Mixed-Membership and Spatial Models for Social Network Data

by

Joseph Thomas Ciminelli

Submitted in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

Supervised by Professor Tanzy Love

Department of Biostatistics and Computational Biology
School of Medicine and Dentistry

University of Rochester
Rochester, New York

2017

Table of Contents

Biographical Sketch

Joseph T. Ciminelli was born in Rochester, New York, United States. He attended the University of Rochester, graduating magna cum laude in 2012 with a Bachelor of Arts degree with high honors in Political Science and minors in Legal Studies and Statistics. In 2013, he began graduate studies in Statistics at the University of Rochester. He graduated with a Master of Arts degree in 2014 and continued on to his doctoral studies thereafter. During his graduate studies, he has been actively involved with statistics education as a teaching assistant for both introductory and advanced statistics courses and as an instructor for an introductory biostatistics course. He has also conducted pedagogical research for statistical education, specifically in the context of student learning in a flipped introductory statistics course. His statistical research is in Bayesian methods for social network data in the contexts of mixed-membership and spatial models, under the direction of Professor Tanzy M. Love.

The following publications were a result of work conducted during his doctoral study:
Ciminelli, J., & Love, T. (2017) "Social Network Topic Model." *Forthcoming.*

Ciminelli, J., & Love, T. (2017) "Social Network Spatial Model." *Forthcoming.*

Maceroli, M., Nikkel, L.E., Mahmood, B., Qiu, X., **Ciminelli, J.**, Messing, S.,& Elfar, J.C. (2016). "Total hip arthroplasty for femoral neck fractures: Improved outcomes with higher hospital volumes." *Journal of Orthopaedic Trauma.*

Acknowledgements

Every adventure in life is made complete thanks to the many people who leave their imprint along the way. In the time leading to the completion of this thesis, I was fortunate to have the support of many loved ones, friends, colleagues, and mentors who made the journey all the more enjoyable and worthwhile. I truly have been blessed and thank God for the wonderful opportunity to pursue my academic passions. Looking back upon my years spent studying at the University of Rochester, I realize how fortunate I have been to be supported by such caring and loving people. Without their encouragement, I undoubtedly would not have been able to complete this work.

First, I thank my parents, Joe and JoAnne Ciminelli. Your love, guidance, and unwavering faith in me in fuels my desire to accomplish any dream. Thank you for inspiring me, for motivating me, for listening to me, and for always caring. Without you, I would not be where I am today. I must also acknowledge my grandparents. In no way would it have been possible for me to complete this thesis without my three biggest supporters. You have always cheered me on, letting me know through your love that every dream is a possibility. My sister, Nicole, and brother-in-law, Pranav, also deserve to be recognized for their support over these past few years. Without them, I would not have had the stability to complete this work. Time and time again, you provided me with the mental breaks I needed to refocus on my studies, always finding ways to put a smile on my face and make the journey all the more fun.

Thank you to Aiden. Being your Godfather motivates me each and every day to do my absolute best. You inspire me to set an example that you can be proud of. Just as I am finishing this thesis, I know you will accomplish any feat in life you set your heart too. To all my family, friends, and colleagues, thank you. You have played a part in getting me to where I am today. You have inspired the work presented in this thesis, and your love and encouragement have been essential to its completion. I have truly been blessed to have each of you in my life, supporting me through this wonderful journey.

I also extend my gratitude to the many members of the Department of Biostatistics and Computational Biology who helped me succeed throughout my graduate studies. In particular, thanks to Karin Gasaway and Christine Brower, I have never had to worry about anything other than my work itself—they have done a fantastic job making everything run so seamlessly. I would also like to thank the members of my thesis advisory committee: Professor Sally Thurston, Professor Anthony Almudevar, and Professor Yonathan Shapir. Your feedback, comments, and support have helped me grow as a researcher and statistician, and I am grateful for your dedication to my success. I would like to particularly recognize Sally, who continually sets an example of what it means to be a great researcher and teacher. She is an inspiration, and I am quite thankful for having had the chance to work alongside her these past few years.

