameters $\alpha_{1:n}$ evaluated at $r_{1:n}$, and $\pi(r_{1:n}|Y_{1:T}, \mathcal{X}_{1:T}, \tau^2, \sigma^2, \beta_{IN}, \beta_{OUT})$ is given in (2.10).

One last note is that the posterior will be invariant to rotations, reflections, and translations of the latent positions. Hence any inference must take into account the non-uniqueness of the estimates. Similar to the approach described in Hoff et al. (2002), we perform a Procrustes transformation to reorient the sampled trajectories. We set an $(nT) \times p$ reference trajectory matrix $\boldsymbol{\mathcal{X}}_0$, and after drawing new $\mathbf{X}_{it}$ for all i and t, we construct from these new draws the new trajectory matrix $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1', \ldots, \mathcal{X}_T')'$. In practice we used the initial latent positions to construct $\boldsymbol{\mathcal{X}}_0$. The Procrustes transformation on $\boldsymbol{\mathcal{X}}$ using $\boldsymbol{\mathcal{X}}_0$ as the target matrix finds

$$\underset{\boldsymbol{\mathcal{X}}^*}{\operatorname{argmin}} \quad tr\left(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{X}}^*\right)'\left(\boldsymbol{\mathcal{X}}_0 - \boldsymbol{\mathcal{X}}^*\right),$$

where $\boldsymbol{\mathcal{X}}^*$ is some rotation and translation of $\boldsymbol{\mathcal{X}}$; see, e.g., Borg (2005). By performing the Procrustes transformation on the trajectory matrix, we obtain a single rotation matrix A with which we use to set $\boldsymbol{\mathcal{X}}^{(\ell)} = \boldsymbol{\mathcal{X}}A$, where the superscript $^{(\ell)}$ denotes the stored values for the ℓ^{th} iteration; that is, we set $\mathbf{X}_{it}^{(\ell)} = A'\mathbf{X}_{it}$. By so doing we are preserving the distances between any actors at any time points, i.e., $\|\mathbf{X}_{it}^{(\ell)} - \mathbf{X}_{js}^{(\ell)}\| = \|\mathbf{X}_{it} - \mathbf{X}_{js}\|$ for any actors i and j and any time points t and s.

2.2.3 Scalability

Scalability is an issue for latent space models for network data. For static networks, this issue has been addressed through using variational Bayes (Salter-Townshend and Murphy, 2012) and also by using case-control principles from epidemiology (Raftery et al., 2012). This latter method reduced the computational cost (for static networks) of computing the log likelihood from $O(n^2)$ to $O(n)$.

The general strategy of the case-control log likelihood approximation is to write the log likelihood as two summations. Assuming that the network becomes sparser as n gets larger, the computational cost of the first of the two summations is linear with respect to n, and the cost of the second is quadratic. The second summation is then replaced by a Monte Carlo estimate obtained from a subsequence of the actors, thus making the overall cost of computing the log likelihood linear in n.

This method as described by Raftery et al. (2012), however, cannot be directly extended to longitudinal network data containing missing edge values which is often the case, especially in social networks. This is because all the links need to be known a priori. By modifying how the log likelihood is decomposed into two summations, we can apply this same method without knowing all y_{ijt} beforehand. I now give the details on this approximation method, modified slightly to allow for missing data.

One can reduce the computational cost involved in an iteration of the MCMC algorithm by approximating the acceptance ratios corresponding to the Metropolis-Hastings steps. Consider the latent positions where, for actor i at time t, it is necessary to compute $\prod_{j:j\neq i} p_{ijt}p_{jit}$. We can write the log of this quantity as three summations:

$$\begin{aligned}\sum_{j:j\neq i}\log(p_{ijt}p_{jit}) &= \sum_{j:j\neq i}\left[y_{ijt}\eta_{ijt}+y_{jit}\eta_{jit}-\log(1+\exp(\eta_{ijt}))-\log(1+\exp(\eta_{jit}))\right]\\ &= \sum_{j:y_{ijt}=1}\eta_{ijt}+\sum_{j:y_{jit}=1}\eta_{jit}-\sum_{j:j\neq i}\left[\log(1+\exp(\eta_{ijt}))+\log(1+\exp(\eta_{jit}))\right].\end{aligned} \tag{2.13}$$

From this expression we see that for each MCMC iteration it is necessary to compute $O(Tn^2)$ terms for updating $\mathcal{X}_{1:T}$. It is reasonable, though, to assume that as n gets larger the adjacency matrices become sparser, and so the first two summations are not growing at an alarming rate; we can formalize this by assuming that either the maximum node degree is fixed or is of $o(n)$. The final summation in (2.13), however, requires $n-1$ calculations, and hence is responsible for the computational cost being quadratic with respect to n. To ease the computational burden, we replace this term with a summation of only a fixed number n_0 of terms. Specifically, we randomly draw for each i a subsequence $\{j_k\}_{k=1}^{n_0}$. Then the last term in (2.13) can be approximated as

$$\begin{aligned}&\sum_{j:j\neq i}\left[\log(1+\exp(\eta_{ijt}))+\log(1+\exp(\eta_{jit}))\right]\\ \approx&\frac{n-1}{n_0}\sum_{k=1}^{n_0}\left[\log(1+\exp(\eta_{ij_kt}))+\log(1+\exp(\eta_{j_kit}))\right].\end{aligned} \tag{2.14}$$

A similar approximation can be used in the acceptance ratios for β_{IN}, β_{OUT} and $r_{1:n}$. For each of these it is necessary to compute $\prod_{t=1}^{T}\prod_{i=1}^{n}\prod_{j:j\neq i}p_{ijt}$. For each i and t, $\log\left(\prod_{j:j\neq i}p_{ijt}\right)$ can be written as two summations:

$$\begin{aligned}\sum_{j:j\neq i}\log(p_{ijt}) &= \sum_{j:j\neq i}\left[y_{ijt}\eta_{ijt}-\log(1+\exp(\eta_{ijt}))\right]\\ &= \sum_{j:y_{ijt}=1}\eta_{ijt}-\sum_{j:j\neq i}\log(1+\exp(\eta_{ijt})).\end{aligned} \tag{2.15}$$

Here again we see that the number of terms to be summed in updating β_{IN}, β_{OUT} and $r_{1:n}$ is of $O(Tn^2)$. Using the same subsequences as before, the second summation in (2.15) can be approxi-

mated as

$$\sum_{j:j\neq i} \log(1+\exp(\eta_{ijt})) \approx \frac{n-1}{n_0} \sum_{k=1}^{n_0} \log(1+\exp(\eta_{ij_kt})). \tag{2.16}$$

By using the approximations in (2.14) and (2.16), the computational cost of each MCMC iteration is reduced from $O(Tn^2)$ to $O(Tn)$, i.e., the computational cost is now linear with respect to n.

It is not necessary to know the values of the y_{ijt}'s a priori in order to draw the subsequences $\{j_k\}_{k=1}^{n_0}$. Further, (2.13) can still be calculated because at each iteration of the MCMC algorithm there will be some estimate of y_{ijt}. Therefore this approximation can be implemented in the context of missing data.

One last note is that the sampling of the subsequences $\{j_k\}_{k=1}^{n_0}$ can be drawn in a variety of ways. In the simulations of Section 2.6 and the cosponsorship data analysis of Section 2.7, the subsequences were drawn via stratified sampling. That is, for each i, the $n-1$ other nodes were divided into two groups, $G_1 = \{j \neq i : y_{ijt} = 1 \text{ or } y_{jit} = 1 \text{ for at least one } t\}$ and $G_2 = \{j \neq i : y_{ijt} = y_{jit} = 0, \ \ \forall t\}$. Then $\{j_k\}_{k=1}^{c}$ was a random sample from G_1 and $\{j_k\}_{k=c+1}^{n_0}$ was a random sample from G_2, where $c = \lfloor n_0|G_1|/(n-1) + 0.5 \rfloor$.

2.3 Missing Data

Missing data in social networks is not uncommon, and can come in various forms, such as boundary specification, non-response, and censoring by vertex degree (Kossinets, 2006). Here we specifically focus on non-responses, i.e., missing edge values. For static networks there have been a number of methods proposed (see, e.g., Robins et al., 2004; Huisman, 2009). For dynamic networks, Huisman and Steglich (2008) compared several methods to handle missing edges in the context of a stochastic actor oriented model. Handcock and Gile (2010) developed a theoretical framework for networks in which only a subset of the dyads are observed; we use this framework in our discussion and refer the reader to Handcock and Gile's paper for more details.

Let $\mathcal{D}$ denote the sampling pattern; that is, $\mathcal{D}$ is the set of $n \times n$ matrices $\{D_1, \ldots, D_T\}$ where D_{ijt} equals 1 if the dyad y_{ijt} is observed and equals 0 otherwise. Letting $\mathcal{Y}^{(obs)}$ and $\mathcal{Y}^{(mis)}$ denote the collection of observed edges and missing edges respectively, the complete data is $(\mathcal{Y}^{(obs)}, \mathcal{Y}^{(mis)}, \mathcal{D})$, and the incomplete (observed) data is $(\mathcal{Y}^{(obs)}, \mathcal{D})$. The unobserved edges $\mathcal{Y}^{(mis)}$ are considered missing completely at random (MCAR) if $\mathbb{P}(\mathcal{D}|\mathcal{Y}^{(obs)}, \mathcal{Y}^{(mis)}, \xi) = \mathbb{P}(\mathcal{D}|\xi)$, where ξ is some set of parameters corresponding to the sampling pattern. If, however, $\mathbb{P}(\mathcal{D}|\mathcal{Y}^{(obs)}, \mathcal{Y}^{(mis)}, \xi) = \mathbb{P}(\mathcal{D}|\mathcal{Y}^{(obs)}, \xi)$, then the unobserved edges are considered missing at random (MAR). The case where the pattern of

unobserved edges depends on the unobserved edges themselves (called non-ignorable missing data) is a difficult scenario which is beyond the scope of this chapter; thus we will continue the discussion assuming that the missing edges are either MCAR or MAR.

Rubin (1976) discussed weak conditions for which it is possible to ignore the process that causes missing data. In our context, we are interested in the posterior distribution $\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mathcal{Y}^{(mis)}|\mathcal{Y}^{(obs)}, \mathcal{D})$; hence if the sampling pattern is ignorable, we may make inference based on the posterior distribution $\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mathcal{Y}^{(mis)}|\mathcal{Y}^{(obs)})$, i.e., $\mathcal{X}_{1:T}$, $\boldsymbol{\psi}$ and $\mathcal{Y}^{(mis)}$ are independent of $\mathcal{D}$ given $\mathcal{Y}^{(obs)}$. There are two sufficient conditions that must be satisfied in order for the sampling pattern to be ignorable (Rubin, 1976). First, the sampling pattern parameters ξ are *a priori* independent with the data $(\mathcal{Y}^{(mis)}, \mathcal{Y}^{(obs)})$, latent positions $\mathcal{X}_{1:T}$ and model parameters $\boldsymbol{\psi}$, i.e., $\pi(\mathcal{Y}^{(mis)}, \mathcal{Y}^{(obs)}, \mathcal{X}_{1:T}, \boldsymbol{\psi}, \xi) = \pi(\mathcal{Y}^{(mis)}, \mathcal{Y}^{(obs)}, \mathcal{X}_{1:T}, \boldsymbol{\psi})\pi(\xi)$. Second, the space of $(\xi, \mathcal{X}_{1:T}, \boldsymbol{\psi})$ is a product space, i.e., if $\xi \in \Xi$, $\mathcal{X}_{1:T} \in \boldsymbol{\mathcal{X}}$ and $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ then $(\xi, \mathcal{X}_{1:T}, \boldsymbol{\psi}) \in \Xi \times \boldsymbol{\mathcal{X}} \times \boldsymbol{\Psi}$. If these two conditions are met and the missing edges are either MCAR or MAR, we have

$$\begin{aligned}
\pi(\mathcal{Y}^{(mis)}, \mathcal{X}_{1:T}, \boldsymbol{\psi}|\mathcal{Y}^{(obs)}, \mathcal{D}) &= \frac{\int \pi(\mathcal{D}|\mathcal{Y}^{(obs)}, \xi)\pi(\mathcal{Y}^{(mis)}, \mathcal{Y}^{(obs)}, \mathcal{X}_{1:T}, \boldsymbol{\psi})\pi(\xi)d\xi}{\int \pi(\mathcal{D}|\mathcal{Y}^{(obs)}, \xi)\pi(\mathcal{Y}^{(obs)})\pi(\xi)d\xi} \\
&= \frac{\pi(\mathcal{Y}^{(mis)}, \mathcal{Y}^{(obs)}, \mathcal{X}_{1:T}, \boldsymbol{\psi})}{\pi(\mathcal{Y}^{(obs)})} \\
&= \pi(\mathcal{Y}^{(mis)}, \mathcal{X}_{1:T}, \boldsymbol{\psi}|\mathcal{Y}^{(obs)}). \qquad (2.17)
\end{aligned}$$

Handling the missing data is easy when using the MH within Gibbs sampling scheme of Section 3. Using the observed data and the current values for the missing data, the full conditionals for $\mathcal{X}_{1:T}$ and $\boldsymbol{\psi}$ are unchanged. The full conditional of $\mathcal{Y}^{(mis)}$ is determined by, for any $y_{ijt} \in \mathcal{Y}^{(mis)}$,

$$\pi(y_{ijt}|\mathcal{X}_{1:T}, \boldsymbol{\psi}) = \frac{\exp(y_{ijt}\eta_{ijt})}{1 + \exp(\eta_{ijt})}, \qquad (2.18)$$

where η_{ijt} is given in (2.4). That is, including the missing data in the MH within Gibbs sampling amounts to an additional draw for each missing y_{ijt} from a Bernoulli distribution with probability determined by (2.4).

2.4 Prediction

Predicting future links is an important and interesting problem. Applications include recommender systems, terrorist networks, protein interaction networks, prediction of friendship networks, and

more (Wang et al., 2007; Kashima and Abe, 2006; Hopcroft et al., 2011; Liben-Nowell and Kleinberg, 2007).

When considering prediction in the latent space context, it is of interest to predict for time $T+1$ both the edges of the adjacency matrix Y_{T+1} and the latent space positions $\mathcal{X}_{T+1}$. It is simple to find point estimates of the latter since

$$\begin{aligned}\pi(\mathcal{X}_{T+1}|Y_{1:T}) &= \int \pi(\mathcal{X}_{T+1}|\mathcal{X}_T,\boldsymbol{\psi})\pi(\mathcal{X}_{1:T},\boldsymbol{\psi}|Y_{1:T})d\mathcal{X}_{1:T}d\boldsymbol{\psi} \\ &\approx \frac{1}{L}\sum_{\ell=1}^{L}\prod_{i=1}^{n} N(\mathbf{X}_{i(T+1)}|\mathbf{X}_{iT}^{(\ell)},\sigma^{2^{(\ell)}}I_p), \end{aligned} \tag{2.19}$$

where the superscript $^{(\ell)}$ indicates the ℓ^{th} draw from the posterior. Hence

$$\widehat{\mathcal{X}}_{T+1} \triangleq \mathbb{E}(\mathcal{X}_{T+1}|Y_{1:T}) \approx \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{X}_T^{(\ell)}. \tag{2.20}$$

It is assumed that an appropriate burn-in period for the chain has been accounted for.

A simple way to compute a point estimate of the probability of an edge between i and j at time $T+1$, $\mathbb{P}(y_{ij(T+1)}=1)$, would be to plug in $\widehat{\mathcal{X}}_{T+1}$ along with the posterior means of the parameters into the observation equation (2.4). We can, however, do a little better by not conditioning on the posterior means of the model parameters, hence eliminating some unnecessary uncertainty. We aim, then, to find $\mathbb{P}(Y_{T+1}|Y_{1:T},\mathcal{X}_{T+1}=\widehat{\mathcal{X}}_{T+1})$. Since conditional on $\mathcal{X}_{T+1}$ we still assume that the $y_{ij(T+1)}$'s are independent, we need only find $\mathbb{P}(y_{ij(T+1)}|Y_{1:T},\widehat{\mathbf{X}}_{i(T+1)},\widehat{\mathbf{X}}_{j(T+1)})$. This can be estimated as follows:

First, we approximate the joint distribution.

$$\begin{aligned}&\mathbb{P}(y_{ij(T+1)},\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)}|Y_{1:T}) \\ =&\int \pi(y_{ij(T+1)}|\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)},\boldsymbol{\psi})\pi(\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)}|\mathcal{X}_{1:T},\boldsymbol{\psi})\pi(\mathcal{X}_{1:T},\boldsymbol{\psi}|Y_{1:T})d\mathcal{X}_{1:T}d\boldsymbol{\psi} \\ \approx&\frac{1}{L}\sum_{\ell=1}^{L}\pi(y_{ij(T+1)}|\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)},\boldsymbol{\psi}^{(\ell)})\pi(\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)}|\mathcal{X}_T^{(\ell)},\boldsymbol{\psi}^{(\ell)}). \end{aligned} \tag{2.21}$$

Next the marginal distribution of $(\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)})|Y_{1:T}$ is found as in (2.19) and approximated by

$$\frac{1}{L}\sum_{\ell=1}^{L}\pi(\mathbf{X}_{i(T+1)},\mathbf{X}_{j(T+1)}|\mathcal{X}_T^{(\ell)},\boldsymbol{\psi}^{(\ell)}) = \frac{1}{L}\sum_{\ell=1}^{L}N(\mathbf{X}_{i(T+1)}|\mathbf{X}_{iT}^{(\ell)},\sigma^{2^{(\ell)}}I_p)N(\mathbf{X}_{j(T+1)}|\mathbf{X}_{jT}^{(\ell)},\sigma^{2^{(\ell)}}I_p). \tag{2.22}$$

Thus the conditional distribution of $y_{ij(T+1)}|Y_{1:T}, \widehat{\mathcal{X}}_{T+1}$ is estimated as a weighted average:

$$\begin{aligned}\mathbb{P}(y_{ij(T+1)}|Y_{1:T}, \widehat{\mathcal{X}}_{T+1}) &= \frac{\pi(y_{ij(T+1)}, \widehat{\mathbf{X}}_{i(T+1)}, \widehat{\mathbf{X}}_{j(T+1)}|Y_{1:T})}{\pi(\widehat{\mathbf{X}}_{i(T+1)}, \widehat{\mathbf{X}}_{j(T+1)}|Y_{1:T})} \\ &\approx \sum_{\ell=1}^{L} w_\ell \pi(y_{ij(T+1)}|\widehat{\mathbf{X}}_{i(T+1)}, \widehat{\mathbf{X}}_{j(T+1)}, \boldsymbol{\psi}^{(\ell)}), \end{aligned} \tag{2.23}$$

where

$$w_\ell = \frac{N(\widehat{\mathbf{X}}_{i(T+1)}|\mathbf{X}_{iT}^{(\ell)}, \boldsymbol{\psi}^{(\ell)})N(\widehat{\mathbf{X}}_{j(T+1)}|\mathbf{X}_{jT}^{(\ell)}, \boldsymbol{\psi}^{(\ell)})}{\sum_{\ell'=1}^{L} N(\widehat{\mathbf{X}}_{i(T+1)}|\mathbf{X}_{iT}^{(\ell')}, \boldsymbol{\psi}^{(\ell')})N(\widehat{\mathbf{X}}_{j(T+1)}|\mathbf{X}_{jT}^{(\ell')}, \boldsymbol{\psi}^{(\ell')})} \tag{2.24}$$

and $\pi(y_{ij(T+1)}|\widehat{\mathbf{X}}_{i(T+1)}, \widehat{\mathbf{X}}_{j(T+1)}, \boldsymbol{\psi})$ is defined in (2.3) and (2.4).

This method outperforms the simpler plug in method mentioned earlier, as will be seen in Section 2.6 when the simulation results are given. The intuition as to why this is so is that we are using fewer estimated parameters to make predictions, hence introducing less uncertainty into the prediction estimates.

2.5 Nodal Influence

The latent space approach gives rise to a novel way of detecting and viewing nodal, or social, influence. Anagnostopoulos et al. (2008) defined social influence as "the phenomenon that the actions of a user can induce his/her friends to behave in a similar way." Many authors attempt to use social influence to track the propagation of ideas or behaviors through a network (e.g., Kempe et al. (2003), Leskovec et al. (2006)). Tang et al. (2009) described a method of determining which nodes will influence which other nodes on a variety of topics. Goyal et al. (2010) proposed a method of labeling each edge by an influence probability, assuming undirected binary edges. These works all require data outside of the network, such as, as phrased by Goyal et al., an "action log." Here we consider nodal influence as the attracting influence one node has on another node's friendships; i.e., how one person draws another person into their own social circle. Hence our new definition of nodal influence is how one node affects the edges of another node.

We assume that the way in which one node can affect another node's movements in the network is manifested in an increased tendency for the influenced node to move in the direction of the influencing node in the social space. To detect the tendency for a node i to move through the social space in the direction of another node j, the transition equation for the latent nodal positions is extended by considering a new parameter to describe the nodal influence between two nodes. We will then carefully define this parameter, implement an appropriate prior, and then look at posterior

probabilities that will help the user determine whether or not there is nodal influence. We will show that this can be done by using the same MCMC output as when the transition equation was assumed to be a random walk. Throughout the next two sections we consider looking at the nodes pairwise, i.e., we look at whether or not a specific node i is influenced by another node j.

2.5.1 Nodal Influence Detection

Consider an extension of the transition equation (2.2) such that $\mathbf{X}_{it} = \mathbf{X}_{i(t-1)} + \boldsymbol{\epsilon}_{it}$ where $\boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{\mu}_t, \sigma^2 I_p)$. In the following, assume that $p = 2$. Let θ_t equal the angle $\text{atan2}(\mathbf{X}_{jt} - \mathbf{X}_{i(t-1)})$, where atan2 is the common variation of the arctangent function which preserves the angle's quadrant, taking a vector rather than a ratio as its argument. Let $\mathcal{R}_t$ be the rotation matrix associated with θ_t. We then let $\boldsymbol{\mu}_t$ be of the form

$$\boldsymbol{\mu}_t = \mathcal{R}_t \begin{pmatrix} \mu \\ 0 \end{pmatrix} = \begin{pmatrix} \cos(\theta_t) & -\sin(\theta_t) \\ \sin(\theta_t) & \cos(\theta_t) \end{pmatrix} \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \tag{2.25}$$

where $\mu = \|\mathbb{E}(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})\|$ is some unknown parameter taking non-negative values. No nodal influence is equivalent to the case where $\mu = 0$, and if there does exist some nodal influence then this will be reflected in some $\mu > 0$. Figure 2.4 gives an illustration of this type of nodal influence. The idea here is that node i will aim towards wherever node j is within the latent characteristic space. If we let the prior on the parameters $(\boldsymbol{\psi}, \mu)$ be independent, i.e., $\pi(\boldsymbol{\psi}, \mu) = \pi(\boldsymbol{\psi})\pi(\mu)$, then the posterior samples obtained from Section 2.2.2 can be equivalently viewed as having come from $\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi} | Y_{1:T}, \mu = 0)$. This is important because, as will be seen later, we can use these same draws to make inference regarding the nodal influence existing between nodes i and j. Also note that under the extended transition equation, the Markov property still holds for the latent positions, i.e., $\pi(\mathcal{X}_t | \mathcal{X}_{1:(t-1)}, \boldsymbol{\psi}, \mu) = \pi(\mathcal{X}_t | \mathcal{X}_{t-1}, \boldsymbol{\psi}, \mu)$.

The prior distribution of μ is chosen to be a mixture of a point mass on 0 and a continuous component over the positive reals:

$$\pi(\mu) = \begin{cases} p_0 & \text{if } \mu = 0 \\ (1 - p_0) f(\mu) & \text{for } \mu > 0, \end{cases} \tag{2.26}$$

where f is some proper continuous density on $(0, \infty)$. Here f will be assumed for convenience to be

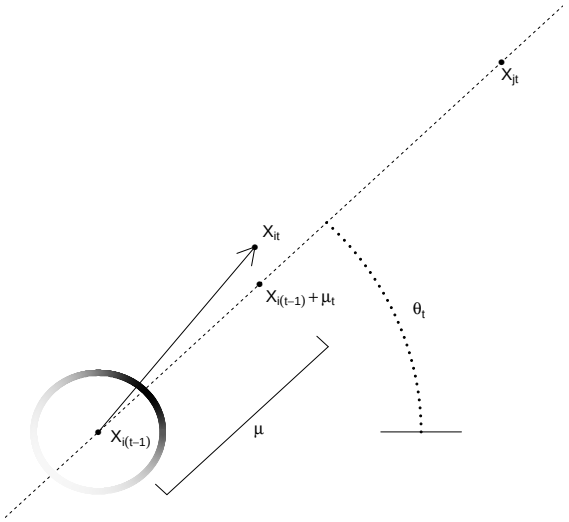


Figure 2.4: The extension of the transition equation to allow for node j's influence on node i. Node i is more likely to move *toward* node j. The circle around $\mathbf{X}_{i(t-1)}$ represents a von Mises distribution for the angle component of $\boldsymbol{\epsilon}_{it}$'s spherical coordinates, where dark values indicate high probability regions and light values indicate low probability regions.

the exponential distribution with mean λ. Then the posterior density is

$$\pi(\mu|Y_{1:T}) = \frac{\pi(Y_{1:T}|\mu)\pi(\mu)}{\pi(Y_{1:T})} = \frac{\pi(Y_{1:T}|\mu)p_0}{\pi(Y_{1:T})}\mathbf{1}_{\{\mu=0\}} + \frac{\pi(Y_{1:T}|\mu)}{\pi(Y_{1:T})}(1-p_0)f(\mu)\mathbf{1}_{\{\mu>0\}}. \tag{2.27}$$

For notation, let $\pi_0(\mu = 0|Y_{1:T}) = \pi(Y_{1:T}|\mu = 0)p_0/\pi(Y_{1:T})$ and let $\pi_+(\mu|Y_{1:T}) = \pi(Y_{1:T}|\mu)(1 - p_0)f(\mu)/\pi(Y_{1:T})$. Then $\pi_0(\mu = 0|Y_{1:T})$ is the point mass posterior probability that $\mu = 0$. If our prior probability $p_0 = 1/2$ and we find that the posterior probability is less than $1/2$ then this implies the data is pulling the posterior probability towards the conclusion that node i is influenced by node j.

Since $1 = \pi_0(\mu = 0|Y_{1:T}) + \int_0^\infty \pi_+(\mu|Y_{1:T})d\mu$, we have that

$$\pi_0(\mu = 0|Y_{1:T}) = \frac{1}{1 + \int_0^\infty \kappa(\nu)d\nu}, \tag{2.28}$$

where $\kappa(\nu) = \pi_+(\mu = \nu|Y_{1:T})/\pi_0(\mu = 0|Y_{1:T})$. So if we can find $\int_0^\infty \kappa(\nu)d\nu$ then we can compute $\pi_0(\mu = 0|Y_{1:T})$. To this end, we have the following lemma:

Lemma 2.5.1. *For $\kappa(\nu)$ as defined above,*

$$\int_0^\infty \kappa(\nu)d\nu = \mathbb{E}(h(\mathcal{X}_{1:T}, \boldsymbol{\psi})|Y_{1:T}, \mu = 0), \tag{2.29}$$

where

$$\begin{aligned} h(\mathcal{X}_{1:T}, \boldsymbol{\psi}) &= \frac{(1-p_0)}{p_0\lambda}\sqrt{\frac{2\pi\sigma^2}{T-1}}\boldsymbol{\Phi}\left(\frac{\lambda\sum_{t=2}^{T}(\mathbf{X}_{it}-\mathbf{X}_{i(t-1)})'\begin{pmatrix}\cos(\theta_t)\\ \sin(\theta_t)\end{pmatrix}-\sigma^2}{\lambda\sqrt{\sigma^2(T-1)}}\right) \\ &\quad \cdot\exp\left\{\frac{\left(\lambda\sum_{t=2}^{T}(\mathbf{X}_{it}-\mathbf{X}_{i(t-1)})'\begin{pmatrix}\cos(\theta_t)\\ \sin(\theta_t)\end{pmatrix}-\sigma^2\right)^2}{2\lambda^2\sigma^2(T-1)}\right\}, \end{aligned} \tag{2.30}$$

and $\boldsymbol{\Phi}$ is the standard normal cumulative distribution function.

The expectation in (2.29) is taken with respect to the posterior $\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}|Y_{1:T}, \mu = 0)$, and hence we can use the posterior draws already obtained from Section 2.2.2 to utilize the following approximation:

$$\int_0^\infty \kappa(\nu)d\nu \approx \frac{1}{N}\sum_{\ell=1}^{N} h(\mathcal{X}_{1:T}^{(\ell)}, \boldsymbol{\psi}^{(\ell)}). \tag{2.31}$$

Combining (2.28) with (2.31) we are able to compute the posterior probability $\pi_0(\mu = 0|Y_{1:T})$.

The quantity in (2.30) that appears in both the normal cumulative distribution function and in the exponent is interesting in that $(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})'\begin{pmatrix}\cos(\theta_t)\\ \sin(\theta_t)\end{pmatrix}$ is the scalar projection of $(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})$ onto the unit vector whose direction is determined by $(\mathbf{X}_{jt} - \mathbf{X}_{i(t-1)})$ (see Figure 2.4). Intuition tells us that if these scalar projections are consistently large then node j is influencing the way node i moves through the social space; the posterior probabilities reflect this intuition in that large scalar projections lead to small values of $\pi_0(\mu = 0|Y_{1:T})$.

One last note of practical value is that for nodal influence to exist in a meaningful way, we must require that the influencing node has during at least one observation period brought the influenced node within his social circle. This becomes easy to evaluate by means of the social reaches by requiring that for nodal influence to exist between i and j, $\{t : d_{ijt} < r_i\}\bigcup\{t : d_{ijt} < r_j\} \neq \emptyset$.

2.5.2 Visualizing the Nodal Influence

Suppose there is evidence from the posterior that $\mu \neq 0$. Then, as mentioned earlier, $\boldsymbol{\epsilon}_{it} = \mathbf{X}_{it} - \mathbf{X}_{i(t-1)} \sim N(\boldsymbol{\mu}_t, \sigma^2 I_p)$. We can visualize and further interpret this influence by considering the polar

coordinates of $\boldsymbol{\epsilon}_{it}$, $d_{it} = \|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|$ and $\phi_{it} = \text{atan2}(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})$. The following lemma gives the distribution of (d_{it}, ϕ_{it}).

Lemma 2.5.2. *Let Z and W be independent random variables such that $Z \sim N(\mu_z, \sigma^2)$ and $W \sim N(\mu_w, \sigma^2)$, and let $d = \|(Z, W)\|$ and $\phi = atan2((Z, W))$ be the polar coordinates of (Z, W). Then*

$$d \sim \textit{Rice}\left(\|(\mu_z, \mu_w)\|, \sigma\right), \tag{2.32}$$

$$\phi|d \sim \textit{von Mises}\left(\frac{d\|(\mu_z, \mu_w)\|}{\sigma^2}, \textit{atan2}((\mu_z, \mu_w))\right). \tag{2.33}$$

Using this lemma, we see that the polar coordinates of $\mathcal{R}_t'\boldsymbol{\epsilon}_{it}$, (d_{it}, ϕ_{it}), follow

$$d_{it} \sim \text{Rice}\left(\mu, \sigma\right), \tag{2.34}$$

$$\phi_{it}\Big|d_{it} \sim \text{von Mises}\left(\frac{d_{it}\mu}{\sigma^2}, 0\right). \tag{2.35}$$

In other words, we can think of the transition from $\mathbf{X}_{i(t-1)}$ to $\mathbf{X}_{it}$ as a two step process, where the distance to move is determined first, and then the angle is chosen. To aid the visualization of the nodal influence we focus on the von Mises distribution that determines this angle. *Hence we are visualizing the extent of the nodal influence from j on i by looking at the propensity of node i to aim towards node j.* The circle around $\mathbf{X}_{i(t-1)}$ in Figure 2.4 represents such a von Mises distribution (with mean θ_t rather than 0), where dark values indicate high probability regions and light values indicate low probability regions. Note that if $\mu = 0$ then $\phi_{it} \sim \text{von Mises}(0, 0) \overset{\mathcal{D}}{=} \text{Unif}(-\pi, \pi)$. That is, any angle with respect to node j is as likely as any other angle and hence node i does not tend to angle towards node j more than any other direction in the latent social space.

We can use the posterior mean latent positions to estimate these von Mises distributions, thus obtaining a good visualization of the nodal influence. First get $\hat{\mu}$, the estimate of μ, by averaging over time the scalar projection of $(\widehat{\mathbf{X}}_{it} - \widehat{\mathbf{X}}_{i(t-1)})$ onto $(\widehat{\mathbf{X}}_{jt} - \widehat{\mathbf{X}}_{i(t-1)})$. Then we can further estimate the $T-1$ concentration parameters (the concentration parameter in (2.33) being $d\|(\mu_z, \mu_w)\|/\sigma^2$) by multiplying this $\hat{\mu}$ by $\|\widehat{\mathbf{X}}_{it} - \widehat{\mathbf{X}}_{i(t-1)}\|/\hat{\sigma}^2$. One can then plot these estimated von Mises distributions wrapped around the nodes being influenced, such as Figure 2.8. This type of plot may become overcrowded when there are multiple influencing nodes; in such a case, for each of, say, m influencing nodes, one could average these $T-1$ concentration parameters over time to obtain one concentration parameter for a summary von Mises distribution. These m von Mises distributions can then be plotted to get an overview of the various influences on the influenced node. An example of this type of plot can be seen in Figures 2.9 and 2.13.

2.6 Simulations

Twenty data sets were simulated, each with the number of nodes $n = 100$ and the number of time points $T = 10$. For each of the twenty simulations the following parameter values were set to be $\beta_{IN} = 1$, $\beta_{OUT} = 2$, and $\sigma^2 = 1/(5n)^2$. The initial latent positions $\mathcal{X}_1$ were drawn according to a mixture of normals with 5 components, equal mixture weights (1/5), and equal spherical covariances $\sigma^2 I_p$, where $p = 2$. The means of the clusters were drawn from a normal distribution with mean zero and covariance matrix $(2/n)^2 I_p$, hence some clustering existed at least at the initial time point. The radii were then drawn randomly from a Dirichlet distribution. The i^{th} parameter of this Dirichlet was computed by taking $n\|\mathbf{X}_{i1}\|^{-1}/\max_j\{\|\mathbf{X}_{j1}\|^{-1}\}$; the motivation for this was for the nodes with the larger social reach to also be the nodes in the center of the social space. For ten of the twenty simulations, the latent positions at subsequent time periods $\mathcal{X}_{2:T}$ were drawn according to (2.2). For the remaining ten, these subsequent latent positions were drawn in a slightly different manner to allow nodal influence to exist in the network. Twenty-five nodes were randomly selected to be influenced, each of which was accompanied by another randomly selected node to do the influencing, and we let $\mu = 0.02$. Then $\mathcal{X}_{2:T}$ were drawn by

$$\mathbf{X}_{it} \sim \begin{cases} N(\mathbf{X}_{i(t-1)}, \sigma^2 I_p) & \text{if node } i \text{ is not influenced,} \\ N\left(\mathbf{X}_{i(t-1)} + \mathcal{R}_t \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \sigma^2 I_p\right) & \text{if node } i \text{ is influenced.} \end{cases} \tag{2.36}$$

The prior of σ^2 was chosen to be $IG(9, 1.5)$ to keep the possibility of a large σ^2. The prior of τ^2 was $IG(2.05, 1.05\sum_{i=1}^{n}\|\mathbf{X}_{i1}^{(1)}\|^2/(np))$, following the suggestions in Section 3.1. The priors of the coefficients β_{IN} and β_{OUT} were normals centered at the initial estimates and variances equal to 100. The prior of μ was an exponential distribution with mean 0.03. When using a normal random walk proposal for the latent space positions and the β coefficients, the variance components were 0.0075 and 0.1 respectively. The tuning parameter κ used in the Dirichlet proposal for the radii was 175,000. For each simulation, the chain length was 50,000, and we then removed from the chain a burn-in of 15,000 samples.

The results from the simulations were compared in several ways: First, for each simulation the posterior means of β_{IN} and β_{OUT} were computed as well as the correlation between the posterior means of $r_{1:n}$ and the truth for each of the 20 simulations. The mean (sd) over all 20 simulations of $\widehat{\beta}_{IN}$ was 0.9172 (0.06207) and for $\widehat{\beta}_{OUT}$ was 2.045 (0.1438). The mean (sd) correlation between $\widehat{r}_{1:n}$

and $r_{1:n}^{(true)}$ was 0.9298 (0.06402). We see then that the posterior means did quite well at estimating the true values of β_{IN} (1), β_{OUT} (2) and the radii.

Second, the area under the ROC curve (AUC) was computed. This was accomplished by plugging in the posterior means of the model parameters and the latent positions into the observation equations (2.3) and (2.4), and then comparing these with the simulated data $Y_{1:T}$. Hence this can be considered as a measure of how well the model fits the data. For each simulation, the directed graphs were also converted to undirected graphs by letting $y_{ijt} = \max\{y_{ijt}, y_{jit}\}$ in order that we might apply the method of Sarkar and Moore (2005). The AUC values for both undirected and directed networks were then computed using the estimates from Sarkar and Moore's method. We similarly computed the AUC values for the directed network using our method, and, again using those same estimates, computed the AUC values for the undirected network by using $\mathbb{P}(\{y_{ijt} = 1\} \cup \{y_{jit} = 1\})$. These results are given in Figure 2.5a, where asterisks indicate the directed networks and triangles indicate undirected. We see that all our values are extremely high, implying that the model fits the data quite well, and also we see that our method uniformly outperformed that of Sarkar and Moore on both the directed and undirected networks.

Third, the pairwise distances from the estimated latent positions were compared to the pairwise distances from the true latent positions. That is, for each triple (i, j, t) we can look at $\|\widehat{\mathcal{X}}_{it} - \widehat{\mathcal{X}}_{jt}\|/\|\mathcal{X}_{it} - \mathcal{X}_{jt}\|$, giving us $Tn(n-1)/2$ such ratios for each simulation. Figure 2.5b gives, for each of the twenty simulations, a smoothed curve of the distribution of these ratios. Notice that all these distributions are narrow and centered around 1, implying that the latent positions from the posterior means are close to the truth.

Fourth, the capability of predicting Y_{T+1} was evaluated. To this end one could simulate a new $\mathcal{X}_{T+1}$ from (2.2) and subsequently simulate a new Y_{T+1} from (2.3) and (2.4), and then compare the predicted edges with these simulated edges. This would, however, introduce unnecessary variation in Y_{T+1}, possibly making good predictions seem bad, or vice versa. Hence we used $\mathbb{P}(Y_{T+1}|\mathcal{X}_{T+1} = \mathbb{E}(\mathcal{X}_{T+1}|\mathcal{X}_T, \boldsymbol{\psi}), \boldsymbol{\psi})$ as the "true" probabilities at time $T+1$. We first wanted to make sure that the more sophisticated prediction method worked better than plugging the posterior means into the observation equations. Hence we compared the sum of squared differences between the true probabilities and the predicted probabilities given by (2.23) to the sum of squared differences between the true probabilities and the probabilities obtained by plugging in the posterior means into (2.3) and (2.4). Specifically, we looked at

$$\begin{aligned}&\textstyle\sum_{i\neq j}\left[\mathbb{P}(y_{ij(T+1)} = 1|\mathcal{X}_{T+1} = \mathbb{E}(\mathcal{X}_{T+1}|\mathcal{X}_T, \boldsymbol{\psi}), \boldsymbol{\psi}) - \mathbb{P}(y_{ij(T+1)} = 1|\mathcal{X}_{T+1} = \widehat{\mathcal{X}}_{T+1}, \widehat{\boldsymbol{\psi}})\right]^2 \\ &\textstyle-\sum_{i\neq j}\left[\mathbb{P}(y_{ij(T+1)} = 1|\mathcal{X}_{T+1} = \mathbb{E}(\mathcal{X}_{T+1}|\mathcal{X}_T, \boldsymbol{\psi}), \boldsymbol{\psi}) - \mathbb{P}(y_{ij(T+1)} = 1|Y_{1:T}, \mathcal{X}_{T+1} = \widehat{\mathcal{X}}_{T+1})\right]^2.\end{aligned} \tag{2.37}$$

In all but one simulation this turned out to be positive, with a mean (sd) over the 20 simulations

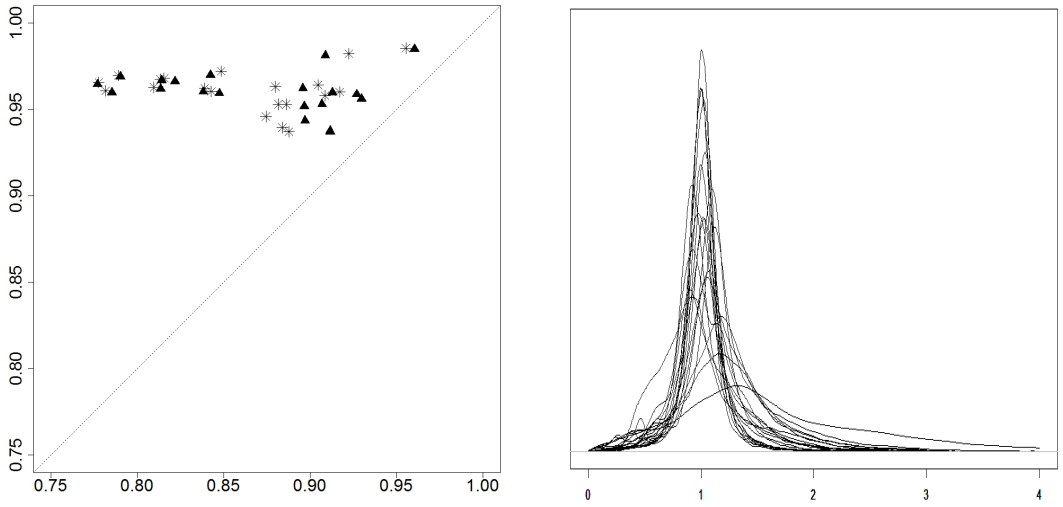
of 0.9725 (0.8679). This implies that better prediction estimates can be obtained when we do not condition on the model parameters. We then let $y_{ij(T+1)} = 1$ if the true probability was greater than 1/2. Using the new Y_{T+1} we computed the AUC for the predicted probabilities, leading to a mean (sd) over the simulations of 0.9546 (0.04003). Finally, making hard predictions of 1 if the predicted probability was greater than 1/2 we computed the sensitivity and specificity. We also compared these values to simply using Y_T to make hard predictions. The results are given in Figures 2.5c and 2.5d, where the vertical axis corresponds to our predictions and the horizontal axis corresponds to Y_T. Clearly by using our prediction method we obtain much better sensitivity, and though the specificity is much more comparable, our method still does better in all but three of the 20 simulations.

For those ten cases where nodal influence was part of the simulation, we computed the sensitivity of detecting nodal influence on those nodes which were in truth influenced, and in all 20 simulations we computed the specificity of not detecting influence on those nodes which were in truth not influenced. The mean (sd) sensitivity and specificity for the ten simulations with nodal influence were 0.952 (0.0316) and 0.832 (0.129). The mean (sd) specificity for the ten simulations without nodal influence was 0.868 (0.116). We see from this that the Bayesian estimation does a very good job at detecting nodal influence without giving many false positives when no such influence exists.

Sensible priors have been outlined in Section 2.2.1 for all model parameters except σ^2. It is important then to determine the sensitivity of the MCMC algorithm to the values of θ_σ and ϕ_σ. To this end we reran the above simulations where the shape and scale parameters of $\pi(\sigma^2)$ were drawn from a uniform distribution ranging from 3 to 15 for the shape parameter and from 0.01 to 2 for the scale parameter. The AUC for rerunning these 20 simulations in this fashion yielded very high AUC values, ranging from 0.9407 to 0.9858, averaging 0.9621. Thus it appears that the estimation is quite robust to the hyperparameters for the prior of σ^2.

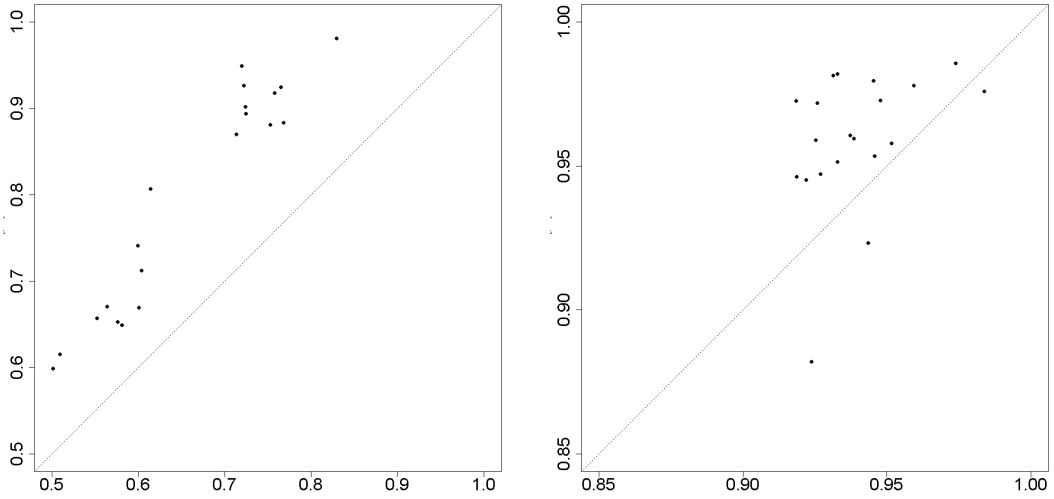
In addition to the simulations described above, five larger data sets were simulated where $n = 500$ and $T = 10$. Estimation was performed both using and not using the approximations outlined in Section 2.2.3, letting $n_0 = 100$, and the AUC was computed to evaluate model fit. Simulations were analyzed on a UNIX machine with a 2.40 GHz processor. The mean (sd) time to perform the MCMC analysis with 50,000 iterations using the approximation was, in minutes, 716 (24), and to perform the MCMC with 50,000 iterations not using the approximation was, in minutes, 2281 (20). Hence by using the approximations there was a mean (sd) decrease in computational time of 68.6% (0.835). The mean (sd) AUC using the approximation was 0.9618 (0.0048), and not using the approximation was 0.9679 (0.0109). This was a mean (sd) decrease in AUC of 0.00618 (0.0120).

Thus by using the approximations of Section 2.2.3 there is a drastic decrease in computational time with very little loss in model fit.



(a) AUC

(b) Distance



(c) Prediction Sensitivity

(d) Prediction Specificity

Figure 2.5: Results for 20 simulations. (a) AUC using Sarkar and Moore's method (horizontal axis) and our method (vertical axis) on both undirected (triangles) and directed (asterisks) networks; (b) Distribution of pairwise distance ratios, comparing estimated latent positions with true latent positions; (c)-(d) Sensitivity and Specificity for predicting edges at time $T + 1$, using Y_T to make predictions (horizontal axis) and our method of prediction (vertical axis).

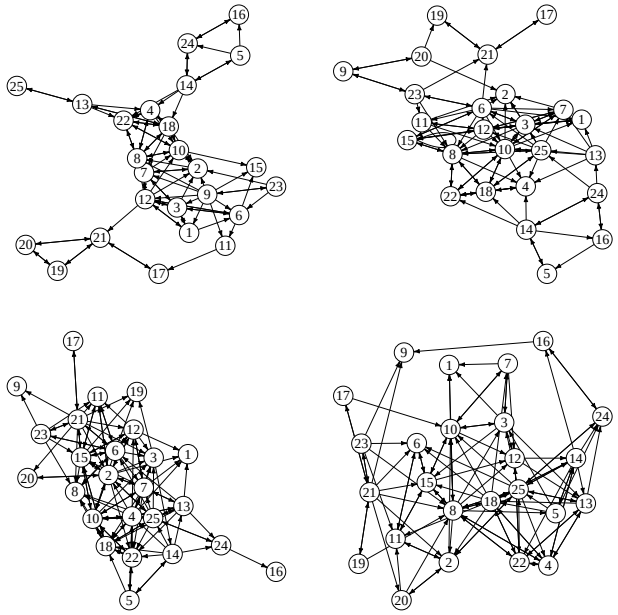


Figure 2.6: Graphs of Dutch classroom data at, from left to right and top to bottom, times 1, 2, 3, and 4.

2.7 Real Data Analyses

2.7.1 Dutch Classroom Data

Knecht (2008) conducted a longitudinal study in which students aged 11 to 13 years in a Dutch class were surveyed over four time points, yielding four asymmetric adjacency matrices where the $(i, j)^{th}$ entry denotes whether student i claims student j as a friend. Figure 2.6 shows the graphs from these adjacency matrices. Demographic and behavioral data were also collected on these individuals. Twenty six students were recorded, although one student left the class before the study was completed; this student was left out of the analysis. Missing edges exist in the data due to some students not being present during a survey. This was dealt with as previously described in Section 2.3.

A burn-in of 15,000 iterations was removed, leaving a chain of length 85,000. We compared our method with that found in Sarkar and Moore (2005) by AUC values. Our method yielded an AUC value of 0.917 vs. 0.8456 from Sarkar and Moore's method.

The posterior means of β_{IN} and β_{OUT} were 1.29 and 1.00 respectively, implying that popularity was more important in edge formation. The posterior means of the latent locations are given in Figure 2.7. Some interesting features can be noticed by comparing the latent positions with

demographic information. Figure 2.7 differentiates the nodes' gender by males as dotted lines and females as solid lines. This plot corroborates results shown by Snijders et al. (2010) in that the friendships between two nodes of the same gender are more prevalent. Also, only two of the students are of non-Dutch ethnicity, circled in Figure 2.7, and these two nodes are very close together in the social space. One last interesting feature seen in Figure 2.7 regards an interesting link between academic capability and social behavior. Each student was assessed and ranked from 4 (lowest) to 8 (highest) based on academic capabilities at the end of primary school. There was only one student (node 9) who was ranked a 4 and one (node 25) who was ranked an 8. The social behavior of these two individuals, as seen via the latent space positions, are complete opposites. The student with the highest ranked capability moves from outside the social network to the center of the social space, while the student with the lowest ranked capabilities moves directly away from the center of the social space. The reason node 25 started outside of the network may be explained in part by the fact that he had only gone to primary school with one other student and hence did not start the school term knowing the others.

Nodal influence was detected in four of the nodes. The posterior mean of the latent positions of these nodes and those doing the influencing are given in Figure 2.8, where each plot shows the influenced node in a circle and the influencing nodes in triangles. A wrapped von Mises distribution is plotted around the influenced node, visually indicating the strength of the influence and in which direction. In cases where there were multiple influencing nodes, that which had the strongest influence was used to depict the von Mises distribution. From these plots we see how certain nodes were drawn towards certain other nodes, and the strength of the influence. This yields much more detailed information on the local scale of how the network evolved over time. Of note is the fourth such influenced node (node 25), who was unique in that he was influenced by many of the other nodes (9 others). As mentioned earlier, node 25 began by only knowing one other node (determined by having or not having gone to the same primary school as the others), and so it is intuitive that he would be more susceptible to being pulled into others' existing social circles, bringing him to the center of the social space. The von Mises distributions in Figure 2.8 match this intuition in that the nodal influence wanes as time progresses and node 25 becomes a part of his own social circle. Figure 2.9 shows the von Mises distributions for each of the influencing nodes averaged over time. From this figure we better see how strongly each node is affecting the direction in which node 25 aims within the social space. Lastly, the results from the analysis of nodal influence fell in line with the overall gender and ethnic separation, in that the first three instances of nodal influence all occurred

within the ethnic Dutch girls, and of the nine nodes influencing node 25 (a male) only one was of the opposite gender.

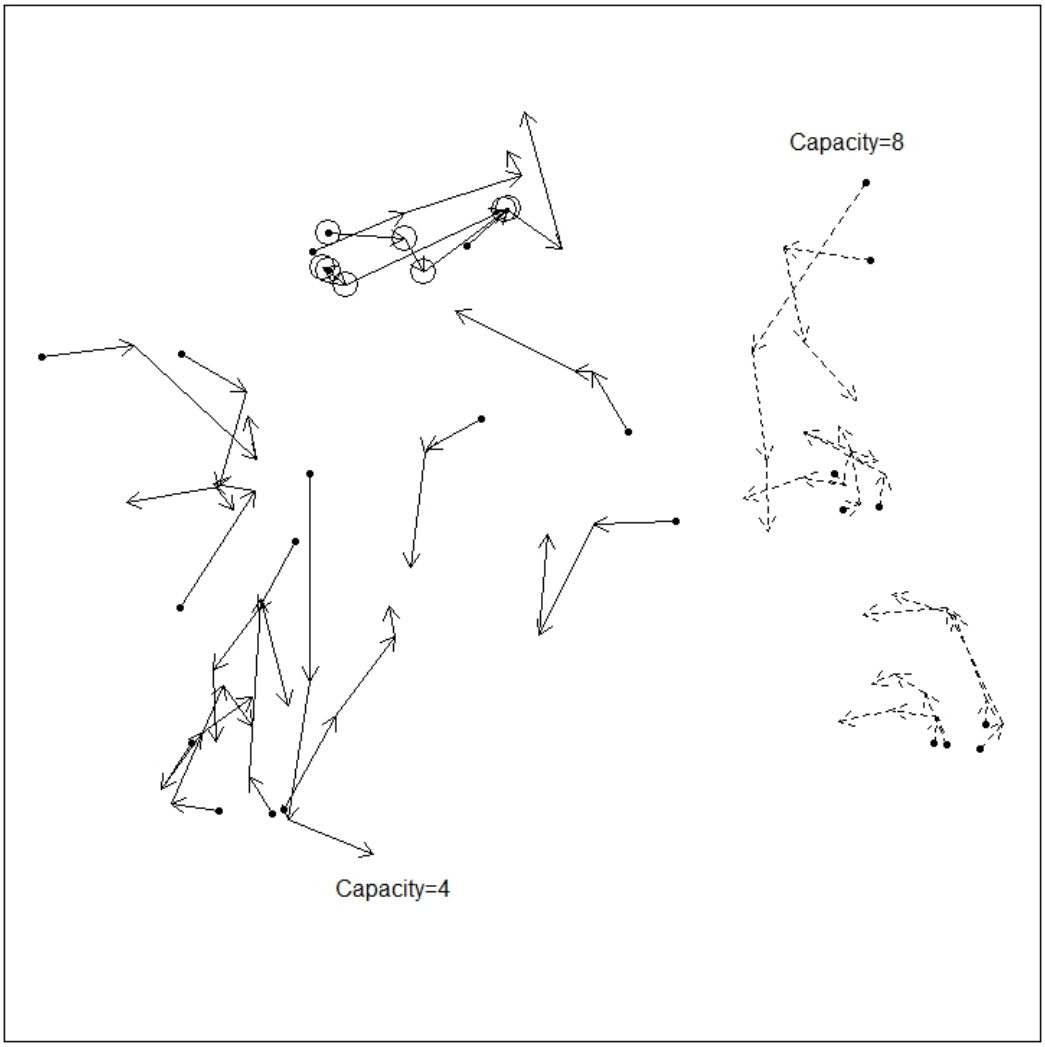
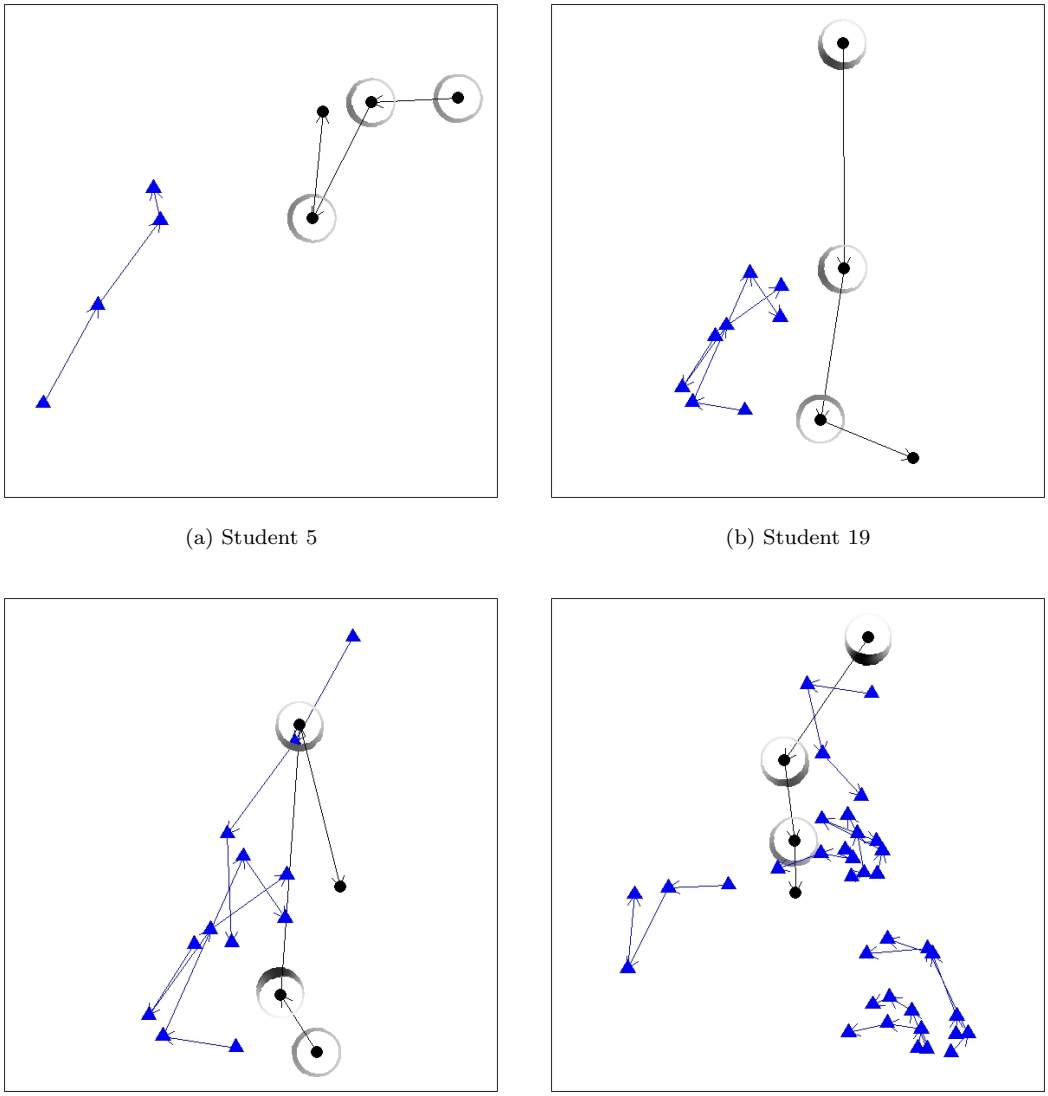


Figure 2.7: Posterior means of latent nodal positions for the Dutch classroom data, arrows indicating the temporal direction of the trajectories. Males' trajectories are in dotted lines, females' in solid lines. Also, two of the students are of non-Dutch ethnicity, and these two nodes' latent positions are circled. The two students with the lowest (4) and highest (8) ranked academic capabilities, nodes 9 and 25 respectively, are also marked as such.

2.7.2 Cosponsorship Data

We analyzed data collected by James Fowler on bill cosponsorship of Congressmen in the U.S. House of Representatives for the 97^{th} to 101^{st} Congresses (see Fowler (2006a) and Fowler (2006b)). There



(a) Student 5

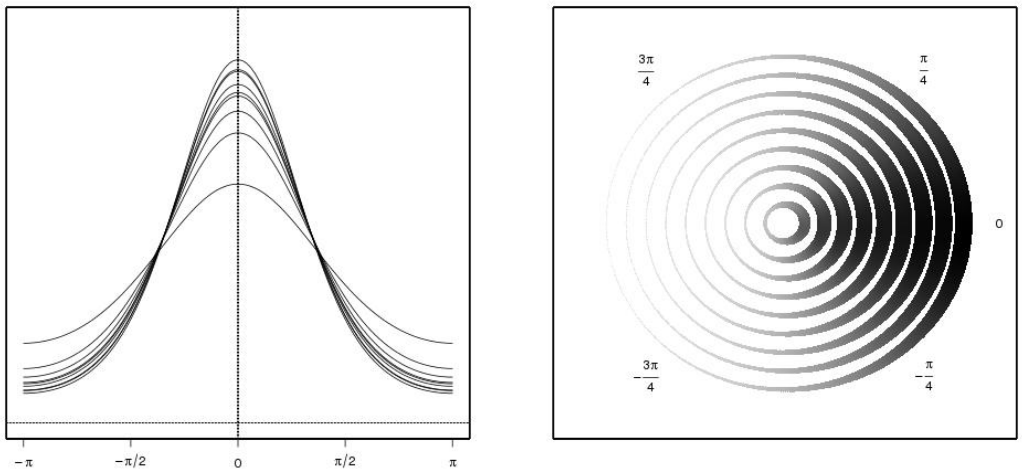
(b) Student 19

(c) Student 20

(d) Student 25

Figure 2.8: Corresponding to the Dutch classroom data, each plot is zooming in on the posterior means of the latent positions of the influenced nodes (circles) and those nodes doing the influencing (triangles). The circles around the influenced nodes are the von Mises distributions from (2.35) that help to visualize the influence being exerted in terms of the direction the influenced node moves (corresponding to the node exerting the strongest influence in cases where multiple nodes are exerting influence). Wider and darker areas on the rings indicate higher probability regions.

were a total of 644 members of Congress (MC's) who served during these five terms. However, at each time point around 30% of MC's were not represented (the actual values ranged from 30.1% to 31.1%). These large proportions of unrepresented MC's leads to even larger proportions of missing edges (from 51.2% to 52.5% missing edges). The data were analyzed by letting $y_{ijt} = 1$ if node j sponsored a bill and node i cosponsored it, hence showing support for node j. Figure 2.10 shows

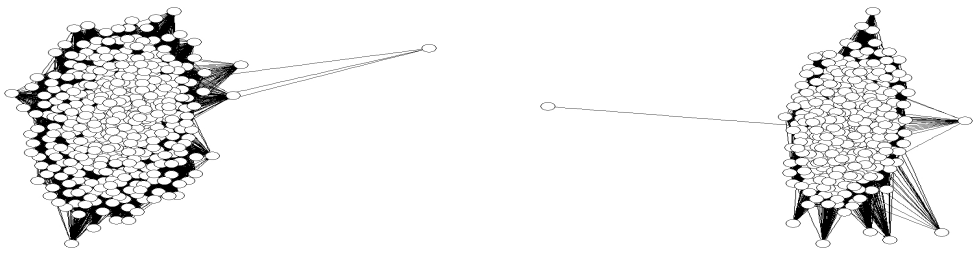


(a) (b)

Figure 2.9: Corresponding to the Dutch classroom data, these von Mises distributions, obtained by averaging the concentration parameters of (31), illustrate the extent of nodal influence on node 25. Each curve in (a) is a von Mises distribution corresponding to an influencing node, demonstrating the propensity for node 25 to aim towards the influencing node. The rings in (b) give the same distributions but wrapped from 0 to 2π, where wider and darker areas on the rings indicate higher probability regions. The radii and order of the rings are for visual appeal only.

the graphs from these adjacency matrices at times 1 (97^{th} Congress) and 5 (101^{st} Congress) sans missing data.

Due to the large amount of missing data, some care was needed in initializing the missing edges in the Markov chain. We modified the preferential attachment method of imputation described by Huisman and Steglich (2008) by doing the following. We first form an aggregated adjacency matrix Y whose entries y_{ij} are set to one if for any t there is a link from i to j and set to zero otherwise. Next, for each missing node i (at a particular time point t), we assign the probability of a link from



(a) (b)

Figure 2.10: Graphs of cosponsorship data at times 1 (left) and 5 (right).

i to j to be proportional to the indegree (averaged over t) of node j and inversely proportional to the shortest path length between i and j in Y. That is, letting k_j denote the average indegree of node j and n_{ij} denote the shortest path length between i and j in Y, the probability of a link from i to j is set to be

$$\mathbb{P}(y_{ijt} = 1) = \frac{k_j/n_{ij}}{\sum_{\ell \neq i} k_\ell / n_{i\ell}}. \tag{2.38}$$

We then look at the average outdegree of i (rounded to the nearest integer), denoted d_i, and randomly draw d_i nodes from $\{1, \ldots, n\} \setminus \{i\}$ according to (2.38); the corresponding y_{ijt}'s are set to be 1. In this way we obtain initial values for the missing data.

A burn-in of 250,000 iterations was removed, leaving a chain of length 1,250,000. Thinning was done by recording only every tenth iteration. Using the posterior means to make predictions on $Y_{1:T}$ led to an AUC value of 0.787 vs. 0.7148 from Sarkar and Moore's method; when applying Sarkar and Moore's method we used $Y_{1:T}$ constructed from the observed edges and imputed edges. From these AUC values we see that our model fits the data quite well and again outperforms the existing method.

Many of the congressmen (328 MC's) analyzed were reelected into the 102^{nd} Congress, and so it was possible to compare predictions with the truth. In addition to the predictions obtained through the methods described in Section 2.4, we also considered prediction by using Y_T to predict Y_{T+1} as well as using $\sum_{t=1}^{T} y_{ijt}/T$ to estimate $\mathbb{P}(y_{ij(T+1)} = 1)$. For both averaging $Y_{1:T}$ and applying our method, hard predictions were made by letting $\widehat{y}_{ij(T+1)} = 1$ if $\widehat{\mathbb{P}}(y_{ij(T+1)} = 1) > 0.5$ and 0 otherwise. Table 2.1 gives the results. From this we see that while using a more naïve prediction method yields higher specificity, it does so at the expense of correctly detecting the future edges. Our method can better find the future edges, and also gives the lowest mean squared error (MSE).

The posterior means of the coefficients, $\beta_{IN} = 0.974, \beta_{OUT} = 0.147$, indicate that popularity was dramatically more responsible for creating edges than activity level; i.e., the probability of a cosponsorship is determined mostly by the MC who sponsors the bill rather than the MC who is contemplating cosponsoring it. Figure 2.11 shows the posterior mean latent positions of the MC's. Unsurprisingly we see the Republicans (hollow circles) and Democrats (solid circles) occupy different

Method	Specificity	Sensitivity	MSE
Averaging $Y_{1:T}$	0.8014	0.5125	0.2492
$\mathbb{P}(Y_{T+1} \mid Y_{1:T}, \widehat{\mathcal{X}}_{T+1})$	0.5962	0.6984	0.2151

Table 2.1: Prediction results for 328 MC's in our analysis who also served in the 102^{nd} Congress.

halves of the network space. Both parties seem to have the majority of their members in the center of the network space along with a scattering of members along the edge of the network space, implying that both parties have active central members which associate with members from both parties, as well as less active outlying members which interact less with members of the opposite party.

It is of interest to study the dynamics of the network. To evaluate the stability of the network we consider the distance each MC moves during each of the four transitions. Figure 2.12 gives a boxplot for these distances, and from this we can see that the distances corresponding to each transition fall within a similar range, though the transition to the 99^{th} Congress involves somewhat larger moves. There were a few MC's (ranging from 4.4% to 7.7% of the MC's) who were above the top whisker, but these typically were different MC's every transition; only 11 of the MC's were beyond the top whisker in two of the transitions, 2 of the MC's in three of the transitions, and none more than three. All this indicates that the dynamics of the network remained stable throughout the five terms.

Political ideology, measured from liberal to conservative, is an extremely important aspect of political science. Much literature exists on this topic; for example, Poole and Rosenthal (2011) wrote an entire book on ideology and its effect on Congress. Levitt (1996) discussed various factors' effects on roll-call voting patterns, concluding that personal ideology is the single most important factor. This relationship between voting patterns and personal ideology is seen in a vivid way by comparing the latent positions of the MC's with their ideologies. Specifically, this comparison is shown visually in Figure 2.11, where the latent positions of the MC's are superimposed upon a surface which represents a political ideological landscape. This surface was obtained in the following way. Each MC has a particular Nominate score which is a measure of their political ideology (see Poole and Rosenthal, 1985). This score ranges from -1 (liberal) to 1 (conservative). Using the latent location coordinates, kriging was performed on the absolute value of the Nominate scores using a spherical variogram model. This gives us the surface in Figure 2.11 that reflects how regions of the latent network space correspond to radical ideologies (high values) or moderate ideologies (low values). In the center of the network space where the nodes are most dense (and hence are more active in legislation) is an interesting dividing line between the two parties that reflects a moderate political ideology. We also see that both parties have a less dense (hence less active in legislation) group of MC's which has more radical ideologies.

Nodal influence was detected in 74 of the MC's. It is intuitive that an MC would be influenced more by members of his or her own party than by members of a different party, and indeed this is the case. Of the influenced MC's, only 29% were influenced more by members of the opposite

party than by members of their own party. As an example of an MC influenced by members of the opposite party, consider Lawrence Coughlin, a Republican from Pennsylvania. Only 35% of those who exerted influence on Coughlin were also Republicans, and in fact the average Nominate score for those exerting influence on Coughlin was -0.073, i.e., Coughlin was influenced mostly by slightly liberal politicians. This influence is manifest in the fact that he is often referred to as a moderate Republican (e.g., Downey, 2001); his moderate ideology (0.163) is also quantitatively reflected in having his Nominate score below the first quartile of fellow Republicans, and below the first quartile of the absolute value of the Nominate scores of all MC's. In contrast to Coughlin, consider Sidney Yates, a Democrat from Illinois. 94% of those exerting influence on Yates were also Democrats, and in fact quite liberal Democrats; the mean ideology score of Yates' influencing MC's was -0.301 (recall that a negative Nominate score implies liberal ideology). The influence of these liberal MC's on Yates is reflected in Yates also being liberal, himself having a Nominate score (-0.477) below the first quartile of all Democrats and an absolute score above the third quartile of the absolute values of all Nominate scores. What is left uncertain is whether Yates aimed towards liberal Democrats in the latent space because he himself already had a liberal ideology or whether these liberal Democrats influenced him to become liberal himself. Figure 2.13 gives plots visualizing the nodal influence exerted on Coughlin and on Yates, as described in Section 2.5.2. From these figures we can see the range of the strength of the influences on the MC's by party. We see from this that for both Coughlin and Yates those from their own party (Republican for Coughlin and Democrat for Yates) typically exert stronger influence than those from the opposite party. We also see that, in general, the influence exerted on Coughlin is stronger than that on Yates.

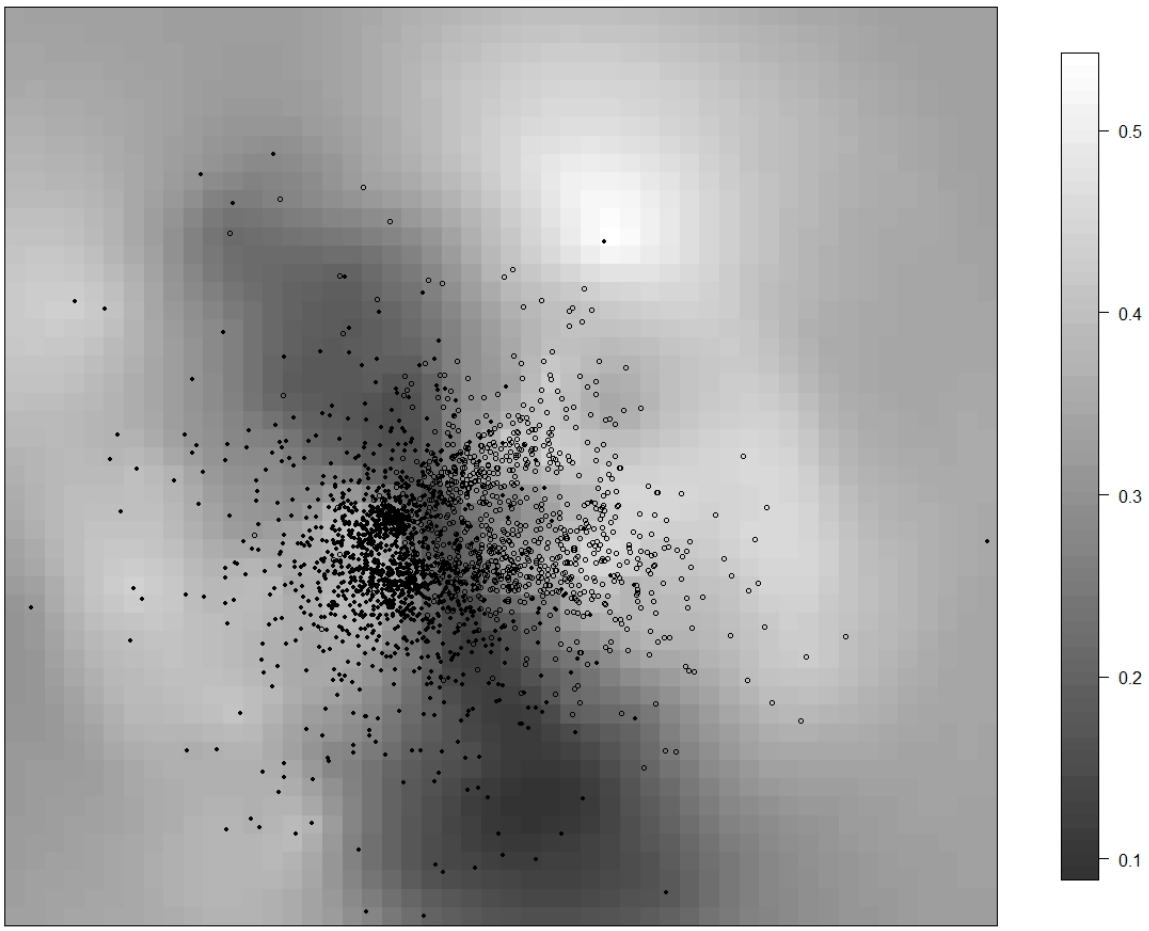


Figure 2.11: Posterior means of latent positions for the cosponsorship data. Latent positions for all 5 Congresses are plotted simultaneously. Hollow circles are republicans and solid circles are democrats. The surface these points lie upon reflects the political ideological landscape within the network space. Darker regions correspond to more moderate ideologies, and lighter regions correspond to more radical ideologies.

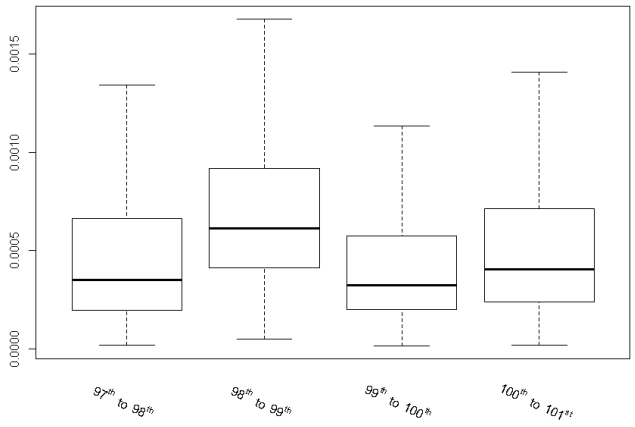


Figure 2.12: Boxplots of the distances MC's traveled within the latent network space during each of the four transitions. The similar ranges imply that the dynamics of the network are fairly constant throughout the five terms.

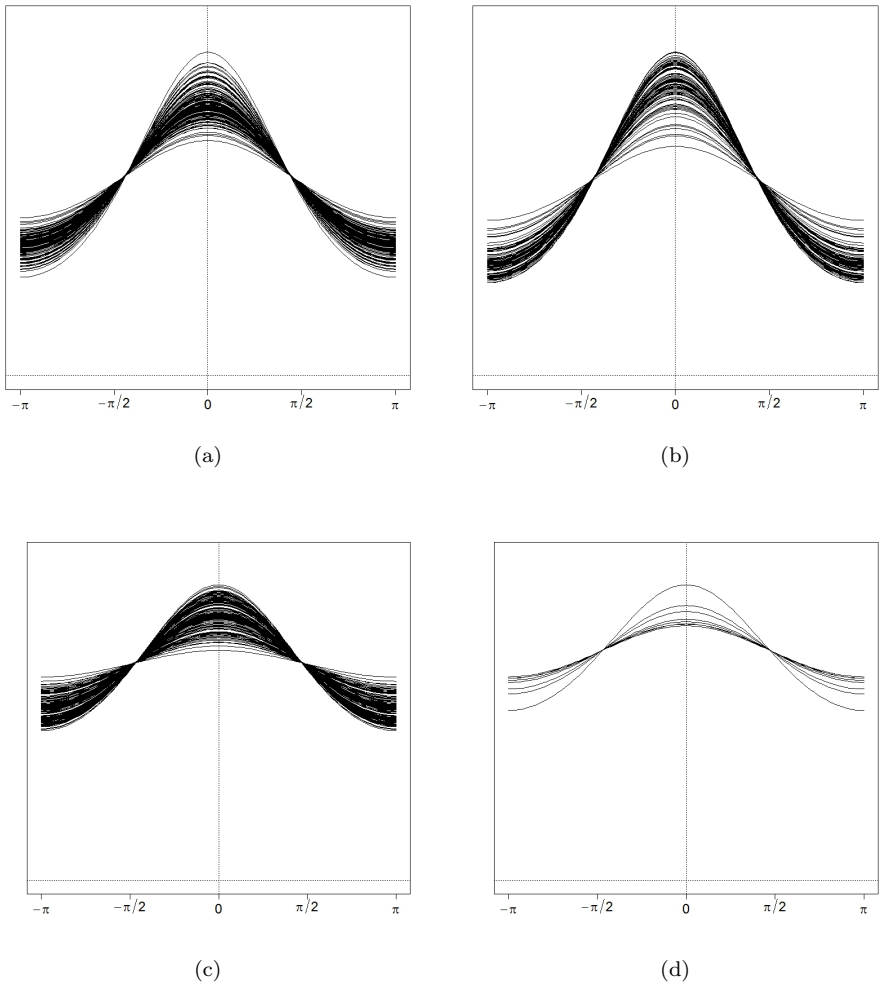


Figure 2.13: Corresponding to the cosponsorship data, these von Mises distributions illustrate the extent of nodal influence on Lawrence Coughlin (top) and Sidney Yates (bottom) from Democrats (left) and Republicans (right). Each curve in each figure is a von Mises distribution corresponds to an influencing MC, demonstrating the propensity for Coughlin or Yates to move towards the influencing MC. Narrow peaked curves correspond to strong influence, flat curves to weak influence.

2.8 Proof of Lemma 2.5.1

Let ν be some positive value. Then

$$\begin{aligned}
& \pi_+(\mu = \nu | Y_{1:T}) \\
= \; & \pi(\mu = \nu | Y_{1:T}) \\
= \; & \int \pi(\mu = \nu, \mathcal{X}_{1:T}, \boldsymbol{\psi} | Y_{1:T}) d\mathcal{X}_{1:T} d\boldsymbol{\psi} \\
= \; & \int \frac{\pi(Y_{1:T} | \mathcal{X}_{1:T}, \boldsymbol{\psi}) \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = \nu)}{\pi(Y_{1:T})} d\mathcal{X}_{1:T} d\boldsymbol{\psi} \\
= \; & \int \frac{\pi(Y_{1:T} | \mathcal{X}_{1:T}, \boldsymbol{\psi}) \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = 0)}{\pi(Y_{1:T})} \frac{\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = \nu)}{\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = 0)} d\mathcal{X}_{1:T} d\boldsymbol{\psi} \\
= \; & \int \frac{\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = \nu)}{\pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = 0)} \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, \mu = 0 | Y_{1:T}) d\mathcal{X}_{1:T} d\boldsymbol{\psi} \\
= \; & \int \frac{\pi(\mathcal{X}_{1:T} | \boldsymbol{\psi}, \mu = \nu) \pi(\boldsymbol{\psi}) \pi(\mu = \nu)}{\pi(\mathcal{X}_{1:T} | \boldsymbol{\psi}, \mu = 0) \pi(\boldsymbol{\psi}) \pi(\mu = 0)} \pi_0(\mu = 0 | Y_{1:T}) \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, | \mu = 0, Y_{1:T}) d\mathcal{X}_{1:T} d\boldsymbol{\psi}. \qquad (2.39)
\end{aligned}$$

Before proceeding, we need to establish that the Markov property still holds for the latent positions. This is seen in that

$$\begin{aligned}
& \pi(\mathcal{X}_t | \mathcal{X}_{1:(t-1)}, \boldsymbol{\psi}, \mu) \\
= \; & \pi(\mathbf{X}_{1t}, \ldots, \mathbf{X}_{(i-1)t}, \mathbf{X}_{it}, \mathbf{X}_{(i+1)t}, \ldots, \mathbf{X}_{nt} | \mathcal{X}_{1:(t-1)}, \boldsymbol{\psi}, \mu) \\
= \; & \pi(\mathbf{X}_{it} | \mathbf{X}_{1t}, \ldots, \mathbf{X}_{(i-1)t}, \mathbf{X}_{(i+1)t}, \ldots, \mathbf{X}_{nt}, \mathcal{X}_{1:(t-1)}, \boldsymbol{\psi}, \mu) \\
& \quad \cdot \pi(\mathbf{X}_{1t}, \ldots, \mathbf{X}_{(i-1)t}, \mathbf{X}_{(i+1)t}, \ldots, \mathbf{X}_{nt} | \mathcal{X}_{1:(t-1)}, \boldsymbol{\psi}, \mu) \\
= \; & \pi(\mathbf{X}_{it} | \mathbf{X}_{jt}, \mathcal{X}_{t-1}, \boldsymbol{\psi}, \mu) \pi(\mathbf{X}_{1t}, \ldots, \mathbf{X}_{(i-1)t}, \mathbf{X}_{(i+1)t}, \ldots, \mathbf{X}_{nt} | \mathcal{X}_{t-1}, \boldsymbol{\psi}, \mu) \\
= \; & N\left(\mathbf{X}_{it} \middle| \mathbf{X}_{i(t-1)} + \mathcal{R}_t \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \sigma^2 I_p \right) \cdot \left[\prod_{k \neq i} N(\mathbf{X}_{kt} \middle| \mathbf{X}_{k(t-1)}, \sigma^2 I_p) \right] \\
= \; & \pi(\mathcal{X}_t | \mathcal{X}_{t-1}, \boldsymbol{\psi}, \mu) \qquad (2.40)
\end{aligned}$$

for $t = 2, 3, \ldots, T$. Thus we have the following closed form expression

$$\begin{aligned}
&\frac{\pi(\mathcal{X}_{1:T}|\boldsymbol{\psi}, \mu = \nu)\pi(\mu = \nu)}{\pi(\mathcal{X}_{1:T}|\boldsymbol{\psi}, \mu = 0)\pi(\mu = 0)} \\
= \quad & \frac{1-p_0}{\lambda p_0} \exp\left\{ -\frac{\nu}{\lambda} - \frac{1}{2\sigma^2} \sum_{t=2}^{T} \left(\left\| \mathbf{X}_{it} - \mathbf{X}_{i(t-1)} - \mathcal{R}_t \begin{pmatrix} \nu \\ 0 \end{pmatrix} \right\|^2 - \|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|^2 \right) \right\} \\
= \quad & \frac{1-p_0}{\lambda p_0} \exp\left\{ -\frac{\nu}{\lambda} - \frac{1}{2\sigma^2} \left((T-1)\nu^2 - 2\nu \sum_{t=2}^{T} (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} \right) \right\} \\
= \quad & \frac{1-p_0}{\lambda p_0} \exp\left\{ -\frac{T-1}{2\sigma^2} \left(\nu - \frac{\lambda \sum_{t=2}^{T} (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} - \sigma^2}{(T-1)\lambda} \right)^2 \right. \\
& \qquad \left. + \frac{\left(\lambda \sum_{t=2}^{T} (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} - \sigma^2 \right)^2}{2\lambda^2\sigma^2(T-1)} \right\}.
\end{aligned}$$

Now by dividing both sides of (2.39) by $\pi_0(\mu = 0|Y_{1:T})$ and integrating over ν we obtain

$$\begin{aligned}
&\int_0^\infty \kappa(\nu) d\nu \\
&= \int_0^\infty \int \frac{\pi(\mathcal{X}_{1:T}|\boldsymbol{\psi}, \mu = \nu)\pi(\mu = \nu)}{\pi(\mathcal{X}_{1:T}|\boldsymbol{\psi}, \mu = 0)\pi(\mu = 0)} \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, |\mu = 0, Y_{1:T}) d\mathcal{X}_{1:T} d\boldsymbol{\psi} d\mu_+ \\
&= \int h(\mathcal{X}_{1:T}, \boldsymbol{\psi}) \pi(\mathcal{X}_{1:T}, \boldsymbol{\psi}, |\mu = 0, Y_{1:T}) d\mathcal{X}_{1:T} d\boldsymbol{\psi} \\
&= \mathbb{E}(h(\mathcal{X}_{1:T}, \boldsymbol{\psi}) \big| \mu = 0, Y_{1:T}),
\end{aligned}$$

where

$$\begin{aligned}
h(\mathcal{X}_{1:T}, \boldsymbol{\psi}) \quad = \quad & \frac{(1-p_0)}{p_0\lambda} \sqrt{\frac{2\pi\sigma^2}{T-1}} \boldsymbol{\Phi} \left(\frac{\lambda \sum_{t=2}^{T} (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} - \sigma^2}{\lambda \sqrt{\sigma^2(T-1)}} \right) \\
& \cdot \exp\left\{ \frac{\left(\lambda \sum_{t=2}^{T} (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \begin{pmatrix} \cos(\theta_t) \\ \sin(\theta_t) \end{pmatrix} - \sigma^2 \right)^2}{2\lambda^2\sigma^2(T-1)} \right\}
\end{aligned}$$

and $\boldsymbol{\Phi}$ is the standard normal cumulative distribution function.

2.9 Proof of Lemma 2.5.2

If a random variable X follows a Rice distribution with distance parameter ν and scale parameter σ, then the probability density of X is

$$f(X|\nu,\sigma) = I_0\left(\frac{X\nu}{\sigma^2}\right)\frac{X}{\sigma^2}\exp\left(-\frac{X^2+\nu^2}{2\sigma^2}\right), \tag{2.41}$$

where I_0 is the modified Bessel function of order 0. If X follows a von Mises distribution with mean μ and concentration parameter κ, then the probability density of X is

$$f(X|\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)}\exp\left(\kappa\cos(X-\mu)\right). \tag{2.42}$$

By assumption $(Z,W) \sim N\left((\mu_z,\mu_w),\sigma^2 I_2\right)$, so $Z = d\cos(\phi)$ and $W = d\sin(\phi)$. Hence

$$\begin{aligned}
f_{d,\phi}(d,\phi) &= f_{x,y}(d\cos(\phi), d\sin(\phi))|J| && (2.43)\\
&= \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(d\cos(\phi)-\mu_z)^2+(d\sin(\phi)-\mu_w)^2}{2\sigma^2}\right\}\cdot d && (2.44)\\
&= \frac{d}{2\pi\sigma^2}\exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}\right.\\
&\qquad \left.+\frac{d}{\sigma^2}\cdot\mu_w\sqrt{\frac{\|(\mu_z,\mu_w)\|^2}{\mu_w^2}}\cos\left(\phi-\left(\frac{\pi}{2}-\tan^{-1}\left(\frac{\mu_z}{\mu_w}\right)\right)\right)\right\}, && (2.45)
\end{aligned}$$

where $|J|$ is the Jacobian. Consider the four cases corresponding to where (μ_z,μ_w) is in each of the four quadrants. If (μ_z,μ_w) is in quadrant I or II then

$$\begin{aligned}
f_{d,\phi}(d,\phi) &= \frac{d}{2\pi\sigma^2}\exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}+\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\cos\left(\phi-\tan^{-1}\left(\frac{\mu_w}{\mu_z}\right)\right)\right\}\\
&= \frac{d}{2\pi\sigma^2}\exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}+\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\cos\left(\phi-\operatorname{atan2}\left(\mu_z,\mu_w\right)\right)\right\}. && (2.46)
\end{aligned}$$

If (μ_z,μ_w) is in quadrant III then consider $\phi^* = \phi+\pi = \tan^{-1}(-Z,-W)$. Then $\|(-\mu_z,-\mu_w)\| = \|(\mu_z,\mu_w)\|$, $\|(-x,-y)\| = d$, $|d\phi^*/d\phi| = 1$ and $(-\mu_z,-\mu_w)$ is in quadrant I. Hence

$$f_{d,\phi^*}(d,\phi^*) = \frac{d}{2\pi\sigma^2}\exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}+\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\cos\left(\phi^*-\tan^{-1}\left(\frac{\mu_w}{\mu_z}\right)\right)\right\}$$

and so

$$f_{d,\phi}(d,\phi) = \frac{d}{2\pi\sigma^2}\exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}+\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\cos\left(\phi-\operatorname{atan2}\left(\mu_z,\mu_w\right)\right)\right\}. \tag{2.47}$$

Finally, if (μ_z,μ_w) is in quadrant IV then consider $\phi^* = -\phi = \tan^{-1}(Z,-W)$. Then $\|(\mu_z,-\mu_w)\| =$

$\|(\mu_z, \mu_w)\|$, $\|(x, -y)\| = d$, $|d\phi^*/d\phi| = 1$ and $(\mu_z, -\mu_w)$ is in quadrant I. Hence we have

$$\begin{aligned} f_{d,\phi^*}(d, \phi^*) &= \frac{d}{2\pi\sigma^2} \exp\left\{-\frac{d^2 + \|(\mu_z, \mu_w)\|^2}{2\sigma^2} + \frac{d\|(\mu_z, \mu_w)\|}{\sigma^2} \cos\left(\phi^* - \tan^{-1}\left(\frac{-\mu_w}{\mu_z}\right)\right)\right\} \\ &= \frac{d}{2\pi\sigma^2} \exp\left\{-\frac{d^2 + \|(\mu_z, \mu_w)\|^2}{2\sigma^2} + \frac{d\|(\mu_z, \mu_w)\|}{\sigma^2} \cos\left(\phi^* + \tan^{-1}\left(\frac{\mu_w}{\mu_z}\right)\right)\right\} \end{aligned}$$

and hence

$$f_{d,\phi}(d, \phi) = \frac{d}{2\pi\sigma^2} \exp\left\{-\frac{d^2 + \|(\mu_z, \mu_w)\|^2}{2\sigma^2} + \frac{d\|(\mu_z, \mu_w)\|}{\sigma^2} \cos\left(\phi - \text{atan2}\left(\mu_w, \mu_z\right)\right)\right\}. \tag{2.48}$$

So (2.46), (2.47) and (2.48) show that regardless of where (μ_z, μ_w) is located, $f_{d,\phi}(d, \phi)$ is of the same form. By multiplying (2.46) by $I_0\left(\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\right) / I_0\left(\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\right)$ the result of Lemma 6.2 follows immediately:

$$\begin{aligned} f_{d,\phi}(d, \phi) = &\left[\frac{d}{\sigma^2} I_0\left(\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\right) \exp\left\{-\frac{d^2+\|(\mu_z,\mu_w)\|^2}{2\sigma^2}\right\}\right] \\ &\cdot \left[\frac{1}{2\pi I_0\left(\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2}\right)} \exp\left\{\frac{d\|(\mu_z,\mu_w)\|}{\sigma^2} \cos\left(\phi - \text{atan2}\left(\mu_z, \mu_w\right)\right)\right\}\right]. \end{aligned} \tag{2.49}$$

Chapter 3

Latent Space Models for Dynamic Networks with Weighted Edges

Representing relational data by networks is extremely useful and widely implemented. The dyadic relations which compose these networks are viewed as a set of actors and a set of edges between the actors. The edges can vary in many ways, such as being directed or undirected, static or temporal, binary or weighted. Binary networks, where between each actor an edge either does or does not exist, are encountered more often in the literature, although many such networks are by nature weighted. Weighted networks, also referred to as valued networks, consist of actors connected by edges which can take more than two values. By accounting for the weight, or strength, of the edges, the richness of the data can be better exploited. Examples of analyses of real world weighted networks include food webs (Krause et al., 2003), gene expression data (Zhang and Horvath, 2005), airline networks (Barrat et al., 2005), mobile phone networks (Onnela et al., 2007), and many more.

Often in binary networks it is of interest to compute various network measures, and recently there has been increasing work in extending these measures to weighted networks. Opsahl et al. (2010) derived for weighted networks measures for degree, closeness, and betweenness. Yang and Knoke (2001) derived a method for computing path length in the case of weighted edges. Opsahl and Panzarasa (2009) developed a method for analyzing the clustering that exists within a network with weighted edges. Other interesting works include Kunegis et al. (2009), which analyzed the case where edges took values in $\{-1, 0, 1\}$, and Newman (2004), which showed how to model networks whose edges are counts by representing them as multigraphs. To fully model the network, Krivitsky (2012) extended the commonly used exponential random graph model (ERGM) to account for networks whose dyads are counts; Krivitsky and Butts (2012) extended the ERGM to account for networks whose dyads are rankings.

Network data are most often inherently dynamic, even though it is frequently the case that the data are simply aggregated over time into one static network. Many popular static networks have been extended to longitudinal network data. Examples of this include the temporal exponential random graph model developed by Hanneke et al. (2010), the mixed membership stochastic blockmodel

for dynamic networks by Xing et al. (2010), and the latent space model for dynamic networks given in the previous chapter.

This chapter is focused on network data that is dynamic, weighted, and possibly directed. There are few resources available to the researcher investigating such data. Some existing approaches focus on latent space models for dynamic undirected binary networks. Latent space models assume the dependence of the network is induced by a set of latent variables. Such approaches are typically intuitive and have the advantage of producing meaningful visualizations, allowing the researcher to better understand the network structure as well as the behavior of individual actors.

Sarkar et al. (2007) extended the CODE model of Globerson et al. (2004) for dynamic undirected networks. This method is an approximate filtering algorithm which models the longitudinal count networks, embedding the actors in a latent space. This method is not easily generalizable to other sorts of co-occurence data besides counts, however, and cannot handle directed edges. Hoff (2011) described a multilinear model for undirected longitudinal networks. In this work, Hoff shows how to model undirected edges or ranked edges, where each dyad is an element from a finite ordered set, though it should be feasible to extend their approach to other types of dyads.

The latent space model presented in the previous chapter handles dynamic network data with directed binary edges. Our purpose is to extend this work in order to handle weighted edges. The remainder of the chapter is organized as follows: Section 3.1 extends the latent space model for dynamic networks with valued edges. Section 3.2 gives a method of estimation. Section 3.3 describes an approximation to reduce computational cost for large networks. Section 3.4 gives simulation results. Section 3.5 gives the results for analyzing Congressional cosponsorship data and world trade data. Section 3.6 gives the full conditional distributions referenced earlier in the chapter.

3.1 Models

We assume here that each actor exists within some latent space which can be interpreted as a characteristic space, or a social space. When actors are closer together in this latent space, the probability of a stronger edge is increased (where a "stronger edge" means a stronger relationship, though the actual form of this is context specific).

First is some general notation to be used throughout. Assume we have a set of actors $\mathcal{N}$ and a set of edges $\mathcal{E}$. Let $n = |\mathcal{N}|$ be the number of actors, and let Y_t be the $n \times n$ adjacency matrix of the observed network at time t whose entries y_{ijt} correspond to the weight of the edge from actor i

to actor j for $t \in \{1, 2, \ldots, T\}$. Let $\mathbf{X}_{it} \in \Re^p$ be the position vector of the i^{th} actor at time t within the p dimensional latent space. Let $\mathcal{X}_t$ be the matrix whose i^{th} row is $\mathbf{X}_{it}$. Finally, let $\mathbf{\Psi}$ be the vector of unknown parameters (which will vary depending on dyadic type).

We assume the latent actor positions transition according to a Markov process, where the initial distribution is

$$\pi(\mathcal{X}_1|\mathbf{\Psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{i1}|\mathbf{0}, \tau^2 I_p), \tag{3.1}$$

and the transition equation is

$$\pi(\mathcal{X}_t|\mathcal{X}_{t-1}, \mathbf{\Psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_p), \tag{3.2}$$

for $t = 2, 3, \ldots, T$, where I_p is the $p \times p$ identity matrix, and $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal probability density function with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$.

In most cases it can be assumed that the dependence structure of the network is fully induced by the latent positions of the actors. This assumption, along with the Markovian properties of the latent positions, leads to the state space temporal dependence structure given in Figure 2.1, as well as the conditional independence of each dyad within a time period. The ranked networks of the form analyzed by Krivitsky and Butts (2012) and in the next chapter are a counter example of where there is an extra dependency constraint in the data, but we will not further discuss here these rare data types. What remains then is to derive an appropriate conditional likelihood function, $\pi(Y_1, \ldots, Y_T|\mathcal{X}_1, \ldots, \mathcal{X}_T, \mathbf{\Psi}) = \prod_{t=1}^{T} \prod_{i \neq j} \pi(y_{ijt}|\mathcal{X}_t, \mathbf{\Psi})$.

Most latent space approaches have the conditional likelihood constructed by writing the logit of the edge probability as a linear form of covariates and a function of the latent variables, i.e., $\text{logit}(\pi(y_{ijt}|\cdot)) = \boldsymbol{\alpha}'\mathbf{w}_{ijt} + f_{\mathbf{\Psi}}(\boldsymbol{X}_{it}, \boldsymbol{X}_{jt})$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters, $\mathbf{w}_{ijt}$ is a vector of dyad specific covariates, and $f_{\mathbf{\Psi}} : \Re^p \times \Re^p \to \Re$ is a function taking as its arguments two actors' latent variables. Our generalization of this has the basic form

$$g(\mathbb{E}(y_{ijt})) = \boldsymbol{\alpha}'\mathbf{w}_{ijt} + f_{\mathbf{\Psi}}(\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}), \tag{3.3}$$

for some link function g. We can utilize the same types of link functions found in generalized linear mixed models. For example if our dyads are in the form of continuous data, we may set g to be the identity; this may arise in, for instance, proximity networks (see, e.g., Olguın et al., 2009), where the distance between individuals is recorded on a regular basis. The common case of modeling binary

dyads through the logit link function is yet another example. In Section 3.1.1 we will go into detail for the context of count data, using a log link function.

In some cases, however, the dyads cannot be modeled directly through a link function as in (3.3). Instead we can introduce additional latent variables, and then adopt a similar strategy. For example, we may consider a zero inflated model. The zero inflated model is a two component mixture model, where one could introduce additional latent indicator variables which determine whether the observation is coming from the component which is a point mass at zero or the component that has some other density function π^* (e.g., π^* is Poisson). We could then model $g(\mathbb{E}_{\pi^*}(y_{ijt}))$ as in (3.3). This situation may arise in large sparse weighted network data, such as company wide email count networks. Zero-inflated models are certainly not the only possibility of this type of data augmentation, as we will see in Section 3.1.2.

For the remainder of the paper we will focus on count data and non-negative continuous edges. We will furthermore utilize the conditional likelihood given in the previous chapter, determined by

$$f_{\boldsymbol{\Psi}}(\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}) = \beta_{IN}\left(1 - \frac{d_{ijt}}{r_j}\right) + \beta_{OUT}\left(1 - \frac{d_{ijt}}{r_i}\right), \tag{3.4}$$

where $d_{ijt} = \|\mathbf{X}_{it} - \mathbf{X}_{jt}\|$ is the distance between actors i and j at time t within the latent space, and $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ is a vector of positive actor specific parameters constrained such that $\sum_{i=1}^{n} r_i = 1$ for model identifiability.

Each r_i can be thought of as the i^{th} actor's social reach. That is, a larger value of r_i implies that it is more likely for an edge, either $y_{i\cdot t}$ or $y_{\cdot it}$, to take a larger value. These r_i's also hold a geometric interpretation within the latent space, specifically a radius. For example, in the context of binary networks, this radius can be understood to imply that actors inside of each others' radii have a greater than $1/2$ probability of an edge, and actors are outside of each other's radii have a smaller than $1/2$ probability of an edge. The coefficients β_{IN} and β_{OUT} can help in understanding the global structure of the network, insofar as telling us whether activity (tendency to send stronger edges) or popularity (tendency to receive stronger edges) is more important in forming high strength edges. Specifically, $\beta_{IN} > \beta_{OUT}$ implies popularity is more important than activity in the edge formation process, and $\beta_{OUT} > \beta_{IN}$ implies the opposite. If the edges are undirected, then setting $\mathbb{P}(y_{ijt}|\cdot) = \mathbb{P}(y_{jit}|\cdot)$ is equivalent to constraining $\beta_{IN} = \beta_{OUT}$.

3.1.1 Counts

A commonly encountered dyadic type which can be modeled by (3.3) is where y_{ijt} is a count. This context may exist in the form of counting the number of phone calls, the number of emails, the number of cosponsored legislative bills, the number of passengers or of flights in airline data, etc. We can use the canonical link for a Poisson random variable to determine the likelihood function in the following way:

$$\mathbb{P}(y_{ijt}|\mathcal{X}_t, \mathbf{\Psi}) = \frac{\lambda_{ijt}^{y_{ijt}} \exp(-\lambda_{ijt})}{y_{ijt}!}, \qquad y_{ijt} = 0, 1, 2, \ldots \tag{3.5}$$

where

$$\log(\lambda_{ijt}) = \beta_{IN}\left(1 - \frac{d_{ijt}}{r_j}\right) + \beta_{OUT}\left(1 - \frac{d_{ijt}}{r_i}\right). \tag{3.6}$$

Here $\mathbf{\Psi} = (\beta_{IN}, \beta_{OUT}, \boldsymbol{r}, \tau^2, \sigma^2)$ is the vector of parameters. Thus the likelihood is

$$\mathbb{P}(Y_1, Y_2, \ldots, Y_T|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T, \mathbf{\Psi}) = \prod_{t=1}^{T}\prod_{i\neq j} \frac{\lambda_{ijt}^{y_{ijt}} \exp(-\lambda_{ijt})}{y_{ijt}!}. \tag{3.7}$$

3.1.2 Non-Negative Continuous Edges

Here we consider the case of non-negative continuous real valued edges. These types of networks can occur in many biological contexts, in economic contexts, in length of phone calls, etc. The latent space framework provides a natural way to think about such a weighted network, in that we can consider two actors with large weighted edges between them as very close in the latent space, and two actors with smaller weighted edges as more separated within the latent space.

By embedding the network into a latent space we can better differentiate between zero valued edges. Consider as an example a longitudinal sequence of social networks, where the dyadic variable measured is the amount of time two individuals spent speaking with each other: Suppose at a particular time, person i has a weighted edge of zero with two others, persons j and k. Now persons i and j are potential friends though they have not currently met; meanwhile, persons i and k know each other already and strongly dislike each other. In both cases the measured edges between i and j and between i and k will be the same (zero), but we can differentiate them in two ways. First and foremost we can compare edge probabilities (e.g., $\mathbb{P}(y_{ij} = 0) << \mathbb{P}(y_{ik} = 0)$). Second, viewing the latent variables as unobserved actor attributes, we can determine the dissimilarity between each pair (e.g., $d_{ij} << d_{ik}$). The key point here is that we are using all the data, not just the data from

the pairs (i, j) and (i, k), to learn more about such observed zeros; i.e., we are letting all dyads help inform us as to the position of each actor within the latent space. This can be better understood by considering if i and j have many links to the same actors, then the geometric constraints within the latent space imply that i and j will be close together, whereas the same would not be true if, say, i and k do not have many links to the same actors.

Network data with non-negative continuous edges is a context where there is not an obvious link function g to be applied to the mean of y_{ijt}, but by introducing an additional latent variable we can adopt a similar strategy. In particular, we apply a tobit model when formulating the likelihood function, letting $y_{ijt} = y^*_{ijt} 1_{\{y^*_{ijt} > 0\}}$, where $1_{\{\cdot\}}$ is the indicator function and y^*_{ijt} is a continuous normal random variable. This type of approach may be most appropriate when the weighted dyads we observe are really proxies for some underlying relationship between the two actors, but we can only observe the effects from positive relationships (e.g., length of phone calls can only serve as a proxy for a relationship between friends, and not between enemies). We then apply (3.3) to the latent variables y^*_{ijt}, letting g be the identity function, obtaining

$$y^*_{ijt} = \beta_{IN}\left(1 - \frac{d_{ijt}}{r_j}\right) + \beta_{OUT}\left(1 - \frac{d_{ijt}}{r_i}\right) + \epsilon_{ijt}, \tag{3.8}$$

$$\epsilon_{ijt} \big| (\mathcal{X}_t, \mathbf{\Psi}) \overset{iid}{\sim} N(0, \gamma^2). \tag{3.9}$$

With this we have

$$\pi(y_{ijt}|\mathcal{X}_t, \mathbf{\Psi}) = \left[N(y_{ijt}|\mathbb{E}(y^*_{ijt}|\mathcal{X}_t, \mathbf{\Psi}), \gamma^2)\right]^{1_{\{y_{ijt}>0\}}} \left[1 - \Phi\left(\frac{\mathbb{E}(y^*_{ijt}|\mathcal{X}_t, \mathbf{\Psi})}{\gamma}\right)\right]^{1_{\{y_{ijt}=0\}}}, \tag{3.10}$$

where Φ is the standard normal cumulative distribution function, and $\mathbb{E}(y^*_{ijt}|\mathcal{X}_t, \mathbf{\Psi}) = \beta_{IN}(1 - d_{ijt}/r_j) + \beta_{OUT}(1 - d_{ijt}/r_i)$ is the conditional expectation of y^*_{ijt}. The vector of parameters is now supplemented by γ^2 such that $\mathbf{\Psi} = (\beta_{IN}, \beta_{OUT}, \gamma^2, \boldsymbol{r}, \tau^2, \sigma^2)$. Since the ϵ_{ijt}'s are conditionally i.i.d., we have that the observation equation is

$$\mathbb{P}(Y_1, Y_2, \ldots, Y_T|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T, \mathbf{\Psi}) = \prod_{t=1}^{T} \prod_{i \neq j} \pi(y_{ijt}|\mathcal{X}_t, \mathbf{\Psi}). \tag{3.11}$$

3.2 Estimation

To obtain estimates of the latent space positions and of the unknown parameters, we sample via a Markov chain Monte Carlo (MCMC) algorithm from the posterior

$$\pi(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T, \boldsymbol{\Psi} | Y_1, Y_2, \ldots, Y_T). \tag{3.12}$$

The general strategy is to find reasonable estimates of the latent positions and of the model parameters to initialize the chain, and then use a Metropolis-Hastings (MH) within Gibbs sampling to obtain the posterior samples.

The prior for $\boldsymbol{r}$ was a Dirichlet distribution, with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ equal to the initial estimates of $\boldsymbol{r}$ (given in the next section). The reason for doing so was because the prior expected value of r_j would then be the initial estimate of r_j, which reflects our prior intuition; additionally, as each α_j would be small (averaging $1/n$), the prior variance for each r_j will be large (leading to a "flat" prior). The priors for τ^2, σ^2 and, in the case of continuous data, γ^2 were chosen to be inverse gamma (IG), as these distributions are conjugate for τ^2 and σ^2. The shape and scale parameters for τ^2 were set to be equal to $2 + \delta$ and $(1 + \delta)\tau_0^2$ respectively for some small δ and some positive constant τ_0^2, and were similarly set for σ^2 and γ^2. With this parameterization, the prior variances of τ^2, σ^2 and γ^2 are kept large. Further, we recommend setting τ_0^2 equal to the initial estimate of τ^2 given in the next section (equation (3.16)) to match our intuition. The prior set on β_{IN} was a normal distribution with mean ν_{IN} and (large) variance ξ_{IN}, and similarly for β_{OUT}.

3.2.1 Initialization

We initialized the radii as

$$r_i = \frac{\sum_{t=1}^{T} \sum_{j \neq i} (y_{ijt} + y_{jit})/2}{\sum_{t'=1}^{T} \sum_{j' \neq i'} y_{i'j't'}}. \tag{3.13}$$

To find initial latent positions, we implemented the generalized multidimensional scaling algorithm (GMDS), as described in Sarkar and Moore (2005). GMDS starts by taking a distance matrix at time 1 and performing classical multidimensional scaling. Then, for each subsequent time period t, $t = 2, 3, \ldots, T$, GMDS balances the position matrix from the previous time point with the classical multidimensional scaling result obtained from the distance matrix at time t.

The original distance matrices can be found in a number of ways, but we offer the following

suggestion. We treated the data as binary, where

$$y_{ijt}^{(binary)} = \begin{cases} 1 & \text{if } y_{ijt} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

We then computed each distance d_{ijt} according to

$$d_{ijt} = \begin{cases} \frac{1}{2}\min\{r_i, r_j\} & \text{if } y_{ijt} = y_{jit} = 1 \\ (r_i + r_j)/2 & \text{if } y_{ijt} + y_{jit} = 1 \\ \frac{1}{2}\sum_{l=1}^{n_{ijt}} r_{k_l} & \text{if } y_{ijt} = y_{jit} = 0 \end{cases} \tag{3.15}$$

where the subsequence $\{k_l\}_{l=1}^{n_{ijt}}$ represents the actors along the shortest path between i and j at time t, neglecting directions, and n_{ijt} is the number of actors in that shortest path. The general idea here is that when there is an edge from one node i to another node j, it is more likely that these two nodes are within r_i, r_j, or both. With the T distance matrices computed, we can then implement GMDS to obtain initial latent positions.

The initial estimate for τ^2 was computed (using the initial estimates of $\mathcal{X}_1$) as

$$\frac{1}{nD}\sum_{i=1}^{n} \|\mathbf{X}_{i1}\|^2. \tag{3.16}$$

We recommend using a large initial estimate of σ^2 in order to initially allow large movements of $\mathcal{X}_t$, $t = 2, 3, \ldots, T$, and thus help reach convergence faster. The initial estimates for γ, β_{IN} and β_{OUT} did not significantly affect the number of iterations required to reach convergence.

3.2.2 Posterior Sampling

We implement a MH within Gibbs sampling scheme. The algorithm is

0. Set the initial values of the latent positions and parameters as given in Section 3.1.
1. For $t = 1, 2, \ldots, T$ and for $i = 1, 2, \ldots, n$, draw $\mathbf{X}_{it}$ via MH.
2. Draw τ^2 from $\pi(\tau^2|\mathcal{X}_1)$.
3. Draw σ^2 from $\pi(\sigma^2|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T)$.
4. Draw $\boldsymbol{r}$ via MH.
5. Draw β_{IN} via MH.
6. Draw β_{OUT} via MH.
If data is non-negative continuous
7. Draw γ^2 via MH.

Repeat steps 1-7.

The full conditional distributions needed for steps 2-7 are given at the end of this chapter. Regarding the proposal distributions, $\mathbf{X}_{it}$, β_{IN}, and β_{OUT} can come from a symmetric proposal (e.g., normal random walk). For γ^2, however, some asymmetric proposal such as a log-normal or an inverse gamma distribution ought to be used to ensure positive valued proposals; this asymmetric proposal will then need to be accounted for in the acceptance probability. Because of the constraint on $\boldsymbol{r}$, a Dirichlet proposal is suggested for the radii, which also will be an asymmetric proposal. Suggested parameters for this Dirichlet proposal are $\kappa \boldsymbol{r}^{curr}$, where $\boldsymbol{r}^{curr}$ are the current values for $\boldsymbol{r}$ and κ is some large value.

One final note is that, as is the case for any such latent space model, the posterior is invariant under rotations, reflections and translations of the latent positions $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$. Hence after each iteration of steps 1-7, a Procrustean transformation will be performed on the T trajectories. The Procrustean transformation finds a set of rotations, reflections and translations to minimize the difference between a given matrix and some target matrix. In our case, the target matrix is always chosen to be the initialized latent position trajectories.

3.3 Scalability

Although the two data sets that we will analyze in Section 3.5 are relatively small data sets, it is likely that researchers will come across network data which is too large for the MCMC algorithm of Section 3.2.2 to be a viable option. Here we extend the case control log likelihood approximation method by Raftery et al. (2012) for models whose dyads can be described by an exponential family of distributions.

For the MCMC algorithm, the MH steps required in updating the latent positions, $\boldsymbol{r}$, β_{IN}, β_{OUT} and other likelihood related parameters (e.g., γ^2 in the case of non-negative real dyads) all require $O(Tn^2)$ terms to be summed. In this discussion we will assume here that a non-relationship between two actors implies that $y_{ijt} = 0$, otherwise y_{ijt} is some positive value; the principles discussed next ought to hold even if this is not the case. Generalizing the approximation method first proposed by Raftery et al. (2012), we can reduce this computational cost to $O(Tn)$.

Suppose that, conditional on the latent positions, the y_{ijt}'s are independent with

$$\pi(y_{ijt}|\cdot) = h(y_{ijt}) \exp(\boldsymbol{\eta}'_{ijt}\mathbf{T}(y_{ijt}) - A(\boldsymbol{\eta}_{ijt})), \tag{3.17}$$

where $\boldsymbol{\eta}_{ijt}$ is a vector valued function of $(\mathcal{X}_t, \boldsymbol{\Psi})$ and $\mathbf{T}(y_{ijt})$ is a vector of sufficient statistics. Then we can rewrite the loglikelihood of $(Y_1, \ldots, Y_T)$ as

$$\ell(\mathcal{X}_1, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) = \sum_{t=1}^{T}\sum_{i=1}^{n}\left[\sum_{j:y_{ijt}>0}\left(\boldsymbol{\eta}_{ijt}'\mathbf{T}(y_{ijt}) + A(\boldsymbol{\eta}_{ijt})\right) + \sum_{j:y_{ijt}=0}\left(\boldsymbol{\eta}_{ijt}'\mathbf{T}(y_{ijt}) + A(\boldsymbol{\eta}_{ijt})\right)\right] + \text{constant}. \tag{3.18}$$

It is reasonable to assume that as the network gets larger and larger, the number of edges of each node does not grow at the same rate (the network gets sparser). Hence we make the assumption that either the maximum degree is fixed or is of $o(n)$. If this is the case, then we can, for each i and t, take a subsample $\{j_k\}_{k=1}^{N_{i,t,0}}$ from the set $\{j : y_{ijt} = 0\}$ and use a simple Monte Carlo estimate of the final summation of (3.18) to reduce the computational cost to linear with respect to n. Then the approximation we use of the log likelihood is

$$\ell(\mathcal{X}_1, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) \approx \sum_{t=1}^{T}\sum_{i=1}^{n}\left[\sum_{j:y_{ijt}>0}\left(\boldsymbol{\eta}_{ijt}'\mathbf{T}(y_{ijt}) + A(\boldsymbol{\eta}_{ijt})\right) + \frac{n_{i,t,0}}{N_{i,t,0}}\sum_{k=1}^{N_{i,t,0}}\left(\boldsymbol{\eta}_{ij_kt}'\mathbf{T}(y_{ij_kt}) + A(\boldsymbol{\eta}_{ij_kt})\right)\right] + \text{constant}, \tag{3.19}$$

where $n_{i,t,0} = |\{j : y_{ijt} = 0\}|$. In most cases, $\mathbf{T}(y_{ijt}) = 0$ if $y_{ijt} = 0$ and hence the above can be simplified such that the second summation is only $\frac{n_{i,t,0}}{N_{i,t,0}}\sum_{k=1}^{N_{i,t,0}} A(\boldsymbol{\eta}_{ij_kt})$. Also, there could potentially be multiple methods of selecting the subsequences $\{j_k\}_{k=1}^{N_{i,t,0}}$; see Raftery et al. (2012) for more details.

For $T = 1$ and $y_{ijt} \in \{0, 1\}$, this leads to Raftery et al.'s approximation. For the context presented in Section 3.1.1, we can approximate the log likelihood as

$$\ell(\mathcal{X}_1, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) \approx \sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\sum_{j:y_{ijt}>0}\left[-\frac{1}{2}\log(\gamma^2) - \frac{1}{2\gamma^2}(y_{ijt} - \mathbb{E}(y_{ijt}^*|\mathcal{X}_t, \boldsymbol{\Psi}))^2\right] + \frac{n_{i,t,0}}{N_{i,t,0}}\sum_{k=1}^{N_{i,t,0}}\log\left(1 - \Phi(\mathbb{E}(y_{ijt}^*|\mathcal{X}_t, \boldsymbol{\Psi})/\gamma)\right)\right\} + \text{constant}. \tag{3.20}$$

.

An interesting point is that if we assume that the network becomes more sparse as n grows, then it may be more appropriate to utilize a zero-inflated model, such as was mentioned in Sec-

tion 3.1. Suppose we can augment the data by component indicator variables $z_{ijt} \in \{1, 2\}$, such that $\pi(y_{ijt}|z_{ijt} = 1, \cdot) = \delta(y_{ijt})$, $\pi(y_{ijt}|z_{ijt} = 2, \cdot)$ can be constructed according to (3.3), $\pi(z_{ijt} = 1) = \alpha$, and δ is the Dirac delta function. Then we can write the complete log likelihood (i.e., $\pi(Y_1, \ldots, Y_T, Z_1, \ldots, Z_T|\cdot)$) as

$$\begin{aligned}
&\ell(\mathcal{X}_1, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) \\
=&\sum_{t=1}^{T}\sum_{i=1}^{n}\Bigg\{\sum_{j:y_{ijt}>0}\big[\log(1-\alpha) + \boldsymbol{\eta}_{ijt}'\mathbf{T}(y_{ijt}) + A(\boldsymbol{\eta}_{ijt})\big] \\
&+\frac{n_{i,t,0}}{N_{i,t,0}}\sum_{k=1}^{N_{i,t,0}}\Big[1_{\{z_{ij_kt}=1\}}\log(\alpha) + 1_{\{z_{ij_kt}=2\}}\Big(\log(1-\alpha) + \boldsymbol{\eta}_{ij_kt}'\mathbf{T}(y_{ij_kt}) + A(\boldsymbol{\eta}_{ij_kt})\Big)\Big]\Bigg\} + \text{constant},
\end{aligned} \tag{3.21}$$

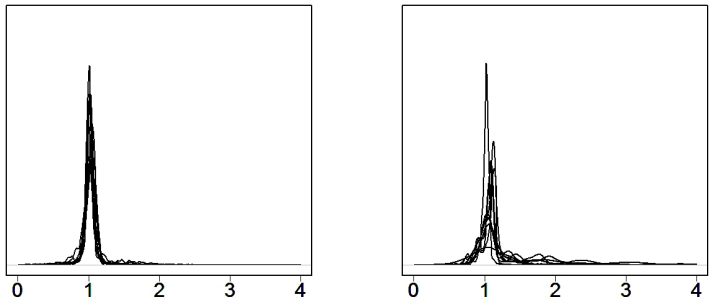
where $[Z_t]_{ij} = z_{ijt}$. Since the z_{ijt}'s are nuisance parameters, we need not sample all of them in the Gibbs sampler, but rather only the z_{ijt}'s corresponding to each of the n subsequences $\{j_k\}_{k=1}^{N_{i,t,0}}$, thus maintaining the computational cost of $O(Tn)$.

3.4 Simulations

3.4.1 Simulated Count Data

Ten data sets were simulated, where the number of actors was 100 and the number of time points was 10. For each of the simulations, the parameter values were set at $\beta_{IN} = 2$, $\beta_{OUT} = 1$, and $\sigma^2 = 5 \times 10^{-7}$. The latent positions at time 1 were drawn from a mixture of 12 normals with equal mixture component weights, where the cluster means were drawn randomly from a multivariate normal distribution with mean zero and covariance $(1 \times 10^{-5})I_p$, and $p = 2$ is the dimension of the latent space. After the initial latent positions $\mathcal{X}_1$ were drawn, the radii $\boldsymbol{r}$ were drawn from a Dirichlet distribution whose i^{th} parameter was equal to $n\|\mathbf{X}_{i1}\|/\max_k\{\|\mathbf{X}_{k1}\|\}$. Subsequent latent positions $\mathcal{X}_t$, $t \geq 2$, were drawn according to (3.2). The adjacency matrices Y_1 to Y_T were then generated according to (3.5) and (3.6).

The priors for σ^2 and τ^2 were inverse gamma distributions with parameters formulated according to the description in Section 3, with $\delta = 0.05$, $\sigma_0^2 = 1 \times 10^{-4}$ and $\tau_0^2 = 1/(np)\sum_{i=1}^{n}\|\mathbf{X}_{i1}\|^2$, using the initial positions of $\mathbf{X}_{i1}$. The prior for β_{IN} was $N(2, 100)$, for β_{OUT} was $N(2, 100)$, and for $\boldsymbol{r}$ was Dirichlet with parameters equal to that given in (3.13). A normal random walk proposal was used for the latent positions $\mathbf{X}_t$ and also for both β_{IN} and β_{OUT}. The proposal for $\boldsymbol{r}$ was a Dirichlet



(a) Count edges (b) Non-negative real edges

Figure 3.1: Distributions of ratios of pairwise distances for simulated data with count, non-negative real, and rank edges. Each curve corresponds to a simulation.

distribution with parameters equal to $\kappa \boldsymbol{r}^{curr}$, where $\boldsymbol{r}^{curr}$ represents the current value of $\boldsymbol{r}$.

To evaluate the simulation results, we compared the estimates of the coefficients β_{IN} and β_{OUT} with the truth, evaluated the pseudo R^2, and evaluated the pairwise ratios of estimated distances to true distances corresponding to the latent positions. The pseudo R^2 value is the deviance based pseudo R^2 for count data found, and recommended, in Cameron and Windmeijer (1996). This is calculated as

$$R^2 = \frac{\sum_{t=1}^{T} \sum_{i \neq j} y_{ijt} \log(\hat{\lambda}_{ijt}/\bar{y}) - (\hat{\lambda}_{ijt} - \bar{y})}{\sum_{t'=1}^{T} \sum_{i' \neq j'} y_{i'j't'} \log(y_{i'j't'}/\bar{y})} \tag{3.22}$$

where $\bar{y} = \sum_{t=1}^{T} \sum_{i \neq j} y_{ijt}$ and $\hat{\lambda}_{ijt}$ is found by plugging in the posterior mean estimates in (3.6). To clarify what is meant by the distance ratios, note that for each simulation there are $Tn(n-1)/2$ distances within the latent space. We calculate all these pairwise distances using the posterior mean latent positions as well as using the true latent positions. So for each simulation we can plot a curve corresponding to the distribution of these ratios. We would hope for this curve to be narrow and centered at 1.

The posterior mean estimate, averaged over the ten simulations, for β_{IN} (β_{OUT}) whose true value was 2 (1), was 2.013 (0.9859), ranging from 2.003 to 2.026 (0.9500 to 0.9998). We see that the posterior mean estimates are very close to the true value in every simulation. The pseudo R^2 values' average was 0.9866, ranging from 0.9309 to 0.9945, implying that the posterior means fit the data extremely well. The distributions of the ratios of pairwise distances are given in Figure 3.1a, where each curve corresponds to a simulation. From this figure we see that the distances from the posterior mean latent positions are very close to the true distances.

3.4.2 Simulated Continuous Data

Ten data sets were simulated, where the number of actors was 100 and the number of time points was 10. For each of the simulations, the parameter values were set at $\beta_{IN} = 10$, $\beta_{OUT} = 2$, $\gamma^2 = 0.2$, and $\sigma^2 = 5 \times 10^{-6}$. The latent positions at time 1 were drawn from a mixture of 12 normals with equal mixture component weights, where the cluster means were drawn randomly from a multivariate normal distribution with mean zero and covariance $(5 \times 10^{-5})I_p$, and $p = 2$ is the dimension of the latent space. After the initial latent positions $\mathcal{X}_1$ were drawn, the radii $\boldsymbol{r}$ were drawn from a Dirichlet distribution whose i^{th} parameter was equal to $n\|\mathbf{X}_{i1}\|/\max_k\{\|\mathbf{X}_{k1}\|\}$. Subsequent latent positions $\mathcal{X}_t$, $t \geq 2$, were drawn according to (3.2). The adjacency matrices Y_1 to Y_T were then constructed by generating y^*_{ijt} according to (3.8) and letting $y_{ijt} = y^*_{ijt}1_{\{y^*_{ijt}>0\}}$.

The priors for σ^2, τ^2 and γ^2 were inverse gamma distributions with parameters formulated according to the description in Section 3, with $\delta = 0.05$, $\sigma_0^2 = 1 \times 10^{-4}$, $\tau_0^2 = 1/(nD)\sum_{i=1}^n \|\mathbf{X}_{i1}\|^2$ using the initial positions of $\mathbf{X}_{i1}$, and $\gamma_0^2 = 3$. The priors for β_{IN} and for β_{OUT} were $N(2, 100)$, and for $\boldsymbol{r}$ was Dirichlet with parameters equal to that given in (3.13). A normal random walk proposal was used for the latent positions $\mathbf{X}_t$ and for both β_{IN} and β_{OUT}. The proposal used for γ^2 was a log-normal distribution with parameter log-mean equal to $\log((\gamma^2)^{curr})$, where $(\gamma^2)^{curr}$ represents the current value of γ^2. The proposal for $\boldsymbol{r}$ was a Dirichlet distribution with parameters equal to $\kappa\boldsymbol{r}^{curr}$, where $\boldsymbol{r}^{curr}$ represents the current value of $\boldsymbol{r}$.

To evaluate the simulation results, we compared the estimates of the coefficients β_{IN} and β_{OUT} with the truth, evaluated the pseudo R^2, and evaluated the pairwise ratios of estimated distance to true distance. In this context of continuous non-negative data, we used the pseudo R^2 value recommended in Veall and Zimmermann (1994), originally derived by McKelvey and Zavoina (1975). This is calculated as

$$R^2 = \frac{\sum_{t=1}^T \sum_{i\neq j}(\hat{y}^*_{ijt} - \hat{\bar{y}}^*)^2}{\sum_{t'=1}^T \sum_{i'\neq j'}(\hat{y}^*_{i'j't'} - \hat{\bar{y}}^*)^2 + Tn(n-1)\hat{\gamma}^2}, \tag{3.23}$$

where $\hat{y}^*_{ijt} = \hat{\beta}_{IN}(1 - \hat{d}_{ijt}/\hat{r}_j) + \hat{\beta}_{OUT}(1 - \hat{d}_{ijt}/\hat{r}_i)$ and $\hat{\bar{y}}^* = 1/(Tn(n-1))\sum_{t=1}^T \sum_{i\neq j} \hat{y}^*_{ijt}$. The $\hat{}$ symbol over the model parameters implies the posterior mean estimate.

The posterior mean estimate, averaged over the ten simulations, for β_{IN} (β_{OUT}) whose true value was 10 (2), was 9.973 (1.970), ranging from 9.633 to 10.07 (1.900 to 2.104). We see that the posterior mean estimates are very close to the true value in every simulation. The pseudo R^2 values' average was 0.9993, ranging from 0.9983 to 0.99999, implying that the posterior means fit the data extremely well. The distributions of the ratios of pairwise distances are given in Figure 3.1b, where

each curve corresponds to a simulation. From this figure we see that the distances from the posterior mean latent positions are very close to the true distances.

3.5 Data Analysis

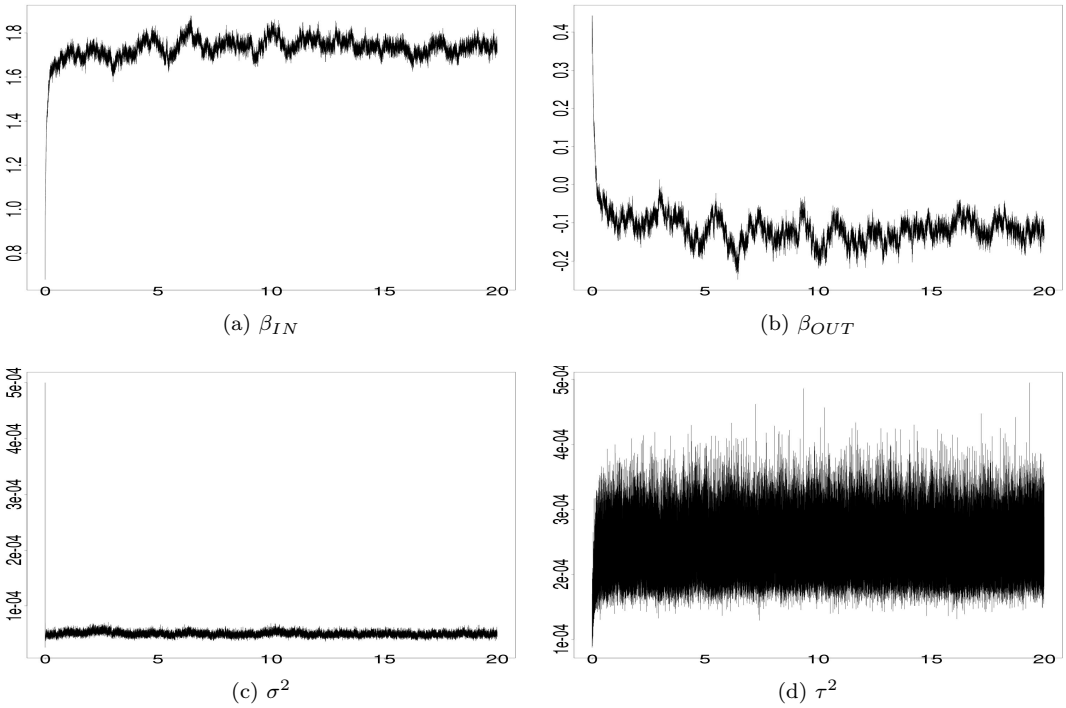
3.5.1 Cosponsorship Data

We consider the cosponsorship data for the 93^{rd} to 110^{th} U.S. Congresses, collected by James Fowler (see Fowler, 2006a,b). Specifically, we consider the same 49 Congressmen that serve in consecutive House of Representatives for 12 terms (the 98^{th} to the 109^{th} Congresses). We let y_{ijt} be the number of bills sponsored by Congressman j and cosponsored by Congressman i, and hence the network's edges consist of counts. Using these counts rather than a more simplified network yields richer and more reliable data.

The priors for σ^2 and τ^2 were inverse gamma distributions with parameters formulated according to the description in Section 3, with $\delta = 0.005$, $\sigma_0^2 = 5 \times 10^{-4}$ and $\tau_0^2 = 1/(np)\sum_{i=1}^{n} \|\mathbf{X}_{i1}\|^2$ using the initial positions of $\mathbf{X}_{i1}$. The prior for the coefficients were $N(0.75, 100)$ for β_{IN} and $N(0.25, 100)$ for β_{OUT}. The prior for $\boldsymbol{r}$ was Dirichlet with parameters equal to that given in (3.13). The variance components of the normal random walk proposals for the latent space positions and the β coefficients were 2.5×10^{-5} and 7.5×10^{-4} respectively. The tuning parameter κ used in the proposal for the radii was set to be 2×10^5. A burn-in of 100000 samples was removed from the chain before analysis, leaving a chain of length 100000. Figure 3.2 gives the trace plots for β_{IN}, β_{OUT}, σ^2 and τ^2.

The pseudo R^2 value was 0.9233, implying a very good fit of the data. The posterior means of the coefficients were $\beta_{IN} = 1.74$ and $\beta_{OUT} = -0.12$, implying that, in the formation of a sponsorship/cosponsorship event, who sponsors a bill is more important than who cosponsors a bill, i.e. popularity is more important in the network structure than activity. This corroborates the results from the previous chapter when the larger binary version of this data was analyzed.

Figure 3.3 gives a plot of the posterior mean latent positions. Triangles indicate republicans and circles indicate democrats. Temporal direction is shown via arrows. The size of a actor's symbol corresponds to his/her social reach, where a larger social reach implies that there is an increased probability of that actor receiving cosponsorships from surrounding actors. From this figure we see several things. First and most expected, there is a clear split between the two parties. Second, each party has a central cluster, and these two clusters are relatively close to each other. In other words, within each party there is a central group of Congressmen, and these two central groups are



(a) β_{IN} (b) β_{OUT}

(c) σ^2 (d) τ^2

Figure 3.2: MCMC trace plots for the model parameters corresponding to the cosponsorship data. Horizontal axis is in iterations $\times 10^4$.

moderate rather than radical. Third, each party has a few outlying Congressmen, who happen to be both far from their opposite party and also have small social reach (receive few cosponsorships). An example of this is Hal Rogers, a Republican from Kentucky, circled in Figure 3.3, who has received much criticism (e.g., Vrazilek, 2009; Dickinson, 2006; Shannon, 2007).

Figure 3.3 can be helpful in understanding a Congressman's career. As an example of this, consider Jim Leach, a Republican from Iowa, whose locations are in boxes. He first was elected by defeating a Democrat incumbant, and we can see that his early associations were with Democrats. We can see his transition over his career towards his fellow Republicans, although he maintained close proximity to the main cluster of Democrats. This last fact was reflected in his stance on several key issues, such as support for stem cell research, voting against the 2003 extension of the Bush tax cuts (one of only three Republicans to do so), voting against the Iraq war, and others (Ericson, 2012; clerk.house.gov, 2003; Leach, 2012).

3.5.2 World Trade Data

World trade data, measuring annual exports/imports between countries in the years 1991-2000, was analyzed. The data, given in millions of US dollars, was obtained through the Economic Web Insti-

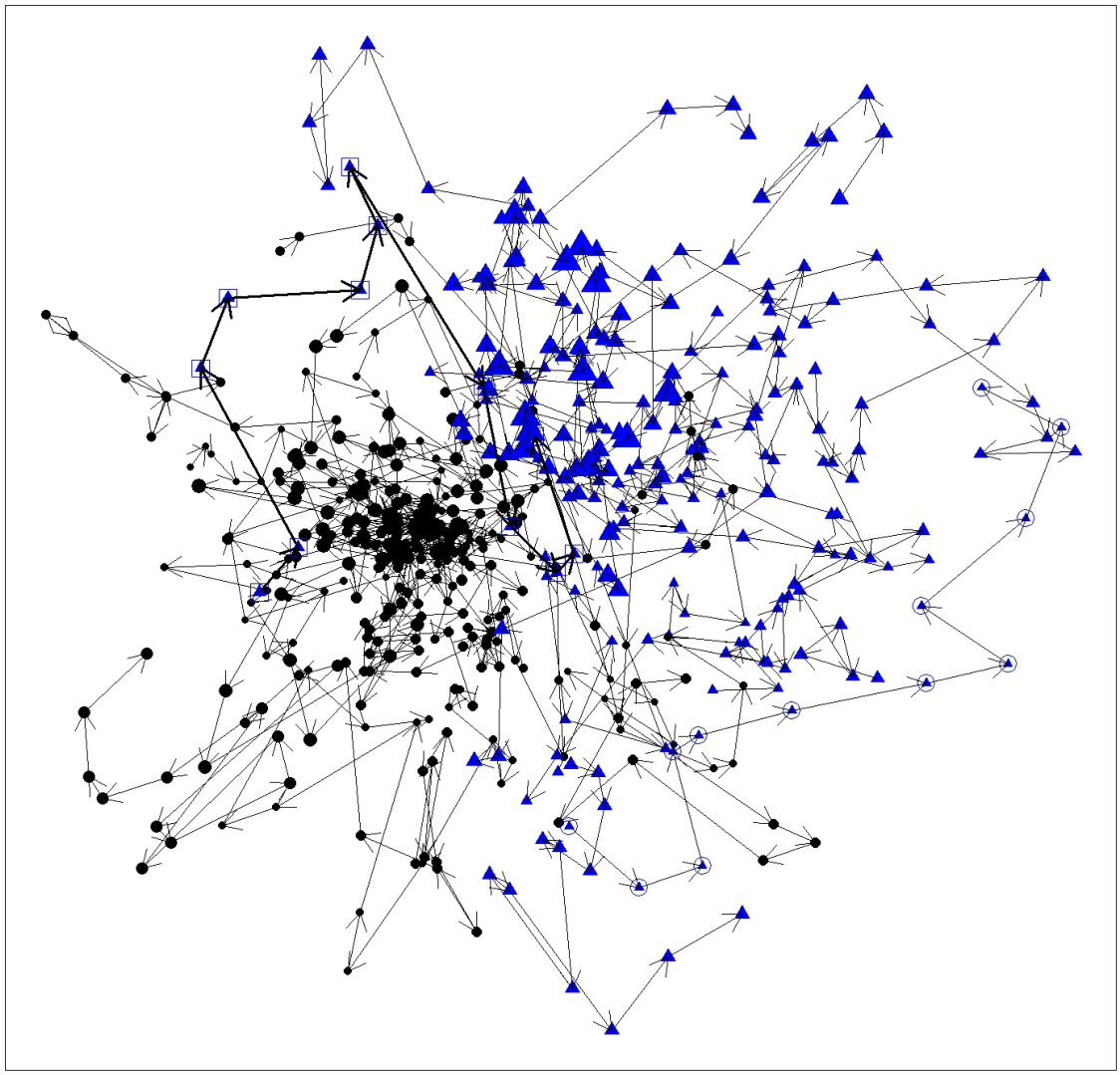


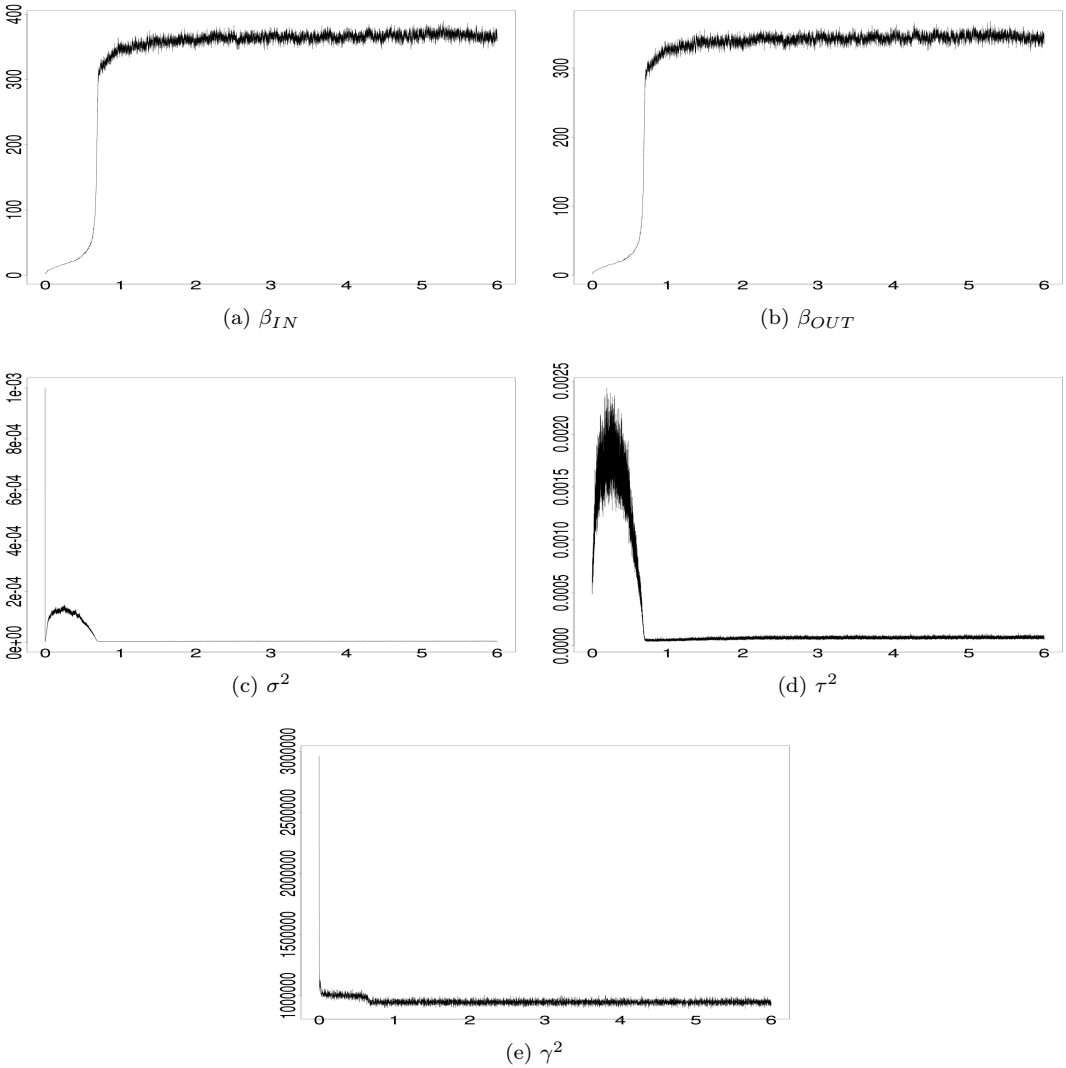
Figure 3.3: Posterior means of latent positions for the cosponsorship data, arrows indicating the temporal direction of the trajectories. Triangles are republicans and circles are democrats. The size of a actor's symbol corresponds to his/her social reach. Hal Rogers, an often criticized Congressman, is circled, and the career path of Jim Leach, a moderate Republican, is in boxes.

tute at *http://www.economicswebinstitute.org/worldtrade.htm*, originally obtained through the IMF Direction of Trade (DOT) Yearbook. The edges here are non-negative reals. As with the cosponsorship data, we analyzed a subset of the data. Included in the analysis was 107 countries who were all involved in world trade through the years 1991 to 2000. To account for global inflation/deflation and any other non-relational economic effects, the data was rescaled such that $\sum_{i \neq j} y_{ijt} = \sum_{i \neq j} y_{ijs}$ for any $t, s = 1, 2, \ldots, T$; i.e., the total quantity of annual world trade is constant.

The priors for σ^2, τ^2 and γ^2 were inverse gamma distributions with parameters formulated according to the description in Section 3. For σ^2, we set $\delta = 0.000005$ and $\sigma_0^2 = 1 \times 10^{-3}$; for τ^2 we set $\delta = 0.05$ and $\tau_0^2 = 1/(np) \sum_{i=1}^{n} \|\mathbf{X}_{i1}\|^2$ using the initial positions of $\mathbf{X}_{i1}$; for γ^2 we set $\delta = 0.0005$ and $\gamma_0^2 = 1 \times 10^5$. The prior for the coefficients were $N(15, 500)$ for β_{IN} and $N(10, 500)$ for β_{OUT}. The prior for $\boldsymbol{r}$ was Dirichlet with parameters equal to that given in (3.13). A normal random walk proposal with variance of 5×10^{-7} was used for the latent positions $\mathbf{X}_t$ and with variance of 100 for both β_{IN} and β_{OUT}. The proposal used for γ^2 was a log-normal distribution with log mean equal to $\log((\gamma^2)^{curr})$ and log standard deviation equal to 0.01, where $(\gamma^2)^{curr}$ represents the current value of γ^2. The proposal for $\boldsymbol{r}$ was a Dirichlet distribution with parameters equal to $(5 \times 10^6)\boldsymbol{r}^{curr}$, where $\boldsymbol{r}^{curr}$ represents the current value of $\boldsymbol{r}$. Figure 3.4 gives the trace plots for β_{IN}, β_{OUT}, σ^2, τ^2 and γ^2. A burn-in of 30000 was used, leaving a chain of length 30000.

The pseudo R^2 value was 0.8233, indicating a very good fit of the data. The estimates for β_{IN} and β_{OUT} were 366 and 344 respectively, implying that the amount of trade is determined more by the importing country than the exporting country, but only slightly so.

Figure 3.5 gives plots of the posterior mean latent positions, broken up into three time periods: from 1991 to 1993, from 1994 to 1996, and from 1997 to 2000. Temporal direction is shown via arrows. The size of the actor corresponds to its r_i value. Each color represents a geographical region, where green is Africa, yellow is Asia, dark red is Eurasia, blue is Europe, red is North America and the Caribbean, sea green is Oceania, and brown is South America. It is apparent that the actors move within the latent space much less during each of these three periods than during the transition from 1993 to 1994 and from 1996 to 1997. These two major shrinkage events occurring both coincide with major events in world trade. In 1993, the General Agreement on Tariffs and Trade was updated, which would later lead to the creation of the World Trade Organization (WTO) (see http://www.wto.org). Looking at Figure 3.5, we can see that there is already some shrinkage happening during the year 1993 which then continues going into 1994. Specifically we see that certain continents (Africa, Asia, and Europe) come together during this time. Europe is the clearest



(a) β_{IN} (b) β_{OUT}

(c) σ^2 (d) τ^2

(e) γ^2

Figure 3.4: MCMC trace plots for the model parameters corresponding to the world trade data. Horizontal axis is in iterations $\times 10^4$.

case, and it turns out there is a good reason for this: the European Economic Area was established on January 1, 1994. Regarding the second shrinkage event in 1997, a publication from the WTO states that "the volume of world merchandise exports grew by 9.5 per cent in 1997." This is seen visually in Figure 3.5. However, since the original data had been scaled to account for such growth, we conclude that the reason for this growth in world exports is not due to existing relationships getting stronger, but rather to the formation of many more trading relationships.

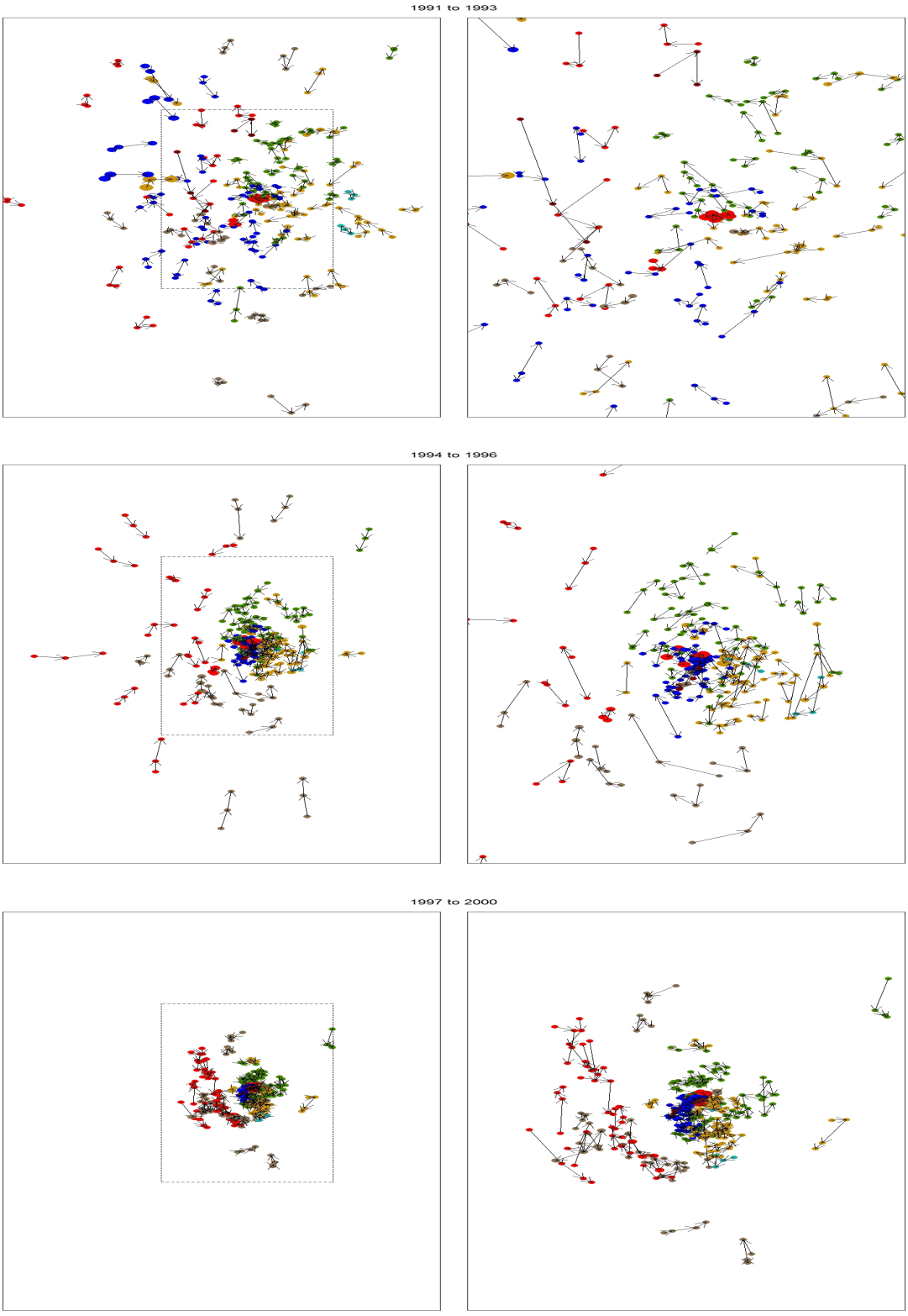


Figure 3.5: World trade import/export relational data. Each column is scaled equally. Each figure on the right is the zoomed in figure of the dotted box in the figures to their left.

3.6 Full Conditional Distributions

The full conditional distributions for τ^2 and σ^2 are respectively

$$\pi(\tau^2|\mathcal{X}_1) \sim \text{IG}(2+\delta+nD/2, (1+\delta)\tau_0^2 + \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{X}_{i1}\|^2), \tag{3.24}$$

$$\pi(\sigma^2|\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T) \sim \text{IG}(2+\delta+\frac{nD(T-1)}{2}, (1+\delta)\sigma_0^2 + \frac{1}{2}\sum_{t=2}^{T}\sum_{i=1}^{n}\|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|^2), \tag{3.25}$$

for δ, τ_0^2, $\sigma_0^2 > 0$.

We let $\pi_{y_{ijt}} \triangleq \pi(y_{ijt}|\mathcal{X}_t, \mathbf{\Psi})$. Then the full conditional distribution for $\mathbf{X}_{it}$ in these two cases is

$$\pi(\mathbf{X}_{it}|Y_{1:T}, \mathbf{\Psi}) \propto \begin{cases} \left(\prod_{j\neq i}\pi_{y_{ijt}}\pi_{y_{jit}}\right) \cdot N(\mathbf{X}_{it}|\mathbf{0}, \tau^2 I_D) \cdot N(\mathbf{X}_{i(t+1)}|\mathbf{X}_{it}, \sigma^2 I_D), & \text{if } t = 1 \\ \left(\prod_{j\neq i}\pi_{y_{ijt}}\pi_{y_{jit}}\right) \cdot N(\mathbf{X}_{i(t+1)}|\mathbf{X}_{it}, \sigma^2 I_D) \cdot N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_D), & \text{if } 1 < t < T \\ \left(\prod_{j\neq i}\pi_{y_{ijt}}\pi_{y_{jit}}\right) \cdot N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, \sigma^2 I_D), & \text{if } t = T. \end{cases} \tag{3.26}$$

The full conditional distribution for each of the model parameters follows the form

$$\pi(\psi|Y_{1:T}, \mathcal{X}_{1:T}, \mathbf{\Psi}\backslash\{\psi\}) \propto \left[\prod_{t=1}^{T}\pi(Y_t|\mathcal{X}_t, \mathbf{\Psi})\right] \cdot \pi(\psi) \tag{3.27}$$

where $\mathbf{\Psi}\backslash\{\psi\}$ is the set of parameters excluding ψ, and for count data $\psi \in \{\beta_{OUT}, \beta_{IN}, \boldsymbol{r}\}$ and for non-negative real data $\psi \in \{\beta_{OUT}, \beta_{IN}, \gamma^2, \boldsymbol{r}\}$.

Chapter 4

Analysis of the Formation of the Structure of Social Networks using Latent Space Models for Ranked Dynamic Networks

The formation and evolution of interpersonal relationships are highly studied in the social sciences. These interpersonal relationships can most easily be thought of in the context of a social network in which we observe how a certain number of actors interact. By analyzing such a network over time, one can hope to quantify the construction and stabilization of the network and its structures. In 1954 T. Newcomb began an observational study using a college fraternity for this purpose, and a very large number of researchers have relied on this study to help understand how social networks form and stabilize. This fraternity data set gives social scientists the unique opportunity to study the evolution and formation of the structure of social networks from a nonexistent state to a stabilized form. The overall goal of the original study was to "improve our understanding of the development of stable interpersonal relationships" (Newcomb, 1961).

Some authors have used Newcomb's fraternity data as an example with which to illustrate new methodology, e.g., Snijders (1996) states "our treatment of Newcomb's fraternity data in this paper is not more than an example . . . " Other authors have analyzed this data set in more depth, utilizing it for its worth in helping to understand social networks and how they form. A notable example includes Doreian et al. (1996), who studied this data to determine how reciprocity, transitivity and group balance, as determined by how well the actors can be partitioned, vary over time. Another such example can be found in Krackhardt and Handcock (2007), where the authors used this data to determine the significance of Heiderian triads and Simmelian triads.

We have three questions in particular we attempt to answer in this chapter regarding Newcomb's fraternity data. First, does the network stabilize, and if so, when does this happen? Second, how do subgroups form and stabilize? That is, do some or all of the actors naturally fall into a small number of groups, and if so when do these groups form? Third, is there a relationship between the popularity of an individual and the social position of that individual? We desire a unifying

framework with which we can answer all three of these questions.

Most of the past analyses of Newcomb's fraternity data have needed to simplify the data to complete their analyses. For example, Breiger et al. (1975) only considered the top two and the bottom two rankings for each individual during the final week of the study; Arabie et al. (1978) similarly used the top two and the bottom three rankings for each individual during the final week. Wasserman (1980) tried using the top four and the top eight rankings to transform the ranked network into a binary network; as may be expected, Wasserman found that the network structure is affected by the binary cutoff. Doreian et al. (1996) used in parts of their analysis only the top four rankings, and in other parts used the top four and the bottom three. More recently, Moody et al. (2005) used only the top four rankings, and Krackhardt and Handcock (2007) used the top eight. Using the methods of Chapter 2, we analyzed the fraternity data using the top four as edges and using the top eight as edges. The resulting two visualizations of the network differed considerably from each other, and both gave different visualizations than that obtained in our final model (see Figures 4.4 and 4.5 for the visualizations obtained from our proposed model). This suggests that, rather than selecting some arbitrary cutoff value we ought to try to model the full data. For more on this topic see Thomas and Blitzstein (2011). One last note is that a common theme among the analyses of Newcomb's fraternity data is that the network inference is based on ad hoc measures. While these methods can still be useful, it is clearly more preferable to have a more rigorous model and estimation method which can elicit more confidence in the estimates and quantify uncertainties.

In this chapter, we propose a latent space model for ranked dynamic network data. Our approach avoids deciding on an arbitrary cutoff for binarizing the network by appropriately modeling the rank data. Our approach also models the temporal dependence structure involved in observing the network over time. Using a latent space approach to dynamic network data allows us to obtain an intuitive visualization of the network and its evolution, giving us a better understanding of the network and allowing us to make qualitative inference. Further, by using a latent space approach we have an intuitive way to think about the network stability by linking network stability with how stable the actors' social positions are. That is, if the network is not stable, then the actors' social positions ought to vary considerably from one time point to the next; however, as the network stabilizes, the social positions in turn ought to stabilize and vary less over time. Our model allows us to measure the statistical precision of the movements of these social positions over time. Our proposed model and estimation method allows us to quantify the uncertainty of the latent positions. This uncertainty allows us to analyze group structure emergence. Finally, our model also incorporates

popularity measures, thereby capturing some of the local structure. These popularity measures, together with the latent positions, can tell us about the relationship between individual popularity and individual stability.

While the main purpose of this chapter is to analyze Newcomb's fraternity data, developing tools for rank-order network data is important in its own right. Ranked networks should inherently contain more information than binary networks. While it is true that rank-order network data is much rarer than binary data, it seems likely that this is due to a lack of analytical tools available. This work adds to the current analytical toolbox, thereby encouraging researchers to collect and analyze ranked network data.

The remainder of the chapter is as follows. Section 4.1 describes the data; Section 4.2 describes the proposed model; Section 4.3 gives the estimation algorithm; Section 4.4 gives a simulation study; Section 4.5 gives the results of analyzing Newcomb's fraternity data; Section 4.6 gives a sensitivity analysis for the initialization strategy in our estimation algorithm.

4.1 Newcomb's Fraternity Data

In 1955, seventeen unacquainted students took part in a semester long study at the University of Michigan. These students were selected in such a way that they were all unknown to each other before the study began. Thus the data on a social network would be collected over time, beginning in its most nascent state and observed as the network evolves and stabilizes to its final form. This purposeful capturing of the emergence of a social network is why this data is still of such interest nearly six decades later. For fifteen out of sixteen weeks in the semester (no responses were recorded for week 9), each student would then rank the sixteen other students from most to least favored. See Newcomb (1961) Chapter 2 for details on the selection of the students and the data acquisition process.

Thus the data come in the form of a sequence of adjacency matrices Y_t for $t = 1, \ldots, 15$. For each time point t, the i^{th} row of Y_t, denoted as $\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{int})$, is a permutation of $\{1, 2, \ldots, n-1\}$ with a 0 inserted into the i^{th} position. The rankings go, in order of most favored to least favored, from 1 to $n-1$.

4.2 Models

We first describe our proposed model in Section 4.2.1. This methodology allowed us to gain insight into the stability of the network, as well as to investigate subgroup formation and the relationship between individual stability and individual popularity. In Section 4.2.2 we review the model derived by Hoff (2011). We used this model to investigate the stability of the fraternity network over time, and while this did not detect all of the stability patterns that our proposed approach detected, it corroborated our main results on the timing of the network stability.

4.2.1 Latent Space Hierarchical Model for Ranked Dynamic Networks

Due to the lack of existing methods for our context, we develop a latent space model for handling ranked longitudinal network data with which to answer our research questions. We assume here that each actor exists within a latent space which can be interpreted as a characteristic space, or a social space. This is the underlying concept of the latent space: a smaller distance between two actors within this space corresponds to a larger probability of receiving a favorable ranking. Therefore if two nodes are far apart in the latent space we would expect them to rank each other unfavorably, whereas if two nodes are close together we would expect them to view each other quite favorably.

First is some general notation to be used throughout. Assume we have a set of actors $\mathcal{N}$ and a set of edges $\mathcal{E}$; let $n = |\mathcal{N}|$ be the fixed number of actors and T the total number of time points at which the network is observed. Often it will be more convenient to work with the ordering of $\mathbf{y}_{it}$ rather than $\mathbf{y}_{it}$ itself. We will let $\boldsymbol{\omega}_{it} = (\omega_{i1t}, \omega_{i2t}, \ldots, \omega_{i(n-1)t})$ be the $(n-1) \times 1$ vector which is the ordering of the rank vector $\mathbf{y}_{it}$ (e.g., if $\mathbf{y}_{1t} = (0, 3, 1, 4, 2)$ then $\boldsymbol{\omega}_{1t} = (3, 5, 2, 4)$). Let $\mathbf{X}_{it} \in \Re^p$ be the position vector of the i^{th} actor at time t within the p dimensional latent space. Let $\mathcal{X}_t$ be the matrix whose i^{th} row is $\mathbf{X}_{it}$. Finally, let $\boldsymbol{\Psi}$ be the vector of unknown parameters to be defined later.

We assume the actors' latent positions transition according to a Markov process, where the initial distribution is

$$\pi(\mathcal{X}_1|\boldsymbol{\Psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{i1}|\mathbf{0}, I_p/\tau_0), \tag{4.1}$$

and the transition equation is

$$\pi(\mathcal{X}_t|\mathcal{X}_{t-1}, \boldsymbol{\Psi}) = \prod_{i=1}^{n} N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, I_p/\tau_t), \tag{4.2}$$

for $t = 2, 3, \ldots, T$, where I_p is the $p \times p$ identity matrix, and $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal probability density function with mean $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$.

The precision parameters τ_t, $t = 2, \ldots, T$, give us the information we need to evaluate the stability of the network. A larger precision implies that the latent positions are moving less and therefore implies the actors' positions are more stabilized, whereas a smaller precision implies that the latent positions are moving more and therefore implies less stable social positions. The network's stability at time t ought to be in some sense smooth over time; that is, one would not expect the stability of the network at time t to be drastically different from the stability at $t-1$ and $t+1$. For this reason we further model the precision parameters τ_t, $t \geq 2$, as a random walk involving gamma distributed random variables. Specifically we have for $t \geq 2$ that

$$\tau_t = \tau_{t-1}\eta_t, \tag{4.3}$$

where $\eta_t \overset{iid}{\sim} \Gamma(\theta, \theta)$, and $\Gamma(a, b)$ indicates a gamma distribution with shape parameter a and rate parameter b. This is equivalent to having the prior

$$\pi(\tau_2, \ldots, \tau_T) \overset{\mathcal{D}}{=} \prod_{t=2}^{T} \Gamma(\tau_t|\theta, \theta/\tau_{t-1}), \tag{4.4}$$

where $\Gamma(x|a, b)$ is the gamma density function with shape a and rate b evaluated at x. With this specification, τ_t conditional on τ_{t-1} has an expected value equal to τ_{t-1} and variance equal to τ_{t-1}^2/θ. Note that τ_1 is a hyperparameter that defines the mean of τ_2 (and therefore the unconditional mean for any τ_t, $t \geq 2$).

The choice of p, the dimension of the latent space, is a topic that is beyond the scope of this chapter. As visualization of the network is a motivation for using the latent space approach to modeling networks, typically p is set to two or three. In our analysis we set $p = 2$.

Many methods, such as the temporal exponential graph model by Hanneke et al. (2010) or the stochastic actor oriented models originated by Snijders (1996), construct the dependence structure through modeling specific dependency structures; latent space approaches, such as our proposed model, assume that the dependency within the network has been induced by the latent variables. Specifically, we assume that the observed networks at differing time points are conditionally independent given the latent positions, and that the observed network at time t depends only on the latent space positions at time t. Figure 2.1 illustrates this dependence structure. We also assume that, conditioning on $(\mathcal{X}_t, \boldsymbol{\Psi})$, $\mathbf{y}_{it}$ is independent of $\mathbf{y}_{i't}$, $i \neq i'$.

We now describe the likelihood component of the model that relates the distances between the latent positions and the observed network. To this end we utilize the Plackett-Luce model for ranked data (see Plackett, 1975). The Plackett-Luce model can be thought of as drawing from a vase. Every member of the set $\{1,\ldots,n\}\setminus\{i\}$ being ranked by i has a particular proportion of the tickets with their name on it in the vase. At time t, i randomly draws a ticket and the name on the ticket determines who is ranked first, i.e., ω_{i1t}. For the second rank, i draws until a new name is drawn and then ranks that name second, ω_{i2t}. This continues until all elements in the set are ranked. Notice that the second rank is obtained according to the same probability distribution as if i was deciding the first rank with the smaller set of $n-2$ elements, i.e., $\{1,\ldots,n\}\setminus\{i,\omega_{i1t}\}$. In other words, i ranks j above k with the same probability with and without ℓ included in the set to be ranked; this condition is called Luce's Choice Axiom. It is reasonable to assume that this axiom holds; if Newcomb had only asked a subset of the students living within the fraternity to rank each other, we would not expect the resulting network to look different than a subnetwork of the full data we actually have, where all the students are included in the network. Using this framework we can write the distribution for $\mathbf{y}_{it}$ as a product of conditional probabilities given as

$$\mathbb{P}(\mathbf{y}_{it}) = \mathbb{P}(\boldsymbol{\omega}_{it}) = \prod_{j=1}^{n-1} \mathbb{P}(\omega_{ijt}|\omega_{i1t},\omega_{i2t},\ldots,\omega_{i(j-1)t}) = \prod_{j=1}^{n-1} \frac{\nu_{i\omega_{ijt}t}}{\sum_{\ell=j}^{n-1}\nu_{i\omega_{i\ell t}t}}, \tag{4.5}$$

where, following the explanation given above, ν_{ijt} corresponds to the proportion of tickets with j's name on it in i's vase at time t.

As mentioned previously, we desire that the greater the distance between actor i and actor j the smaller the probability of each giving the other a favorable ranking. Further, even within a common social circle there will still be more popular and less popular actors, and so it is important to capture this local structure in the model. Therefore it is intuitive to model the ν_{ijt}'s as functions of the latent positions and actor specific parameters. The parameterization is chosen such that

$$\nu_{ijt} = r_j \exp(-d_{ijt}), \tag{4.6}$$

where $d_{ijt} = \|\mathbf{X}_{it} - \mathbf{X}_{jt}\|$ and $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ is the vector of positive actor specific parameters constrained such that $\sum_{i=1}^{n} r_i = 1$ for model identifiability. These r_i's can be interpreted as each actor's social reach, where a larger value implies a higher probability of receiving a favorable ranking from others. Thus if an actor is generally well liked they will have a large r_i value. This parameterization is similar to that of Gormley and Murphy (2007), who applied the Plackett-Luce model to a

bipartite network, though here we also incorporate the popularity measures into the likelihood.

From (4.5) and (4.6) we have that the conditional likelihood of $(Y_1, Y_2, \ldots, Y_T)$ is

$$\mathbb{P}(Y_1, Y_2, \ldots, Y_T | \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) = \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{j=1}^{n-1} \frac{r_{\omega_{ijt}} \exp(-d_{i\omega_{ijt}t})}{\sum_{\ell=j}^{n-1} r_{\omega_{i\ell t}} \exp(-d_{i\omega_{i\ell t}t})}, \tag{4.7}$$

where $\boldsymbol{\Psi} = (\boldsymbol{r}, \tau_0, \tau_1, \tau_2, \ldots, \tau_T)$.

Further motivation for the parameterization in (4.6) is that we can consider the Thurstonian model interpretation of the Plackett-Luce model. Thurstone (1927) described the following model: For a vector of ranked data $\mathbf{y} = (y_1, y_2, \ldots, y_m)$, there is a vector of latent random variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$ and a vector of scalars $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ such that $Z_j - \mu_j \overset{iid}{\sim} F$ for some continuous distribution function F. Then $\mathbb{P}(\mathbf{y}) = \mathbb{P}(Z_{\omega_1} > Z_{\omega_2} > \cdots > Z_{\omega_m})$, where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)$ is the ordering of $\mathbf{y}$. Yellott Jr (1977) showed that the Plackett-Luce model is equivalent to the Thurstone model if and only if F is the Gumbel distribution. They further showed that if F is a Gumbel distribution with location parameter equal to zero and scale parameter equal to 1, then the relationship between the two models is that $\nu_j = \exp(\mu_j)$. Coming back to our context, we let $\mathbf{Z}_{it} = (Z_{i1t}, Z_{i2t}, \ldots, Z_{i(n-1)t})$ be a vector of latent random variables which measure how actor i regards the strength of his/her relationship with the other $n-1$ actors. We define these measures such that

$$Z_{ijt} = \mu_{ijt} + \epsilon_{ijt} \tag{4.8}$$

where $\mu_{ijt} = \log(r_j) - d_{ijt}$, $\epsilon_{ijt} \overset{iid}{\sim} F = \text{Gumbel}(-\gamma_{EM}, 1)$, and γ_{EM} is the Euler-Mascheroni constant (≈ 0.5772); the location shift is because a Gumbel(0,1) random variable has mean γ_{EM} and thus by including the location shift we set the mean of ϵ_{ijt} to be zero. Note also that the non-zero location parameter of F does not change the relationship between the Thurstonian model and the Plackett-Luce model. To see why this is so, it is necessary to recognize that the Plackett-Luce model is invariant to rescaling the ν_{ijt}'s, and hence we can rescale by $\exp(-\gamma_{EM})$. By the relationship mentioned above, we have that $\nu_{ijt} = \exp(\mu_{ijt} - \gamma_{EM})$, hence $Z_{ijt} - (\mu_{ijt} - \gamma_{EM}) \overset{iid}{\sim} \text{Gumbel}(0, 1)$, which is equivalent to (4.8). This meets our intuition that the ranking of the Z_{ijt}'s should not be affected by a location shift of F. The actual reason we desire this non-zero location parameter of F is so that we have

$$\mathbb{E}(Z_{ijt} | \mathcal{X}_t, \boldsymbol{r}) = \log(r_j) - d_{ijt}. \tag{4.9}$$

Therefore the Plackett-Luce model in (4.7) can be thought of as, for individual i at time t, obtaining

a set of variables Z_{ijt}, $j \neq i$, whose mean is determined by the social reach of the actor being ranked and by the social distance between the ranking actor and the ranked actor, which measures on a continuous scale the relationship between individual i and the rest (as perceived by i). Then the vector $\mathbf{y}_{it}$ is the ranking of the realizations z_{ijt} of Z_{ijt}.

4.2.2 Multilinear Model for Multiway Data

Hoff (2011) developed a latent space approach for analyzing multiway data, which he then demonstrated how to apply the model on dynamic network data. In particular, he applied his model to a dynamic network whose edges y_{ijt} consist of ranking the relationship on the constant set $\{-5, -4, \ldots, 2\}$. This type of ranked network is different than the fraternity data, where there is the added constraint on the response variables that the rows of the response array must be a permutation of $\{0, 1, \ldots, n-1\}$. In applying this model to the fraternity data set, we relax this extra constraint, thereby allowing the model to predict networks that violate the permutation constraint. This can be thought of as another form of simplifying the network at some cost to the information contained therein, much like, and arguably to a much lesser degree than, the information lost associated with transforming the network from weighted to binary according to some arbitrary cutoff. Hoff's model utilized an ordered probit model, which we now briefly describe within the context of the fraternity data.

Let z_{ijt} be latent variables such that $y_{ijt} = \max\{k : z_{ijt} > c_k, k \in \{1, \ldots, n-1\}\}$, where the c_k's are unknown cutoff points to be estimated. These latent variables are assumed to be normally distributed whose mean can be written as the following factor model:

$$\mathbb{E}(z_{ijt}) = \sum_{\ell=1}^{p} u_{i\ell} u_{j\ell} v_{t\ell}. \tag{4.10}$$

The p dimensional vectors $\mathbf{u}_i = (u_1, \ldots, u_p)$ are student specific vectors that can be equated to the latent positions $\mathbf{X}_{it}$ in the model of Section 4.2.1, though instead of being time dependent, in Hoff's model the temporal aspect of the data is accounted for by the p dimensional vectors $\mathbf{v}_t = (v_{t1}, \ldots, v_{tp})$. The $\mathbf{u}_i$'s can then be thought of as the time invariant latent positions of the students, and the $\mathbf{v}_t$'s can be thought of as stretching or compressing the p axes to alter the closeness of the students at different time points. There are no structural constraints placed on the $\mathbf{u}_i$'s and $\mathbf{v}_t$'s beyond the regularization that the Bayesian framework imposes via the prior distributions. Note also that the closeness between the actors is not measured via Euclidean distance, as in Section

4.2.1, but rather by the cosine of the angle between the two students, more akin to the dot product graph model (see, e.g., Young and Scheinerman, 2007). The dimension p, just as in our proposed approach, is assumed to be 2, though this is in actuality an unknown quantity. Hoff suggested using the Deviance Information Criterion (Spiegelhalter et al., 2002), though determining the optimal p could and should be a topic of future research.

The usefulness of this model within our context lies in the values of the $\mathbf{v}_t$'s. These vectors give us a good sense as to the stability of the network, as conceptualized by how much the students' social positions are changing over time. For example, if the network is completely stabilized over a set of time points $\mathcal{T}$ then the students' positions are static, and thus $\mathbf{v}_t = \mathbf{v}_s$ for $s, t \in \mathcal{T}$. If, on the other hand, the network is quite unstable, then we would expect to see these $\mathbf{v}_t$'s to vary considerably from week to week during the unstable time period.

Estimation for this model was performed by first running a Markov chain Monte Carlo (MCMC) algorithm to initialize the unknown quantities, and then applying an alternating least squares algorithm to obtain point estimates of the $\mathbf{u}_i$'s and $\mathbf{v}_t$'s. Section 4.3 gives the details on the estimation procedure for our proposed model given in Section 4.2.1.

4.2.3 Pseudo-R^2

In the context of linear regression, one can determine how well the model explains the data by using the R^2 or adjusted R^2 value. For standard ranked data, there exist some measures that are approximately equivalents (see, e.g., Marden, 1995). However, we cannot apply these measures to our context due to having each actor ranking a different set, i.e., each i ranks the set $\{1, 2, \ldots, n\} \setminus i$. For the ordinal probit model, McKelvey and Zavoina (1975) devised a goodness of fit measure; Veall and Zimmermann (1992) showed that McKelvey and Zavoina's pseudo R^2 is closest to the R^2 corresponding to the underlying continuous (latent) data. We developed a pseudo-R^2 with a similar flavor by using the Thurstonian model specification outlined at the end of Section 4.2.1. Specifically, we note that

$$(z_{ijt} - \bar{z})^2 = (\mu_{ijt} - \bar{z})^2 + \epsilon_{ijt}^2 + 2\epsilon_{ijt}(\mu_{ijt} - \bar{z}),$$

where $\mu_{ijt} = \log(r_j) - d_{ijt}$ and $\bar{z} = 1/(Tn(n-1)) \sum_t \sum_{i \neq j} z_{ijt}$. Since $\epsilon_{ijt} \overset{iid}{\sim} \text{Gumbel}(-\gamma_{EM}, 1)$, $\mathbb{E}(\epsilon_{ijt}) = 0$ and $Var(\epsilon_{ijt}) = \pi^2/6$; thus we have that

$$\begin{aligned} & \sum_{t=1}^{T} \sum_{i \neq j} \mathbb{E}(\epsilon_{ijt}^2 + 2\epsilon_{ijt}(\mu_{ijt} - \bar{z})) \\ = & \sum_{t=1}^{T} \sum_{i \neq j} \mathbb{E}(\epsilon_{ijt}^2) - 2 \sum_{t=1}^{T} \sum_{i \neq j} \mathbb{E}\left(\epsilon_{ijt} \frac{1}{Tn(n-1)} \sum_{t'=1}^{T} \sum_{i' \neq j'} (\mu_{i'j't'} + \epsilon_{i'j't'}) \right) \\ = & \frac{\pi^2}{6}(Tn(n-1) - 2). \end{aligned} \tag{4.11}$$

We can then, similarly to the method used by McKelvey and Zavoina, approximate the total sum of squares by

$$\sum_{t=1}^{T} \sum_{i \neq j} (z_{ijt} - \bar{z})^2 \approx (\hat{\mu}_{ijt} - \hat{\bar{\mu}})^2 + \frac{\pi^2}{6}(Tn(n-1) - 2), \tag{4.12}$$

where $\hat{\mu}_{ijt} = \log(\hat{r}_j) - \hat{d}_{ijt}$, $\hat{\bar{\mu}} = 1/(Tn(n-1)) \sum_t \sum_{i \neq j} \hat{\mu}_{ijt}$, and the $\hat{}$ symbol over the model parameters implies the posterior mean estimate. Therefore we define the pseudo R^2 to be

$$R^2 = \frac{\sum_{t=1}^{T} \sum_{i \neq j} (\hat{\mu}_{ijt} - \hat{\bar{\mu}})^2}{\sum_{t'=1}^{T} \sum_{i' \neq j'} (\hat{\mu}_{i'j't'} - \hat{\bar{\mu}})^2 + \pi^2(Tn(n-1) - 2)/6}. \tag{4.13}$$

This R^2 value can be interpreted to be the approximate proportion of the variability of the underlying latent variables z_{ijt} explained by the model; hence, all other things equal, we desire to have a higher R^2 value.

4.3 Estimation

Estimation is done within a Bayesian framework; thus we desire to make inference based on the posterior distribution $\pi(\mathcal{X}_1, \ldots, \mathcal{X}_T, \mathbf{\Psi} | Y_1, \ldots, Y_T)$. The strategy is to find reasonable initial estimates of the latent positions and of the model parameters, and use these estimates to initialize a Metropolis-Hastings (MH) within Gibbs Markov chain Monte Carlo. From the samples from the Markov chain we can then obtain posterior inference of the latent positions and of $\mathbf{\Psi}$.

To perform the Bayesian estimation, we first need priors on the model parameters. We use the

following:

$$\pi(\boldsymbol{r}) \stackrel{\mathcal{D}}{=} Dir(\alpha_1, \ldots, \alpha_n), \tag{4.14}$$

$$\pi(\tau_0) \stackrel{\mathcal{D}}{=} Exp(\lambda_0), \tag{4.15}$$

$$\pi(\tau_1) \stackrel{\mathcal{D}}{=} \Gamma^{-1}(\lambda_1/2, 1/2), \tag{4.16}$$

$$\pi(\theta) \stackrel{\mathcal{D}}{=} LN(\mu, \sigma^2), \tag{4.17}$$

where $Dir(\alpha_1, \ldots, \alpha_n)$ is the Dirichlet distribution, $Exp(a)$ is the exponential distribution with rate a, $\Gamma^{-1}(a/2, 1/2)$ is the inverse gamma distribution with shape $a/2$ and scale $1/2$ (this is also the inverse-χ^2 distribution with degrees of freedom a), and $LN(a, b)$ is the log-normal distribution with log-mean a and log-variance b. The Dirichlet is a natural prior for such constrained parameters as $\boldsymbol{r}$, the priors for τ_0 and τ_1 were chosen based on conjugacy, and the prior for θ was chosen to be able to put a flat prior on θ and also for ease of sampling.

4.3.1 Initialization

In a complicated hierarchical model such as ours, it is difficult to know how to reasonably choose initial values of the Markov chain estimation algorithm or how to specify the hyperparameters of the prior distributions. We attempt to address both these issues simultaneously via an approach which is similar in concept to empirical Bayes methods. That is, we use the data to determine the initial values and the hyperparameters of the prior distributions. The way in which we use the data is through a preliminary, and admittedly somewhat ad hoc, analysis of the data. Therefore we make the priors flat and uninformative where possible, otherwise we use this preliminary analysis to determine the values of the hyperparameters. In so doing we naturally obtain initial values for the Markov chain estimation algorithm.

Since the social reaches should reflect the popularity of the individuals, we initialized the social reaches as

$$r_i^{(1)} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{n} 2(n - y_{jit})}{n^2(n-1)T}, \tag{4.18}$$

where the superscript (1) denotes the initial estimate. These values account for how favorable student i was with respect to all other students over all time points. One could use $\boldsymbol{r}^{(1)}$ as the hyperparameters $\alpha_1, \ldots, \alpha_n$; in this case, however, we can make the prior distribution flat and uninformative by setting these hyperparameters all equal to one. This also has the beneficial effect

of reducing the computational complexity of the algorithm.

To find the initial latent positions we used classic multidimensional scaling (MDS) at each time point. To implement this, we first needed a dissimilarity matrix for each time point. We constructed this by setting

$$d_{ijt} \propto \frac{r_j^{(1)}}{n - y_{ijt}} + \frac{r_i^{(1)}}{n - y_{jit}}. \tag{4.19}$$

The logic behind this choice is that the more favorable i and j rank each other, the closer they ought to be in the latent space. The latent social positions in our latent space model account for popularity, however, and so we use the initial values of the social reaches $\boldsymbol{r}^{(1)}$ to determine d_{ijt}. The idea is that even if i gives j a favorable ranking, this may not imply that i and j are particularly close if j has a large social reach. If, however, i gives j a favorable ranking and j has a very small social reach then this implies that i and j should be very close together in the latent social space.

With the T dissimilarity matrices computed, we can then implement MDS to obtain initial latent positions. In many contexts it would be more appropriate to initialize using the generalized multidimensional scaling derived by Sarkar and Moore (2005), which implements MDS while accounting for the longitudinal aspect of the dissimilarity matrices. However, this method implicitly assumes that $\tau_2 = \tau_3 = \cdots = \tau_T$, which we do not assume here; thus we have used a simpler MDS approach to initialize the latent positions, i.e., we use MDS on each of the T dissimilarity matrices. After each dissimilarity matrix has been used to embed the actors within a p-dimensional latent space, we used a Procrustes transformation to orient the latent positions at time t as closely as possible to those at time $t-1$. The Procrustes transformation finds a set of rotations, reflections and translations to minimize the difference between a given matrix and some target matrix (see, e.g., Borg, 2005). Lastly, we needed to know how to scale the latent positions. To this end we maximized the likelihood using a simple line search to find

$$c_0 = \underset{c}{\operatorname{argmax}}\ \pi(Y_1, \ldots, Y_T | c\mathcal{X}_1^*, \ldots, c\mathcal{X}_T^*, \boldsymbol{r}^{(1)}),$$

and then we set $\mathcal{X}_t^{(1)} = c_0 \mathcal{X}_t^*$ for $t = 1, \ldots, T$, where $\mathcal{X}_t^*$ is the t^{th} latent positions found by using MDS.

The prior mean of τ_0 and the initial estimate $\tau_0^{(1)}$ was computed as

$$\left[\frac{1}{np} \sum_{i=1}^{n} \|\mathbf{X}_{i1}^{(1)}\|^2 \right]^{-1}. \tag{4.20}$$

We then set $\lambda_0 = 1/\tau_0^{(1)}$, thereby matching the prior expected value of τ_0 to $\tau_0^{(1)}$. Similarly, for $t \geq 2$, $\tau_t^{(1)}$ was computed as

$$\left[\frac{1}{np} \sum_{i=1}^{n} \|\mathbf{X}_{it}^{(1)} - \mathbf{X}_{i(t-1)}^{(1)}\|^2 \right]^{-1}. \tag{4.21}$$

We set $\tau_1^{(1)}$ to equal $\tau_2^{(1)}$. Matching the prior expected value of τ_1 to equal $\tau_1^{(1)}$ implies setting $\lambda_1 = 2 + 1/\tau_1^{(1)}$. Looking at (4.3), we see that the variance of η_t $(= \tau_t/\tau_{t-1})$ equals $1/\theta$. Therefore we can set the initial estimate $\theta^{(1)}$ equal to the inverse of the sample variance of $\{\tau_t^{(1)}/\tau_{t-1}^{(1)}, t \geq 2\}$. We then set $\mu = \log(\theta^{(1)})$ and set σ^2 to be some large value, thereby making the prior flat.

We checked the sensitivity to this initialization scheme on our analysis of the fraternity data. We checked this sensitivity by choosing two alternative methods of initialization, each of which reflects some incorrect concept behind the latent space model (a misinterpretation of the latent positions and an assumption of constant network stability over time). In neither case did the conclusions based on the samples from the posterior, which will be discussed in Section 4.5, change. The details of this sensitivity analysis are given at the end of the chapter.

4.3.2 Posterior Sampling

To sample from the posterior distribution, we use a MH within Gibbs sampling scheme. For this algorithm we need the full conditional distributions. For the latent positions these are given as

$$\begin{aligned} &\pi(\mathbf{X}_{it}|\cdot) \\ &\propto \begin{cases} \pi(Y_1|\mathcal{X}_1, \mathbf{\Psi}) N(\mathbf{X}_{i1}|\mathbf{0}, I_p/\tau_0) N(\mathbf{X}_{i2}|\mathbf{X}_{i1}, I_p/\tau_2) & \text{if } t = 1 \\ \pi(Y_t|\mathcal{X}_t, \mathbf{\Psi}) N(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, I_p/\tau_t) N(\mathbf{X}_{i(t+1)}|\mathbf{X}_{it}, I_p/\tau_{t+1}) & \text{if } 2 \leq t < T \\ \pi(Y_T|\mathcal{X}_T, \mathbf{\Psi}) N(\mathbf{X}_{iT}|\mathbf{X}_{i(T-1)}, I_p/\tau_T) & \text{if } t = T, \end{cases} \end{aligned} \tag{4.22}$$

and for the parameters are given as

$$\pi(\boldsymbol{r}|\cdot) \propto \pi(Y_1, \ldots, Y_T|\mathcal{X}_1, \ldots, \mathcal{X}_T, \boldsymbol{\Psi}) \tag{4.23}$$

$$\pi(\tau_2, \ldots, \tau_T|\cdot) = \prod_{t=2}^{T} \Gamma\Big(\tau_t|\theta + \frac{np}{2}, \frac{\theta}{\tau_{t-1}} + \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|^2\Big) \tag{4.24}$$

$$\pi(\tau_0|\cdot) \stackrel{\mathcal{D}}{=} \Gamma\Big(1 + \frac{np}{2}, \lambda_0 + \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{X}_{i1}\|^2\Big) \tag{4.25}$$

$$\pi(\tau_1|\cdot) \stackrel{\mathcal{D}}{=} \Gamma^{-1}\Big(\frac{\lambda_1}{2} + \theta, \frac{1}{2} + \theta\tau_2\Big) \tag{4.26}$$

$$\pi(\theta|\cdot) \propto \left[\prod_{t=2}^{T} \Gamma(\tau_t|\theta, \theta/\tau_{t-1})\right] \cdot LN(\theta|\mu, \sigma^2). \tag{4.27}$$

The algorithm is

0. Set the initial values of the latent positions and parameters as given in Section 4.3.1.
1. For $t = 1, 2, \ldots, T$ and for $i = 1, 2, \ldots, n$, draw $\mathbf{X}_{it}$ from (4.22) via MH.
2. Draw τ_0 from (4.25).
3. Draw τ_1 from (4.26).
4. For $t = 2, \ldots, T$, draw τ_t from its conditional distribution in (4.24).
5. Draw θ from (4.27) via MH.
6. Draw $\boldsymbol{r}$ from (4.23) via MH.
Repeat steps 1-6.

Regarding the proposal distributions, $\mathbf{X}_{it}$, β_{IN}, and β_{OUT} can come from a symmetric proposal (e.g., normal random walk). Because of the constraint on $\boldsymbol{r}$, a Dirichlet proposal is suggested for the radii, which also will be an asymmetric proposal. Suggested parameters for this Dirichlet proposal are $\kappa\boldsymbol{r}^{curr}$, where $\boldsymbol{r}^{curr}$ are the current values for $\boldsymbol{r}$ and κ is some large value (e.g., we set $\kappa = 10,000$).

One final note is that, as is the case for any such latent space model, the posterior is invariant under rotations, reflections and translations of the latent positions $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$. Hence after each iteration of steps 1-6, a Procrustean transformation will be performed on the n trajectories; that is, the transformation is performed on the $nT \times p$ matrix $(\mathcal{X}_1', \mathcal{X}_2', \ldots, \mathcal{X}_T')'$. In our context, the target matrix is chosen to be constructed from the first MCMC draw of the latent positions after the burn-in. In so doing we find a rotation matrix A such that for any i and t, $\mathbf{X}_{it}^{(\ell)} = A'\mathbf{X}_{it}^{*}$, where $\mathbf{X}_{it}^{(\ell)}$ is the stored latent positions for the ℓ^{th} iteration and $\mathbf{X}_{it}^{*}$ is the newly drawn latent positions.

4.4 Simulation Study

Two sets of 20 simulations were run, where the simulated data in the first set was from a properly specified model, i.e., the data came from the model of Section 4.2.1, and the second simulated data was from a misspecified model, the details of which will be explained shortly. In each simulation we desired to evaluate the correlation between the estimated and the true popularity measures and the relationship between the estimated and true τ_t's. For each simulation we drew 100,000 MCMC samples, and on a UNIX machine with a 2.40 GHz processor the computation time averaged close to one hour.

For each simulation we set $n = 17$, $T = 15$ and $p = 2$. We set $\tau_0 = 3.16$, $\tau_1 = 5.57$, and $\theta = 4.50$, which are the posterior means from the analysis of the fraternity data; these values were chosen to make the simulated data similar to the real data set we analyzed. We then drew $\boldsymbol{r}$ from $Dir(1, \ldots, 1)$ and for $t \geq 2$ drew τ_t from $\Gamma(\theta, \theta/\tau_{t-1})$. For the simulations from the correctly specified model we then drew $\mathcal{X}_1, \ldots, \mathcal{X}_T$ according to (4.1) and (4.2) in the main text. For the 20 simulations from the misspecified model we generated the data in the following way. We drew the initial latent positions from a mixture of normals with three components. The means of the three normals were drawn from $N(\mathbf{0}, I_p/\tau_0)$ and the covariances were all set to equal $I_p/(5\tau_0)$. Then for $i = 1, \ldots, n$, $\mathbf{X}_{i1}$ was drawn from this mixture of normals. If we write the transition equation as $\mathbf{X}_{it} = \mathbf{X}_{i(t-1)} + \boldsymbol{\epsilon}_{it}$, then the correctly specified model would have the correlation matrix of $(\boldsymbol{\epsilon}_{i2}, \ldots, \boldsymbol{\epsilon}_{iT})$ equal $I_{p(T-1)}$. However, we set

$$corr(\boldsymbol{\epsilon}_{it}, \boldsymbol{\epsilon}_{is}) = \begin{bmatrix} 1 & \rho^{|t-s|} \\ \rho^{|t-s|} & 1 \end{bmatrix}$$

where ρ was drawn from a uniform distribution with support over $(0.25, 0.75)$. Finally, the data $Y_1, \ldots, Y_T$ were drawn according to the model described in Section 4.2.1 (using the Thurstonian interpretation of the model proved most convenient).

Figure 4.1 gives the boxplots of the correlation between the true social reaches $\boldsymbol{r}$ and the posterior means of the social reaches $\widehat{\boldsymbol{r}}$. From this we see that these popularity measures are estimated well, with the estimates corresponding to the correctly specified model performing similarly to, though somewhat better than, those corresponding to the misspecified model. Figure 4.2 shows the scatterplot of the true τ_t's vs. the posterior mean estimates. From this we see that we can be reasonably certain that our estimates of the τ_t's will be indicative of the truth. It is also important to note that while we obtain reasonable estimates of the τ_t's, there is a clear pattern in Figure 4.2

which implies that the variance of the estimates, perhaps unsurprisingly, increases as the true τ_t's values increase. Indeed, this pattern is reflected in the estimates of the fraternity data, as we have seen.

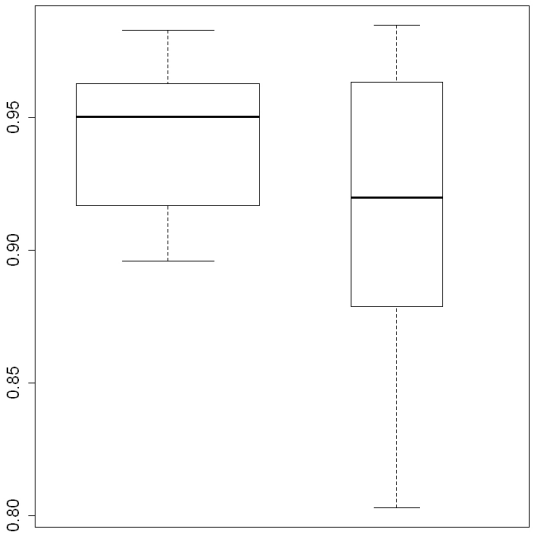


Figure 4.1: Correlation between true $\boldsymbol{r}$ and $\widehat{\boldsymbol{r}}$ values for the simulations. The wide boxplot corresponds to the data generated from the correctly specified model, the narrow boxplot to the data generated from the misspecified model.

4.5 Results

We applied our method to Newcomb's (1961) fraternity data . We let the MCMC algorithm run for 250,000 iterations, including a burn in period of 50,000 iterations. Figure 4.3 gives the trace plots for selected parameters, namely θ and τ_t for $t = 0, 1, 2, 9, 15$. From this we see that the MCMC algorithm converges. The hyperparameters $\alpha_1, \ldots, \alpha_{17}$ were all set to 1, σ, the log standard deviation of $\pi(\theta)$, was set to equal 5, and all other hyperparameters were chosen as described in Section 4.3.

The pseudo-R^2 value was 0.622 (this was equal up to three decimal places of the mean pseudo-R^2 values obtained from analyzing 20 data sets simulated from the model of Section 4.2.1 whose parameters were set to be equal to those learned from this data set; see the Supplementary Materials for details on the simulation study). As this value approximates the amount of the variation in the underlying process explained by our model, we get some sense as to the noisiness of the data. Our model has explained more than half of the variation of the latent process, though there is still some inherent unexplained noise in the network data. Figures 4.4 and 4.5 give the posterior means of

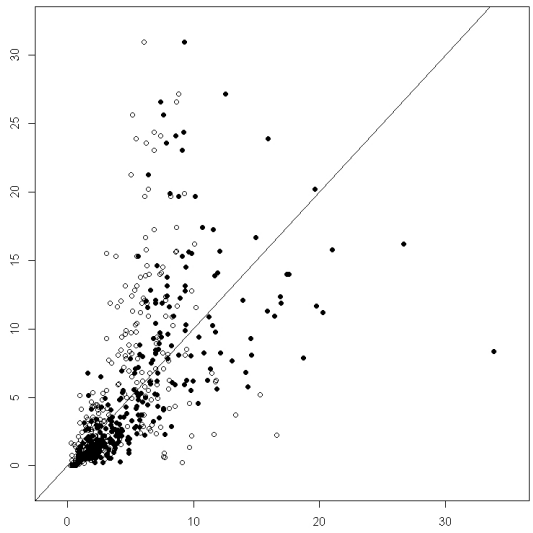


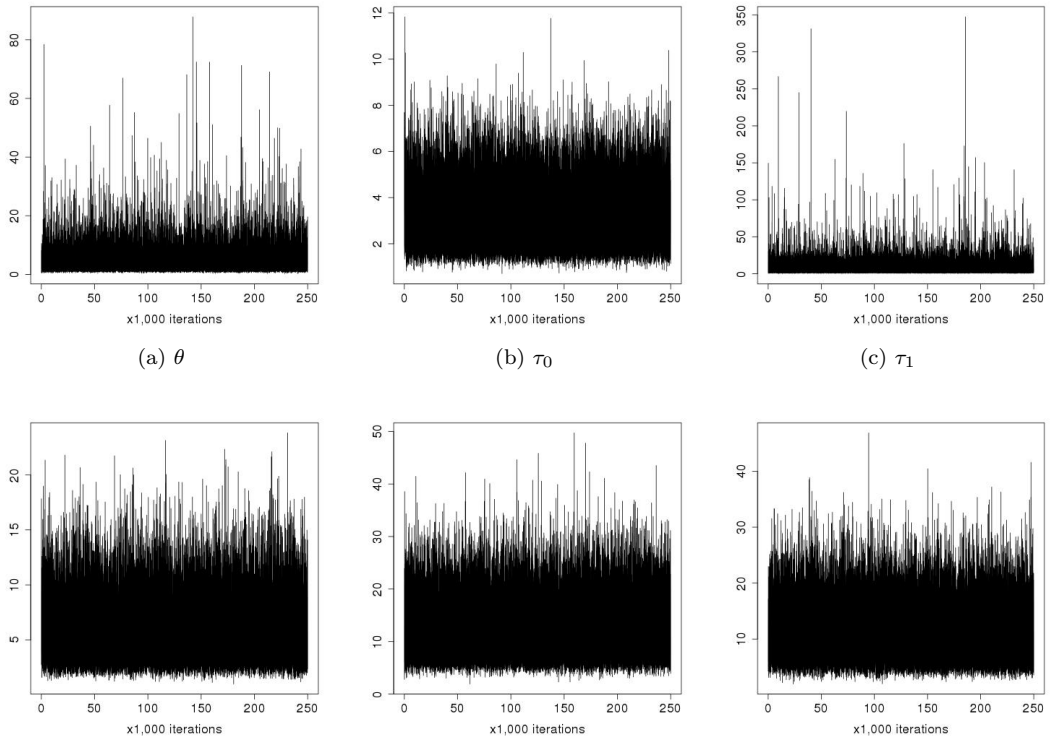
Figure 4.2: True values of $\tau_2, \ldots, \tau_T$ correspond to the vertical axis, posterior means $\widehat{\tau}_2, \ldots, \widehat{\tau}_T$ correspond to the horizontal axis. Solid circles indicate that the model was correctly specified, and the hollow circles indicate that the model was misspecified.

the latent trajectories of the 17 students through the 15 weeks of the study. From this we get a better understanding of what the network looks like, what groupings exist, and which actors find their social positions early and which find their social positions late. The details are given in the following sections.

4.5.1 Network Stability

Newcomb (1961) and Nakao and Romney (1993) both measured the stability of the network by comparing each individual's rankings from week to week. Newcomb claimed that the stability sharply increases in the first three weeks, and the network is essentially stable after this point. Nakao and Romney claimed that the network is stable after week five. Much more recently, Krivitsky and Butts (2012) extended the exponential random graph for ranked network data. Krivitsky and Butts used this model to analyze Newcomb's fraternity data, determining the stability of the network through ranking inconsistencies, showing that according to this measure the stability of the network increases over time with a decrease at week 15. We wish to use our model to conduct a formal analysis, giving quantitative answers to how the stability of the network evolves. In so doing we verify the general trends discovered earlier, as well as discovering a new pattern in the stability of the network.

In a latent space approach to modeling the network, network stability is considered to be how constant the actor's social positions become. Before applying our model from Section 4.2.1 for ranked



(a) θ (b) τ_0 (c) τ_1

(d) τ_2 (e) τ_9 (f) τ_{15}

Figure 4.3: Trace plots for select parameters corresponding to the analysis of Newcomb's fraternity data.

dynamic networks, we first use Hoff's multilinear model to obtain a visualization of the evolution of the stability of the network. Figure 4.6 gives the resulting figures from the analysis. Keep in mind that the interpretation of the latent positions from Hoff's model is different than that of the latent positions from our proposed method, in that a smaller angle, not a smaller distance, between the actors increases the probability of a favorable ranking. The plots of $\mathbf{v}_1$ and $\mathbf{v}_2$ give the scalar time effects (which stretch the g^{th} axis at time t if $v_{tg} > 1$ or contract if $v_{tg} < 1$, $g = 1, 2$) for each of the two dimensions in the latent space. It is these two plots which indicate how much the latent positions are moving over time. During the first six weeks we see from Figure 4.6 that the axes are being scaled by different (increasing) factors, whereas from week six to the end of the study the axes are being scaled by a nearly constant factor. This implies that for the first six weeks the latent positions are varying and thus the network is not stabilized, but after week six the latent positions are mostly static and hence the network is stable. This result implies that both Newcomb and Nakao and Romney underestimated the time at which the network stabilised.

We next apply our proposed model for ranked dynamic networks to obtain more quantitative

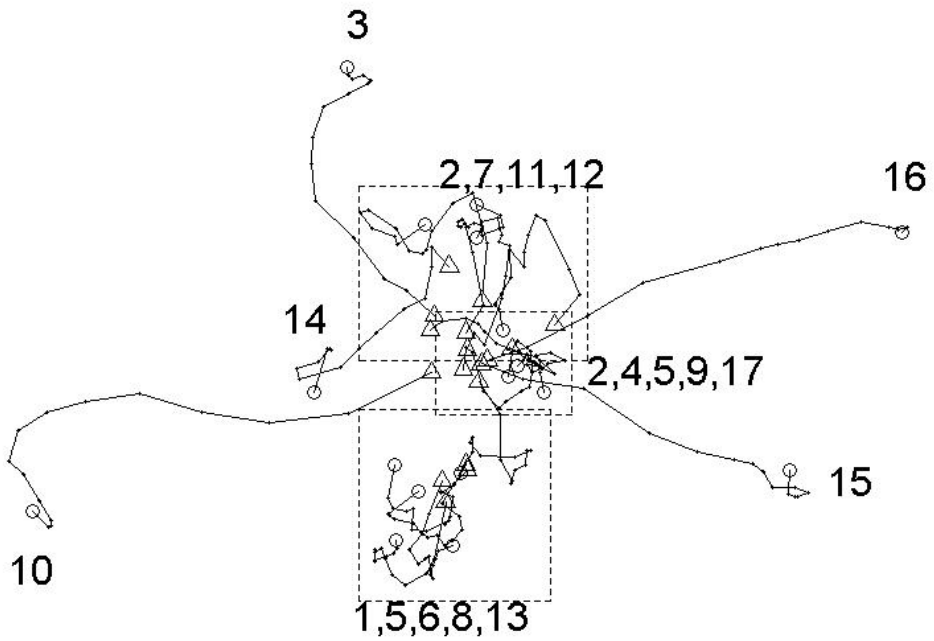
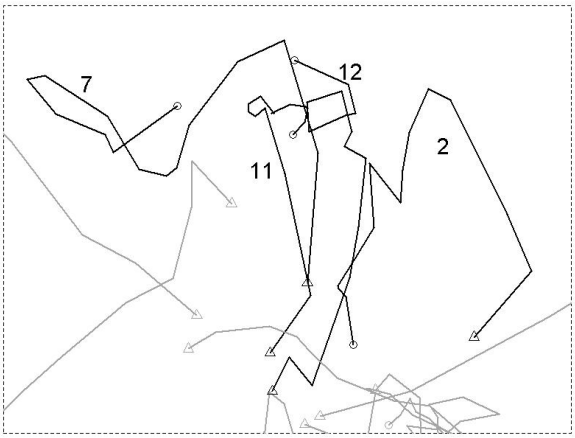
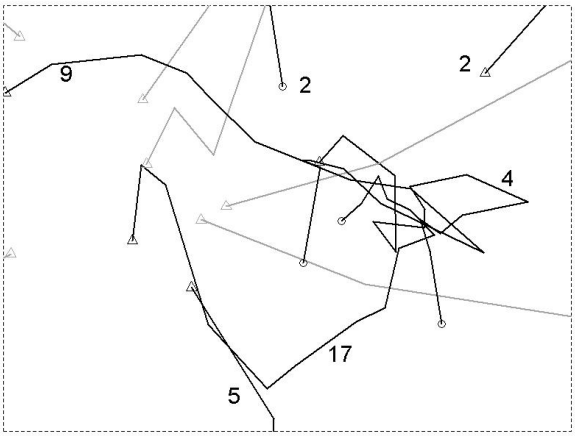


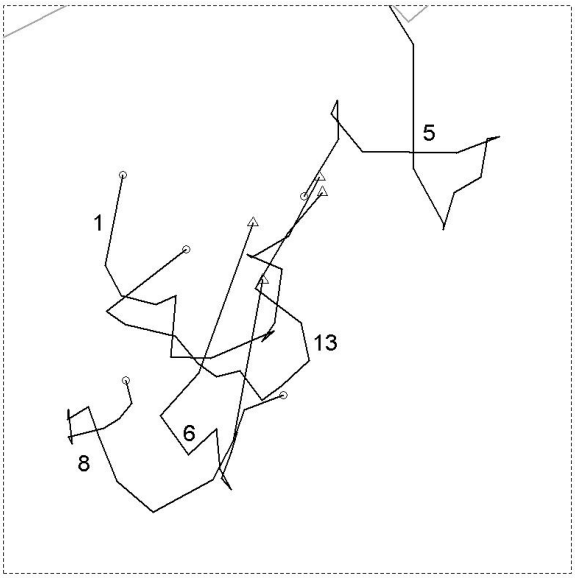
Figure 4.4: Posterior means of the latent positions of the students in Newcomb's fraternity study. Triangles indicate the beginning of the trajectory (week 1) and circles indicate the end of the trajectory (week 15). When students' trajectories are obfuscated by each other, the students forming the group is given, rather than labeling each individual trajectory.



(a)



(b)



(c)

Figure 4.5: Latent positions of the students; (a)–(c) zoom in on the top, central and bottom dashed boxes respectively of Figure 4.4, where the obfuscated student trajectories, in dark, are labeled.

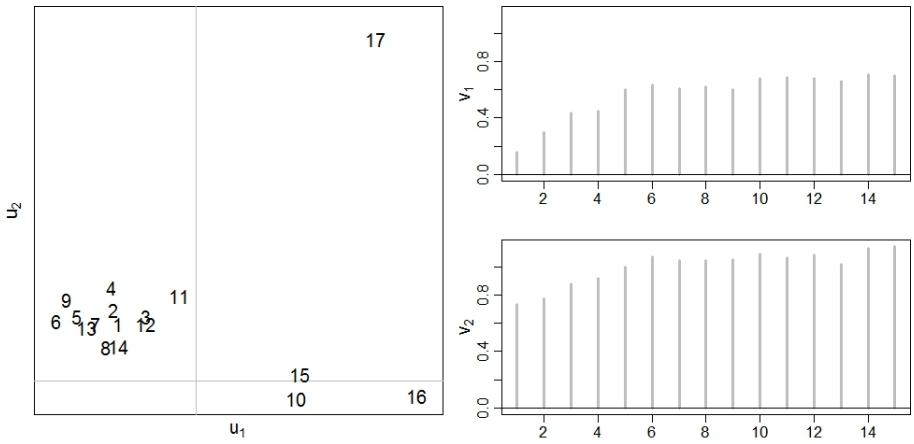


Figure 4.6: Application of the multilinear model to the fraternity data. The latent positions of the actors are given by $\mathbf{u}_1$ and $\mathbf{u}_2$, and the time effects are given by $\mathbf{v}_1$ and $\mathbf{v}_2$.

results on the evolution of the network stability. Again, in using a latent space approach to modeling the network we consider the network stability to be equivalent to the stability of the actors' social positions. While the stability of the social positions is an intuitive way of measuring network stability, we can understand even better how the actors' social positions are accurate measures of stability by considering the fact that the variability of the latent positions directly affects the variability over time of the probability distribution of the rankings. Thus the stability of the network can be characterized in the proposed model by the precision variables τ_t, $t \geq 2$.

Our method gives both quantitative point estimates of the network stability as well as uncertainty estimates. Figure 4.7 gives the posterior means of the τ_t's and their 95% credible intervals based on the posterior samples. The higher the precision the more stable the network. The credible intervals in Figure 4.7 give a good idea as to what values the precision parameters may take, but the intervals cannot be directly compared, i.e., they are not simultaneous credible intervals. Table 1 is given to compare the τ_t's directly. The r^{th} row c^{th} column entry of this table is the posterior probability that $\tau_{c+1} > \tau_{r+1}$. From Figure 4.7 we can see the overall pattern of the stability of the network over time, and by using Table 1 we can have more confidence in our inference about the pattern in stability of the network. For example, looking at Table 1 we see that the posterior probability that $\tau_7 > \tau_6$ is 0.85, that $\tau_7 > \tau_5$ is 0.96, and that $\tau_7 > \tau_2, \tau_3, \tau_4$ each is 0.99, verifying the pattern we see in Figure 4.7 that the network transitions from week 6 to a more stable form in week 7.

Our results echo that found by using Hoff's multilinear model in that the first few weeks are particularly unstable until around week 6. Our model also captures the behavior mentioned by Krivitsky and Butts that the network had a downturn of stability heading into the final week of the

semester, which is not present in the output of Hoff's model. We see that even though there is a drastic downturn in network stability, the stability still seems to be above that found in the first five weeks (the probability of the stability being higher in week 15 ranges from 0.75 to 0.91). This artifact in the data may be due to, as Nakao and Romney suggest, the students becoming distracted during the final week of the semester and of the experiment.

We also detect a new phenomenon in the stability of the network currently unremarked upon by previous analyses of the fraternity data. From Figure 4.7 we can see that there is a minor decrease in network stability transitioning from week 8 to week 9. From Table 1 we see that there is a posterior probability of 0.79 that there is a decrease in stability compared to the previous week, though only a 0.42 probability of having less stability than that observed in week 6 and 0.16 or smaller probability of having less stability than that observed in weeks 1-5. This is exactly the time when one week of data was not recorded, and one can only conjecture what occurred during this time to decrease the network stability.

The emerging stability within the network implies that the students are making progressively smaller movements over time within the social space. Looking at Figure 4.4, the movements of actors 3, 10, 14, 15 and 16 move progressively towards the edges of the social space, but this is not the same concept as what has been discussed in regards to network stability. In fact, using our notions of stability, a network could in theory be considered stable while some nodes are moving continually in one direction; in our context we do not in fact see this, but rather most of the actors seem to reach their social position, wherever it may be, and maintain it.

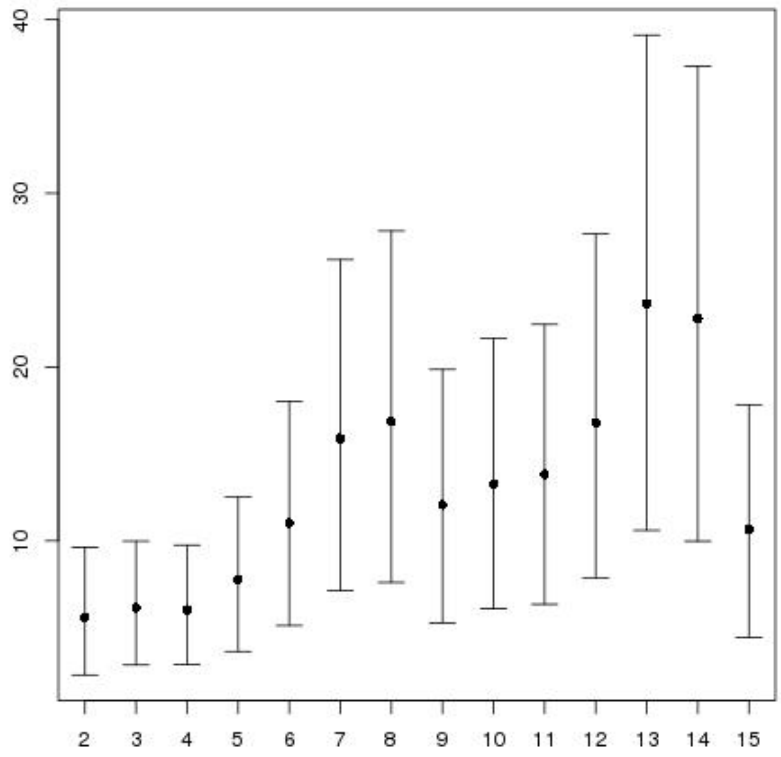


Figure 4.7: Estimates of the precision parameters τ_t, $t = 2, \ldots, 15$, for the fraternity data. 95% credible intervals are also given.

Table 4.1: Posterior probabilities that $\tau_{c+1} > \tau_{r+1}$ corresponding to the r^{th} row and c^{th} column of the table.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0	0.61	0.58	0.78	0.93	0.99	0.99	0.95	0.97	0.98	0.99	1.00	1.00	0.91
3		0	0.49	0.72	0.91	0.99	0.99	0.93	0.96	0.97	0.99	1.00	1.00	0.88
4			0	0.76	0.93	0.99	0.99	0.94	0.97	0.97	0.99	1.00	1.00	0.89
5				0	0.84	0.96	0.96	0.84	0.89	0.91	0.96	1.00	0.99	0.75
6					0	0.85	0.85	0.58	0.67	0.70	0.84	0.96	0.95	0.47
7						0	0.58	0.28	0.35	0.38	0.56	0.82	0.80	0.20
8							0	0.21	0.29	0.33	0.50	0.78	0.75	0.16
9								0	0.61	0.63	0.78	0.93	0.92	0.39
10									0	0.55	0.71	0.91	0.89	0.31
11										0	0.71	0.90	0.88	0.28
12											0	0.83	0.77	0.16
13												0	0.48	0.04
14													0	0.03
15														0

4.5.2 Subgroups

From early on, researchers have attempted to find well connected subgroups within the overall network; see, e.g., Breiger et al. (1975) and Arabie et al. (1978). These efforts at what is referred to as community detection were aimed more at demonstrating a new methodology than obtaining any real meaning from the data, making very limited use of the richness in the data. However, Nakao and Romney (1993) performed a more serious analysis of Newcomb's fraternity data. The authors embedded Newcomb's fraternity data into a Euclidean space using an ad hoc method of comparing the correlation between actors' rankings and then applying MDS on the resulting similarity matrices; thus two actors would be close together in this space if they ranked the other actors similarly. Nakao and Romney then used this visualization to determine two subgroups consisting of actors $(1, 5, 6, 8, 13)$ in group one and $(2, 4, 7, 9, 11, 12, 17)$ in group two. After fitting our model for ranked dynamic networks, we see similar groupings in Figures 4.4 and 4.5. Nakao and Romney's group one seems to be identically grouped in our visualization, and group two is similarly grouped in our visualization with the exception that actors 4, 9 and 17 seem to form a third, more central, group which bridges group one and group two. Also, actors 5 and 2 seem to bridge the central group with group one and group two respectively.

The remaining actors, $(3, 10, 14, 15, 16)$, were labeled by Nakao and Romney as "outliers," by which the authors meant that these actors did not find their social positions during the course of the study. Their visualization has these five actors moving all over the latent space. However, in our visualization we see that rather than roaming aimlessly, these nodes simply moved farther towards the edge of the social space; this implies deteriorating friendships rather than allegiance swapping. Moody et al. (2005) were also able to discover this move towards the edge of the social space in actors 10 and 15 through their visualization methods.

The question remains as to when these subgroups formed. Nakao and Romney simply state that the subgroups form early in the study and remain stable afterwards. Using blockmodeling on the binarized network at week 15 to obtain blocks and comparing the proportion of edges between and within blocks at each time point, Arabie et al. (1978) claimed that the subgroup formation became stable at week 5. By partitioning the actors at each time point according to their top four rankings and bottom three rankings and then comparing the partitions over time, Doreian et al. (1996) claimed that the subgroup formation reached a stable form at week 7. These methods while all somewhat reasonable are nevertheless rather ad hoc and typically do not make full use of the ranked data.

By using a formal statistical framework to model the fraternity data, we obtain what the other methods do not have: uncertainty estimates. We utilize these uncertainty estimates to evaluate the timing of the subgroup formation. From the MCMC output, we can obtain Bayesian credible regions for the latent positions. If the subgroups have not yet formed we would expect to see these credible regions to be overlapping considerably, i.e., groups of actors are not well separated with high probability, whereas after the subgroups have stabilized we would expect to see overlap in credible regions only in actors belonging to the same subgroup, i.e., low probability that actors of two differing subgroups would be near.

Figure 4.8 gives, for $t = 1, 4, 6, 7, 9, 10$, the latent position plots with the 95% posterior probability regions, using a bivariate density estimation to estimate the boundaries of the regions. At week 1 we see that there is no subgroup structure at all. However, by week 4 we see that the top and bottom subgroups have begun to form and are already separated, and also that student 10 and to a lesser degree student 15 are already making their way to the edge of the social space. At week 6 all three subgroups have started to separate, and at week 7 this structure becomes even more clear. At week 6 we also see that actor 5 is bridging the bottom and middle subgroups and that actors 3, 10, 15 and 16 have departed from the three main subgroups; at week 7 actor 14 also seems to depart from the three subgroups. At weeks 9 and 10 the subgroup structure is quite clear, with the final change taking place; this change is due to actor 2 becoming a bridge between the top and middle subgroups. Although there are some small local changes, it is this structure at week 10 that is in place for the remainder of the study. We have indicated the top subgroup by a light solid gray shading, the bottom subgroup by a dark solid gray shading, the central subgroup by speckling, the outlying students by horizontal stripes, and the bridging students by diagonal stripes. Note that at each time point there may be several students who do not belong to any subgroup, in which case there is no shading.

Corresponding to Figure 4.8, we have constructed T adjacency matrices, $A_1, \ldots, A_T$, where the i^{th} row and j^{th} column of A_t equals one if actors i and j have overlapping credible regions at time t. Thus if two subgroups have separated, we would expect to see blocks of ones along the diagonal corresponding to the closeness of the subgroups and blocks of zeros in the off-diagonals corresponding to the separation between the subgroups. While we have not experimented on large data, it would be interesting to develop algorithms which would take these matrices and find the well separated subgroups. It may be possible to utilize standard clustering methods on these adjacency matrices to determine the subgroups. For example, on the fraternity data we performed k-means clustering on

the $n \times (Tn)$ matrix $(A_1, A_2, \cdots, A_T)$, which accurately returned the subgroups discussed previously (assigning the two bridging students 2 and 5 to the central and bottom subgroup respectively).

Tables 4.2 through 4.7 give the adjacency matrices corresponding to Figure 4.8 in the paper. While still slightly variable over time, these adjacency matrices help verify the timing of the formation of the subgroups within the network. The shadings and lines drawn are to aid the reader in viewing the subgroups, but some explanation is needed. Keep in mind throughout that not all students belong to one of the distinct subgroups at all time points; such a student is implied by a white background. The symmetric blocks that are a solid shaded gray indicate the subgroups that exist; the (possibly asymmetrical) off diagonal blocks that show whether these subgroups are separated are indicated by stippled gray shading. The outlying students, when they are moving towards the edge of the social space, are grouped together and are also indicated by stippled gray shading. At time 6 and beyond, when we desire to point out the bridging actors we use dotted lines rather than solid lines to separate the bridging actors from the distinct subgroups that are being bridged. Thus, for example, at week 6 we see that there is a top subgroup consisting of students 7, 11 and 12 (corresponding to the lightly shaded students in Figure 4.8 of the paper), a central subgroup consisting of students 4 and 17 (corresponding to the speckled students in Figure 4.8), a bottom subgroup consisting of students 1, 6, 8 and 13 (corresponding to the darkly shaded students in Figure 4.8), student 5 bridges the central and bottom subgroup (corresponding to the diagonally striped student in Figure 4.8), and students 3, 10, 15 and 16 are outlying (corresponding to the horizontally striped students in Figure 4.8).

Table 4.2: Subgroup adjacency matrix at $t = 1$.

		Top			Bridge		Central		Bridge		Bottom					Outlying			
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
	11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	12	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Bridge	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Central	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bridge	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Bottom	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	6	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
	8	0	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1
	13	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
Outlying	14	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1
	3	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
	10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4.3: Subgroup adjacency matrix at $t = 4$.

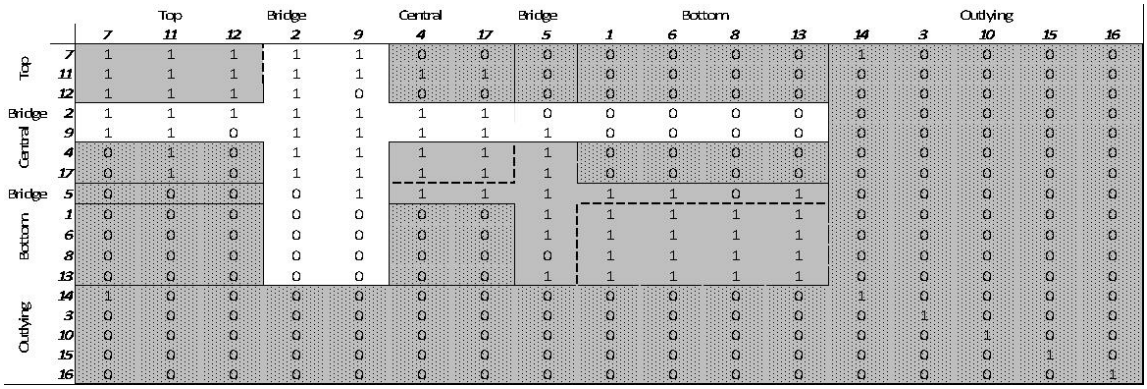
		Top			Bridge		Central		Bridge		Bottom					Outlying			
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
	11	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0	0	1
	12	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0
Bridge	2	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	1
Central	9	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0
	4	0	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1
	17	0	1	0	0	1	1	1	1	1	0	0	1	1	0	0	1	0
Bridge	5	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	1	0
Bottom	1	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0
	6	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	13	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0
Outlying	14	1	1	1	1	1	1	1	0	0	0	0	0	1	1	0	0	1
	3	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	15	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1
	16	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	1	1

Table 4.4: Subgroup adjacency matrix at $t = 6$.

		Top			Bridge		Central		Bridge		Bottom					Outlying			
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
	11	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0
	12	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
Bridge	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Central	9	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0
	4	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	17	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0
Bridge	5	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0
Bottom	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0
	6	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
	13	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
Outlying	14	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
	3	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

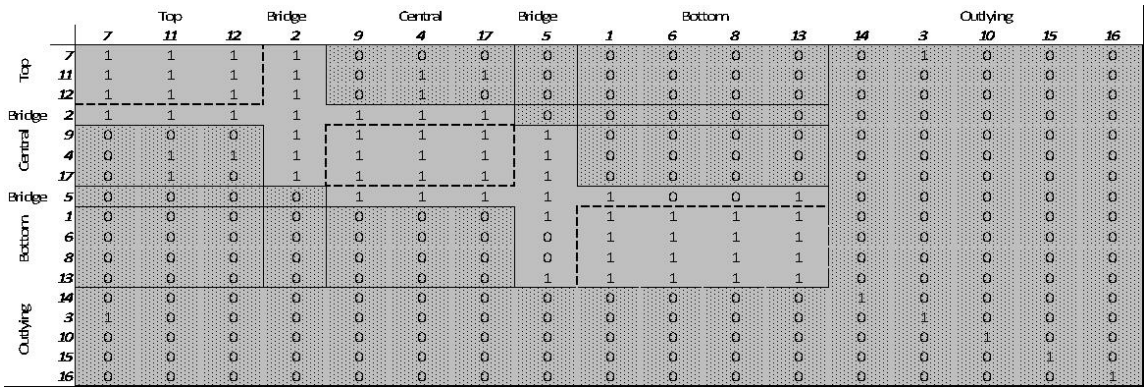
It seems reasonable to expect that not all groups would form and stabilize at the same time, and this is what we see here. The top and bottom groups form first around week 4, the third group forms at week 6 or 7. Meanwhile, over the first half of the semester certain individuals fail to join a subgroup, moving farther toward the edge of the social space.

Table 4.5: Subgroup adjacency matrix at $t = 7$.



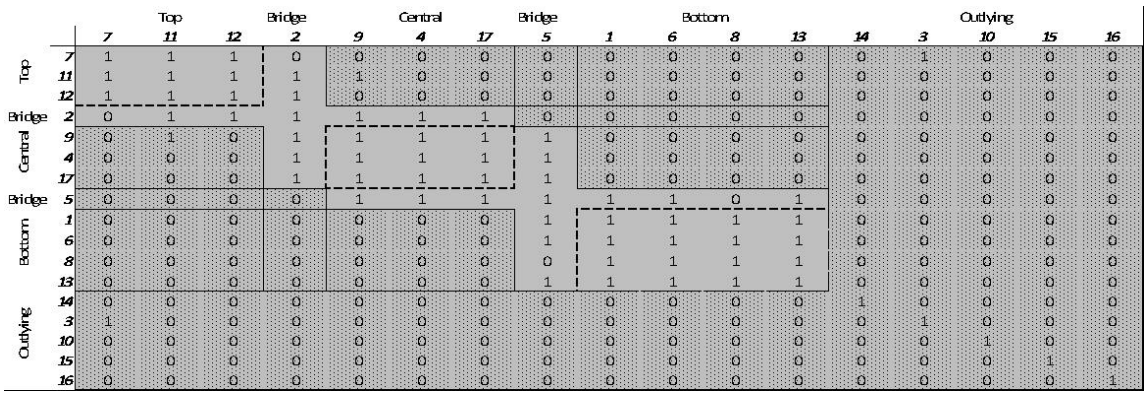
		Top			Bridge		Central			Bridge		Bottom					Outlying				
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0
	11	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	12	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Bridge	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Central	9	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	17	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Bridge	5	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0
Bottom	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	6	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
	13	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
Outlying	14	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 4.6: Subgroup adjacency matrix at $t = 9$.



		Top			Bridge		Central			Bridge		Bottom					Outlying				
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
	11	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0
	12	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Bridge	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Central	9	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	17	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Bridge	5	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	0	0
Bottom	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
	13	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
Outlying	14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 4.7: Subgroup adjacency matrix at $t = 10$.



		Top			Bridge		Central			Bridge		Bottom					Outlying				
		7	11	12	2	9	4	17	5	1	6	8	13	14	3	10	15	16
Top	7	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	11	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	12	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Bridge	2	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Central	9	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	17	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Bridge	5	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	0
Bottom	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	6	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
	13	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
Outlying	14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

(a) $t = 1$ (b) $t = 4$

(c) $t = 6$ (d) $t = 7$

(e) $t = 9$ (f) $t = 10$

Figure 4.8: Plots of 95% credible regions for latent positions. The overlap/nonoverlap of the credible regions gives information on the timing of the subgroup formation. The light, dark and speckled shadings indicate the top, bottom and central subgroups respectively; the horizontal stripes indicate outlying students; diagonal stripes indicate students who bridge two subgroups.

4.5.3 Popularity and Individual Stability

We now address the question of whether or not an individual's popularity has any effect on that individual's personal stability within the network. Nakao and Romney (1993), by embedding the fraternity data in a latent Euclidean space, claimed that individual stability can be used to predict the individual's position in the final subgroup structure. In other words, this statement by Nakao and Romney says that actors who have difficulty finding their social position will not find their social position within one of the subgroups. This is not telling us too much since the subgroup structure was determined by actors which stayed close together in the latent social space, and hence actors that have large movements in the latent space would not tend to stay close to any one particular region of the latent space. Here we are more interested in discovering whether an individual's popularity, i.e., how well liked an individual is, is related to the individual's stability within the network structure. That is, does a more popular actor find their social position more effectively than a less popular individual? Wasserman (1980) used his proposed method to analyze Newcomb's fraternity data to claim that popular individuals remain popular while less popular individuals become even less so over time. This statement implies some of the movements we see in Figures 4.4 and 4.5, where some individuals stay in the center and others move farther over time towards the edge of the social space. However, if we take popularity to be an intrinsic time-independent quality of how likeable an individual is, then we still have not answered the question of whether or not popularity is related to individual stability.

Using our proposed model for ranked dynamic networks, we frame our question in terms of finding a relationship between average step size, i.e., $\sum_{t\geq 2} \|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|/(T-1)$ (we will denote this quantity by s_i), with the social reach r_i. A key understanding in this approach is that by including $\boldsymbol{r}$ in the model, the step sizes are already accounting for the popularity of the individuals. Hence there is no forced relationship between the step sizes and $\boldsymbol{r}$ in the model; that is, if $\boldsymbol{r}$ was not included in the model then an unpopular individual would be forced to move around the outside of the network to maintain low probabilities of receiving favorable rankings, but here that is not the case since we have already accounted for the intrinsic popularity of the individuals. Therefore any relationship we see between step size and $\boldsymbol{r}$ is indicative of some fundamental relationship between individual stability and popularity.

To make sure that $\boldsymbol{r}$ held the intended meaning of intrinsic likability of an individual, we computed the correlation between the posterior mean of the log of $\boldsymbol{r}$ with the mean ranking for each individual, averaged over all other nodes at all time points; this correlation was -0.949 (recall that

a lower ranking is a more favorable ranking), implying that the interpretation of the social reaches is valid. We then used the posterior means of the latent positions to compute the average step size and the posterior means of the social reaches to estimate the correlation between $\mathbf{s} = (s_1, \ldots, s_n)$ and $\log(\boldsymbol{r})$. We did comparisons with the $\log(\boldsymbol{r})$ because from plotting $\mathbf{s}$ vs. $\log(\boldsymbol{r})$ we see a strong linear relationship (see Figure 4.9); this is not surprising since the means of z_{ijt} equal $\log(r_j) - d_{ijt}$ (see Section 4.2.1), and hence we might have expected to see a linear relationship between the step sizes with the $\log(\boldsymbol{r})$. The correlation was -0.819. Hence we see that there is a strong positive relationship between an individual's intrinsic popularity and the individual's ability to stabilize his social position.

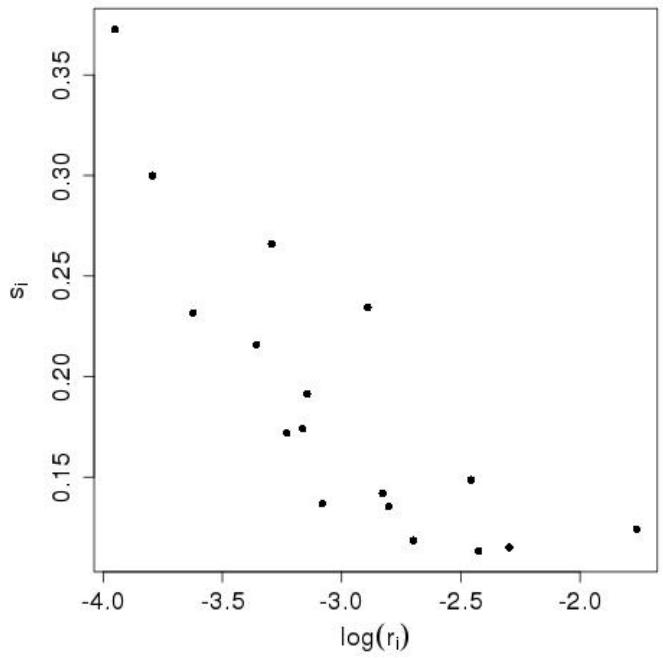


Figure 4.9: Plot of posterior means of the step sizes vs. the log of the posterior means of the social reaches.

4.6 Sensitivity Analysis

It is of interest to determine how sensitive our MCMC results are to the initialization method given in Section 4.1. We checked this sensitivity by choosing two alternative methods of initialization, each of which reflects some incorrect concept behind the latent space model (a misinterpretation of the latent positions and an assumption of constant network stability over time). In neither case did the conclusions corresponding to the three substantive questions (labeled Q1, Q2 and Q3 below)

change.

1. We altered the dissimilarity matrix used to determine the initial latent positions to reflect an incorrect interpretation of the latent space, in order to see whether this would lead to different conclusions to the three main substantive questions. In particular, d_{ijt} in equation (4.19) were computed as the correlations between $\mathbf{y}_{it}^{-(i,j)}$ and $\mathbf{y}_{jt}^{-(i,j)}$, where $\mathbf{y}_{it}^{-(i,j)}$ is the vector $\mathbf{y}_{it}$ excluding the i^{th} and j^{th} entries. In other words, we are setting the initial dissimilarities to reflect how similarly actors i and j view the other $n-2$ actors rather than how favorably actors i and j view each other after accounting for their respective popularities.

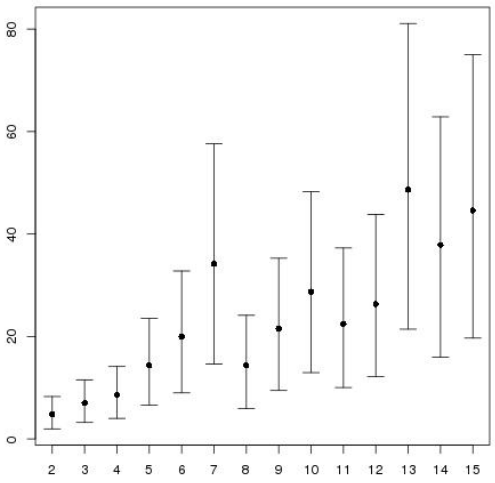
   (Q1) Figure 4.10a shows the estimates of τ_t, $t \geq 2$. We see that while the scale of the τ_t's are somewhat larger, the overall pattern is the same as found before, as seen in Figure 4.7. That is, the precision is very low in the first 5 weeks, has a dip around 8-9 weeks, and remains relatively high afterwards. (Q2) While the latent positions seem somewhat different (not unexpected, considering the dimension of the latent space), the subgroup pattern seems to be very consistent with what was found before as described in Section 4.5.2. That is, we find in our original analysis (Figures 4.4, 4.5 & 4.8) and that from the alternative dissimilarity matrix (Figure 4.10b) that actors 1, 6, 8 and 13 are close as are 2, 4, 9 and 17 with actor 5 bridging these two groups; also actors 7, 11 and 12 are close together with actor 2 bridging them; finally, actors 3, 10, 15 and 16 all make their way to the edge of the social space while actor 14 stays somewhat closer without seeming to join one of the three subgroups. (Q3) The correlation between the step sizes and social reaches (using posterior means) is -0.64, which leads to the same conclusion that there is a positive relationship between individual popularity and individual stability (small step sizes imply more stability).

2. We next tried to initialize the latent positions and τ_t's in such a way as to incorrectly reflect the assumptions on the stability of the network. That is, we assume that the stability will change over time, but we initialize as though we assume that the stability ought to be constant. To this end we used Sarkar and Moore's (2005) generalized multidimensional scaling, which implicitly assumes that the latent positions at each time point t should be within relatively the same distance from the latent positions at the previous time point $t-1$ for all t. We then set all $\tau_t^{(1)}$, $t \geq 2$ equal to the value

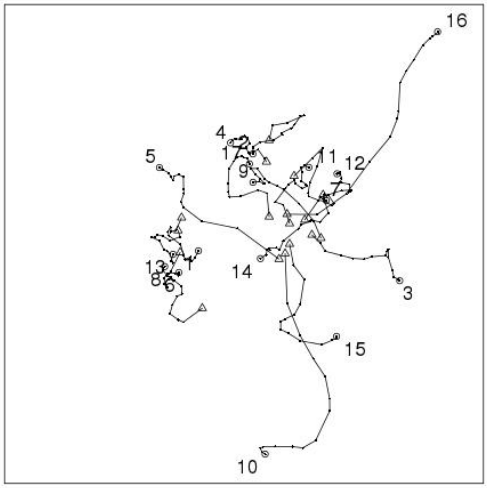
$$\tau_t = \left[\frac{1}{(T-1)np} \sum_{t \geq 2} \sum_{i=1}^{n} \|\mathbf{X}_{it} - \mathbf{X}_{i(t-1)}\|^2 \right]^{-1} \quad \text{for } t \geq 2.$$

We finished by setting $\theta^{(1)}$ arbitrarily equal to 1. The other parameters were initialized as discussed in Section 4.3.1.

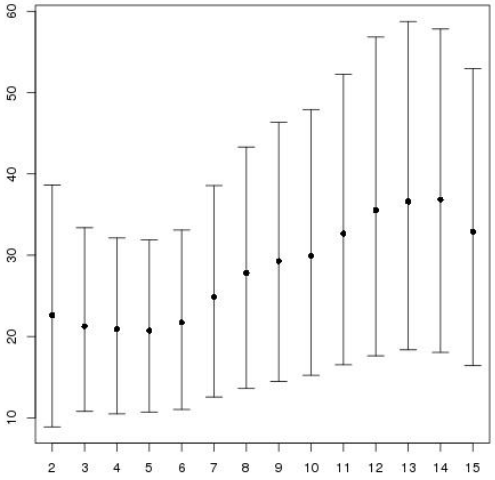
(Q1) Figure 4.10c shows the estimates of τ_t, $t \geq 2$. We again see the same overall trend of having small precision values in the first few weeks with an increasing trend afterwards. (Q2) We see in Figure 4.10d the same subgroupings as before, with actors 1, 6, 8 and 13 close together and actors 4, 9 and 17 close with 5 bridging the two groups; also actors 7, 11 and 12 are close with actor 2 bridging this group and the middle group of 4, 9 and 17; we finally see that actors 3, 10, 15 and 16 move toward the edge of the social space, with 14 a little closer to the center without clearly belonging to any of the three subgroups. (Q3) The correlation between the step sizes and social reaches (using posterior means) is -0.790, which again leads to the same conclusion that there is a positive relationship between individual popularity and individual stability.



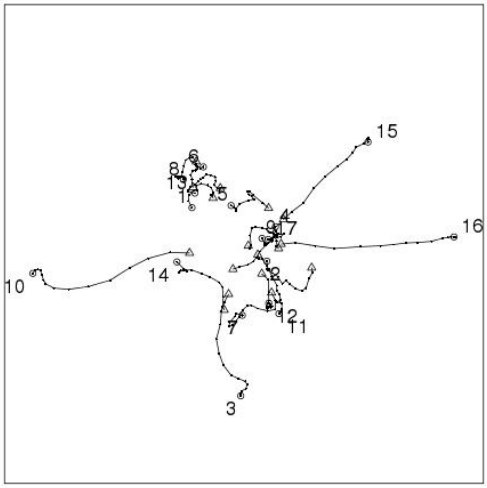
(a) τ_t, $t \geq 2$ for alternative dissimilarity matrix



(b) Latent trajectories for alternative dissimilarity matrix



(c) τ_t, $t \geq 2$ for alternative MDS algorithm



(d) Latent trajectories for alternative MDS algorithm

Figure 4.10: Results based on the posterior samples using the two alternative initialization strategies.

Chapter 5

Model-based longitudinal clustering

Numerous applications exist in which it is of interest to partition the data into homogeneous groups, or clusters. With longitudinal data in which we observe many objects over a series of time points, one often wishes to better understand how and why these objects are grouped and how they transition from one group to another over time. Longitudinal clustering is an important topic in many fields such as genomics, clinical research, political science, etc. In this chapter, there is a brief departure from the broader context of dynamic networks in order to introduce a new mixture model for longitudinal data, though the material to be presented will be used in the following chapter discussing community detection in dynamic networks.

Much of the previous work on clustering longitudinal data focuses on curve clustering, also called trajectory clustering. Curve clustering methods view the time series as curves or sometimes as random functions, and intends to give cluster assignments of the objects based on shape similarity. For example, Ray and Mallick (2006) used Bayesian wavelets to model the curves, and Luan and Li (2003) used B-splines to model the cluster means; often this type of longitudinal clustering is achieved by regression based approaches (see, e.g., Lou et al., 1993; Gaffney and Smyth, 1999).

More related to our work is latent transition analysis (LTA) (see, e.g., Graham et al., 1991; Collins and Wugalter, 1992; Collins et al., 1994), which is commonly used in the social sciences. LTA assumes that observed categorical responses are noisy measurements of sequential latent stages, i.e., we imperfectly can measure over time a subject's latent stage (at each time point). Vermunt et al. (1999) and Chung et al. (2005) were able to incorporate extra covariates in their model to help estimate the latent stages. However, these methods are focused on categorical panel data and sequential stages hypothesized a priori which are strongly associated with the measured categorical variables.

We are concerned with clustering vectors of continuous data observed over time, and where the clusters are based on locations rather than the shape of the trajectories, i.e., the observed values rather than the shape of the time series. A simple algorithm for this context is the extension of

the k-means method for longitudinal data (Genolini and Falissard, 2010). This does not, however, account for the correlation between each object at different time points, but simply vectorizes the time series, and then performs k-means based on the Euclidean distance (or some other metric) of the vectorized time series. Further, this method requires the time series to be of the same length, which may not be the case in practice. A model based approach which accounts for correlation across time points was given by De la Cruz-Mesía et al. (2008), who extended the model consisting of mixture of normal distributions (see Banfield and Raftery, 1993) to univariate time series data. Further work in the context of mixture of normal distributions was done by McNicholas and Murphy (2010), who derived estimates for specific covariance constraints corresponding to univariate time series data. Anderlucci and Viroli (2014) built on these works to handle multivariate data.

Scott et al. (2005) used a hidden Markov model to estimate the cluster assignments which were allowed to differ for each object at each time point, where the estimation was done using Markov chain Monte Carlo methods. While a vast improvement, this work still assumes that for each object, conditioning on the cluster assignment at the current time point, the observed value at the current time is independent of the observed values at previous (and future) time points. This assumption ignores individual effects which may be important in modeling the data. Moreover, while the model of Scott et al. allows for multiple treatments to affect the cluster transition probabilities, it is not flexible enough to handle multiple explanatory variables of varying types.

The context which is considered in this chapter is where we have recorded a multivariate response at many (and possibly at a varying number of) time points from many objects. We wish to cluster these objects at each time point based on the similarity of their responses. The number of clusters is assumed to be a fixed constant, and the clusters themselves are assumed to be static, in the sense that the structures that define each cluster are constant over time. To this end we propose a novel model-based clustering algorithm for longitudinal data. We allow cluster assignments to change over time, borrowing information across each object's time series when making these assignments, thus allowing the researcher to better understand how these objects transition from one group to another. We incorporate temporal dependence into the model by modeling each object's current location as a blending of the current (unobserved) cluster assignment and the object's previous values. Our model allows explanatory variables of any form to be incorporated into the model in order to explain the clustering by incorporating multinomial logistic regression into the model. This is fundamentally different than other clustering models which incorporate covariates, such as that of De la Cruz-Mesía et al. (2008) and Anderlucci and Viroli (2014), in that rather than directly predicting the response

with the covariates, the group membership of the objects are predicted by the covariates. This results in clustering being performed purely on the response variables while simultaneously learning how the covariates explain the clustering results. Estimation is accomplished by using the generalized EM algorithm. The computational cost of a straightforward implementation of this algorithm can be, however, prohibitively high. To address this, we derive several recursive relationships which greatly reduce the computational cost from exponential to linear with respect to the number of time points.

The rest of the chapter is organized as follows. Section 5.1 describes the proposed model. Section 5.2 outlines the details of the generalized EM algorithm. Section 5.3 provides the results of a simulation study. Section 5.4 gives the analysis of longitudinal data collected on Democrats serving in the U.S. House of Representatives. Section 5.5 provides a proof of a result stated earlier in the chapter. Sections 5.6 and 5.7 provide important derivations in the estimation algorithm.

5.1 Models

5.1.1 Non-regressed cluster probabilities

Denote the value of the i^{th} object at time t as $\mathbf{X}_{it}$, and let $\mathcal{X}_i = (\mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{iT_i})'$, for $i = 1, \ldots, n$, denote the data to be clustered, where n is the number of distinct objects, T_i is the length of the i^{th} time series, and the dimension of each $\mathbf{X}_{it}$ is p. We assume that each object is initially assigned to one of K clusters according to the $1 \times K$ probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ (K is a fixed positive integer assumed to be known). The notation to be used throughout lets Z_{it} be the i^{th} object's cluster assignment at time t. Hence $\mathbb{P}(Z_{i1} = k) = \alpha_k$. For subsequent time points, these objects are assigned to a cluster according to the $K \times K$ transition matrix $\boldsymbol{\beta}$; that is, $\mathbb{P}(Z_{it} = k | Z_{i(t-1)} = h) = \beta_{hk}$.

Hidden Markov models are widely used in many areas such as bioinformatics, biology, economics, finance, hydrology, marketing, medicine and speech recognition (Frühwirth-Schnatter, 2006). The usefulness of hidden Markov models in such a wide range of fields motivates us to borrow principles from this class of models. As we will see shortly, each cluster in our context is defined by a specific static probability distribution. A hidden Markov model applied to our setting would then imply that if we knew which cluster a particular object belonged to at each time point, the observed temporal observations from the object would look like random samples from the clusters' corresponding distributions. It seems more reasonable to assume that in most real life contexts these observations would still be dependent in some way, even after accounting for the clustering. For example, in the context of clustering human behaviors, we might expect a person's future behavior to depend on that

person's past behavior as well as on the influence exerted on him/her by the group to which he/she belongs. Therefore the current expected behavior of a particular person should be some blending of the previous behavior (individual effect) and the overall expected behavior from someone coming from the group to which the person belongs (cluster effect). This more general type of dependency is known as a Markov switching model.

With this motivation, we make two assumptions about the dependence structure of the model. First, given all past cluster assignments, the current cluster assignment depends only on the cluster assignment at the previous time point. Second, given the entire history of both the cluster assignments and the object's values, the current value of the object depends only on the value of the object at the previous time point as well as the current cluster assignment. These requirements lead to the following two dependency assumptions

$$Z_{it}|Z_{i1},\ldots,Z_{i(t-1)} \stackrel{\mathcal{D}}{=} Z_{it}|Z_{i(t-1)} \triangleq \beta_{Z_{i(t-1)}Z_{it}}, \tag{5.1}$$

(hence the cluster assignments follow a Markov process) and

$$\mathbf{X}_{it}|\mathbf{X}_{i1},\ldots,\mathbf{X}_{i(t-1)},Z_{i1},\ldots,Z_{iT_i} \stackrel{\mathcal{D}}{=} \mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it}. \tag{5.2}$$

One of the most widely used clustering methods is to fit a normal mixture model (and the commonly used k-means is but a special case of this). This approach performs well in a wide range of settings and so we make the distributional assumptions that

$$\pi(\mathbf{X}_{i1}|Z_{i1}) = N(\mathbf{X}_{i1}|\boldsymbol{\mu}_{Z_{i1}},\Sigma_{Z_{i1}}), \tag{5.3}$$

$$\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it}) = N(\mathbf{X}_{it}|\lambda\boldsymbol{\mu}_{Z_{it}}+(1-\lambda)\mathbf{X}_{i(t-1)},\Sigma_{Z_{it}}), \tag{5.4}$$

where $\boldsymbol{\mu}_k$ and Σ_k are the mean vector and covariance matrix for the k^{th} cluster for $k=1,\ldots,K$, $\lambda \in (0,1)$, and $N(\mathbf{x}|\boldsymbol{\mu},\Sigma)$ denotes the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The transition distribution in (5.4) blends the role of the current cluster with the individual effect. That is, in this framework one can look at the distribution of the current position of the i^{th} individual as being influenced by both where the individual has been and by a sense of belonging to a particular cluster. This model can be thought of as extending the widely used normal mixture model clustering to longitudinal data.

Assuming independence between the objects to be clustered, the complete-data likelihood (the

distribution of the observed data and unobserved cluster assignments) can be written as

$$\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}, Z_{i1}, \ldots Z_{iT_i}) = \prod_{i=1}^{n} \alpha_{Z_{i1}} \pi(\mathbf{X}_{i1}|Z_{i1}) \prod_{t=2}^{T_i} \beta_{Z_{i(t-1)}Z_{it}} \pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}). \tag{5.5}$$

We can then write the marginal distribution of the data as

$$\prod_{i=1}^{n} \pi(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT_i}) =$$
$$\prod_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \Big(\alpha_{Z_{i1}} \pi(\mathbf{X}_{i1}|Z_{i1}) \sum_{Z_{i2}=1}^{K} \Big(\beta_{Z_{i1}Z_{i2}} \pi(\mathbf{X}_{i2}|\mathbf{X}_{i1}, Z_{i2}) \cdots \sum_{Z_{iT_i}=1}^{K} \Big(\beta_{Z_{i(T_i-1)}Z_{iT_i}} \pi(\mathbf{X}_{iT_i}|\mathbf{X}_{i(T_i-1)}, Z_{iT_i}) \Big) \cdots \Big). \tag{5.6}$$

5.1.2 Regressed cluster probabilities

Suppose there is a set of explanatory variables that may influence how the objects are clustered. That is, we are not interested in clustering the objects based on the values of these explanatory variables, but rather we are interested in how these variables explain the clustering results. Let $\mathbf{w}_{it}$ denote a vector of length d corresponding to these explanatory variables for the i^{th} object at time t. Then the model described in Section 5.1.1 can be slightly modified to allow $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be functions of the explanatory variables, $\mathbf{w}_{it}$, and unknown parameters, denoted as $\delta_{\ell k}$ and $\boldsymbol{\gamma}_{\ell k}$ for $\ell = 0, \ldots, K$ and $k = 1, \ldots, K$. Mimicking the multinomial logistic regression model, for $k = 1, \ldots, K$ we let

$$\log\left(\frac{\alpha_k(\mathbf{w}_{it})}{\alpha_K(\mathbf{w}_{it})}\right) = \delta_{0k} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{0k}, \tag{5.7}$$

where we fix $\delta_{0K} = 0$ and $\boldsymbol{\gamma}_{0K} = \mathbf{0}$; similarly for each $h = 1, \ldots, K$, for $k = 1, \ldots, K$ we let

$$\log\left(\frac{\beta_{hk}(\mathbf{w}_{it})}{\beta_{hK}(\mathbf{w}_{it})}\right) = \delta_{hk} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{hk}, \tag{5.8}$$

where again we fix $\delta_{hK} = 0$ and $\boldsymbol{\gamma}_{hK} = \mathbf{0}$ for $h = 1, \ldots, K$. Hence we can write the initial and transition cluster probabilities respectively as

$$\alpha_k(\mathbf{w}_{it}) = \frac{\exp(\delta_{0k} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{0k})}{1 + \sum_{\ell=1}^{K-1} \exp(\delta_{0\ell} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{0\ell})} \tag{5.9}$$

$$\beta_{hk}(\mathbf{w}_{it}) = \frac{\exp(\delta_{hk} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{hk})}{1 + \sum_{\ell=1}^{K-1} \exp(\delta_{h\ell} + \mathbf{w}_{it}'\boldsymbol{\gamma}_{h\ell})} \tag{5.10}$$

$$\tag{5.11}$$

for $h, k = 1, \ldots, K$. It is easy to see that setting $\boldsymbol{\gamma}_{0k} = \boldsymbol{\gamma}_{hk} = \mathbf{0}$ for all h, k yields a model equivalent to the non-regressed transition model of Section 5.1.1.

The model of Section 5.1.1 then changes in that we are conditioning on the explanatory variables $\mathbf{w}_{it}$. Specifically we have that the conditional distribution of the cluster assignments changes from (5.1) to become

$$Z_{it}|Z_{i1},\ldots,Z_{i(t-1)},\mathbf{w}_{i1},\ldots,\mathbf{w}_{iT_i} \stackrel{\mathcal{D}}{=} Z_{it}|Z_{i(t-1)},\mathbf{w}_{it} \triangleq \beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it}), \tag{5.12}$$

while (5.2) does not change. The complete-data likelihood becomes

$$\begin{aligned}&\prod_{i=1}^{n}\pi(\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT_i},Z_{i1},\ldots Z_{iT_i}|\mathbf{w}_{i1},\ldots,\mathbf{w}_{iT_i})\\ =&\prod_{i=1}^{n}\alpha_{Z_{i1}}(\mathbf{w}_{i1})\pi(\mathbf{X}_{i1}|Z_{i1})\prod_{t=2}^{T_i}\beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it})\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it}),\end{aligned} \tag{5.13}$$

and the marginal likelihood becomes

$$\begin{aligned}&\prod_{i=1}^{n}\pi(\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT_i}|\mathbf{w}_{i1},\ldots,\mathbf{w}_{iT_i}) =\\ &\prod_{i=1}^{n}\sum_{Z_{i1}=1}^{K}\Big(\alpha_{Z_{i1}}(\mathbf{w}_{i1})\pi(\mathbf{X}_{i1}|Z_{i1})\sum_{Z_{i2}=1}^{K}\Big(\beta_{Z_{i1}Z_{i2}}(\mathbf{w}_{i2})\pi(\mathbf{X}_{i2}|\mathbf{X}_{i1},Z_{i2})\\ &\cdots\sum_{Z_{iT_i}=1}^{K}\Big(\beta_{Z_{i(T_i-1)}Z_{iT_i}}(\mathbf{w}_{iT_i})\pi(\mathbf{X}_{iT_i}|\mathbf{X}_{i(T_i-1)},Z_{iT_i})\Big)\cdots\Big).\end{aligned} \tag{5.14}$$

5.1.3 Computational issues

The estimation procedure to be outlined in Section 5.2 is an iterative procedure aimed to maximize the likelihood, i.e., the marginal distribution of the data (5.6), and hence we must be able to compute the likelihood to know if we have reached convergence. However, in computing the likelihood we run into a potentially debilitating problem — if computed directly, the number of terms to be summed grows exponentially with T_i. To address this we have the following lemma and theorem, suppressing the subscript i for ease of notation.

Lemma 5.1.1. *The current latent cluster assignments are conditionally independent of the history of the data up to the previous time point given the previous cluster assignments; i.e.,*

$$\pi(Z_t|Z_{t-1},\mathbf{X}_1,\ldots,\mathbf{X}_{t-1}) = \beta_{Z_{t-1}Z_t}.$$

The proof is given in Section 5.5.

Theorem 5.1.1. *The marginal distribution of the data $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ can be computed in $O(nT)$ terms.*

Proof. Using lemma 5.1.1, we have that

$$\begin{aligned}\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t, Z_t) &= \sum_{Z_{t-1}=1}^{K} \pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t)\pi(Z_{t-1}, Z_t|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \\ &= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \sum_{Z_{t-1}=1}^{K} \pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t). \quad (5.15)\end{aligned}$$

Computing the marginal likelihood $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ can be accomplished by using this recursion in two steps, first obtaining $\pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$ and then $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t, Z_t)$. More specifically, we first compute and normalize

$$\pi(Z_1|\mathbf{X}_1) \propto \pi(\mathbf{X}_1|Z_1)\alpha_{Z_1}. \quad (5.16)$$

Then for $t = 3, \ldots, T$, from (5.15) we can compute and normalize

$$\begin{aligned}&\pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \\ \propto &\sum_{Z_{t-2}=1}^{K} \pi(Z_{t-2}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-2})\beta_{Z_{t-2}Z_{t-1}}\pi(\mathbf{X}_{t-1}|\mathbf{X}_{t-2}, Z_{t-1}). \quad (5.17)\end{aligned}$$

Second, we can compute $\pi(\mathbf{X}_1) = \sum_{Z_1=1}^{K} \pi(\mathbf{X}_1|Z_1)\alpha_{Z_1}$, and then use the normalized results from (5.16) and (5.17) to compute, for $t = 2, \ldots, T$,

$$\begin{aligned}&\pi(\mathbf{X}_1, \ldots, \mathbf{X}_t) \\ = \ &\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \sum_{Z_t=1}^{K} \sum_{Z_{t-1}=1}^{K} \pi(Z_{t-1}|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})\beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t). \quad (5.18)\end{aligned}$$

By utilizing these recursive relationships, the number of terms required to be summed in the computation of the likelihood density grows linearly, rather than exponentially, in time. □

Regarding the context of regressed cluster assignment probabilities, the previous theorem still holds simply by exchanging $\alpha_{Z_{i1}}$ for $\alpha_{Z_{i1}(\mathbf{w}_{i1})}$ and $\beta_{Z_{i(t-1)}Z_{it}}$ for $\beta_{Z_{i(t-1)}Z_{it}}(\mathbf{w}_{it})$ in equations (5.15) to (5.18).

5.2 Estimation

We aim to find the maximum likelihood estimators (MLE's) for the model parameters, henceforth denoted as Θ. In the case of non-regressed cluster probabilities,

$$\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K, \lambda, \alpha_1, \ldots, \alpha_K, \beta_{11}, \ldots, \beta_{KK}\},$$

and for the case of regressed cluster probabilities,

$$\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K, \lambda, \delta_{01}, \ldots, \delta_{K(K-1)}, \boldsymbol{\gamma}_{01}, \ldots, \boldsymbol{\gamma}_{K(K-1)}\}.$$

To find the MLE's, we employ the generalized EM algorithm (Dempster et al., 1977; Wu, 1983) to obtain parameter estimates, as well as cluster assignment probabilities. Heretofore the notation for probability densities have not included the parameter set Θ, this parameter set being implicit in the density itself; in the forthcoming discussion, however, the dependency on the parameter set Θ will be explicitly written as $\pi(\cdot|\Theta)$ since it is necessary to make it clear which value of Θ is being used to compute the density in the iterative algorithm. We first derive the solutions for the non-regressed case. Letting $\widehat{\Theta}$ denote the current estimate of Θ, we iteratively find

$$Q(\Theta, \widehat{\Theta}) = \mathbb{E}\left(\log(L^c)|\mathcal{X}_1, \ldots, \mathcal{X}_n, \widehat{\Theta}\right), \tag{5.19}$$

(E step) where L^c is the complete-data likelihood, as given in (5.5), and find Θ^*, where $Q(\Theta^*, \widehat{\Theta}) \geq Q(\widehat{\Theta}, \widehat{\Theta})$ (M step), subsequently setting $\widehat{\Theta} = \Theta^*$. Now $Q(\Theta, \widehat{\Theta})$ can, with some algebra (see Section 5.6), be written in the tractable form

$$\begin{aligned} &Q(\Theta, \widehat{\Theta}) = \\ &\sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\alpha_{\ell_1}) + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t})\right] \mathbb{P}(Z_{i1} = \ell_1|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s|Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}) \\ &+ \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\pi(\mathbf{X}_{i1}|Z_{i1} = \ell_1, \Theta)) + \sum_{t=2}^{T_i} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta))\right] \\ &\qquad \cdot \mathbb{P}(Z_{i1} = \ell_1|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s|Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}). \end{aligned} \tag{5.20}$$

In computing a valid Θ^* which increases the value of $Q(\widehat{\Theta}, \widehat{\Theta})$, we again run into the same sort of trouble as in the computation of (5.6), in that the number of terms to be summed grows

exponentially with respect to time. To address this, we have the following theorem and corollary, again suppressing the subscript i for ease of notation.

Theorem 5.2.1. *The conditional posterior distributions of the cluster assignments* $\mathbb{P}(Z_t|Z_{(t-1)}, \mathcal{X}, \widehat{\Theta})$ *can be written recursively up to a proportionality constant, specifically*

$$\mathbb{P}(Z_t|Z_{t-1}, \mathcal{X}, \Theta) \propto q(Z_t|Z_{t-1}), \tag{5.21}$$

where

$$q(Z_T|Z_{T-1}) = \beta_{Z_{T-1}Z_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T, \Theta), \tag{5.22}$$

and for $2 \leq t < T$,

$$q(Z_t|Z_{t-1}) = \beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t, \Theta)\sum_{\ell_{t+1}=1}^{K} q(Z_{t+1} = \ell_{t+1}|Z_t). \tag{5.23}$$

and

$$\mathbb{P}(Z_1|\mathcal{X}) \propto \alpha_{Z_1}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t)\sum_{\ell_2} q(Z_2 = \ell_2|Z_1). \tag{5.24}$$

Proof. We start by noticing that

$$\begin{aligned}\mathbb{P}(Z_T|Z_{T-1}, \mathcal{X}) &\propto \pi(\mathcal{X}|Z_T, Z_{T-1})\mathbb{P}(Z_T|Z_{T-1}) \\ &\propto \beta_{Z_{T-1}Z_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T).\end{aligned} \tag{5.25}$$

We can then find the iterative relationship for $2 \leq t < T$

$$\begin{aligned}&\mathbb{P}(Z_t|Z_{t-1}, \mathcal{X}) \\ &\propto \pi(\mathbf{X}_t, \ldots, \mathbf{X}_T|\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_t, Z_{t-1})\mathbb{P}(Z_t|Z_{t-1}) \\ &\propto \beta_{Z_{t-1}Z_t}\sum_{\ell_{t+1}}\cdots\sum_{\ell_T}\pi(\mathbf{X}_t, \ldots, \mathbf{X}_T, Z_{t+1} = \ell_{t+1}, \ldots, Z_T = \ell_T|\mathbf{X}_{t-1}, Z_t) \\ &\propto \beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t) \\ &\cdot \sum_{\ell_{t+1}}\left(\beta_{Z_t\ell_{t+1}}\pi(\mathbf{X}_{t+1}|\mathbf{X}_t, Z_{t+1} = \ell_{t+1})\cdots\sum_{\ell_T}\left(\beta_{\ell_{T-1}\ell_T}\pi(\mathbf{X}_T|\mathbf{X}_{T-1}, Z_T = \ell_T)\right)\cdots\right) \\ &\propto \beta_{Z_{t-1}Z_t}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t)\sum_{\ell_{t+1}} q(Z_{t+1} = \ell_{t+1}|Z_t).\end{aligned} \tag{5.26}$$

Similarly, $\mathbb{P}(Z_1|\mathcal{X}) \propto \alpha_{Z_1}\pi(\mathbf{X}_t|\mathbf{X}_{t-1}, Z_t)\sum_{\ell_2} q(Z_2 = \ell_2|Z_1)$. □

Corollary 5.2.1 (Theorem 5.2.1)**.** *The marginal posterior distributions and the conditional posterior distributions of the cluster assignments,* $\mathbb{P}(Z_t|\mathcal{X},\widehat{\Theta})$ *and* $\mathbb{P}(Z_t|Z_{(t-1)},\mathcal{X},\widehat{\Theta})$ *respectively, can be computed in* $O(T)$ *terms.*

Proof. From Theorem 5.2.1 we immediately have that $\mathbb{P}(Z_t|Z_{(t-1)},\mathcal{X},\widehat{\Theta})$ can be computed in $O(T)$ terms. To obtain the marginals, we start by computing and normalizing

$$\mathbb{P}(Z_1|\mathcal{X},\Theta) \propto \alpha_{Z_1}\pi(\mathbf{X}_1|Z_1,\Theta)\sum_{\ell_2=1}^{K} q(Z_2=\ell_2|Z_1), \tag{5.27}$$

and recursively obtain

$$\mathbb{P}(Z_t|\mathcal{X},\Theta) = \sum_{\ell_{t-1}=1}^{K} \mathbb{P}(Z_{t-1}=\ell_{t-1}|\mathcal{X},\Theta)\mathbb{P}(Z_t|Z_{t-1}=\ell_{t-1},\mathcal{X},\Theta). \tag{5.28}$$

□

Using the method of Lagrange multipliers, we can find that the value of α_k that maximizes $Q(\Theta,\widehat{\Theta})$ is

$$\alpha_k^* = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(Z_{i1}=k|\mathcal{X}_i,\widehat{\Theta}), \tag{5.29}$$

and similarly the optimizing value of β_{hk} is

$$\beta_{hk}^* = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T_i}\mathbb{P}(Z_{i(t-1)}=h|\mathcal{X}_i,\widehat{\Theta})\mathbb{P}(Z_{it}=k|Z_{i(t-1)}=h,\mathcal{X}_i,\widehat{\Theta})}{\sum_{i=1}^{n}\sum_{t=2}^{T_i}\mathbb{P}(Z_{i(t-1)}=h|\mathcal{X}_i,\widehat{\Theta})}. \tag{5.30}$$

To update λ, $\boldsymbol{\mu}_k$ and Σ_k, $k=1,\ldots,K$, we employ a coordinate ascent method. That is, we initialize $\lambda^*=\hat{\lambda}$, $\boldsymbol{\mu}_k^*=\hat{\boldsymbol{\mu}}_k$ and $\Sigma_k^*=\widehat{\Sigma}_k$ and then find

$$\begin{aligned}\lambda^* &= \arg\max_{\lambda} Q(\{\lambda,\Theta^*\setminus\{\lambda^*\}\},\widehat{\Theta})\\ &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{T_i}\sum_{\ell_t=1}^{K}\mathbb{P}(Z_{it}=\ell_t|\mathcal{X}_i,\widehat{\Theta})(\mathbf{X}_{it}-\mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k-\mathbf{X}_{i(t-1)})}{\sum_{i=1}^{n}\sum_{t=2}^{T_i}\sum_{\ell_t=1}^{K}\mathbb{P}(Z_{it}=\ell_t|\mathcal{X}_i,\widehat{\Theta})(\boldsymbol{\mu}_k-\mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k-\mathbf{X}_{i(t-1)})},\end{aligned} \tag{5.31}$$

$$\begin{aligned}\boldsymbol{\mu}_k^* &= \arg\max_{\boldsymbol{\mu}_k} Q(\{\boldsymbol{\mu}_k,\Theta^*\setminus\{\boldsymbol{\mu}_k^*\}\},\widehat{\Theta})\\ &= \frac{\sum_{i=1}^{n}\left\{\mathbb{P}(Z_{i1}=k|\mathcal{X}_i,\widehat{\Theta})\mathbf{X}_{i1}+\lambda\sum_{t=2}^{T_i}\mathbb{P}(Z_{it}=k|\mathcal{X}_i,\widehat{\Theta})(\mathbf{X}_{it}-(1-\lambda)\mathbf{X}_{i(t-1)})\right\}}{\sum_{i=1}^{n}\left\{\mathbb{P}(Z_{i1}=k|\mathcal{X}_i,\widehat{\Theta})+\lambda^2\sum_{t=2}^{T_i}\mathbb{P}(Z_{it}=k|\mathcal{X}_i,\widehat{\Theta})\right\}},\end{aligned} \tag{5.32}$$

and

$$\Sigma_k^* = \arg\max_{\Sigma_k} Q(\{\Sigma_k, \Theta^* \setminus \{\Sigma_k^*\}\}, \widehat{\Theta})$$
$$= \frac{\sum_{i=1}^n \left\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{i1} - \boldsymbol{\mu}_k)(\mathbf{X}_{i1} - \boldsymbol{\mu}_k)' + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{H}_{it}\mathbf{H}_{it}'\right\}}{\sum_{i=1}^n \left\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})\right\}} \tag{5.33}$$

for $k = 1, \ldots, K$, where $\mathbf{H}_{it} = \mathbf{X}_{it} - \lambda\boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}$. See Section 5.7 for proofs of these solutions.

In the case of regressed cluster probabilities, the complete-data likelihood in $Q(\Theta, \widehat{\Theta})$ is that found in (5.13). To obtain the correct updates for λ, $\boldsymbol{\mu}_k$ and Σ_k, $k = 1, \ldots, K$, simply replace in the above work $\mathbb{P}(Z_{it}|\mathcal{X}_i, \Theta)$ with $\mathbb{P}(Z_{it}|\mathcal{X}_i, \Theta, \mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT_i})$; these latter distributions are easily computed by first replacing α_k with $\alpha_k(\mathbf{w}_{i1})$ in (5.27) and β_{hk} with $\beta_{hk}(\mathbf{w}_{it})$ in (5.22) and (5.23), and then proceeding with the recursive relationships previously outlined. To obtain updates for δ_{0k}, $\boldsymbol{\gamma}_{0k}$, δ_{hk} and $\boldsymbol{\gamma}_{hk}$, $h = 1, \ldots, K$, $k = 1, \ldots, K-1$, one can use a numerical optimization method (we implemented the well known quasi-Newton BFGS method) to find

$$\arg\max_{\{\delta_{0k}, \boldsymbol{\gamma}_{0k}: k=1,\ldots,K\}} = \sum_{i=1}^n \sum_{k=1}^K \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) \log(\alpha_k(\mathbf{w}_{i1})) \tag{5.34}$$

and

$$\arg\max_{\{\delta_{hk}, \boldsymbol{\gamma}_{hk}: k=1,\ldots,K\}} = \sum_{i=1}^n \sum_{t=2}^{T_i} \sum_{k=1}^K \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})\mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta}) \log(\beta_{hk}(\mathbf{w}_{it})) \tag{5.35}$$

for $h = 1, \ldots, K$. Since (5.34) and (5.35) are both concave, finding the global maximum at each M-step is assured upon convergence of the BFGS algorithm.

5.3 Simulation Study

We simulated two illustrative data sets, one without and one with regressed cluster probabilities, where p, the dimension of each $\mathbf{X}_{it}$, is two. We then simulated 100 data sets where $p = 5$, 50 without and 50 with regressed cluster probabilities. The resulting clustering was compared in both the variation of information (VI) (Meilă, 2003) and in the corrected Rand index (CRI) (Hubert and Arabie, 1985). The VI, a true metric on clusterings, yields a value of 0 for identical clusterings; higher values indicate more disparate clustering assignments. The corrected Rand index, adjusted to account for chance, yields a value of 1 for perfect cluster agreement and has an expected value of

0 for random cluster assignments. The results are both in the text as well as in Table 5.1.

Each simulated data set (when $p = 2$ and when $p = 5$) without regressed cluster probabilities was achieved according to the following. The number of objects was 100, the maximum number of time steps was 10, and the number of clusters was 5. To check whether our method can accurately group the objects, the simulations consisted of a wide range of possible parameter values. We obtained the simulated data in the following way. The quantities T_i were uniformly sampled from 1 to 10; λ was drawn from a beta distribution with both parameters equal to 10 (the mean of the distribution is 0.5); $\boldsymbol{\alpha}$ was drawn from a Dirichlet distribution with parameters all equalling 10; $\boldsymbol{\mu}_k$'s were drawn from a multivariate normal distribution with mean $\mathbf{0}$ and diagonal covariance matrix with diagonal entries equal to 20; Σ_k's were drawn from an inverse Wishart distribution with $p + 4$ degrees of freedom and diagonal scale matrix whose diagonal entries equal 3. The h^{th} row of β was set to be proportional to

$$\left(\frac{1}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_h\|}, \ldots, \frac{1}{\|\boldsymbol{\mu}_{h-1} - \boldsymbol{\mu}_h\|}, 15 \times \max_{k \neq h} \left\{ \frac{1}{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_h\|} \right\}, \frac{1}{\|\boldsymbol{\mu}_{h+1} - \boldsymbol{\mu}_h\|}, \ldots, \frac{1}{\|\boldsymbol{\mu}_K - \boldsymbol{\mu}_h\|} \right).$$

This formulation of β was chosen to put most of the probability on remaining in the same cluster and to force the probability of jumping to a different cluster to decrease as the distance to that different cluster increases. The Z_{it}'s were then generated according to $\boldsymbol{\alpha}$ and β, and the $\mathcal{X}_i$'s were generated according to (5.3) and (5.4).

The simulations with regressed cluster probabilities were generated in a similar manner, but we allowed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to depend on a positive valued covariate which affects the probability of belonging to cluster one. A high value of the covariate implied a tendency to belong to cluster one while a low value of the covariate implied a tendency to avoid belonging to cluster one. The other clusters were treated equally with respect to the covariate. Just as in the simulation without a covariate, there was a stronger tendency to remain in the same cluster than to switch. This was accomplished in the following way. For $i = 1, \ldots, n$, the explanatory variable w_{i1} $(d = 1)$ was drawn from a chi-squared distribution with five degrees of freedom, and for $t = 2, \ldots, T_i$, w_{it} was drawn from a normal distribution with mean $w_{i(t-1)}$ and standard deviation 0.5. The parameters γ_{0k} and, for $h = 1, \ldots, K$, γ_{hk} were set to be $\log(1.5)$ if $k = 1$ and 0 otherwise. We set δ_{0k} and, for $h = 1, \ldots, K$, δ_{hk} equal to $-5\log(1.5)$ if $k = 1$ and 0 otherwise; in addition, $\log(15)$ was added to δ_{hh}, $h < K$, and $-\log(15)$ was added to δ_{Kk}, $k < K - 1$. This leads to the situation described above where there is an increased chance of staying in the same cluster than switching to another cluster. If w_{it} is 5

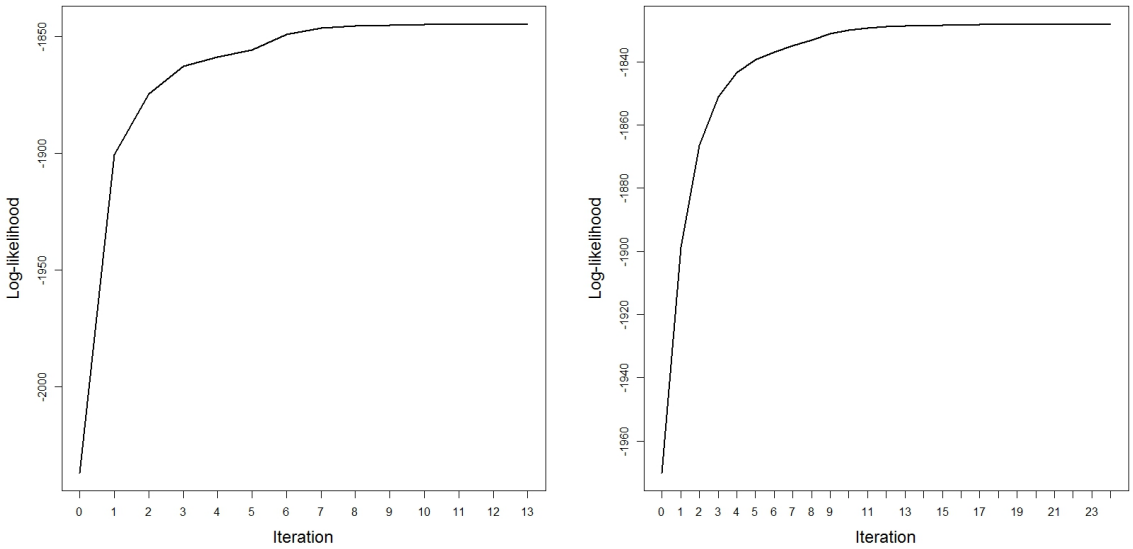
(the mean of w_{it}) then there is no particular tendency or avoidance of belonging to cluster one, if $w_{it} > 5$ then there is an increased probability of belonging to cluster one, and if $w_{it} < 5$ then there is a decreased probability of belonging to cluster one.

For the illustrative simulation without regressed cluster probabilities ($p = 2$), the VI from using the proposed approach was 0.2679 and CRI was 0.9232. Simply using k-means and choosing the best k-means result from 15 starting positions (which was used to initialize the generalized EM algorithm), led to a VI of 1.031 and CRI of 0.6615. Figure 5.1a gives the plot of the log-likelihood over the iterations of the generalized EM algorithm of the proposed method, starting with the initialized values. We see that the algorithm converged within a small number of iterations. Figure 5.2 shows the simulated data, where the numbers correspond to the true cluster assignments, and the shapes correspond to the estimated hard cluster assignments. Regarding the 50 simulations without regressed cluster probabilities ($p = 5$), the mean (median) VI was 0.0180 (0) and CRI was 0.9953 (1). Using k-means, the mean (median) VI was 0.5516 (0.5175) and CRI was 0.8426 (0.8609). Using a UNIX machine with a 2.40 GHz processor, the mean (median) elapsed time for our proposed method was 0.5507 sec (0.5120 sec). These values of VI and CRI imply that our model performs extremely well, and much better than the naïve k-means approach.

For the illustrative simulation with regressed cluster probabilities ($p = 2$), the VI from using the proposed approach was 0.5080 and the CRI was 0.8232. Simply using k-means resulted in a VI of 1.342 and a CRI of 0.5147. Figure 5.1b gives the plot of the log-likelihood over the iterations of the generalized EM algorithm in the proposed method, starting with the initialized values. From this we again see that the algorithm reached convergence within a small number of iterations. Figure 5.3 shows the simulated data, where the numbers correspond to the true cluster assignments, and the shapes correspond to the estimated hard cluster assignments. Regarding the 50 simulations with regressed cluster probabilities ($p = 5$), the mean (median) VI was 0.0531 (0) and CRI was 0.9780 (1). Using k-means, the mean (median) VI was 0.7413 (0.7429) and CRI was 0.7729 (0.7953). The mean (median) elapsed time for our proposed method was 41.82 sec (35.41 sec). As in the case

	Without regressed cluster probabilities		With regressed cluster probabilities	
	Proposed method	k-means	Proposed method	k-means
VI	0.0180 (0)	0.5516 (0.5175)	0.0531 (0)	0.7413 (0.7429)
CRI	0.9953 (1)	0.8426 (0.8609)	0.9780 (1)	0.7729 (0.7953)

Table 5.1: From the simulations run without and with regressed cluster probabilities, the estimated clustering results are compared to the true cluster assignments for both k-means and our proposed approach using both variation of information (VI) and the corrected Rand index (CRI). Means over 50 simulations are given with medians in parentheses.



(a) Without regressed cluster probabilities

(b) With regressed cluster probabilities

Figure 5.1: Log-likelihood over the iterations for the two illustrative simulations.

for non-regressed cluster probabilities, these values of VI and CRI imply that our model performs extremely well, and much better than the naïve k-means approach. We can also investigate whether the model captured the effect of the explanatory variable. An effective visualization of this is to plot the cluster probabilities $\alpha_k(w_{i1})$ and $\beta_{hk}(w_{it})$ for $h, k = 1, \ldots, K$ over the range of w_{it}. This is much more intuitive than trying to look at the individual δ_{hk}'s and γ_{hk}'s, allowing the user to see exactly how the explanatory variables affect the cluster probabilities. Figure 5.4 gives these plots, where each line corresponds to probabilities of belonging to a particular cluster. Here we can see that the estimated probability curves (solid lines) correspond very closely to the true probability curves (dashed lines), implying that the estimation method captured the effect of the explanatory variable very effectively.

5.4 U.S. Congressional Data

We collected 15 variables measuring legislative activity for Democrats in the House of Representatives in the 101^{st}-110^{th} Congresses (1989-2008). Because there are multiple variables for a single concept, we combined these indicators into eight indices — paying attention to the home district; showboating (e.g., making speeches and writing editorials); voting with the party on roll calls; giving campaign funds to the party; specializing in particular policy issues; building bipartisan coalitions; overall

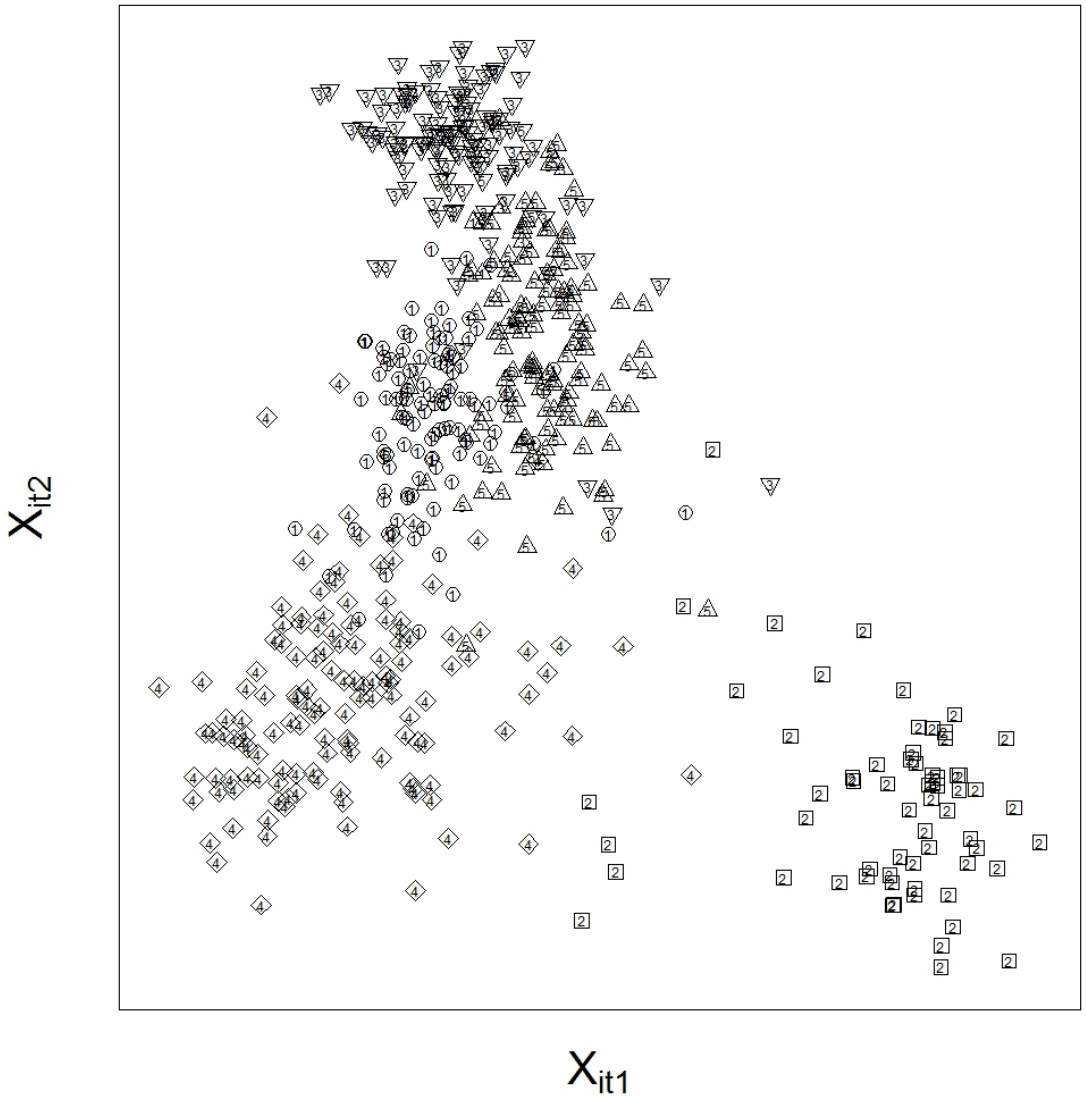


Figure 5.2: Plot of simulated data without regressed cluster probabilities, where the horizontal and vertical axes correspond to the two variables on which clustering is performed. Numbers correspond to true clustering, shapes correspond to estimated hard clustering assignments.

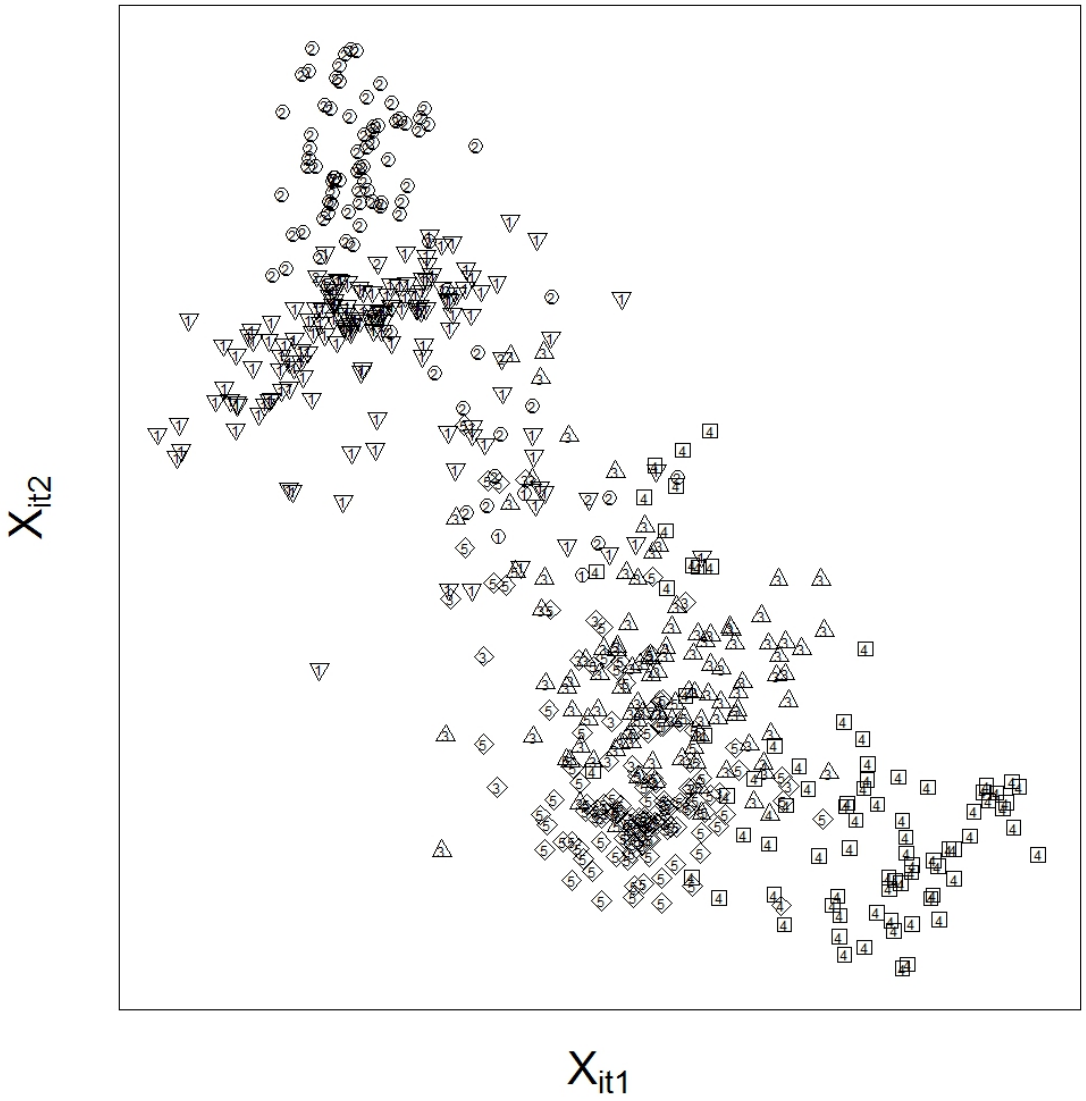
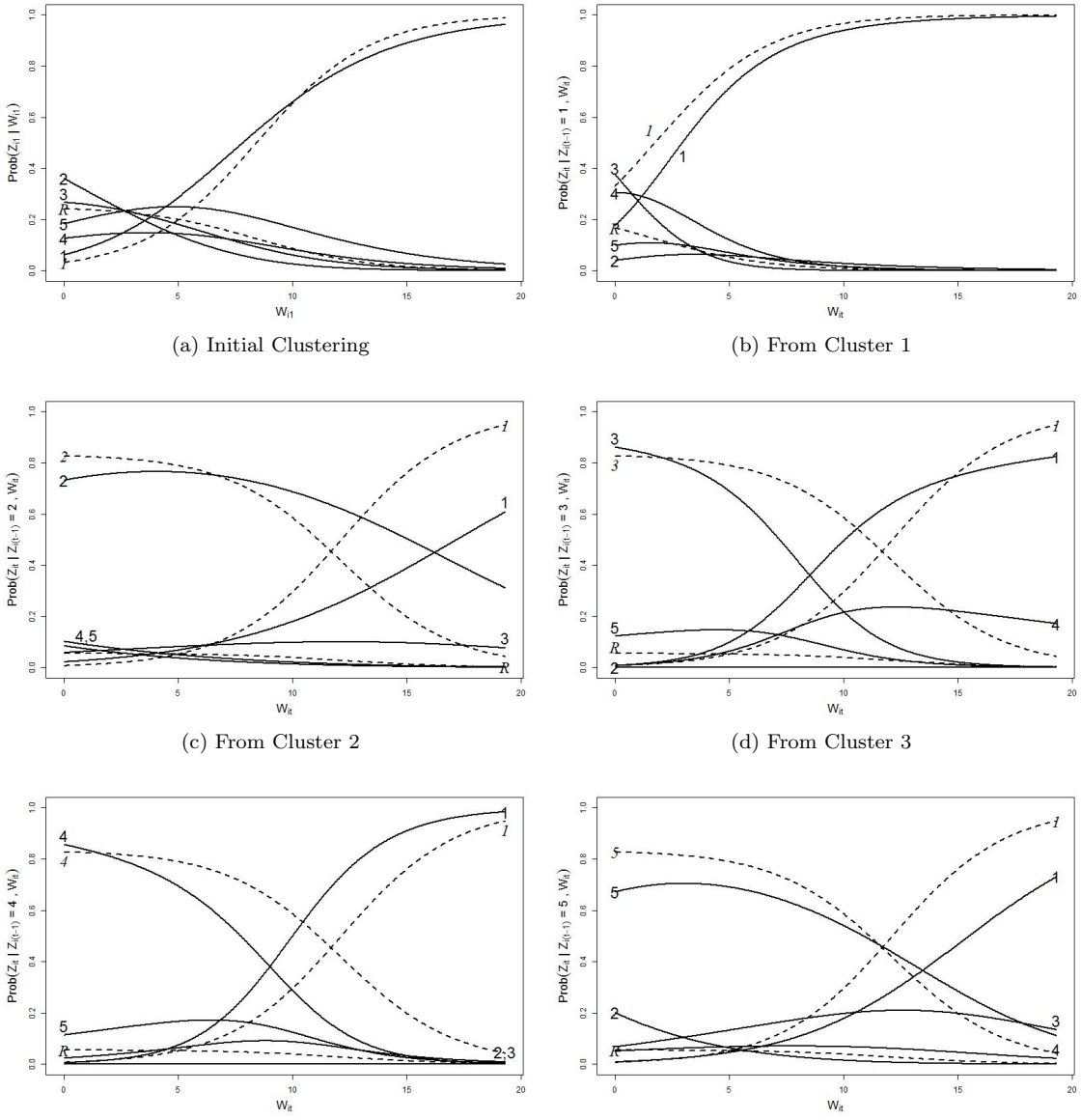


Figure 5.3: Plot of simulated data with regressed cluster probabilities, where the horizontal and vertical axes correspond to the two variables on which clustering is performed. Numbers correspond to true clustering, shapes correspond to estimated hard clustering assignments.



(a) Initial Clustering

(b) From Cluster 1

(c) From Cluster 2

(d) From Cluster 3

(e) From Cluster 4

(f) From Cluster 5

Figure 5.4: Cluster probabilities for illustrative simulated data with regressed cluster probabilities. Solid lines are estimated probability curves, dashed lines are true curves. Each curve represents, for the varying values of the explanatory variable, the probability of belonging to the cluster whose number is adjacent to the curve. There are only two or three (depending on the figure) unique true probability curves which are labeled in italics, hence "R" indicates the probability curve corresponding to those remaining cluster numbers which are not unique, e.g., clusters 2 to 5 in the initial clustering plot.

fundraising; and lawmaking (e.g., introducing legislation)*. Thus $\mathbf{X}_{it}$ is the eight-dimensional vector corresponding to the scores on these eight indices measured for the i^{th} member of Congress (MC) at his or her t^{th} term. There were $n = 539$ unique MC's that served one or more terms over the 20 years included in the study.

We wished to determine the impact of the MCs' ideology on the MCs' behaviors. To this end we included a covariate for MC ideology to help predict initial cluster assignments for individual MCs, as well as transitions between clusters across time. Our measure of ideology was the common-space NOMINATE score. These scores are based on a multidimensional scaling algorithm for roll call voting developed by Poole and Rosenthal (1997) and are a function of how often each individual MC makes the same vote choice as each other MC. The algorithm allows all legislators to be arrayed on a scale of about -1 to 1, with negative scores indicating liberal ideology and positive scores indicating conservative ideology. Since the sample is solely Democrats, the scores are skewed toward the negative end of the scale, ranging from -0.725 to 0.190. In addition to this ideology variable, we also included time dummy variables in order to account for behavioral shifts across the entire party that occur during the various Congresses (each Congress has its own dummy variable). Lastly, for the 101^{st} Congress we differentiated between true freshmen politicians as well as those MCs who have served previous terms not included in the data by allowing the true freshmen (for all Congresses) and the first observed MCs (of the 101^{st} Congress) to have different initial clustering coefficients δ_{0k} and $\boldsymbol{\gamma}_{0k}$. This was accomplished by letting $\mathbf{w}_{i1} = (ideology, ideology \cdot 1_{\{\text{First Observed}\}}, 1_{\{t_1<101\}}, 1_{\{t_1=101\}}, \ldots, 1_{\{t_1=109\}})$, where *ideology* is the NOMINATE score, $1_{\{\text{First Observed}\}}$ is 1 if not a true freshman and 0 otherwise, $1_{\{t_1<101\}}$ if the MC's first term was before the 101^{st} Congress and $1_{\{t_1=s\}}$ is 1 if the i^{th} MC's first term (of the study) was served during the s^{th} Congress and 0 otherwise, for $s = 101, 102, \ldots, 109$.

To determine the number of clusters we used the average silhouette statistic (Rousseeuw, 1987).

*Our variables include the number of district offices operated by the MC, the proportion of legislative staff assigned to the district (rather than Capitol Hill), the number of bill introductions, cosponsorships, and amendments made by the MC, the number of one minute speeches the MC gives on the House floor, the number of editorials and opinion pieces he or she writes for state and national papers, the total amount of campaign money he or she raises, the total amount of money he or she contributes to the party campaign committee (the Democratic Congressional Campaign Committee, which collects and redistributes funds for election campaigns), the total amount of money he or she contributes directly to colleagues, the percentage of the time he or she votes with the party on party votes (defined as votes where a majority of one party votes against a majority of the other), the percentage of the time he or she votes with the party leadership on leadership votes (defined as votes where the leadership of one party votes one way and the leadership of the other party the other), the number of issue areas (out of eighteen) in which he or she introduces legislative bills, the proportion of his or her cosponsorships that are bipartisan (i.e., for measures introduced by an MC from the Republican party), and the percentage of his or her introduced bills that are referred to a committee on which he or she sits.

Each index used in the clustering algorithm was constructed in the following way. First, each raw variable was standardized at each time point to eliminate global temporal patterns or shifts. Second, each variable was associated with one index. Finally, each index was constructed by averaging the associated variables.

We chose this rather than a statistic that is dependent on the number of model parameters, such as AIC or BIC, in order to avoid the number of covariates determining the number of clusters. That is, the number of clusters ought not to depend on the number of covariates used to explain the clustering. The results indicated four clusters of legislator types.

We applied our proposed method and obtained the following results. Figure 5.5 gives the cluster means $\boldsymbol{\mu}_k$. Cluster one, the largest cluster (50% of observations, i.e., 50% of Z_{it}'s equal 1), consists of MCs that we call party soldiers. This group is marked by its very high score on party voting, but relatively low scores on lawmaking, showboating, and other activities. These are the rank-and-file legislators who toe the party line, but do not distinguish themselves in other ways. Cluster two, the second largest cluster comprising 26% of the observations, represents district advocates — these are legislators who devote their efforts to their home districts, but are not particularly active in the lawmaking process and are not committed partisans. Cluster three, with 16% of the observations, are the elites — MCs who are publicly visible, strongly support the party both in voting and giving of contributions, and are very active in the lawmaking process. Across the sample period, the actual party leadership (e.g., Richard Gephardt and Nancy Pelosi) fall into this cluster. The final cluster, the "conscientious objectors," is the smallest, with just 8% of observations. These are MCs who are publicly visible (scoring high on the showboat index), are policy specialists, and are very bipartisan in their coalitions and also have the lowest mean of the four clusters on party support in voting. These MCs chart their own courses, pursuing their policy goals with less regard for the party leadership's preferences.

The results are in line with intuitions about the relationship between ideology and legislative styles. For example, as shown in Figure 5.6, the probability of falling in the party soldier cluster decreases as the ideology score increases: strong liberals are more likely than their moderate counterparts to be party soldiers. On the other hand, the probability of being a conscientious objector increases as the ideology score increases: moderates are more likely than liberals to fall into this category. The probabilities associated with belonging to the district advocates or elites clusters are less linear. The likelihood of being in the elite category peaks just to the left of the mean ideology score — leaders tend to have more pragmatic outlooks, and, while solidly liberal, are seldom ideologically extreme. The probability of being a district advocate, on the other hand, is greatest to the right of the mean ideology score. MCs who focus on their districts often come from heterogeneous constituencies, and, as such, must devote special time and attention to cultivating the district in order to win reelection.

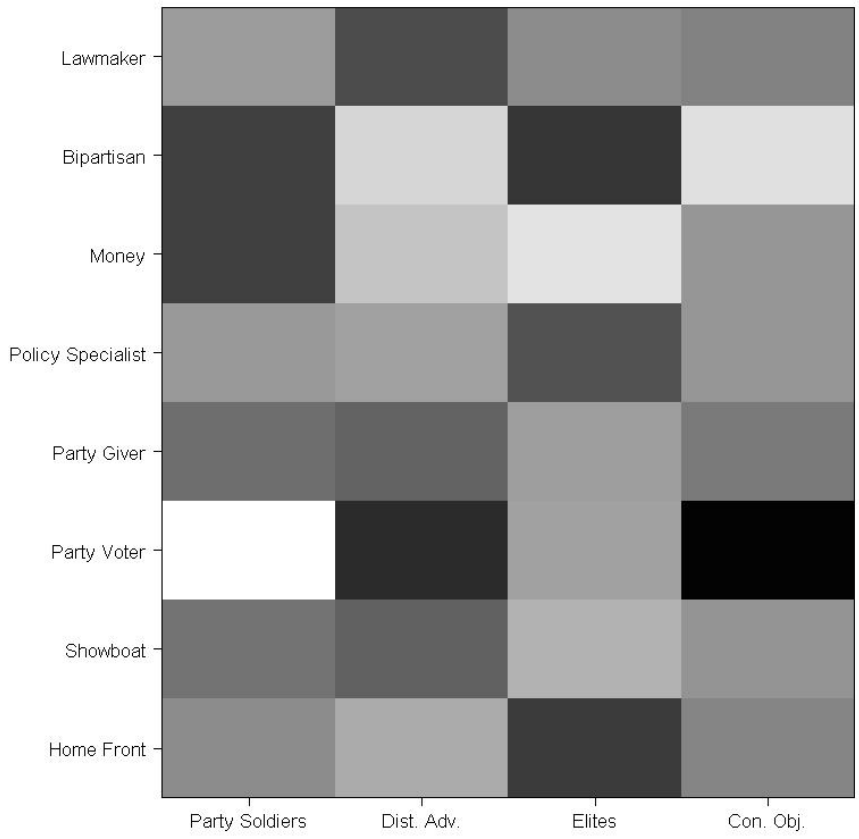


Figure 5.5: Cluster means for the Democrat MCs. The vertical axis corresponds to the behavioral variables on which the MCs were clustered, and the horizontal axis corresponds to the clusters. Lighter hues indicate higher values, darker hues indicate lower values.

Figure 5.7 demonstrates the relationship between the MCs' ideology and their probability of falling in a particular cluster at time $t+1$, given their cluster assignment at time t. These results largely correspond to intuitions as well. For example, party soldiers at time t are more likely to remain as such at $t+1$ if they are very liberal. In contrast, district advocates, conscientious objectors and elites are the most likely to maintain their cluster assignments across time if they are moderate, although the probability of remaining an elite drops off quickly if the MC becomes too moderate.

5.5 Proof of $\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) = \beta_{Z_{t-1}Z_t}$

We wish to find a tractable form of $\pi(Z_t|Z_{t-1}, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1})$. Note that the subscript i is suppressed for ease of notation. The key equality to show is that $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1}, Z_t) = \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|Z_{t-1})$.

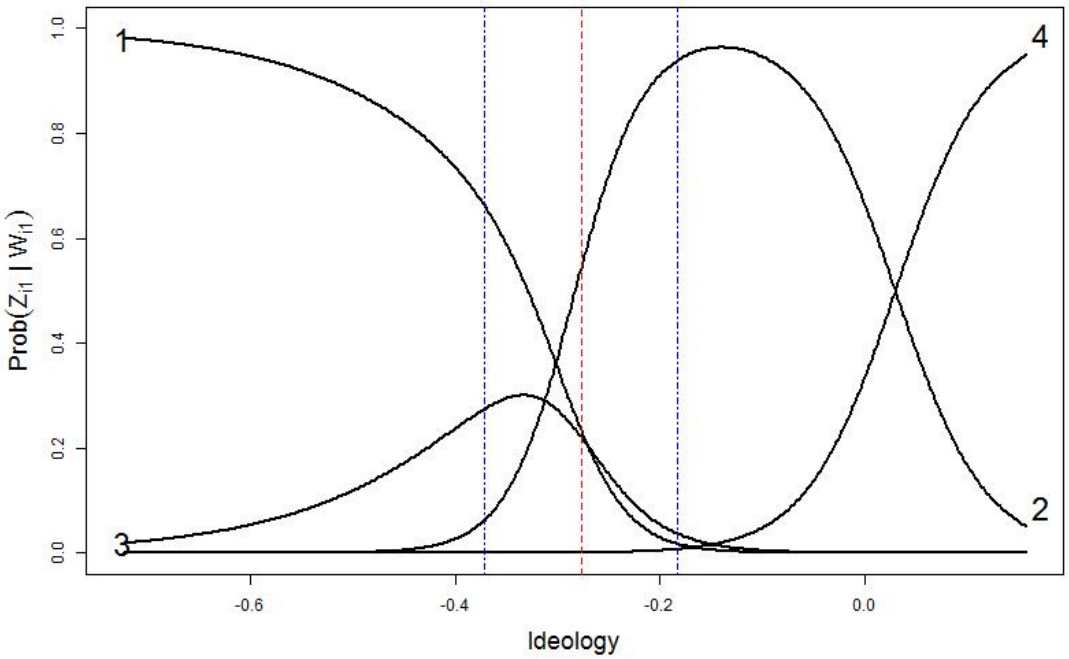
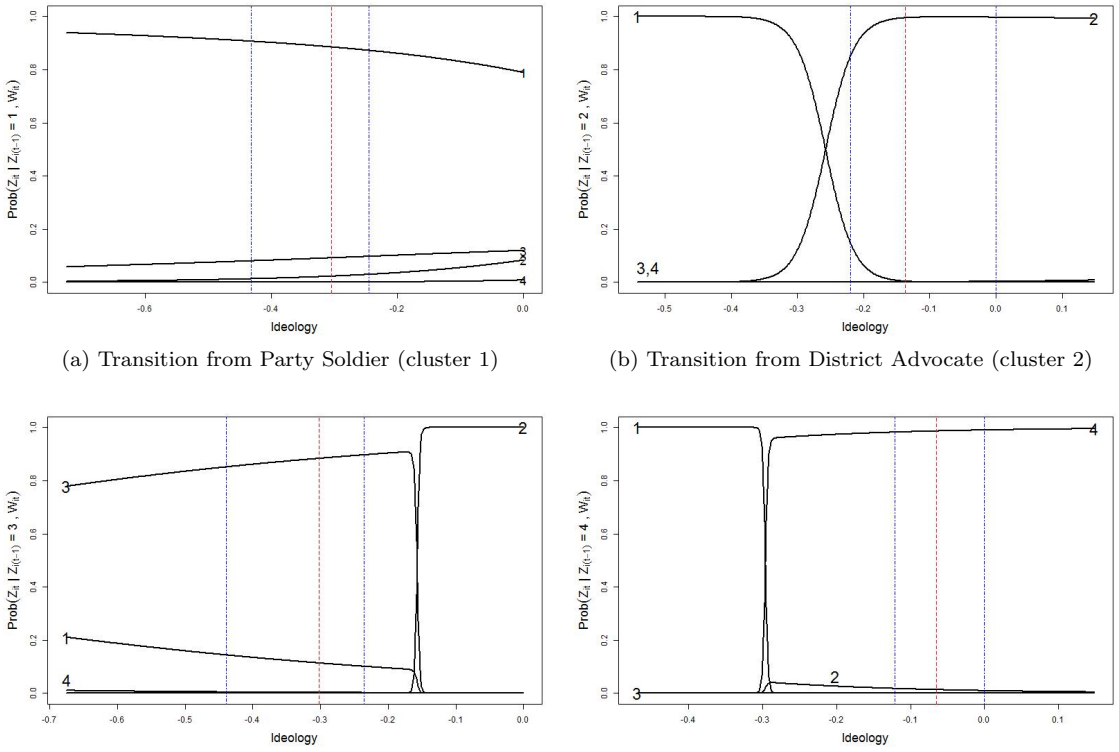


Figure 5.6: Initial cluster probabilities for true freshmen in the Congressional data. Each curve represents, for the varying values along the horizontal axis of the explanatory variable (ideology), the probability of belonging to the cluster whose number is adjacent to the curve. The horizontal axis spans the range of all MCs ideology; the dot-dashed lines give the 25% and 75% quantiles of ideology, and the long dashed line gives the mean ideology.

This is shown by

$$
\begin{aligned}
&\pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1} | Z_{t-1}, Z_t) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_1, \ldots, Z_t) / \pi(Z_{t-1}, Z_t) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1} | Z_1, \ldots, Z_t) \pi(Z_1, \ldots, Z_t) / \big(\pi(Z_{t-1}) \beta_{Z_{t-1} Z_t}\big) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \alpha_{Z_1} \pi(\mathbf{X}_1 | Z_1) \beta_{Z_1 Z_2} \pi(\mathbf{X}_2 | \mathbf{X}_1, Z_2) \cdots \beta_{Z_{t-2} Z_{t-1}} \pi(\mathbf{X}_{t-1} | \mathbf{X}_{t-2}, Z_{t-1}) / \pi(Z_{t-1}) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1} | Z_1, \ldots, Z_{t-1}) \pi(Z_1, \ldots, Z_{t-1}) / \pi(Z_{t-1}) \\
&= \sum_{Z_1, \ldots, Z_{t-2}} \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_1, \ldots, Z_{t-1}) / \pi(Z_{t-1}) \\
&= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}, Z_{t-1}) / \pi(Z_{t-1}) \\
&= \pi(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1} | Z_{t-1}).
\end{aligned}
$$



(a) Transition from Party Soldier (cluster 1)

(b) Transition from District Advocate (cluster 2)

(c) Transition from Elite (cluster 3)

(d) Transition from Conscientious Objector (cluster 4)

Figure 5.7: Transition cluster probabilities for Congressional data. Each curve represents, for the varying values along the horizontal axis of the explanatory variable (ideology), the probability of belonging to the cluster whose number is adjacent to the curve given that they currently belong to cluster (a) 1 (party soldiers) (b) 2 (district advocates) (c) 3 (elites) (d) 4 (conscientious objectors). The horizontal axis spans the range of all MCs ideology; the dot-dashed lines give the 25% and 75% quantiles of ideologies of MCs who have belonged to the corresponding cluster, and the long dashed line gives the mean.

From this it is straightforward to obtain the result in the following way

$$\begin{aligned}\pi(Z_t|Z_{t-1},\mathbf{X}_1,\ldots,\mathbf{X}_{t-1}) &= \frac{\pi(\mathbf{X}_1,\ldots,\mathbf{X}_{t-1},Z_t|Z_{t-1})}{\pi(\mathbf{X}_1,\ldots,\mathbf{X}_{t-1}|Z_{t-1})} \\ &= \frac{\pi(\mathbf{X}_1,\ldots,\mathbf{X}_{t-1}|Z_{t-1},Z_t)\beta_{Z_{t-1}Z_t}}{\pi(\mathbf{X}_1,\ldots,\mathbf{X}_{t-1}|Z_{t-1})} \\ &= \beta_{Z_{t-1}Z_t}.\end{aligned}$$

5.6 Deriving the Tractable Form of $Q(\Theta,\widehat{\Theta})$

Let $\mathbf{Z}_i$ denote the latent cluster assignments $Z_{i1},\ldots,Z_{iT_i}$. Then we have

$$\begin{aligned}&Q(\Theta,\widehat{\Theta}) \\ &= \sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n}\sum_{i=1}^{n}\left[\log(\alpha_{Z_{it}}\pi(\mathbf{X}_{i1}|Z_{i1},\Theta)) + \sum_{t=2}^{T_i}\log(\beta_{Z_{i(t-1)}Z_{it}}\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it},\Theta))\right]\prod_{j=1}^{n}\mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j,\widehat{\Theta}) \\ &= \sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n}\sum_{i=1}^{n}\sum_{\ell_1=1}^{K}\cdots\sum_{\ell_{T_i}=1}^{K}1_{[Z_{i1}=\ell_1]}\cdots 1_{[Z_{iT_i}=\ell_T]} \\ &\cdot\left[\log(\alpha_{\ell_1}\pi(\mathbf{X}_{i1}|Z_{i1}=\ell_1,\Theta)) + \sum_{t=2}^{T_i}\log(\beta_{\ell_{t-1}\ell_t}\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it}=\ell_t,\Theta))\right]\prod_{j=1}^{n}\mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j,\widehat{\Theta}) \\ &= \sum_{i=1}^{n}\sum_{\ell_1=1}^{K}\cdots\sum_{\ell_{T_i}=1}^{K}\left[\log(\alpha_{\ell_1}\pi(\mathbf{X}_{i1}|Z_{i1}=\ell_1,\Theta)) + \sum_{t=2}^{T_i}\log(\beta_{\ell_{t-1}\ell_t}\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)},Z_{it}=\ell_t,\Theta))\right] \\ &\quad\cdot\sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n}1_{[Z_{i1}=\ell_1]}\cdots 1_{[Z_{iT_i}=\ell_{T_i}]}\prod_{j=1}^{n}\mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j,\widehat{\Theta}). \qquad (5.36)\end{aligned}$$

We can simplify this last expression by noting that for a fixed i,

$$\begin{aligned}&\sum_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n}1_{[Z_{i1}=\ell_1]}\cdots 1_{[Z_{iT_i}=\ell_T]}\prod_{j=1}^{n}\mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j,\widehat{\Theta}) \\ =\ &\left[\prod_{j\neq i}\sum_{\mathbf{Z}_j}\mathbb{P}(\mathbf{Z}_j|\mathcal{X}_j,\widehat{\Theta})\right]\mathbb{P}(Z_{i1}=\ell_1,\ldots,Z_{iT_i}=\ell_{T_i}|X_i,\widehat{\Theta}) \\ =\ &\mathbb{P}(Z_{i1}=\ell_1|\mathcal{X}_i,\widehat{\Theta})\prod_{t=1}^{T_i}\mathbb{P}(Z_{it}=\ell_t|Z_{i(t-1)}=\ell_{t-1},\mathcal{X}_i,\widehat{\Theta}). \qquad (5.37)\end{aligned}$$

Thus we can rewrite Q in the more tractable form

$$
\begin{aligned}
&Q(\Theta, \widehat{\Theta}) \\
&= \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\alpha_{\ell_1}) + \sum_{t=2}^{T_i} \log(\beta_{\ell_{t-1}\ell_t}) \right] \mathbb{P}(Z_{i1} = \ell_1 | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s | Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}) \\
&+ \sum_{i=1}^{n} \sum_{\ell_1=1}^{K} \cdots \sum_{\ell_{T_i}=1}^{K} \left[\log(\pi(\mathbf{X}_{i1} | Z_{i1} = \ell_1, \Theta)) + \sum_{t=2}^{T_i} \log(\pi(\mathbf{X}_{it} | \mathbf{X}_{i(t-1)}, Z_{it} = \ell_t, \Theta)) \right] \\
&\qquad \cdot \mathbb{P}(Z_{i1} = \ell_1 | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} = \ell_s | Z_{i(s-1)} = \ell_{s-1}, \mathcal{X}_i, \widehat{\Theta}). \qquad (5.38)
\end{aligned}
$$

5.7 Deriving the Parameter Updates

5.7.1 Update $\boldsymbol{\alpha}$

Letting λ_α be the Lagrange multiplier corresponding to the constraint $\sum_{\ell=1}^{K} \alpha_\ell = 1$, we have

$$
\begin{aligned}
\frac{\partial Q}{\partial \alpha_h} &= \frac{\partial}{\partial \alpha_h} \left\{ \sum_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\alpha_{Z_{i1}}) \mathbb{P}(Z_{i1} | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} | Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) - \lambda_\alpha \left(\sum_{\ell=1}^{K} \alpha_\ell - 1 \right) \right\} \\
&= \frac{1}{\alpha_h} \sum_{i=1}^{n} \mathbb{P}(Z_{i1} = h | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \sum_{Z_{is}=1}^{K} \mathbb{P}(Z_{is} | Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) - \lambda_\alpha, \qquad (5.39)
\end{aligned}
$$

hence

$$
\hat{\alpha}_h = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(Z_{i1} = h | \mathcal{X}_i, \widehat{\Theta}). \qquad (5.40)
$$

5.7.2 Update $\boldsymbol{\alpha}$

Here we have K Lagrange multipliers corresponding to the constraint on each row of the transition matrix, specifically that $\sum_{k=1}^{K} \beta_{jk} = 1$ for $j = 1, \ldots, K$. Denoting these as λ_j, we have

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_{hk}} &= \frac{\partial}{\partial \beta_{hk}} \left\{ \sum_{i=1}^{n} \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \sum_{t=2}^{T_i} \log(\beta_{Z_{i(t-1)} Z_{it}}) \mathbb{P}(Z_{i1} | \mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is} | Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \right. \\
&\qquad \left. - \sum_{j=1}^{K} \lambda_j \left(\sum_{m=1}^{K} \beta_{jm} - 1 \right) \right\}. \qquad (5.41)
\end{aligned}
$$

We can make this expression more tractable by noticing that for any i and $t \geq 2$

$$\begin{aligned}
&\sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \\
&= \sum_{Z_{i1}=1}^{K} \left(\mathbb{P}(Z_{i1}|\mathcal{X}_i, \widehat{\Theta}) \sum_{Z_{i2}=1}^{K} \left(\mathbb{P}(Z_{i2}|Z_{i1}, \mathcal{X}_i, \widehat{\Theta}) \cdots \right. \right. \\
&\quad \left. \left. \cdots \sum_{Z_{it}=1}^{K} \left(\log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}) \cdots \sum_{Z_{iT_i}=1}^{K} \mathbb{P}(Z_{iT_i}|Z_{i(T_i-1)}, \mathcal{X}_i, \widehat{\Theta}) \right) \cdots \right) \right) \\
&= \sum_{Z_{i(t-1)}=1}^{K} \mathbb{P}(Z_{i(t-1)}|\mathcal{X}_i, \widehat{\Theta}) \sum_{Z_{it}=1}^{K} \log(\beta_{Z_{i(t-1)}Z_{it}}) \mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}). \qquad (5.42)
\end{aligned}$$

Thus we can find that

$$\frac{\partial Q}{\partial \beta_{hk}} = \frac{1}{\beta_{hk}} \sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta}) \mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta}) - \lambda_h \qquad (5.43)$$

and hence

$$\hat{\beta}_{hk} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta}) \mathbb{P}(Z_{it} = k|Z_{i(t-1)} = h, \mathcal{X}_i, \widehat{\Theta})}{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \mathbb{P}(Z_{i(t-1)} = h|\mathcal{X}_i, \widehat{\Theta})}. \qquad (5.44)$$

5.7.3 Update λ, $\boldsymbol{\mu}_k$ and Σ_k

Before we proceed, we need a few preliminaries. The partial derivative of (5.3) and (5.4) with respect to $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, is

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_k} \log(\pi(\mathbf{X}_{i1}|k, \Theta)) \qquad (5.45) \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} (-\frac{1}{2}) \left[(\mathbf{X}_{i1} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{X}_{i1} - \boldsymbol{\mu}_k) \right] \\
&= \Sigma_k^{-1} \mathbf{X}_{i1} - \Sigma_k^{-1} \boldsymbol{\mu}_k \qquad (5.46)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_k} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, k, \Theta)) \qquad (5.47) \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} (-\frac{1}{2}) \left[(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)} - \lambda \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)} - \lambda \boldsymbol{\mu}_k) \right] \\
&= \lambda \Sigma_k^{-1} (\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) - \lambda^2 \Sigma_k^{-1} \boldsymbol{\mu}_k, \qquad (5.48)
\end{aligned}$$

with respect to Σ_k^{-1}, $k = 1, \ldots, K$, is

$$\begin{aligned}
& \frac{\partial}{\partial \Sigma_k^{-1}} \log(\pi(\mathbf{X}_{i1}|k, \Theta)) \\
= & \frac{\partial}{\partial \Sigma_k^{-1}} \left[\frac{1}{2} \log |\Sigma_k^{-1}| - \frac{1}{2} tr(\Sigma_k^{-1}(\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)') \right] \\
= & \Sigma_k - \frac{1}{2} diag(\Sigma_k) - (\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)' + \frac{1}{2} diag((\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)') \\
= & \Sigma_k \circ \Big(\frac{1}{2} I_p \Big) - (\mathbf{X}_{it} - \boldsymbol{\mu}_k)(\mathbf{X}_{it} - \boldsymbol{\mu}_k)' \circ \Big(\frac{1}{2} I_p \Big), \qquad (5.49)
\end{aligned}$$

where $\circ$ is the Hadamard product, and similarly

$$\begin{aligned}
& \frac{\partial}{\partial \Sigma_k^{-1}} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, k, \Theta)) \\
& = \Sigma_k \circ \Big(\frac{1}{2} I_p \Big) - \big(\mathbf{X}_{it} - \lambda \boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)} \big) \big(\mathbf{X}_{it} - \lambda \boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)} \big)' \circ \Big(\frac{1}{2} I_p \Big), \qquad (5.50)
\end{aligned}$$

and with respect to λ is

$$\begin{aligned}
& \frac{\partial}{\partial \lambda} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, k, \Theta)) \\
= & \frac{\partial}{\partial \lambda} (-\frac{1}{2}) \left[(\mathbf{X}_{it} - \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\mathbf{X}_{it} - \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \mathbf{X}_{i(t-1)}) \right] \\
= & (\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \lambda (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})' \Sigma_k^{-1} (\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}). \qquad (5.51)
\end{aligned}$$

To make the form of $Q(\Theta, \widehat{\Theta})$ more tractable, we notice that for any i

$$\begin{aligned}
& \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\pi(\mathbf{X}_{i1}|Z_{i1}, \Theta)) \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \\
= & \sum_{Z_{i1}=1}^{K} \log(\pi(\mathbf{X}_{i1}|Z_{i1}, \Theta)) \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \prod_{s=2}^{T_i} \sum_{Z_{is}=1}^{K} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \\
= & \sum_{Z_{i1}=1}^{K} \log(\pi(\mathbf{X}_{i1}|Z_{i1}, \Theta)) \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}), \qquad (5.52)
\end{aligned}$$

and also for any i and $t \geq 2$

$$\begin{aligned}
&\sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{iT_i}=1}^{K} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta))\mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \prod_{s=2}^{T_i} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta}) \\
&= \sum_{Z_{i1}=1}^{K} \cdots \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{i1}|\mathcal{X}_j, \widehat{\Theta}) \left[\prod_{s=2}^{t-1} \mathbb{P}(Z_{is}|Z_{i(s-1)}, \mathcal{X}_i, \widehat{\Theta})\right] \\
&\quad \cdot \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta))\mathbb{P}(Z_{it}|Z_{i(t-1)}, \mathcal{X}_i, \widehat{\Theta}) \\
&\quad \cdot \left[\prod_{u=t+1}^{T_i} \sum_{Z_{iu}=1}^{K} \mathbb{P}(Z_{iu}|Z_{i(u-1)}, \mathcal{X}_i, \widehat{\Theta})\right] \\
&= \sum_{Z_{it}=1}^{K} \log(\pi(\mathbf{X}_{it}|\mathbf{X}_{i(t-1)}, Z_{it}, \Theta))\mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta}). \qquad (5.53)
\end{aligned}$$

With the above it is not difficult to find, for each distribution parameter, the value which maximizes $Q(\Theta, \widehat{\Theta})$ where the other parameters are fixed, hence finding the solutions to the coordinate ascent approach. The solutions are found as follows. The derivative of $Q(\Theta, \widehat{\Theta})$ with respect to λ is

$$\begin{aligned}
\frac{\partial Q}{\partial \lambda} &= \sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta}) \\
&\quad \cdot \left[(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)}) - \lambda(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})\right] \qquad (5.54)
\end{aligned}$$

Hence the update for λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{Z_{it}=1}^{K} \mathbb{P}(Z_{it}|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{i(t-1)})}{\sum_{j=1}^{n} \sum_{s=2}^{T_j} \sum_{Z_{js}=1}^{K} \mathbb{P}(Z_{js}|\mathcal{X}_j, \widehat{\Theta})(\boldsymbol{\mu}_k - \mathbf{X}_{j(s-1)})'\Sigma_k^{-1}(\boldsymbol{\mu}_k - \mathbf{X}_{j(s-1)})}. \qquad (5.55)$$

The derivative of $Q(\Theta, \widehat{\Theta})$ with respect to $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, is

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{\mu}_k} &= \sum_{i=1}^{n} \Bigg\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) \left[\Sigma_k^{-1}\mathbf{X}_{i1} - \Sigma_k^{-1}\boldsymbol{\mu}_k\right] \\
&\quad + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta}) \left[\lambda\Sigma_k^{-1}(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) - \lambda^2\Sigma_k^{-1}\boldsymbol{\mu}_k\right] \Bigg\}. \qquad (5.56)
\end{aligned}$$

Hence the update for $\boldsymbol{\mu}_k$ is

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} \left\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{X}_{i1} + \lambda \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)}) \right\}}{\sum_{i=1}^{n} \left\{ \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) + \lambda^2 \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta}) \right\}}. \qquad (5.57)$$

Let $\mathbf{A}_{ik} = \mathbf{X}_{i1} - \boldsymbol{\mu}_k$ and $\mathbf{B}_{itk} = \mathbf{X}_{it} - \lambda\boldsymbol{\mu}_k - (1-\lambda)\mathbf{X}_{i(t-1)}$. Then the derivative of $Q(\Theta, \widehat{\Theta})$ with respect to Σ_k^{-1}, $k = 1, \ldots, K$, is

$$\begin{aligned}\frac{\partial Q}{\partial \Sigma_k^{-1}} = \sum_{i=1}^{n}\Bigg\{ & \mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) \left[\Sigma_k \circ \left(\frac{1}{2} I_p\right) - \mathbf{A}_{ik}\mathbf{A}_{ik}' \circ \left(\frac{1}{2} I_p\right)\right] \\ & + \sum_{t=2}^{T_i} \mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta}) \left[\Sigma_k \circ \left(\frac{1}{2} I_p\right) - \mathbf{B}_{itk}\mathbf{B}_{itk}' \circ \left(\frac{1}{2} I_p\right)\right]\Bigg\}. \end{aligned} \tag{5.58}$$

Hence the update for Σ_k can be found as

$$\widehat{\Sigma}_k = \frac{\sum_{i=1}^{n}\left\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{A}_{ik}\mathbf{A}_{ik}' + \sum_{t=2}^{T_i}\mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})\mathbf{B}_{itk}\mathbf{B}_{itk}'\right\}}{\sum_{i=1}^{n}\left\{\mathbb{P}(Z_{i1} = k|\mathcal{X}_i, \widehat{\Theta}) + \sum_{t=2}^{T_i}\mathbb{P}(Z_{it} = k|\mathcal{X}_i, \widehat{\Theta})\right\}}.$$

Chapter 6

Community Detection in Dynamic Networks

Researchers are often interested in detecting communities within dyadic data. Clustering these data into communities can lead to better understanding of the organization of the objects in the network, and, for dynamic networks, how this organization evolves over time.

Xing et al. (2010) developed a dynamic mixed membership stochastic blockmodel. This work builds off of the stochastic blockmodel (Holland et al., 1983), further developed into the mixed membership blockmodel (Airoldi et al., 2005). In the work of Xing et al. (2010), each actor has an individual membership probability (time-varying) vector and, based on this probability vector, can choose certain roles with which to interact with other actors. A different approach to be taken in this chapter begins with the work of Hoff et al. (2002) where the actors are embedded either within a latent Euclidean space, referred to as the distance model, or within a hypersphere, referred to as the projection model. Handcock et al. (2007) used their distance model and performed community detection on the latent actor positions. Further, the distance model of Hoff et al. (2002) was extended in Chapters 2 through 4. We build on these latent space models with the aim of developing a method of clustering directed or undirected dynamic network data into communities.

We assume that there is a fixed number of communities that constitute the observed dynamic network. These communities are treated as constant over time, thus allowing the researcher to interpret the clusters; the community membership, however, may vary over time, thereby reflecting realistic community evolution. The clustering, then, is performed at each time point, though our proposed methods borrow information across time in assigning actors to communities.

Applying a latent space model allows the user to capture local and global structure and outputs meaningful visualization of the data, providing rich qualitative information. Unlike a blockmodel approach, the latent space approach naturally accounts for transitivity, reciprocity and, in our models, individual effects which reflect the propensity to send or receive edges (although some degree corrected blockmodels have been developed for static networks to account for such individual effects; see, e.g., Yan et al., 2014; Karrer and Newman, 2011).

The remainder of the chapter is as follows. Section 6.1 gives the model and methodology. Section 6.2 gives estimation methods. Section 6.3 describes a simulation study. Section 6.4 reports the results from analyzing Newcomb's fraternity data (Newcomb, 1956) and world trade data. Section 6.5 gives the full conditional distributions mentioned earlier in the chapter, and Section 6.6 gives the derivations for the variational Bayes estimation algorithm described earlier in the chapter.

6.1 Models

The data we will analyze are of the form $(\mathcal{N}, \{\mathcal{E}_t : t \in \{1, 2, \ldots, T\})$, where $\mathcal{N}$ is the set of all actors (also called by some authors nodes or vertices), and $\mathcal{E}_t \subseteq \{\{i, j\}, i, j \in \mathcal{N}, i \neq j\}$ is the set of edges at time t. The edges $\mathcal{E}_t$ can be viewed as an adjacency matrix Y_t with entries y_{ijt} denoting the edge from actor i to actor j at time t. The latent space approach to modeling networks assumes that there is, for each actor at each time point, a latent position within a network space. We will assume that at each time point, each actor belongs to one of a fixed number G of clusters; this cluster assignment may change over time. We will denote the latent position of actor i at time t as $\mathbf{X}_{it}$ and the cluster assignment for actor i at time t as $\boldsymbol{Z}_{it}$, a G dimensional vector in which one element is 1 and the others are zero. We will also let $\mathcal{X}_t = (\boldsymbol{X}_{1t}', \ldots, \boldsymbol{X}_{nt}')'$ and $\mathcal{Z}_t = (\boldsymbol{Z}_{1t}', \ldots, \boldsymbol{Z}_{nt}')'$. While the dependency structure of the model may vary, we assume throughout the chapter that given the latent positions $\mathcal{X}_t$, Y_t and Y_s, $s \neq t$, are conditionally independent; in many cases (such as binary networks) this assumption can be further extended such that y_{ijt} and $y_{i'j't}$ are conditionally independent given $\mathcal{X}_t$.

6.1.1 Distance Model

Within the context of the distance model, the network is embedded within a latent Euclidean space, where the probability of edge formation increases as the Euclidean distance between actors decreases. Let $D(\mathcal{X}_t)$ denote the $n \times n$ distance matrix constructed such that $\big(D(\mathcal{X}_t)\big)_{ij} \triangleq d_{ijt} = \|\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}\|$. In general we will assume that the density of Y_t can be written as a function of the distance matrix $D(\mathcal{X}_t)$ and some set of likelihood parameters, which we will denote as θ_ℓ. For example, the original likelihood for binary networks in Hoff et al. (2002) is

$$\mathbb{P}(y_{ijt} = 1 | \mathcal{X}_t, \theta_\ell) = \frac{\exp\{y_{ijt}\eta_{ijt}\}}{1 + \exp\{\eta_{ijt}\}}, \qquad \eta_{ijt} = \alpha - d_{ijt}, \tag{6.1}$$

where in this context $\theta_\ell = \{\alpha\}$. Variants of this likelihood have been proposed, such as in Sarkar and Moore (2005), Krivitsky et al. (2009), and Chapter 2. This last was then extended to account for a wide range of weighted networks in Chapter 3. Other likelihoods may be better suited for various other types of weighted edges (see, e.g., Chapter 4).

Handcock et al. (2007) clustered static network data by clustering the latent positions via a normal mixture model. This cannot be directly applied to dynamic network data since the latent positions must have some sort of temporal dependency imposed. Therefore we propose applying the model-based longitudinal clustering model given in the previous chapter to the latent positions. Our focus here is the modeling of the latent positions, which can then be used for whatever likelihood formulation is most appropriate to the data. We will now describe this model for the latent variables.

We make two assumptions on the latent positions and the cluster assignments. First, the cluster assignments are assumed to follow a Markov process, i.e.,

$$\boldsymbol{Z}_{it}|\boldsymbol{Z}_{i1},\ldots,\boldsymbol{Z}_{i(t-1)} \stackrel{\mathcal{D}}{=} \boldsymbol{Z}_{it}|\boldsymbol{Z}_{i(t-1)}.$$

Second, given the current cluster assignment and all previous cluster assignments and latent positions, we assume the current latent positions depend only on the previous latent positions and the current cluster assignments, i.e.,

$$\boldsymbol{X}_{it}|\boldsymbol{X}_{i1},\ldots,\boldsymbol{X}_{i(t-1)},\boldsymbol{Z}_{i1},\ldots,\boldsymbol{Z}_{it} \stackrel{\mathcal{D}}{=} \boldsymbol{X}_{it}|\boldsymbol{X}_{i(t-1)},\boldsymbol{Z}_{it}.$$

The joint density of the latent positions and the cluster assignments is given as

$$\begin{aligned}&\pi(\{\mathcal{X}_t\}_{t=1}^T,\{\mathcal{Z}_t\}_{t=1}^T)\\&=\prod_{i=1}^n\prod_{g=1}^G\left[\beta_{0g}N(\boldsymbol{X}_{i1}|\boldsymbol{\mu}_g,\Sigma_g)\right]^{Z_{i1g}}\prod_{t=2}^T\prod_{h=1}^G\left[\prod_{k=1}^G\left[\beta_{hk}N(\boldsymbol{X}_{it}|\lambda\boldsymbol{\mu}_k+(1-\lambda)\boldsymbol{X}_{i(t-1)},\Sigma_k)\right]^{Z_{itk}}\right]^{Z_{i(t-1)h}},\end{aligned} \tag{6.2}$$

where $N(\boldsymbol{X}|\boldsymbol{\mu},\Sigma)$ is the normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\boldsymbol{X}$. Thus the communities are each modeled as a multivariate normal distribution in the latent space with mean $\boldsymbol{\mu}_g$ and covariance matrix Σ_g. The mean of the latent position $\boldsymbol{X}_{it}$ is then modeled as $\lambda\boldsymbol{\mu}_g + (1-\lambda)\boldsymbol{X}_{i(t-1)}$, $\lambda \in (0,1)$, which is a blending of the current cluster effect $\boldsymbol{\mu}_g$ with the individual temporal effect $\boldsymbol{X}_{i(t-1)}$. The β_{0g}'s determine the probability of initially belonging to the

g^{th} community and the β_{hk}'s determine the probability of transitioning from the h^{th} community to the k^{th} community.

6.1.2 Projection Model

Cox and Cox (1991) and Banerjee et al. (2005) gave many contexts in which there has been empirical evidence that embedding data onto a hypersphere and/or using cosine distances is preferable to Euclidean space/distances. Here we continue this tradition by embedding dynamic network data onto the hypersphere. In this section we assume the more specific, but most commonly encountered, context of directed binary edges (the model to be proposed can be simplified for undirected edges). In the projection model, every actor is embedded within some latent hypersphere; the probability of an edge forming between two actors depends on the angle, rather than the Euclidean distance, between them. Thus it is the angle between any two actors that represents the "closeness" of the actors. Though the latent space is strictly a Euclidean space rather than a hypersphere, it is more helpful to think of the positions within $\Re^p$ as unit vectors on a $p-1$ dimensional hypersphere with individual edge propensities reflected in the magnitude of the latent positions.

Our proposed likelihood of the adjacency matrices adapts the likelihood of the projection model originally proposed by Hoff et al. (2002), and extends Durante and Dunson (2014) to allow for directed edges. The specific form of the likelihood is given as

$$\pi(\{Y_t\}_{t=1}^T|\{\mathcal{X}_t\}_{t=1}^T,\theta_\ell) = \prod_{t=1}^{T}\prod_{i\neq j}\frac{\exp\{y_{ijt}\eta_{ijt}\}}{1+\exp\{\eta_{ijt}\}}, \tag{6.3}$$

$$\eta_{ijt} = \alpha + s_j \boldsymbol{X}_{it}'\boldsymbol{X}_{jt} \tag{6.4}$$

$$= \alpha + \|\boldsymbol{X}_{it}\|\cdot(s_j\|\boldsymbol{X}_{jt}\|)\cdot\cos(\phi_{ijt}), \tag{6.5}$$

where ϕ_{ijt} is the angle between $\boldsymbol{X}_{it}$ and $\boldsymbol{X}_{jt}$; in this context $\theta_\ell = \{\alpha, \boldsymbol{s}\}$, where α reflects a baseline edge propogation rate and $\boldsymbol{s} = (s_1,\ldots,s_n)$ is a vector of actor specific parameters that reflect the tendency of the actors to receive edges. While (6.4) is simpler, (6.5) makes it clear how the probability of an edge from i to j is made up of some constant plus the product of the sending effect of i, the receiving effect of j, and the closeness between i and j in the latent space as measured by the cosine of the angle between the two actors.

The question remains as to how to perform clustering. With the projection model the latent positions are embedded within a hypersphere, and thus the clustering must be done in a fundamentally different way than that done for the distance model. Since we would expect a group of highly

connected actors to have small angles between them all, we propose clustering based on the angles of the actors' latent positions.

We first assume that the latent positions follow a hidden Markov model, with the cluster assignments as the hidden states. That is, the cluster assignments follow a Markov process (i.e., given $\boldsymbol{Z}_{i(t-1)}$, $\boldsymbol{Z}_{it}$ is conditionally independent of $\boldsymbol{Z}_{i(t-s)}$ for any $s > 1$), and given the cluster assignments $\mathcal{Z}_t$, the latent positions $\boldsymbol{X}_t$ are assumed to be conditionally independent of $\boldsymbol{X}_s$ for any $s \neq t$.

The joint density on the latent positions and cluster assignments is given as

$$\begin{aligned}
&\pi(\{\mathcal{X}_t\}_{t=1}^T, \{\mathcal{Z}_t\}_{t=1}^T) \\
&= \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\beta_{0g} N(\boldsymbol{X}_{i1} | r_i \boldsymbol{u}_g, \tau_i^{-1} I_p)\right]^{Z_{i1g}} \prod_{t=2}^{T} \prod_{h=1}^{G} \left[\prod_{k=1}^{G} \left[\beta_{hk} N(\boldsymbol{X}_{it} | r_i \boldsymbol{u}_k, \tau_i^{-1} I_p)\right]^{Z_{itk}}\right]^{Z_{i(t-1)h}}. \qquad (6.6)
\end{aligned}$$

As with the distance model of Section 6.1.1, the communities are modeled as multivariate normal distributions within the latent space. Here $\boldsymbol{r} = (r_1, \ldots, r_n)$ are individual effects representing the individual propensities to send edges, $\boldsymbol{u}_g$ is the unit vector corresponding to the direction of the g^{th} community, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ are the precision parameters, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0G})$ is the vector of initial cluster probabilities, $\boldsymbol{\beta}_h = (\beta_{h1}, \ldots, \beta_{hG})$, $h = 1, \ldots, G$, is the vector of probabilities for transitioning from cluster h, and I_p is the $p \times p$ identity matrix.

From (6.6) we can see how the different aspects of the network are captured in the joint density of $\{\mathcal{X}_t\}_{t=1}^T$ and $\{\mathcal{Z}_t\}_{t=1}^T$. The clusters are completely determined by the unit vectors $\boldsymbol{u}_g$, and it is these unit vectors which determine the direction in which the latent actor positions aim, and hence with whom the actors have the highest probability of forming an edge. The permanence and transience of the clusters are captured in the transition probabilities $\boldsymbol{\beta}_h$, $h = 1, \ldots, G$. The individual effects are captured by the parameters $\boldsymbol{r}$, which determine the expected magnitude of $\boldsymbol{X}_{it}$, and $\boldsymbol{s}$. Note that the parameterization (6.4) of the likelihood (6.3) is not identifiable, as $\boldsymbol{s}$ and $\mathcal{X}_t$ can be scaled arbitrarily. The estimation is done within a Bayesian framework, however, and thus by fixing the hyperparameters corresponding to the priors on the unknown parameters, the posterior distribution is identifiable.

6.2 Estimation

Our estimation is done within the Bayesian framework, with the goal of finding the maximum *a posteriori* (MAP) estimators of the unknown parameters and latent positions.

6.2.1 MCMC for the Distance Model

We propose a Markov chain Monte Carlo (MCMC) method to obtain posterior modes to estimate the latent positions and model parameters of the distance model given in Section 6.1.1. Specifically, we implement a Metropolis-Hastings within Gibbs sampler.

We assign the following priors:

$$\begin{aligned}
\lambda &\sim N_{(0,1)}(\nu_\lambda, \xi_\lambda), && (6.7)\\
\boldsymbol{\mu}_g &\sim N(\mathbf{0}, \tau^2 I_p) \quad \text{for } g = 1, \ldots, G, && (6.8)\\
\Sigma_g &\sim W^{-1}(p+1, diag(\gamma_1, \ldots, \gamma_p)) \quad \text{for } g = 1, \ldots, G, && (6.9)\\
\tau^2 &\sim \Gamma^{-1}(a, b), && (6.10)\\
\gamma_\ell &\sim \Gamma(c, 1/d) \quad \text{for } \ell = 1, \ldots, p, && (6.11)\\
\boldsymbol{\beta}_h &\sim Dir(1, \ldots, 1) \quad \text{for } h = 0, 1, \ldots, G, && (6.12)
\end{aligned}$$

where $N_{(0,1)}(\mu, \sigma^2)$ indicates the normal distribution with mean μ and variance σ^2 truncated to the range of $(0, 1)$, $W^{-1}(a, B)$ indicates the inverse Wishart distribution with degrees of freedom a and scale matrix B, $diag(d_1, \ldots, d_K)$ indicates a $K \times K$ diagonal matrix with $d_1, \ldots, d_K$ on the diagonal, $Dir(a_1, \ldots, a_K)$ indicates the Dirichlet distribution with parameters a_1 to a_K, $\Gamma^{-1}(a, b)$ indicates the inverse gamma distribution with shape and scale parameters a and b respectively, and $\Gamma(a, b)$ indicates the gamma distribution with shape and scale parameters a and b respectively. Additionally, there will be some prior $\pi(\theta_\ell)$ on the likelihood parameters θ_ℓ that will depend on the formulation of the likelihood.

With the exception of the latent positions and θ_ℓ, these priors are conjugate with respect to the full conditional distributions; these distributions are given in Section 6.5. For the latent positions and likelihood parameters, Metropolis-Hastings steps are necessary.

6.2.2 Variational Bayesian Inference for the Projection Model

We apply a mean field variational Bayes (VB) algorithm for estimating the latent positions and model parameters of the projection model given in Section 6.1.2. Variational Bayes procedures have been gaining popularity in large part due to their greatly decreased computational cost in comparison with most sampling methods. Unlike the MCMC approach given for the distance model which obtains samples asymptotically from the posterior distribution, the VB algorithm here iteratively finds an

approximation to the posterior density $\pi(\{\mathcal{X}_t, \mathcal{Z}_t\}_{t=1}^T, \theta_\ell, \theta_p | \{Y_t\}_{t=1}^T)$, where θ_p is all the remaining model parameters corresponding to the prior on $\{\mathcal{X}_t, \mathcal{Z}_t\}_{t=1}^T$. Using the mean field VB implies that we are finding a factorized approximation Q of the posterior which minimizes the Kullback-Liebler divergence between the true posterior and Q. This factorized form will be given shortly.

Salter-Townshend and Murphy (2013) applied VB to the static latent space cluster model for networks given by Handcock et al. (2007) (which is a static form of the distance model). Within this iterative scheme, the factorized distributions of the latent positions and many of the model parameters required a numerical optimization, as a closed form analytical solution was unavailable. By utilizing the projection model as described in Section 6.1.2 we can find closed form solutions for each iteration, thereby reducing the computational cost involved in the estimation algorithm. However, to accomplish this, we need to augment the model with extra latent variables; we now describe this augmentation method.

Polson et al. (2013) gave a data augmentation scheme for logistic models by utilizing the Pólya-Gamma distribution. This scheme starts by introducing a random variable ω_{ijt} which, given η_{ijt}, follows $PG(1, \eta_{ijt})$, where $PG(b, c)$ denotes the Pólya-Gamma distribution with parameters $b > 0$ and $c \in \Re$. This auxiliary variable ω_{ijt} is conditionally independent of y_{ijt} given η_{ijt}. Polson et al. show that the conditional joint density of y_{ijt} and ω_{ijt} can be written as

$$\pi(y_{ijt}, \omega_{ijt} | \eta_{ijt}) = \frac{1}{2} e^{(y_{ijt} - 1/2)\eta_{ijt}} e^{-\omega_{ijt}\eta_{ijt}^2/2} PG(\omega_{ijt} | 1, 0), \tag{6.13}$$

where $PG(\omega | b, c)$ is the Pólya-Gamma density with parameters b and c evaluated at ω. There are two aspects which make (6.13) much more tractable than the original likelihood of $y_{ijt} | \eta_{ijt}$ given in (6.3). The first is the normal kernel of η_{ijt}. The second is that while the Pólya Gamma density $PG(\omega | b, c)$ $\left(\propto \exp\{-c^2\omega/2\} PG(\omega | b, 0) \right)$ has the awkward form of an infinite summation, the expected value has a closed form, namely $(b/2c)\tanh(c/2)$. These useful results from Polson et al. lead to easy and efficient estimation for binary data using Gibbs sampling (Choi and Hobert, 2013), the EM algorithm (Scott and Sun, 2013) and, as we will show here, VB.

We assign the following priors:

$$\omega_{ijt} \sim PG(1,0) \quad \text{for } t = 1,\ldots,T,\ 1 \le i \ne j \le n, \tag{6.14}$$

$$s_i \sim Exp(1) \quad \text{for } i = 1,\ldots,n, \tag{6.15}$$

$$r_i|\tau_i \sim \Gamma(1, c\tau_i^{-1}) \quad \text{for } i = 1,\ldots,n, \tag{6.16}$$

$$\tau_i \sim \Gamma(a_2^*, b_2^*) \quad \text{for } i = 1,\ldots,n, \tag{6.17}$$

$$\alpha \sim N(0, b_3^*), \tag{6.18}$$

$$\pi(\boldsymbol{u}_g) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \quad \text{for } h = 0,1,\ldots,G, \tag{6.19}$$

$$\boldsymbol{\beta}_h \sim Dir(\boldsymbol{\gamma}_h^*) \quad \text{for } h = 0,1,\ldots,G. \tag{6.20}$$

To estimate the posterior $\pi(\{\mathcal{X}_t, \mathcal{Z}_t\}_{t=1}^T, \theta_\ell, \theta_p|\{Y_t\}_{t=1}^T)$, we use the factorized approximation Q, which looks like

$$\begin{aligned} &Q(\Omega, \{\mathcal{X}_t\}_{t=1}^T, \{\mathcal{Z}_t\}_{t=1}^T, \alpha, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\tau}, \boldsymbol{u}, \{\boldsymbol{\beta}_h\}_{h=0}^G) \\ &= q(\Omega)q(\{\mathcal{X}_t\}_{t=1}^T)q(\{\mathcal{Z}_t\}_{t=1}^T)q(\alpha)q(\boldsymbol{s})q(\boldsymbol{r})q(\boldsymbol{\tau})q(\boldsymbol{u})q(\{\boldsymbol{\beta}_h\}_{h=0}^G), \end{aligned} \tag{6.21}$$

where $\Omega = \{\omega_{ijt}\}_{t,i\ne j}$. The factorized distributions on the right hand side of (6.21) all belong to well known families of distributions. The exact forms are given in Section 6.6.

6.2.3 Number of Communities

An implicit challenge underlying the previous discourse is that in practice we do not in general know the number of communities G. We found the strategy given by Handcock et al. (2007) to be quite successful in our simulation study. We briefly summarize this method and refer the interested reader to the original source for more details.

Rather than estimating the integrated likelihood $\pi(\{Y_t\}_{t=1}^T|G)$ as would typically be done, we instead consider the joint distribution of the observed network data and unobserved latent positions, using our MAP estimator as the fixed values of the latent positions, i.e., $\pi(\{Y_t\}_{t=1}^T, \{\widehat{\mathcal{X}}_t\}_{t=1}^T|G)$, where $\{\widehat{\mathcal{X}}_t\}_{t=1}^T$ is the MAP estimators of the latent positions. We can rewrite this as

$$\pi(\{Y_t\}_{t=1}^T, \{\widehat{\mathcal{X}}_t\}_{t=1}^T|G) = \int \pi(\{Y_t\}_{t=1}^T|\{\widehat{\mathcal{X}}_t\}_{t=1}^T, \theta_\ell)\pi(\theta_\ell)d\theta_\ell \int \pi(\{\widehat{\mathcal{X}}_t\}_{t=1}^T|\theta_p)\pi(\theta_p)d\theta_p, \tag{6.22}$$

where all distributions are implicitly conditioning on G. The two integrals on the right hand side of

(6.22) can each be estimated via the Bayesian Information Criterion (BIC), thus allowing us to find the BIC approximation of $2\log(\pi(\{Y_t\}_{t=1}^T, \{\widehat{\mathcal{X}}_t\}_{t=1}^T|G))$ as

$$\text{BIC} = \text{BIC}_1 + \text{BIC}_2,$$

where

$$\text{BIC}_1 = 2\log(\pi(\{Y_t\}_{t=1}^T|\{\widehat{\mathcal{X}}_t\}_{t=1}^T, \hat{\theta}_\ell)) - \dim(\theta_\ell)\log\Big(\sum_{t,i\neq j} y_{ijt}\Big),$$

$$\text{BIC}_2 = 2\log(\pi(\{\widehat{\mathcal{X}}_t\}_{t=1}^T|\hat{\theta}_p)) - \dim(\theta_p)\log(nT).$$

Rather than using maximum likelihood estimators for $\hat{\theta}_\ell$ and $\hat{\theta}_p$ in computing the BIC's, we used the MAP estimators, as was also done in, e.g., Fraley and Raftery (2007). One last note is that we utilized recursive relations identical or similar to those given in Chapter 4 in order for the number of terms required to compute $\pi(\{\mathcal{X}_t\}_{t=1}^T|\hat{\theta}_p)$ to be linear, rather than exponential, with respect to T.

6.3 Simulation Study

We simulated 20 binary networks, each with $n = 100$ actors and $T = 10$ time points. Ten of these were simulated via the distance model, ten via the projection model; the details are given below. For each simulation we evaluated both the model fit and the misclassification of the cluster assignments. For the former, we looked at the actual edges and the predicted edge probabilities and computed the AUC (area under the receiver operating characteristic curve); a value of one implies a perfect fit, whereas a value of 0.5 implies that the predictions are random. For the latter we computed the variation of information (VI) (Meilă, 2003). The VI is a true metric, and hence a smaller VI value implies that the two clusterings being compared are closer to being identical. Thus we desire to have values close to one for AUC and close to zero for the VI.

For the ten data sets simulated via the distance model, we used likelihood formulation found in the dynamic latent space model of Chapter 2. This likelihood is given as

$$\text{logit}(\mathbb{P}(y_{ijt} = 1|\mathcal{X}_t, \beta_{IN}, \beta_{OUT}, s_i, s_j)) = \beta_{IN}\Big(1 - \frac{d_{ijt}}{s_j}\Big) + \beta_{OUT}\Big(1 - \frac{d_{ijt}}{s_i}\Big), \tag{6.23}$$

where β_{IN} and β_{OUT} are global parameters that reflect the relative importance of popularity and

activity respectively, the s_i's are actor specific parameters that reflect the tendency to send and receive edges, and d_{ijt} is the distance between actors i and j within the latent Euclidean space at time t.

We set $\lambda = 0.8$, $G = 6$, $p = 2$, and the likelihood parameters $\beta_{IN} = 0.3$, $\beta_{OUT} = 0.7$. We simulated $\boldsymbol{\mu}_g \sim N(\mathbf{0}, (5 \times 10^{-4})I_2)$, $\Sigma_k \sim W^{-1}(13, (1 \times 10^{-5})I_2)$, $\boldsymbol{\beta}_0 \sim Dir(10, \ldots, 10)$, and for $h = 1, \ldots, 6$, $\boldsymbol{\beta}_h$ was set to be proportional to

$$\left(\frac{1}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_h\|}, \ldots, \frac{1}{\|\boldsymbol{\mu}_{h-1} - \boldsymbol{\mu}_h\|}, 50 \times \max_{k \neq h} \left\{ \frac{1}{\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_h\|} \right\}, \frac{1}{\|\boldsymbol{\mu}_{h+1} - \boldsymbol{\mu}_h\|}, \ldots, \frac{1}{\|\boldsymbol{\mu}_K - \boldsymbol{\mu}_h\|} \right).$$

The cluster assignments $\{\mathcal{Z}_t\}_{t=1}^T$ and latent positions $\{\mathcal{X}_t\}_{t=1}^T$ were drawn according to (6.2), and the individual effects $(s_1, \ldots, s_n) \sim Dir\Big(100\frac{1/\|X_{1,1}\|}{\max_j(1/\|X_{j,1}\|)}, \ldots, 100\frac{1/\|X_{100,1}\|}{\max_j(1/\|X_{j,1}\|)}\Big)$. Finally, the adjacency matrices were simulated according to (6.23).

For the ten data sets simulated via the projection model, we set $G = 6$, $p = 3$, $\alpha = -10$, $\boldsymbol{\beta}_0 = (1/6, \ldots, 1/6)$ and

$$(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_6) = \begin{bmatrix} 0 & 90 & 0 & 0 & 45 & 45 \\ 0 & 0 & 90 & -90 & 45 & -45 \end{bmatrix},$$

where the cluster means $\boldsymbol{u}_g$ are given in the spherical coordinate angles in degrees. For $h = 1, \ldots, 6$, $\boldsymbol{\beta}_h$ was set to be proportional to $(\exp(3\boldsymbol{u}_h'\boldsymbol{u}_1), \ldots, \exp(3\boldsymbol{u}_h'\boldsymbol{u}_6))$. For $i = 1, \ldots, 100$, we simulated $s_i \sim N(1, 0.15)$, $\tau_i \sim \Gamma(16, 1/4)$ and then set $r_i = 8/\sqrt{\tau_i}$. The cluster assignments $\{\mathcal{Z}_t\}_{t=1}^T$ and latent positions $\{\mathcal{X}_t\}_{t=1}^T$ were drawn according to (6.6). Finally, the adjacency matrices were simulated according to (6.3) and (6.4).

We fit both our distance and projection models to all 20 simulated dynamic networks. By so doing we are to some degree evaluating how our models do under misspecification. We also compare our methods to clustering results obtained by applying the clustering models of Handcock et al. (2007) and of Krivitsky et al. (2009). These latter two models cluster static networks via a latent space approach; to apply them to dynamic networks, clustering was performed at each time point and then combined sequentially using the relabeling algorithm given in Stephens (2000).

Table 6.1 gives the simulation results. From the AUC values we see that all models fit the data quite well. However, the VI values imply that our models always outperform the static methods adapted for dynamic data, even when misspecified.

Also of interest is the computational time required for our proposed methods, and in particular

True Model	Method	AUC	VI
Distance	Projection	0.884 (0.0403)	0.923 (0.273)
	Handcock et al.	0.894 (0.0387)	0.965 (0.414)
	Krivitsky et al.	0.903 (0.0370)	1.31 (0.469)
	Distance	0.887 (0.0378)	0.186 (0.146)
Projection	Projection	0.991 (1.40×10^{-3})	0.357 (0.273)
	Handcock et al.	0.975 (3.06×10^{-3})	2.78 (0.181)
	Krivitsky et al.	0.991 (1.27×10^{-3})	2.85 (0.210)
	Distance	0.978 (3.78×10^{-3})	0.522 (0.285)

Table 6.1: Simulation results from data generated according to the distance and projection models. The average values are reported, with standard deviations in parentheses.

how the VB algorithm decreases the computational time required. For our simulations we used a UNIX machine with a 2.40 GHz processor. Using the VB estimation algorithm for the projection model yielded an average (over the 20 simulations) of 19.1 minutes and standard deviation of 5 minutes, and using a chain of length 50,000 using the MCMC algorithm for the distance model yielded an average of 72.4 minutes and a standard deviation of 8.5 minutes.

6.4 Data Analysis

6.4.1 Newcomb's Fraternity Data

Newcomb (1956) discussed data collected on 17 male college students who were previously unknown to each other. These 17 students, as part of Newcomb's study, agreed to live together for sixteen weeks (though the data set excludes the ninth week due to school vacation). For each week, every student ranks the other 16 students from 1 (most favored) to 16 (least favored).

In this context Y_t is the t^{th} $n \times n$ adjacency matrix whose i^{th} row, denoted $\mathbf{y}_{it}$, is how the i^{th} actor ranks the other $n-1$ actors. Without loss of generality, assume that the rankings go, in order of most favored to least favored, from 1 to $n-1$. Then we let $\boldsymbol{o}_{it} = (o_{i1t}, o_{i2t}, \ldots, o_{i(n-1)t})$ denote the $(n-1) \times 1$ vector which is the ordering of the rank vector $\mathbf{y}_{it}$ (e.g., if $\mathbf{y}_{1t} = (0, 4, 3, 1, 2)$ then $\boldsymbol{o}_{1t} = (4, 5, 3, 2)$). We assume that, conditioning on $(\mathcal{X}_t, \boldsymbol{\Psi})$, $\mathbf{y}_{it}$ is independent of $\mathbf{y}_{i't}$, $i \neq i'$.

The likelihood we will use is that used in Chapter 4, given as

$$\mathbb{P}(Y_t|\mathcal{X}_t, \boldsymbol{s}) = \prod_{i=1}^{n} \prod_{j=1}^{n-1} \frac{s_{o_{ijt}} \exp(-d_{io_{ijt}t})}{\sum_{\ell=j}^{n-1} s_{o_{i\ell t}} \exp(-d_{io_{i\ell t}t})}, \tag{6.24}$$

where again $\boldsymbol{s} = (s_1, \ldots, s_n)$ are actor specific parameters which indicate an actor's social reach, and for identifiability $\sum_{i=1}^{n} s_i = 1$. This is a Plackett-Luce model (Plackett, 1975), and as such

satisfies Luce's Choice axiom which can be characterized by having actor i rank actor j over actor k with the same probability whether or not actor ℓ is included in the set to be ranked. See Chaper 4 for further motivation of this model. As this likelihood depends on the latent positions through the distances $D(\mathcal{X}_t)$'s, we implement the distance model of Section 6.1.1. This flexible framework allows us to detect communities through the latent positions of the students.

For $G = 2, \ldots, 10$, we ran 100,000 iterations of the MCMC algorithm of Section 6.2.1, thus having a maximum of ten clusters. For each of the 9 chains, we used a short MCMC chain (the same chain for each G) following the model with no clustering of Chapter 4 to initialize the latent positions $\{\mathcal{X}_t\}_{t=1}^T$ and the actor specific likelihood parameters $\boldsymbol{s}$, and for the remaining prior parameters we used the generalized EM algorithm given by Chapter 5.

There was negligible difference in the BIC values corresponding to four and five communities, where the BIC values were computed as described in Section 6.2.3; therefore, to increase interpretability and reduce model complexity we chose to analyze the data assuming four communities. The goodness of fit was evaluated using the pseudo-R^2 value described in Chapter 4. The pseudo-R^2 takes values in the interval $[0, 1)$, where a higher value implies a better fit of the data. After analyzing the data, we obtained a pseudo-R^2 value of 0.561. This is slightly less than that obtained in Chapter 4 (0.622), which we feel satisfied with since we are imposing much more structure via the clustering on the prior of the latent positions.

Nakao and Romney (1993), when analyzing Newcomb's fraternity data, created similarity matrices for each time point and then performed multidimensional scaling to obtain latent network positions. From these plots the authors visually ascertained two clusters, one consisting of actors 2, 4, 7, 9, 11, 12 and 17 and the other consisting of actors 1, 5, 6, 8 and 13; the actors not in these two subgroups, 3, 10, 14, 15 and 16, were called "outliers". Moody et al. (2005) used various visualization methods and also commented on some clustering that were noticed via visual inspection. Moody et al. similarly concluded that actors 1, 6, 8 and 13 belonged in a cluster together. They also noticed that actors 7 and 12 stayed close together. Both Nakao and Romney and Moody et al. agree that actor 10 is not incorporated into a cluster, though Nakao and Romney attribute this to actor 10 having a "difficult time finding his social position" while Moody et al. describe actor 10 as being on "the edge of the social structure" (they also describe actor 15 in this way). In Chapter 4 we provided a detailed analysis of Newcomb's fraternity data which included an analysis of the subgroup formation. Students 1, 5, 6, 8 and 13 were again found to belong to the same group, and students 2, 7, 11 and 12 were also found to be grouped together. In Chapter 4 we also found students

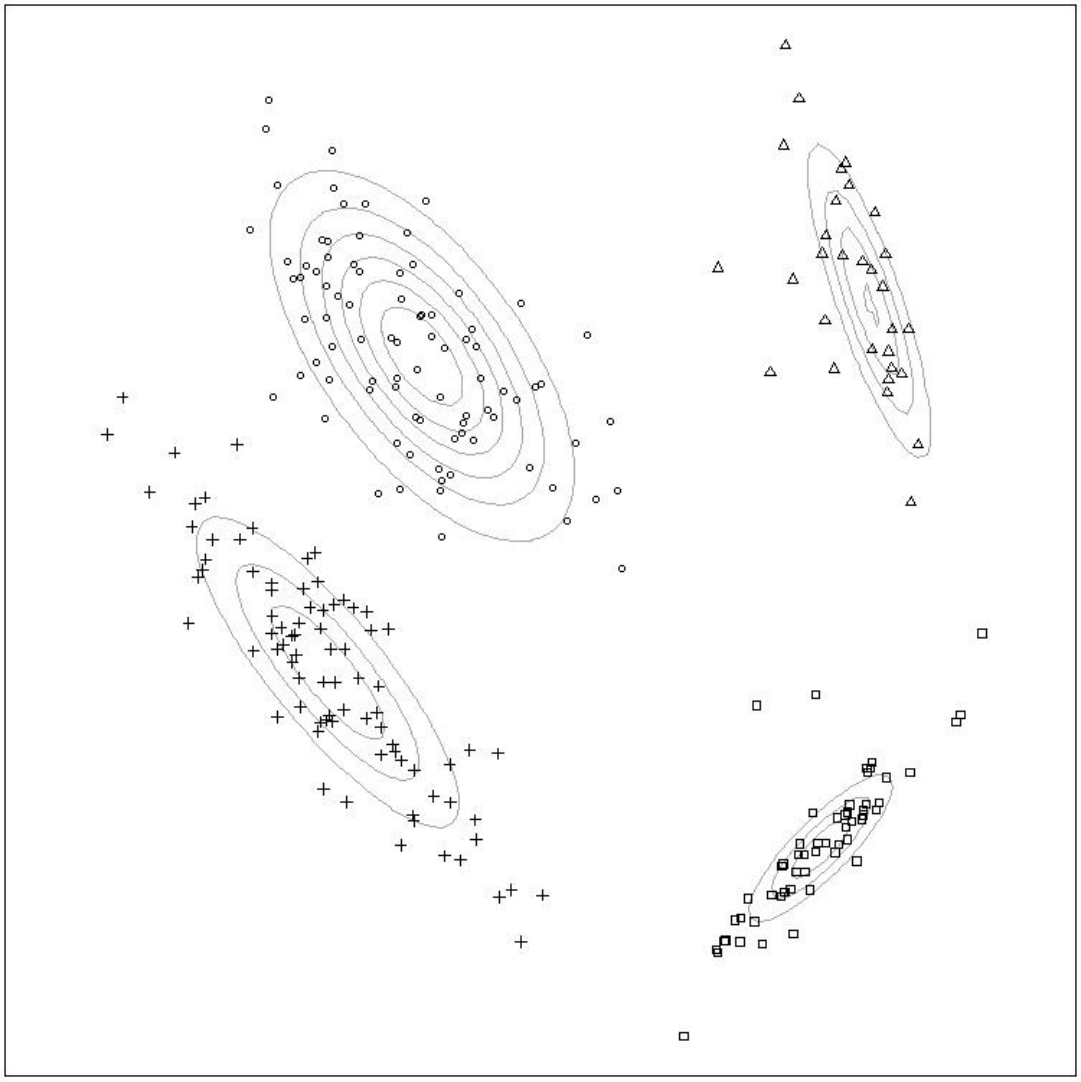
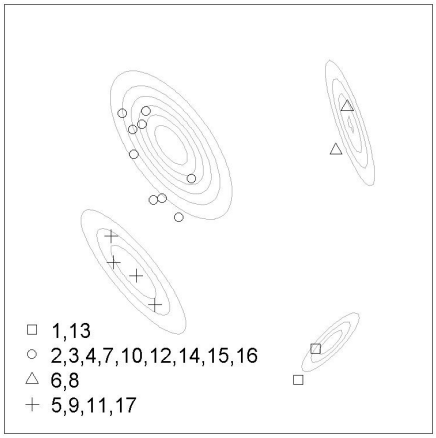
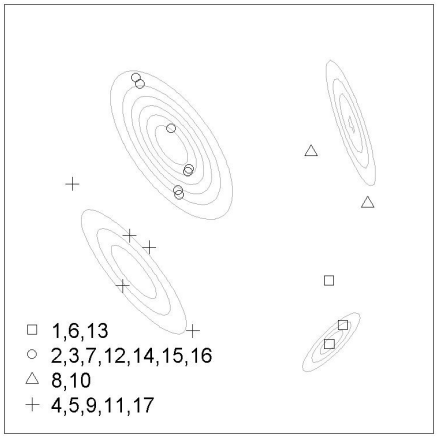


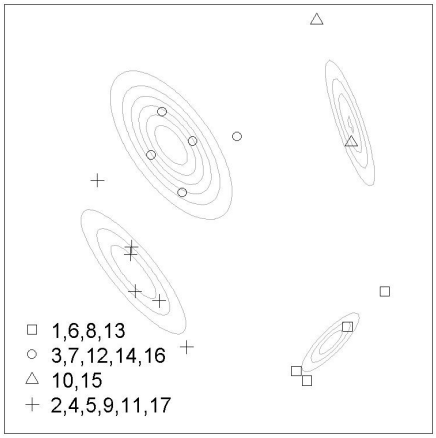
Figure 6.1: Latent positions of all actors at all time points in Newcomb's fraternity data. The contour lines correspond to the normal distributions which characterize the four communities. The symbols correspond to the community assignments given.



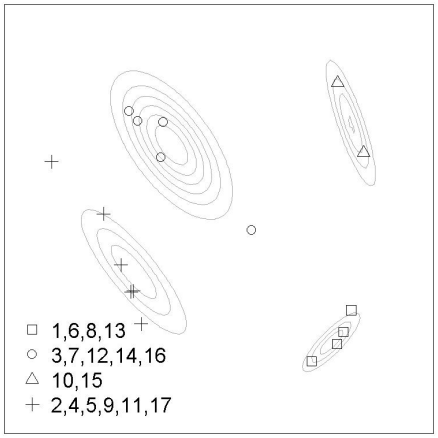
(a) Week 1



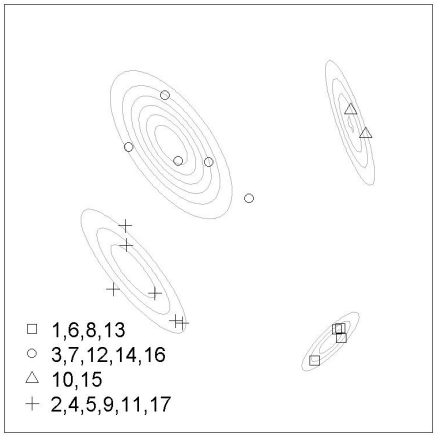
(b) Week 3



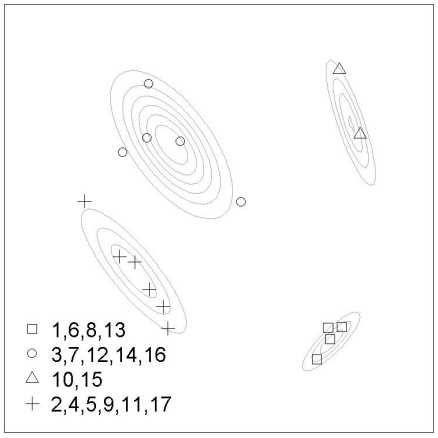
(c) Week 6



(d) Week 9



(e) Week 12



(f) Week 14

Figure 6.2: Latent positions of Newcomb's fraternity data at specific weeks. The contour lines correspond to the normal distributions which characterize the four communities. The symbols correspond to the community assignments given.

4, 9 and 17 to be in a common group, while students 3, 10, 14, 15 and 16 were all found to lie on the outside of the social space. Our formal analysis of the data captures the areas of agreement between these three studies, while giving more detailed information on the cluster composition. We now give the results from our study.

Figure 6.1 shows the latent space with the MAP estimators of the latent positions, thus showing the overall structure of the subgroups of the network. All actors at all time points are shown here. Figure 6.2 shows the latent positions at time points 1, 3, 6, 9, 12 and 14. We can characterize our four communities, referencing these groups using the shapes given in Figures 6.1 and 6.2. The $\square$ community, which consists of students 1, 6, 8 and 13, matches well with communities discovered by Nakao and Romney, Chapter 4, and the main community discovered by Moody. Once all the members eventually joined this community within the first few weeks (none departed the community), it remained constant for the remainder of the study. The $\bigcirc$ community seemed to be the opposite, in that it was the most transient, and all its member changes were departures to other communities. This community was the most spread out in the latent space (as measured by the determinant of the covariance matrix of the community) and the most diverse in terms of popularity. Here we determine popularity by averaging how each student was ranked by the others, using all time points. The $\bigcirc$ community had students 7 and 12 together, which corroborates what all three previously described studies claim. The $\triangle$ community was also somewhat transient compared to the others. It stabilized at week 4, thereafter containing only students 10 and 15, who were the least and fourth from least popular students. The + community was the opposite of the $\triangle$ community, in that this community was the most popular, measured by averaging the community members' popularity; in fact the three most popular students all belonged to the + community. This community, whose members primarily consist of students 2, 4, 5, 9, 11 and 17, matches well with clusters found by Nakao and Romney and in Chapter 4.

As the network was completely nascent at the first week, it is hardly a surprise that there are some actors that switch from one community to another during the beginning of the study. Our model was able to capture this evolution of the network, unlike many clustering algorthims which assume constant cluster assignments over time. In all, there were seven transitions, and five of these were during the first three transition periods. This implies that the subgroup formation of the social network was fairly stable after week four. The other two transitions were between weeks 5 and 6 and between weeks 14 and 15; this last also corroborates statements by various researchers who noticed some fluctuations in the network structure during the final week (Nakao and Romney, 1993;

Krivitsky and Butts, 2012; Chapter 4).

6.4.2 World Trade Data

We consider world trade data with the goals of determining trade blocs and gleaning what information we can from these blocs. We look at annual export and import data between countries during the years 1964 to 1976 (so $T = 13$). A (directed) trade relation is established from country i to country j if country i exports some non-negligible amount of goods to country j. During this time, for a variety of reasons a few countries are not constant throughout, and so we only include the $n = 111$ countries which exist throughout the entirety of the study period. Thus we have thirteen 111×111 binary adjacency matrices. As this is primarily a pedagogical example, we chose these years to strike a balance between a large number of time points with a large number of countries. The data we used was obtained through the Economic Web Institute at *http://www.economicswebinstitute.org/worldtrade.htm*, originally obtained through the IMF Direction of Trade Yearbook.

To detect trade blocs within the binary trade relations data, we implemented the VB estimation procedure for the projection model. We obtained an AUC of 0.981, showing that our model fit the data very well. This result came from choosing five communities according to the method given in Section 6.2.3, letting 20 be the upper bound on the possible number of clusters. Figure 6.3 shows the latent positions of all countries at all time points (nT points plotted), where the five communities have been labeled along segments from the origin to the communities' centers. For ease of viewing we have plotted the countries based only on their directional unit vectors, disregarding the magnitudes of the vectors which correspond to the individual effects.

Bloc 1 is the smallest community, including only four countries. These countries are very active in international trade; three out of the four are in the top ten countries as ranked by number of trading partners averaged over all time points. The United States and Brazil are both in this group, and the fact that they are together may be due to the 1964 coup d'état which led to a pro U.S. military ruled government in Brazil. Blocs 2 and 3 are both central communities within the latent space, yet differ in character. Bloc 2 is the largest community averaging 52 countries per year, and involves, with the exception of Canada, only eastern hemisphere nations. Bloc 3 is of a truly global constitution with all 6 inhabited continents represented, though its average community size is only around a third of the size of bloc 2. This is a very unconnected yet central community and may therefore represent those remaining nations which did not belong to one of the other major trading blocs.

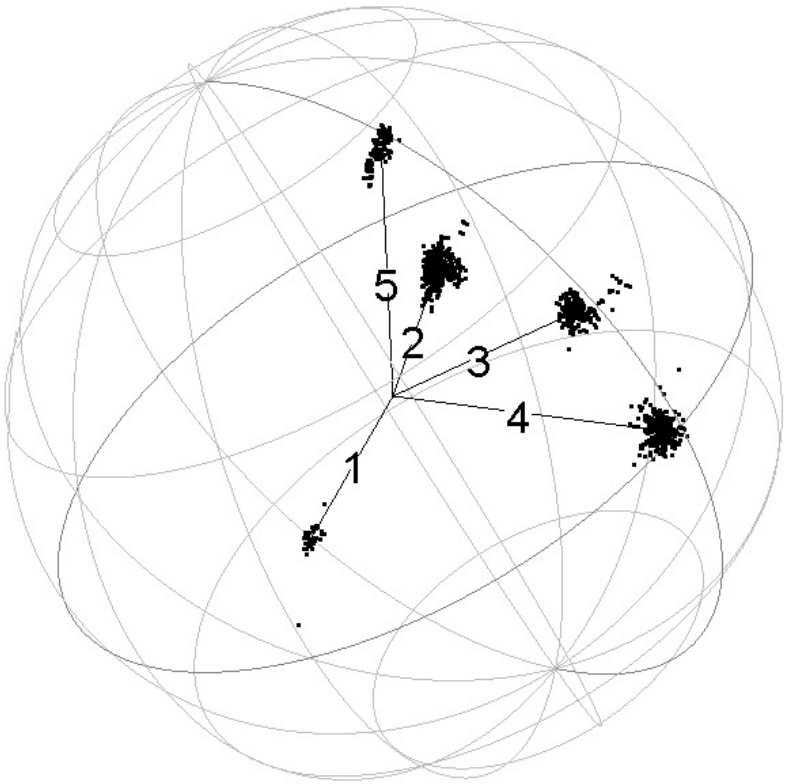


Figure 6.3: Variational Bayes estimates of latent locations (plotting the unit vectors indicating direction and ignoring the magnitude of the vectors that correspond to individual effects) of countries in the international export/import data. The five communities have been labeled along segments from the origin to the communities' centers.

	1	2	3	4	5
1	0.949	0.489	0.501	0.527	0.308
2		0.165	0.127	0.102	0.153
3			0.049	0.103	0.135
4				0.195	0.079
5					0.248

Table 6.2: Densities within each community and between each community averaged over all time points. These densities are computed by dividing the total number of edges by the total possible number of edges.

See Table 6.2 for the densities within and between each trade bloc, averaged over all time points. This table shows the connectedness and interconnectedness of the trade blocs. Bloc 4 consists of the U.S.S.R, several eastern European countries, and most of Latin America (17 countries). This gives quantitative evidence in favor of claims of close ties between U.S.S.R and Latin America and the Soviet influence in the western hemisphere (e.g., Blasier, 1988). Bloc 5 is a well connected community that is indicative of a very interesting vestigial effect from French colonization. Of the 11 countries that belonged to bloc 5 during all time points, France and her former colonies constitute 7 of them. French colonial policy required her colonies to import only from or through France, export only to France, and to ship using French vessels (Grier, 1999). That France and her former colonies form a well connected trading bloc gives evidence that colonial policy established a longer term trend for the continued advantage of France.

Only 13 transitions were made in total, implying that nations switched trading blocs infrequently. These transitions can help make sense of certain world events, as well as lead researchers to future avenues of investigation. One interesting transition was Romania changing from bloc 3 to bloc 2 in 1965, coinciding with a new direction in foreign and domestic policies, as well as a new economic program. For more information, see, e.g., Georgescu and Călinescu (1991, p.242-248). These new policy changes may have led Romania to switch trading blocs. Another interesting transition was West Germany switching from bloc 2 to the same trading bloc as the United States (bloc 1) in 1973, the same year the Basic Treaty between West and East Germany went into effect, and one year after the Four Power Agreement on Berlin occurred. This transition then should lead researchers to ask more questions about how these agreements may have affected the economic policies and foreign relations of West Germany.

6.5 Full Conditional Distributions for the Distance Model

For ease of notation let $\mathcal{I}_{gt} = \{i \in \{1, \ldots, n\} : Z_{itg} = 1\}$ be the set of actors in community g at time t, and let $\mathcal{J}_{hkt} = \mathcal{I}_{h(t-1)} \bigcap \mathcal{I}_{kt}$ be the set of actors transitioning from community h at time $t-1$ to community k at time t. Also, let $\tilde{Z}_{it} = \sum_{g=1}^{G} g 1_{\{Z_{itg}=1\}}$ denote which cluster actor i at time t belongs to. We denote the conditioning of everything but the parameter of interest as $|\cdot$.

The full conditional distributions for the clustering parameters are

$$\lambda|\cdot \sim N_{(0,1)}\left(\frac{\nu_\lambda + \xi_\lambda \sum_{i=1}^{n}\sum_{t\geq 2}(\boldsymbol{\mu}_{\tilde{Z}_{it}} - \mathbf{X}_{i(t-1)})'\Sigma^{-1}_{\tilde{Z}_{it}}(\mathbf{X}_{it} - \mathbf{X}_{i(t-1)})}{1 + \xi_\lambda \sum_{i=1}^{n}\sum_{t\geq 2}(\boldsymbol{\mu}_{\tilde{Z}_{it}} - \mathbf{X}_{i(t-1)})'\Sigma^{-1}_{\tilde{Z}_{it}}(\boldsymbol{\mu}_{\tilde{Z}_{it}} - \mathbf{X}_{i(t-1)})}, \right.$$
$$\left. \frac{\xi_\lambda}{1 + \xi_\lambda \sum_{i=1}^{n}\sum_{t\geq 2}(\boldsymbol{\mu}_{\tilde{Z}_{it}} - \mathbf{X}_{i(t-1)})'\Sigma^{-1}_{\tilde{Z}_{it}}(\boldsymbol{\mu}_{\tilde{Z}_{it}} - \mathbf{X}_{i(t-1)})}\right), \tag{6.25}$$

$$\boldsymbol{\mu}_g|\cdot \sim N(\Sigma_{\mu_g|\cdot}\Sigma_g^{-1}\left(\sum_{i\in\mathcal{I}_{g1}}\mathbf{X}_{i1} + \lambda\sum_{t\geq 2}\sum_{i\in\mathcal{I}_{gt}}(\mathbf{X}_{it} - (1-\lambda)\mathbf{X}_{i(t-1)})\right), \Sigma_{\mu_g|\cdot}), \tag{6.26}$$

$$\text{where } \Sigma_{\mu_g|\cdot} = \left(|\mathcal{I}_{g1}|\Sigma_g^{-1} + \lambda^2\sum_{t\geq 2}|\mathcal{I}_{gt}|\Sigma_g^{-1} + \frac{1}{\tau^2}I_p\right)^{-1} \text{ for } g = 1, \ldots, G, \tag{6.27}$$

$$\Sigma_g|\cdot \sim W^{-1}\left(p + 1 + \sum_{t=1}^{T}|\mathcal{I}_{gt}|, diag(\gamma_1, \ldots, \gamma_p) + \sum_{i\in\mathcal{I}_{g1}}(\mathbf{X}_{i1} - \boldsymbol{\mu}_g)(\mathbf{X}_{i1} - \boldsymbol{\mu}_g)' + \right.$$
$$\left. \sum_{t\geq 2}\sum_{i\in\mathcal{I}_{gt}}(\mathbf{X}_{it} - \lambda\boldsymbol{\mu}_g - (1-\lambda)\mathbf{X}_{i(t-1)})(\mathbf{X}_{it} - \lambda\boldsymbol{\mu}_g - (1-\lambda)\mathbf{X}_{i(t-1)})'\right) \text{ for } g = 1, \ldots, G, \tag{6.28}$$

$$\tau^2|\cdot \sim \Gamma^{-1}(a + G/2, b + \sum_{g=1}^{G}\|\boldsymbol{\mu}_g\|^2/2), \tag{6.29}$$

$$\gamma_\ell|\cdot \sim \Gamma(c + G(p+1)/2, \left[d + \frac{1}{2}\sum_{g=1}^{G}[\Sigma_g^{-1}]_{\ell,\ell}\right]^{-1}) \text{ for } \ell = 1, \ldots, p, \tag{6.30}$$

$$\boldsymbol{\beta}_0|\cdot \sim Dir(1 + |\mathcal{I}_{11}|, \ldots, 1 + |\mathcal{I}_{G1}|), \tag{6.31}$$

$$\boldsymbol{\beta}_h|\cdot \sim Dir(1 + \sum_{t\geq 2}|\mathcal{J}_{h1t}|, \ldots, 1 + \sum_{t\geq 2}|\mathcal{J}_{hGt}|) \text{ for } h = 1, \ldots, G, \tag{6.32}$$

where $[\Sigma_g^{-1}]_{\ell,\ell}$ is the ℓ^{th} diagonal of the inverse of Σ_g. The log of the full conditional distributions

for the latent positions are

$$\begin{aligned}\log(\pi(\boldsymbol{X}_{i1}|\cdot)) &= const + \log(\pi(Y_1|\mathcal{X}_1,\theta_\ell)) - \frac{1}{2}(\boldsymbol{X}_{i1}-\boldsymbol{\mu}_{\tilde{Z}_{i1}})'\Sigma^{-1}_{\tilde{Z}_{i1}}(\boldsymbol{X}_{i1}-\boldsymbol{\mu}_{\tilde{Z}_{i1}}) \\ &\quad - \frac{1}{2}(\boldsymbol{X}_{i2}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{i2}}-(1-\lambda)\boldsymbol{X}_{i1})'\Sigma^{-1}_{\tilde{Z}_{i2}}(\boldsymbol{X}_{i2}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{i2}}-(1-\lambda)\boldsymbol{X}_{i1}), \qquad (6.33)\\ \log(\pi(\boldsymbol{X}_{iT}|\cdot)) &= const + \log(\pi(Y_T|\mathcal{X}_T,\theta_\ell)) \\ &\quad - \frac{1}{2}(\boldsymbol{X}_{iT}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{iT}}-(1-\lambda)\boldsymbol{X}_{i(T-1)})'\Sigma^{-1}_{\tilde{Z}_{iT}}(\boldsymbol{X}_{iT}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{iT}}-(1-\lambda)\boldsymbol{X}_{i(T-1)}), \qquad (6.34)\end{aligned}$$

and for $1 < t < T$,

$$\begin{aligned}\log(\pi(\boldsymbol{X}_{it}|\cdot)) &= const + \log(\pi(Y_t|\mathcal{X}_t,\theta_\ell)) \\ &\quad - \frac{1}{2}(\boldsymbol{X}_{it}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{it}}-(1-\lambda)\boldsymbol{X}_{i(t-1)})'\Sigma^{-1}_{\tilde{Z}_{it}}(\boldsymbol{X}_{it}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{it}}-(1-\lambda)\boldsymbol{X}_{i(t-1)}) \\ &\quad - \frac{1}{2}(\boldsymbol{X}_{i(t+1)}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{i(t+1)}}-(1-\lambda)\boldsymbol{X}_{it})'\Sigma^{-1}_{\tilde{Z}_{i(t+1)}}(\boldsymbol{X}_{i(t+1)}-\lambda\boldsymbol{\mu}_{\tilde{Z}_{i(t+1)}}-(1-\lambda)\boldsymbol{X}_{it}). \qquad (6.35)\end{aligned}$$

Finally the log of the full conditional distribution for the likelihood parameters is

$$\log(\theta_\ell|\cdot) = const + \log(\pi(\{Y_t\}_{t=1}^T|\{\mathcal{X}_t\}_{t=1}^T,\theta_\ell)) + \log(\pi(\theta_\ell)). \qquad (6.36)$$

6.6 VB Distributions for the Projection Model

Suppose we are interested in a posterior distribution $\pi(\theta|Data)$, but cannot find this distribution in closed form. Then the mean field variational Bayes approximation finds a distribution $Q(\theta)$ that minimizes the Kullback-Leibler divergence between Q and the posterior, where Q can be written in a factorized form over subsets $\{\theta_i\}$ of θ, i.e., $Q(\theta) = \prod_i q(\theta_i)$. Each component of Q can be found by

$$q(\theta_i) = \frac{\exp\{\mathbb{E}(\pi(Data,\theta))\}}{\int \exp\{\mathbb{E}(\pi(Data,\theta))\}\, d\theta_i}, \qquad (6.37)$$

where the expectation is taken under Q with respect to $\theta \setminus \theta_i$ (see, e.g., **?**).

We employ this approximation to the projection model and obtain the solutions provided below. To aid in reading these solutions, we first mention the following, where all expectations are taken under Q:

$$
\begin{aligned}
q(\tau_i) &= \Gamma(a_{i2}, b_{i2}),\\
\mathbb{E}(\alpha) &= a_3,\\
Var(\alpha) &= b_3,\\
\mathbb{E}(\boldsymbol{X}_{it}) &= \boldsymbol{\mu}_{it},\\
Cov(\boldsymbol{X}_{it}) &= \Sigma_{it},\\
\mathbb{E}(\boldsymbol{u}_g) &= \boldsymbol{\nu}_g,\\
\mathbb{E}(1_{\{Z_{i1g}=1\}}) &= \beta_{i0g}.\\
\mathbb{E}(1_{\{Z_{ith}=1\}}) &= \bar{\beta}_{ith},\\
\mathbb{E}(1_{\{Z_{itk}=1\}}|Z_{i(t-1)h}=1) &= \beta_{ithk}.
\end{aligned}
$$

Note that $\boldsymbol{\mu}$ and Σ are not referring to similar notation used in the distance model.

(1) Derivation of $q(\omega_{ijt})$

Since

$$
\log(q(\omega_{ijt})) = const + \mathbb{E}\left[-\frac{\omega_{ijt}}{2}(\alpha + s_j \boldsymbol{X}_{it}'\boldsymbol{X}_{jt})^2 + \log(PG(\omega_{ijt}|1,0)\right],
$$

we have

$$
q(\omega_{ijt}) \propto \exp(-\omega_{ijt}c_{ijt}^2/2)\cdot PG(\omega_{ijt}|1,0),
$$

so

$$
q(\omega_{ijt}) \overset{\mathcal{D}}{=} PG(1, c_{ijt}).
$$

Here

$$
c_{ijt}^2 = b_3 + a_3^2 + 2a_3\mathbb{E}(S_j)\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt} + \mathbb{E}(s_j^2)\left(tr(\Sigma_{it}\Sigma_{jt}) + \boldsymbol{\mu}_{jt}'\Sigma_{it}\boldsymbol{\mu}_{jt} + \boldsymbol{\mu}_{it}'\Sigma_{jt}\boldsymbol{\mu}_{it} + (\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})^2\right),
$$

since

$$\begin{aligned}\mathbb{E}(\boldsymbol{X}_{it}'\boldsymbol{X}_{jt}\boldsymbol{X}_{jt}'\boldsymbol{X}_{it}) &= \mathbb{E}(\boldsymbol{X}_{it}'(\Sigma_{jt}+\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}')\boldsymbol{X}_{it}) \\ &= \mathbb{E}((L_j'\boldsymbol{X}_{it})'(L_j'\boldsymbol{X}_{it})) \\ &= tr(L_j'\sigma_{it}L_j) + (L_j'\boldsymbol{\mu}_{it})'(L_j'\boldsymbol{\mu}_{it}) \\ &= tr(\Sigma_{it}\Sigma_{jt}) + \boldsymbol{\mu}_{jt}'\Sigma_{it}\boldsymbol{\mu}_{jt} + \boldsymbol{\mu}_{it}'\Sigma_{jt}\boldsymbol{\mu}_{it} + (\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})^2,\end{aligned}$$

where L_jL_j' was used to represent the Cholesky decomposition of $\Sigma_{jt}+\boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}'$. Therefore

$$\mathbb{E}(\omega_{ijt}) = \frac{1}{2c_{ijt}}\left(\frac{e^{c_{ijt}}-1}{1+e^{c_{ijt}}}\right).$$

(2) Derivation of $q(\{\mathcal{X}_t\}_{t=1}^T)$

We have

$$\begin{aligned}&\log(q(\boldsymbol{X}_i)) \\ &= const + \sum_{t=1}^{T}\sum_{j\neq i}\mathbb{E}\left\{(y_{ijt}-1/2)s_j\boldsymbol{X}_{it}'\boldsymbol{X}_{jt} - \frac{\omega_{ijt}}{2}(s_j^2\boldsymbol{X}_{it}'\boldsymbol{X}_{jt}\boldsymbol{X}_{jt}'\boldsymbol{X}_{it} + 2\alpha s_j\boldsymbol{X}_{it}'\boldsymbol{X}_{jt})\right. \\ &\left.+(y_{jit}-1/2)s_i\boldsymbol{X}_{it}'\boldsymbol{X}_{jt} - \frac{\omega_{jit}}{2}(s_i^2\boldsymbol{X}_{it}'\boldsymbol{X}_{jt}\boldsymbol{X}_{jt}'\boldsymbol{X}_{it} + 2\alpha s_i\boldsymbol{X}_{it}'\boldsymbol{X}_{jt})\right\} \\ &- \frac{\mathbb{E}\tau_i}{2}\mathbb{E}\left\{\sum_{g=1}^{G}Z_{i1g}\|\boldsymbol{X}_{i1}-r_i\boldsymbol{u}_g\|^2 + \sum_{t\geq 2}\sum_{h=1}^{G}\sum_{k=1}^{G}Z_{i(t-1)h}Z_{itk}\|\boldsymbol{X}_{it}-r_i\boldsymbol{u}_k\|^2\right\}.\end{aligned}$$

For $t \geq 2$, we have

$$
\begin{aligned}
&\log(q(\boldsymbol{X}_{it})) \\
&= const + \sum_{j \neq i} \boldsymbol{X}_{it}' \left[((y_{ijt} - 1/2)\mathbb{E}(s_j) + (y_{jit} - 1/2)\mathbb{E}(s_i))\boldsymbol{\mu}_{jt} \right. \\
&\left. -(\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j^2)/2 + \mathbb{E}(\omega_{jit})\mathbb{E}(s_i^2)/2)(\Sigma_{jt} + \boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}')\boldsymbol{X}_{it} - a_3(\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j) + \mathbb{E}(\omega_{ijt})\mathbb{E}(s_i))\boldsymbol{\mu}_{jt} \right] \\
&- \frac{a_{i2}b_{i2}}{2} \left[\sum_{h=1}^{G}\sum_{k=1}^{G} \bar{\beta}_{i(t-1)h}\beta_{ithk}(\boldsymbol{X}_{it}'\boldsymbol{X}_{it} - 2\boldsymbol{X}_{it}'\mathbb{E}(r_i)\boldsymbol{\nu}_k) \right] \\
&= const - \frac{1}{2} \left\{ \boldsymbol{X}_{it}' \left[\sum_{j \neq i} (\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j^2) + \mathbb{E}(\omega_{jit})\mathbb{E}(s_i^2))(\Sigma_{jt} + \boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}') \right. \right. \\
&\left. + a_{i2}b_{i2} \sum_{h=1}^{G}\sum_{k=1}^{G} \bar{\beta}_{i(t-1)h}\beta_{ithk} I_p \right] \boldsymbol{X}_{it} - 2\boldsymbol{X}_{it}' \left[\sum_{j \neq i} \Big((y_{ijt} - 1/2)\mathbb{E}(s_j) + (y_{jit} - 1/2)\mathbb{E}(s_i) \right. \\
&\left. \left. -a_3(\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j) + \mathbb{E}(\omega_{jit})\mathbb{E}(s_i)) \Big)\boldsymbol{\mu}_{jt} + \mathbb{E}(r_i)a_{i2}b_{i2} \sum_{h=1}^{G}\sum_{k=1}^{G} \bar{\beta}_{i(t-1)h}\beta_{ithk}\boldsymbol{\nu}_k \right] \right\}.
\end{aligned}
$$

Therefore $q(\boldsymbol{X}_{it}) \overset{\mathcal{D}}{=} N(\boldsymbol{\mu}_{it}, \Sigma_{it})$ where

$$
\begin{aligned}
\boldsymbol{\mu}_{it} &= \Sigma_{it} \left(\sum_{j \neq i} \Big((y_{ijt} - 1/2)\mathbb{E}(s_j) + (y_{jit} - 1/2)\mathbb{E}(s_i) - a_3(\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j) + \mathbb{E}(\omega_{jit})\mathbb{E}(s_i)) \Big)\boldsymbol{\mu}_{jt} \right. \\
&\quad \left. + \mathbb{E}(r_i)a_{i2}b_{i2} \sum_{h=1}^{G}\sum_{k=1}^{G} \bar{\beta}_{i(t-1)h}\beta_{ithk}\boldsymbol{\nu}_k \right), \\
\Sigma_{it}^{-1} &= \sum_{j \neq i} (\mathbb{E}(\omega_{ijt})\mathbb{E}(s_j^2) + \mathbb{E}(\omega_{jit})\mathbb{E}(s_i^2))(\Sigma_{jt} + \boldsymbol{\mu}_{jt}\boldsymbol{\mu}_{jt}') + a_{i2}b_{i2} \sum_{h=1}^{G}\sum_{k=1}^{G} \bar{\beta}_{i(t-1)h}\beta_{ithk} I_p.
\end{aligned}
$$

Similarly, $q(\boldsymbol{X}_{i1}) \overset{\mathcal{D}}{=} N(\boldsymbol{\mu}_{i1}, \Sigma_{i1})$ where

$$
\begin{aligned}
\boldsymbol{\mu}_{i1} &= \Sigma_{i1} \left(\sum_{j \neq i} \Big((y_{ij1} - 1/2)\mathbb{E}(s_j) + (y_{ji1} - 1/2)\mathbb{E}(s_i) - a_3(\mathbb{E}(\omega_{ij1})\mathbb{E}(s_j) + \mathbb{E}(\omega_{ji1})\mathbb{E}(s_i)) \Big)\boldsymbol{\mu}_{j1} \right. \\
&\quad \left. + \mathbb{E}(r_i)a_{i2}b_{i2} \sum_{g=1}^{G} \beta_{i0g}\boldsymbol{\nu}_g \right), \\
\Sigma_{i1}^{-1} &= \sum_{j \neq i} (\mathbb{E}(\omega_{ij1})\mathbb{E}(s_j^2) + \mathbb{E}(\omega_{ji1})\mathbb{E}(s_i^2))(\Sigma_{j1} + \boldsymbol{\mu}_{j1}\boldsymbol{\mu}_{j1}') + a_{i2}b_{i2} \sum_{g=1}^{G} \beta_{i0g} I_p.
\end{aligned}
$$

(3) Derivation of $q(\{\mathcal{Z}_t\}_{t=1}^T)$

We have

$$\log(q(\mathcal{Z})) = const + \sum_{i=1}^{n} \mathbb{E}\left\{ \sum_{g=1}^{G} Z_{i1g} \left[\log(\beta_{0g}) + \frac{p}{2}\log(\tau_i) - \frac{\tau_i}{2}\|\boldsymbol{X}_{i1} - r_i \boldsymbol{u}_g\|^2 \right] + \sum_{t\geq 2}\sum_{h=1}^{G} Z_{i(t-1)h} \left[\sum_{k=1}^{G} Z_{itk} \left[\log(\beta_{hk}) + \frac{p}{2}\log(\tau_i) - \frac{\tau_i}{2}\|\boldsymbol{X}_{it} - r_i \boldsymbol{u}_k\|^2 \right] \right] \right\}.$$

Therefore, since

$$\log(q(\boldsymbol{Z}_{it}|Z_{i(t-1)h}=1)) = const + \sum_{g=1}^{G} Z_{itg} \left[\psi(\gamma_{hg}) - \psi(\textstyle\sum_k \gamma_{hk}) + \frac{p}{2}\big(\psi(a_{i2}) + \log(b_{i2})\big) - \frac{a_{i2}b_{i2}}{2}\Big(tr(\Sigma_{it}) + \boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{it} - 2\boldsymbol{\mu}_{it}'\mathbb{E}(r_i)\boldsymbol{\nu}_g + \mathbb{E}(r_i^2)\Big) \right]$$

and

$$\log(q(\boldsymbol{Z}_{i1})) = const + \sum_{g=1}^{G} Z_{i1g} \left[\psi(\gamma_{0g}) - \psi(\textstyle\sum_k \gamma_{0k}) + \frac{p}{2}\big(\psi(a_{i2}) + \log(b_{i2})\big) - \frac{a_{i2}b_{i2}}{2}\Big(tr(\Sigma_{i1}) + \boldsymbol{\mu}_{i1}'\boldsymbol{\mu}_{i1} - 2\boldsymbol{\mu}_{i1}'\mathbb{E}(r_i)\boldsymbol{\nu}_g + \mathbb{E}(r_i^2)\Big) \right],$$

where $\psi(\cdot)$ is the digamma function, we have that $\boldsymbol{Z}_{i1} \sim \text{Multinomial}(1, \boldsymbol{\beta}_{i0})$, where

$$\begin{aligned}\beta_{i0g} &\propto \exp\left\{ \psi(\gamma_{0g}) - \psi(\textstyle\sum_k \gamma_{0k}) + \frac{p}{2}\big(\psi(a_{i2}) + \log(b_{i2})\big) - \frac{a_{i2}b_{i2}}{2}\Big(tr(\Sigma_{i1}) + \boldsymbol{\mu}_{i1}'\boldsymbol{\mu}_{i1} - 2\boldsymbol{\mu}_{i1}'\mathbb{E}(r_i)\boldsymbol{\nu}_g + \mathbb{E}(r_i^2)\Big) \right\} \\ &\propto \exp\left\{ \psi(\gamma_{0g}) + \mathbb{E}(r_i) a_{i2}b_{i2}\boldsymbol{\mu}_{i1}'\boldsymbol{\nu}_g \right\}\end{aligned}$$

and $\boldsymbol{Z}_{it}|Z_{i(t-1)h} = 1 \sim \text{Multinomial}(1, \boldsymbol{\beta}_{ith})$, where

$$\begin{aligned}\beta_{ithg} &\propto \exp\left\{ \psi(\gamma_{hg}) - \psi(\textstyle\sum_k \gamma_{hk}) + \frac{p}{2}\big(\psi(a_{i2}) + \log(b_{i2})\big) - \frac{a_{i2}b_{i2}}{2}\Big(tr(\Sigma_{it}) + \boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{it} - 2\boldsymbol{\mu}_{it}'\mathbb{E}(r_i)\boldsymbol{\nu}_g + \mathbb{E}(r_i^2)\Big) \right\} \\ &\propto \exp\left\{ \psi(\gamma_{hg}) + \mathbb{E}(r_i) a_{i2}b_{i2}\boldsymbol{\mu}_{it}'\boldsymbol{\nu}_g \right\}.\end{aligned}$$

The row vector of marginal probabilities can be computed recursively in the following way:

$$\mathbb{P}(\boldsymbol{Z}_{it}) \triangleq \bar{\boldsymbol{\beta}}_{it} = \bar{\boldsymbol{\beta}}_{i(t-1)}\boldsymbol{\beta}_{it},$$

where $\boldsymbol{\beta}_{it}$ is the $G \times G$ transition matrix for $\boldsymbol{Z}_{it}$.

(4) Derivation of $q(\boldsymbol{r})$

Since

$$\begin{aligned}
\log(q(r_i)) &= const + \mathbb{E}\left[\sum_{g=1}^{G} Z_{i1g}\left(-\frac{\tau_i}{2}(r_i^2 - 2r_i\boldsymbol{X}_{i1}'\boldsymbol{u}_g)\right)\right. \\
&\quad \left. + \sum_{t\geq 2}\sum_{h=1}^{G}\sum_{k=1}^{G} Z_{i(t-1)h}Z_{itk}\left(-\frac{\tau_i}{2}(r_i^2 - 2r_i\boldsymbol{X}_{it}'\boldsymbol{u}_k)\right) - \frac{r_i\tau_i}{c}\right] \\
&= const - \frac{a_{i2}b_{i2}}{2}\left\{r_i^2 T - 2r_i\left(\sum_{g=1}^{G}\beta_{i0g}\boldsymbol{\mu}_{i1}'\boldsymbol{\nu}_g + \sum_{t\geq 2}\sum_{h=1}^{G}\bar{\beta}_{i(t-1)h}\sum_{k=1}^{G}\beta_{ithk}\boldsymbol{\mu}_{it}'\boldsymbol{\nu}_k - \frac{1}{c}\right)\right\},
\end{aligned}$$

we have that $q(r_i) \overset{\mathcal{D}}{=} N_{(0,\infty)}(a_{i1}, b_{i1})$, where

$$\begin{aligned}
a_{i1} &= \frac{1}{T}\left(\boldsymbol{\mu}_{i1}'\sum_{g=1}^{G}\beta_{i0g}\boldsymbol{\nu}_g + \sum_{t\geq 2}\boldsymbol{\mu}_{it}'\sum_{h=1}^{G}\bar{\beta}_{i(t-1)h}\sum_{k=1}^{G}\beta_{ithk}\boldsymbol{\nu}_k - \frac{1}{c}\right), \\
b_{i1} &= \frac{1}{Ta_{i2}b_{i2}}.
\end{aligned}$$

The mean of r_i then is

$$\mathbb{E}(r_i) = a_{i1} + \frac{\sqrt{b_{i1}}\phi(a_{i1}/\sqrt{b_{i1}})}{\Phi(a_{i1}/\sqrt{b_{i1}})}$$

and the mean of r_i^2 is

$$\mathbb{E}(r_i^2) = b_{i1}\left(1 + \frac{a_{i1}/\sqrt{b_{i1}}\phi(a_{i1}/\sqrt{b_{i1}})}{\Phi(a_{i1}/\sqrt{b_{i1}})}\right) + a_{i1}^2,$$

where $\phi(x)$ is the standard normal density evaluated at x and $\Phi(x)$ is the cumulative distribution function of a standard normal random variable evaluated at x.

(5) Derivation of $q(\boldsymbol{\tau})$

Since

$$\begin{aligned}\log(q(\tau_i)) = const + \frac{pT}{2}\log(\tau_i) - \frac{\tau_i}{2}\mathbb{E}\Bigg\{&\sum_{g=1}^{G} Z_{ig1}(\boldsymbol{X}_{i1}'\boldsymbol{X}_{i1} - 2r_i\boldsymbol{X}_{i1}'\boldsymbol{u}_g + r_i^2) \\ &+ \sum_{t\geq 2}\sum_{h=1}^{G}\sum_{k=1}^{G} Z_{i(t-1)h}Z_{itk}(\boldsymbol{X}_{it}'\boldsymbol{X}_{it} - 2r_i\boldsymbol{X}_{it}'\boldsymbol{u}_k + r_i^2)\Bigg\} \\ &+ \log(\tau_i) - \tau_i\mathbb{E}(r_i)/c + (a_2^* - 1)\log(\tau_i) - \tau_i/b_2^*,\end{aligned}$$

we have that $q(\tau_i) \overset{\mathcal{D}}{=} \Gamma(a_{i2}, b_{i2})$, where

$$\begin{aligned}a_{i2} &= a_2^* + pT/2 + 1, \\ b_{i2}^{-1} &= \frac{\mathbb{E}(r_i)}{c} + \frac{1}{b_2^*} + \frac{1}{2}\sum_{g=1}^{G}\beta_{i0g}\Big(tr(\Sigma_{i1}) + \boldsymbol{\mu}_{i1}'\boldsymbol{\mu}_{i1} - 2\mathbb{E}(r_i)\boldsymbol{\mu}_{i1}'\boldsymbol{\nu}_g + \mathbb{E}(r_i^2)\Big) \\ &\quad + \frac{1}{2}\sum_{t\geq 2}\sum_{h=1}^{G}\sum_{k=1}^{G}\bar{\beta}_{i(t-1)h}\beta_{ithk}\Big(tr(\Sigma_{it}) + \boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{it} - 2\mathbb{E}(r_i)\boldsymbol{\mu}_{it}'\boldsymbol{\nu}_k + \mathbb{E}(r_i^2)\Big).\end{aligned}$$

(6) Derivation of $q(\boldsymbol{u}_g)$

Since

$$\begin{aligned}\log(q(\boldsymbol{u}_g)) &= const + \sum_{i=1}^{n}\mathbb{E}\left\{Z_{i1g}\tau_i r_i\boldsymbol{X}_{i1}'\boldsymbol{u}_g + \sum_{t\geq 2}\sum_{h=1}^{G} Z_{i(t-1)h}Z_{itg}\tau_i r_i\boldsymbol{X}_{it}'\boldsymbol{u}_g\right\} \\ &= const + \left(\sum_{i=1}^{n} a_{i2}b_{i2}\mathbb{E}(r_i)\left(\beta_{i0g}\boldsymbol{\mu}_{i1} + \sum_{t\geq 2}\sum_{h=1}^{G}\bar{\beta}_{i(t-1)h}\beta_{ithg}\boldsymbol{\mu}_{it}\right)\right)'\boldsymbol{u}_g,\end{aligned}$$

we have that $q(\boldsymbol{u}_g) \overset{\mathcal{D}}{=}$ von Mises-Fisher$(\kappa_g, \boldsymbol{\nu}_g)$, where

$$\begin{aligned}\kappa_g &= \Big\|\sum_{i=1}^{n} a_{i2}b_{i2}\mathbb{E}(r_i)\left(\beta_{i0g}\boldsymbol{\mu}_{i1} + \sum_{t\geq 2}\sum_{h=1}^{G}\bar{\beta}_{i(t-1)h}\beta_{ithg}\boldsymbol{\mu}_{it}\right)\Big\|, \\ \boldsymbol{\nu}_g &= \frac{1}{\kappa_g}\sum_{i=1}^{n} a_{i2}b_{i2}\mathbb{E}(r_i)\left(\beta_{i0g}\boldsymbol{\mu}_{i1} + \sum_{t\geq 2}\sum_{h=1}^{G}\bar{\beta}_{i(t-1)h}\beta_{ithg}\boldsymbol{\mu}_{it}\right),\end{aligned}$$

and therefore $\mathbb{E}(\boldsymbol{u}_g) = \boldsymbol{\nu}_g$.

(7) Derivation of $q(\alpha)$

Since

$$\begin{aligned}\log(q(\alpha)) &= const + \mathbb{E}\left\{\sum_t\sum_{i\neq j}\left(\left(y_{ijt}-\frac{1}{2}\right)\alpha - \frac{\omega_{ijt}}{2}(\alpha^2 + 2\alpha s_j \boldsymbol{X}_{it}'\boldsymbol{X}_{jt})\right)\right\} - \frac{\alpha^2}{2b_3^*}\\ &= const - \frac{1}{2}\left(\alpha^2\left(\sum_t\sum_{i\neq j}\mathbb{E}(\omega_{ijt}) + \frac{1}{b_3^*}\right) - 2\alpha\sum_t\sum_{i\neq j}((y_{ijt}-1/2) - \mathbb{E}(\omega_{ijt})\mathbb{E}(s_j)\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})\right),\end{aligned}$$

we have that $q(\alpha) \overset{\mathcal{D}}{=} N(a_3, b_3)$, where

$$\begin{aligned}a_3 &= \frac{\sum_t\sum_{i\neq j}((y_{ijt}-1/2) - \mathbb{E}(\omega_{ijt})\mathbb{E}(s_j)\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})}{\sum_t\sum_{i\neq j}\mathbb{E}(\omega_{ijt}) + 1/b_3^*},\\ b_3^{-1} &= \sum_t\sum_{i\neq j}\mathbb{E}(\omega_{ijt}) + 1/b_3^*.\end{aligned}$$

(8) Derivation of $q(\boldsymbol{s})$

Since

$$\begin{aligned}&\log(q(s_j))\\ &= const + \mathbb{E}\left\{\sum_{t=1}^T\sum_{i\neq j}\left((y_{ijt}-1/2)s_j\boldsymbol{X}_{it}'\boldsymbol{X}_{jt} - \frac{\omega_{ijt}}{2}(2\alpha s_j\boldsymbol{X}_{it}'\boldsymbol{X}_{jt} + s_j^2\boldsymbol{X}_{it}'\boldsymbol{X}_{jt}\boldsymbol{X}_{jt}'\boldsymbol{X}_{it})\right)\right\} - s_j\\ &= const - \frac{1}{2}\left\{s_j^2\sum_t\sum_{i\neq j}\mathbb{E}(\omega_{ijt})\left[tr(\Sigma_{it}\Sigma_{jt}) + \boldsymbol{\mu}_{jt}'\Sigma_{it}\boldsymbol{\mu}_{jt} + \boldsymbol{\mu}_{it}'\Sigma_{jt}\boldsymbol{\mu}_{it} + (\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})^2\right]\right.\\ &\left.-2s_j\left[\sum_t\boldsymbol{\mu}_{jt}'\sum_{i\neq j}\big(y_{ijt}-1/2-a_3\mathbb{E}(\omega_{ijt})\big)\boldsymbol{\mu}_{it} - 1\right]\right\},\end{aligned}$$

we have that $q(s_j) \overset{\mathcal{D}}{=} N_{(0,\infty)}(a_{j4}, b_{j4})$, where

$$\begin{aligned}a_{j4} &= b_{j4}\left(\sum_t\boldsymbol{\mu}_{jt}'\sum_{i\neq j}\big(y_{ijt}-1/2-a_3\mathbb{E}(\omega_{ijt})\big)\boldsymbol{\mu}_{it} - 1\right),\\ b_{j4}^{-1} &= \sum_t\sum_{i\neq j}\mathbb{E}(\omega_{ijt})\left[tr(\Sigma_{it}\Sigma_{jt}) + \boldsymbol{\mu}_{jt}'\Sigma_{it}\boldsymbol{\mu}_{jt} + \boldsymbol{\mu}_{it}'\Sigma_{jt}\boldsymbol{\mu}_{it} + (\boldsymbol{\mu}_{it}'\boldsymbol{\mu}_{jt})^2\right]\\ &= \sum_{t=1}^T\left\{\sum_{i\neq j}\left[\mathbb{E}(\omega_{ijt})\Big(tr(\Sigma_{it}\Sigma_{jt}) + \boldsymbol{\mu}_{it}'\Sigma_{jt}\boldsymbol{\mu}_{it}\Big)\right] + \boldsymbol{\mu}_{jt}'\Big(\sum_{i\neq j}\mathbb{E}(\omega_{ijt})(\Sigma_{it} + \boldsymbol{\mu}_{it}\boldsymbol{\mu}_{it}')\Big)\boldsymbol{\mu}_{jt}\right\}.\end{aligned}$$

The mean of s_j then is

$$\mathbb{E}(s_j) = a_{j4} + \frac{\sqrt{b_{j4}}\phi(a_{j4}/\sqrt{b_{j4}})}{\Phi(a_{j4}/\sqrt{b_{j4}})},$$

and the mean of s_j^2 is

$$\mathbb{E}(s_j^2) = b_{j4}\left(1 + \frac{a_{j4}/\sqrt{b_{j4}}\phi(a_{j4}/\sqrt{b_{j4}})}{\Phi(a_{j4}/\sqrt{b_{j4}})}\right) + a_{j4}^2.$$

(9) Derivation of $q(\boldsymbol{\beta})$

Since

$$\begin{aligned}\log(q(\boldsymbol{\beta}_0)) &= const + \mathbb{E}\left\{\sum_{g=1}^{G}\left[\sum_{i=1}^{n} Z_{i1g}\log(\beta_{0g}) + (\gamma_{0g}^* - 1)\log(\beta_{0g})\right]\right\} \\ &= const + \sum_{g=1}^{G}\left\{\sum_{i=1}^{n}\beta_{i0g} + \gamma_{0g}^* - 1\right\}\log(\beta_{0g}),\end{aligned}$$

we have that $q(\boldsymbol{\beta}_0) \overset{\mathcal{D}}{=} Dir(\boldsymbol{\gamma}_0)$, where

$$\gamma_{0g} = \gamma_{0g}^* + \sum_{i=1}^{n}\beta_{i0g}.$$

Similarly, for $h = 1, \ldots, G$,

$$\begin{aligned}\log(q(\boldsymbol{\beta}_h)) &= \mathbb{E}\left\{\sum_{i=1}^{n}\sum_{t\geq 2}\left[Z_{i(t-1)h}\sum_{k=1}^{G} Z_{itk}\log(\beta_{hk})\right] + \sum_{k=1}^{G}(\gamma_{hk}^* - 1)\log(\beta_{hk})\right\} \\ &= \sum_{g=1}^{G}\left\{\gamma_{hg}^* + \sum_{i=1}^{n}\sum_{t\geq 2}\bar{\beta}_{i(t-1)h}\beta_{ithg} - 1\right\}\log(\beta_{hg}).\end{aligned}$$

Thus $q(\boldsymbol{\beta}_h) \overset{\mathcal{D}}{=} Dir(\boldsymbol{\gamma}_h)$, where

$$\gamma_{hg} = \gamma_{hg}^* + \sum_{i=1}^{n}\sum_{t\geq 2}\bar{\beta}_{i(t-1)h}\beta_{ithg}.$$

References

Airoldi, E., D. Blei, E. Xing, and S. Fienberg (2005). A latent mixed membership model for relational data. In *Proceedings of the* 3^{rd} *International Workshop on Link Discovery*, pp. 82–89. ACM.

Anagnostopoulos, A., R. Kumar, and M. Mahdian (2008). Influence and correlation in social networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 7–15. ACM.

Anderlucci, L. and C. Viroli (2014). Covariance pattern mixture models for multivariate longitudinal data with application to the health and retirement study. *arXiv preprint arXiv:1401.1301*.

Arabie, P., S. A. Boorman, and P. R. Levitt (1978). Constructing blockmodels: How and why. *Journal of Mathematical Psychology 17*(1), 21–63.

Banerjee, A., I. S. Dhillon, J. Ghosh, and S. Sra (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 1345–1382.

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 803–821.

Bansal, S., J. Read, B. Pourbohloul, and L. A. Meyers (2010). The dynamic nature of contact networks in infectious disease epidemiology. *Journal of Biological Dynamics 4*(5), 478–489.

Barrat, A., M. Barthélemy, and A. Vespignani (2005). The effects of spatial constraints on the evolution of weighted complex networks. *Journal of Statistical Mechanics: Theory and Experiment 2005*(05), P05003.

Blasier, C. (1988). *The Giants Rival: The USSR and Latin America*. University of Pittsburgh Pre.

Borg, I. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Breiger, R. L., S. A. Boorman, and P. Arabie (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology 12*(3), 328–383.

Cameron, A. C. and F. A. Windmeijer (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics 14*(2), 209–220.

Carley, K. M. (2006). A dynamic network approach to the assessment of terrorist groups and the impact of alternative courses of action. Technical report, DTIC Document.

Choi, H. M. and J. P. Hobert (2013). The pólya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics 7*, 2054–2064.

Chung, H., Y. Park, and S. T. Lanza (2005). Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine 24*(18), 2895–2910.

clerk.house.gov (2003, May). Final vote results for roll call 182.

Collins, L. M., J. W. Graham, S. S. Rousculp, P. L. Fidler, J. Pan, and W. B. Hansen (1994). Latent transition analysis and how it can address prevention research questions. *NIDA Research Monograph 142*, 81–111.

Collins, L. M. and S. E. Wugalter (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research 27*(1), 131–157.

Cox, T. F. and M. A. Cox (1991). Multidimensional scaling on a sphere. *Communications in Statistics-Theory and Methods 20*(9), 2943–2953.

De la Cruz-Mesía, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis 52*(3), 1441–1457.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Dickinson, T. (2006). The 10 worst congressmen. *Rolling Stone*.

Doreian, P., R. Kapuscinski, D. Krackhardt, and J. Szczypula (1996). A brief history of balance through time. *Journal of Mathematical Sociology 21*(1-2), 113–131.

Downey, S. (2001, December). R. lawrence caughlin, former u.s. representative. *Philadelphia Inquirer*.

Durante, D. and D. B. Dunson (2014). Bayesian dynamic financial networks with time-varying predictors. *Statistics and Probability Letters 93*, 19–26.

Durrett, R. (2007). *Random graph dynamics*, Volume 20. Cambridge university press.

Ericson, J. (2012). Former iowa congressman jim leach has harsh words for today's politics. *Sioux City Journal*.

Foulds, J., C. DuBois, A. Asuncion, C. Butts, and P. Smyth (2011). A dynamic relational infinite feature model for longitudinal social networks. In *AI and Statistics*, Volume 15, pp. 287–295.

Fowler, J. (2006a). Connecting the congress: A study of cosponsorship networks. *Political Analysis 14*(4), 456–487.

Fowler, J. (2006b). Legislative cosponsorship networks in the us house and senate. *Social Networks 28*(4), 454–465.

Fraley, C. and A. E. Raftery (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification 24*(2), 155–181.

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer.

Gaffney, S. and P. Smyth (1999). Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 63–72. ACM.

Genolini, C. and B. Falissard (2010). Kml: k-means for longitudinal data. *Computational Statistics 25*(2), 317–328.

Geweke, J. and H. Tanizaki (2001). Bayesian estimation of state-space models using the metropolis–hastings algorithm within gibbs sampling. *Computational Statistics & Data Analysis 37*(2), 151–170.

Globerson, A., G. Chechik, F. Pereira, and N. Tishby (2004). Euclidean embedding of co-occurrence data. *Advances in Neural Information Processing Systems 17*, 497–504.

Gormley, I. C. and T. B. Murphy (2007). A latent space model for rank data. In *Statistical Network Analysis: Models, Issues, and New Directions*, pp. 90–102. Springer.

Goyal, A., F. Bonchi, and L. Lakshmanan (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM International Conference on Web Search and Data Mining*, pp. 241–250. ACM.

Graham, J. W., L. M. Collins, S. E. Wugalter, N. Chung, and W. B. Hansen (1991). Modeling transitions in latent stage-sequential processes: a substance use prevention example. *Journal of Consulting and Clinical Psychology 59*(1), 48–57.

Grier, R. M. (1999). Colonial legacies and economic growth. *Public Choice 98*(3-4), 317–335.

Handcock, M., A. Raftery, and J. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A 170*(2), 301–354.

Handcock, M. S. and K. J. Gile (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics 4*(1), 5–25.

Hanneke, S., W. Fu, and E. P. Xing (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics 4*, 585–605.

Hoff, P. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association 100*(469), 286–295.

Hoff, P. (2011). Hierarchical multilinear models for multiway data. *Computational Statistics & Data Analysis 55*(1), 530–543.

Hoff, P., A. Raftery, and M. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*(460), 1090–1098.

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

Hopcroft, J., T. Lou, and J. Tang (2011). Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 1137–1146. ACM.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure 10*(1), 1–29.

Huisman, M. and C. Steglich (2008). Treatment of non-response in longitudinal network studies. *Social Networks 30*(4), 297–308.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(1), 016107.

Kashima, H. and N. Abe (2006). A parameterized probabilistic model of network evolution for supervised link prediction. In *Proceedings of the Sixth International Conference on Data Mining*, pp. 340–349. IEEE.

Kempe, D., J. Kleinberg, and É. Tardos (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146. ACM.

Knecht, A. (2008). *Friendship Selection and Friends' Influence. Dynamics of Networks and Actor Attributes in Early Adolescence*. Ph. D. thesis, University of Utrecht.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks 28*(3), 247–268.

Krackhardt, D. and M. S. Handcock (2007). Heider vs Simmel: Emergent features in dynamic structures. In *Statistical Network Analysis: Models, Issues, and New Directions*, pp. 14–27. Springer.

Krause, A. E., K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor (2003). Compartments revealed in food-web structure. *Nature 426*(6964), 282–285.

Krivitsky, P., M. Handcock, A. Raftery, and P. Hoff (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks 31*(3), 204–213.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics 6*, 1100–1128.

Krivitsky, P. N. and C. T. Butts (2012). Exponential-family random graph models for rank-order relational data. *arXiv:1210.0493*.

Kumar, R., P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal (2000). Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pp. 57–65. IEEE.

Kunegis, J., A. Lommatzsch, and C. Bauckhage (2009). The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 741–750. ACM.

Leach, J. (2012, June 26,). Jim leach: An iowa republican carves a life in public service. *Washington Post*.

Leskovec, J., A. Singh, and J. Kleinberg (2006). Patterns of influence in a recommendation network. *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, 380–389.

Levitt, S. D. (1996). How do senators vote? disentangling the role of voter preferences, party affiliation, and senator ideology. *The American Economic Review*, 425–441.

Liben-Nowell, D. and J. Kleinberg (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology 58*(7), 1019–1031.

Lou, S., J. Jiang, and K. Keng (1993). Clustering objects generated by linear regression models. *Journal of the American Statistical Association 88*(424), 1356–1362.

Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics 19*(4), 474–482.

McKelvey, R. D. and W. Zavoina (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology 4*(1), 103–120.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics 38*(1), 153–168.

Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pp. 173–187. Springer.

Moody, J., D. McFarland, and S. Bender-deMoll (2005). Dynamic network visualization. *American Journal of Sociology 110*(4), 1206–1241.

Nakao, K. and A. Romney (1993). Longitudinal approach to subgroup formation: Re-analysis of newcomb's fraternity data. *Social Networks 15*(2), 109–131.

Newcomb, T. M. (1956). The prediction of interpersonal attraction. *American Psychologist 11*(11), 575.

Newcomb, T. M. (1961). *The Acquaintance Process*. Number 3. Holt, Rinehart and Winston New York.

Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E 70*(5), 056131.

Olguin, D. O., P. A. Gloor, and A. S. Pentland (2009). Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, Volume 9.

Onnela, J.-P., J. Saramäki, J. Hyvönen, G. Szabó, M. A. De Menezes, K. Kaski, A.-L. Barabási, and J. Kertész (2007). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics 9*(6), 179.

Opsahl, T., F. Agneessens, and J. Skvoretz (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks 32*(3), 245–251.

Opsahl, T. and P. Panzarasa (2009). Clustering in weighted networks. *Social Networks 31*(2), 155–163.

Plackett, R. (1975). The analysis of permutations. *Applied Statistics*, 193–202.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, 357–384.

Poole, K. T. and H. L. Rosenthal (2011). *Ideology and Congress*, Volume 1. Transaction Books.

Raftery, A. E., X. Niu, P. Hoff, and K. Y. Yeung (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics 21*(4), 901–919.

Ray, S. and B. Mallick (2006). Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 305–332.

Robins, G., P. Pattison, and J. Woolcock (2004). Missing data in networks: Exponential random graph (p^*) models for networks with non-respondents. *Social Networks 26*(3), 257–283.

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

Robinson, L. F. and C. E. Priebe (2012, December). Detecting Time-dependent Structure in Network Data via a New Class of Latent Process Models. *ArXiv e-prints*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Salter-Townshend, M. and T. B. Murphy (2012). Variational bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis 57*(1), 661–671.

Salter-Townshend, M. and T. B. Murphy (2013). Variational bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis 57*(1), 661–671.

Sarkar, P. and A. Moore (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter 7*(2), 31–40.

Sarkar, P., S. Siddiqi, and G. Gordon (2007). A latent space approach to dynamic embedding of co-occurrence data. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics AISTATS*, Volume 7.

Schweinberger, M. and T. Snijders (2003). Settings in social networks: A measurement model. *Sociological Methodology 33*(1), 307–341.

Scott, J. G. and L. Sun (2013). Expectation-maximization for logistic regression. *arXiv preprint arXiv:1306.0040*.

Scott, S. L., G. M. James, and C. A. Sugar (2005). Hidden markov models for longitudinal comparisons. *Journal of the American Statistical Association 100*(470).

Shannon, R. (2007). Mcconnell, rogers on 'most corrupt' list. *The Richmond Register*.

Snijders, T. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology 21*(1-2), 149–172.

Snijders, T., G. Van de Bunt, and C. Steglich (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks 32*(1), 44–60.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B 64*(4), 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Tang, J., J. Sun, C. Wang, and Z. Yang (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 807–816. ACM.

Thomas, A. C. and J. K. Blitzstein (2011). Valued ties tell fewer lies: Why not to dichotomize network edges with thresholds. *arXiv:1101.0788*.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*(4), 273–286.

Veall, M. R. and K. F. Zimmermann (1992). Pseudo-r^2s in the ordinal probit model. *Journal of Mathematical Sociology 16*(4), 333–342.

Veall, M. R. and K. F. Zimmermann (1994). Goodness of fit measures in the tobit model. *Oxford Bulletin of Economics and Statistics 56*(4), 485–499.

Vermunt, J. K., R. Langeheine, and U. Bockenholt (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics 24*(2), 179–207.

Vivar, J. C. and D. Banks (2011). Models for networks: a cross-disciplinary science. *Wiley Interdisciplinary Reviews: Computational Statistics 4*(1), 13–27.

Vrazilek, J. (2009). Hal rogers: a congressional disgrace. *National Review Online*.

Wang, C., V. Satuluri, and S. Parthasarathy (2007). Local probabilistic models for link prediction. In *Seventh IEEE International Conference on Data Mining*, pp. 322–331. IEEE.

Wasserman, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association 75*(370), 280–294.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of small-worldnetworks. *nature 393*(6684), 440–442.

Wu, C. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics 11*(1), 95–103.

Xing, E. P., W. Fu, and L. Song (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics 4*(2), 535–566.

Yan, X., C. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborová, P. Zhang, and Y. Zhu (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment 2014*(5), P05007.

Yang, S. and D. Knoke (2001). Optimal connections: strength and distance in valued graphs. *Social Networks 23*(4), 285–295.

Yellott Jr, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology 15*(2), 109–144.

Young, S. J. and E. R. Scheinerman (2007). Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph*, pp. 138–149. Springer.

Zhang, B. and S. Horvath (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology 4*(1), Article 17.