Finally, I must extend a heartfelt thank you to my advisor, Professor Tanzy Love. I cannot put in words how grateful I am for all your guidance over the years. Through every question, every project, and every paper, you have always given me encouragement and insightful input. With you, I can ask any question, no matter how simple it may be. You have shown me what it means to be an advisor, and I am forever grateful for all you have given to me. Thank you for caring about me as a student and as a person. I cannot imagine completing this journey without you.

Abstract

Social media outlets have greatly increased the availability, prevalence, and relevance of networking data. Societal norms are characterized by network connections, with people communicating and interacting frequently with other individuals from around the world. As networks become ever more common, there grows a large source of data relating to inter-personal relationships and connections. Much of this data naturally links to additional sources of information that can be used to better understand network models.

Our work is motivated by a desire to incorporate the vast wealth of network data into other related statistical situations. In particular, we present a unified Bayesian approach for hierarchically soft clustering discrete data linked to members of a network. Focusing on text documents authored by members of a social network as our source of discrete data, we introduce a model that simultaneously represents a network and clusters text in a single social space, providing a unified representation of social relationships and topic membership. This method is evaluated through simulation studies, and extensions are introduced to allow for the more general inclusion of network data into discrete-data soft-clustering methods. In particular, we consider social networks that are linked to binary response data for multiple outcomes.

In addition to including social network data in hierarchical mixed-membership models, we also introduce a method for modeling the spatial correlations that exist

over a social network. In particular, we model attributes measured for each member of the network as a continuous process over the social space created by their connections. Our method simultaneously models the network in social space and the spatial process that exists over that network. The model is evaluated through simulation studies and applied to the importance ranking for a network of emergency response organizations. The introduced methods incorporate network data into mixed-membership and spatial frameworks, expanding traditional models to include this often relevant source of additional information.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professors Tanzy Love, Sally Thurston, and Anthony Almudevar of the Department of Biostatistics and Computational Biology and Professor Yonathan Shapir of the Department of Physics and Astronomy. Graduate study was supported by the School of Medicine and Dentistry Graduate Studies Dean Fellowship, a teaching assistantship from the Goergen Institute for Data Science, the Department of Biostatistics and Computational Biology, and the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL116906 (subcontract from University of Cincinnati).

List of Tables

List of Figures

Chapter 1

Introduction

People value their connections with others, if it be with friends who make them laugh, family they love, collaborators they work alongside, or even people they have never met but only interact with online. The growth of social media and technology has made social networks an integral part of life, allowing for people to easily communicate with anyone in the world. With each technological advancement and new social media website, network data becomes ever more plentiful and relevant to describing inter-personal relationships. By analyzing networks, we can better understand how people interact, determine whom they are closest to, and even recommend new connections so that individuals can become better connected to others who are similar to them.

Although the influx of data has been a recent phenomenon, social networks are not new; they have always existed. For as long as there have been groups of people, there have been cliques of those who get along and have interactions. Networks, though, are not limited to inter-personal relationships. In fact, networks span across biological functions, electronic systems, animal kingdoms, and many other areas (Newman, 2010). While we focus our applications on inter-personal social networks, our methods can be extended to other relevant networks.

The work presented in this thesis aims to build upon existing social network analyses and expand their relevance into other statistical fields, specifically topic modeling and spatial analysis. Since we often have more information relating to each person in the network aside from simply their connections, we aim to incorporate these attributes in order to better understand the relationships between the members of a network. This attribute data can provide for a richer understanding of social network interactions and the models associated with them.

1.1 Background

Although the study of networks is relatively new, there is a strong foundation of methods for network analyses. In particular, we base our network methodology on the spatial representation of social connections developed by Hoff, Raftery, and Handcock (2002), where the distances between individuals in a network represent the strengths of their relationships. The abstract space in which social relationships are placed is called *social space*. This two-dimensional Euclidean plane provides a visual representation of the network and is the space in which we situate our work. A two-dimensional specification of social space is not necessary, as higher dimensions can also be used.

1.1.1 Presence of Connections

Let us start by considering a group of N students in the same class year. Each student in this class is called a *character*, which is our generic reference for any *node* in the network. Each pair of characters can either have some connection or be unconnected (Newman, 2010). For any characters i and i', where $i' \neq i$, we let $y_{ii'} = 1$ if i is connected to i' and otherwise $y_{ii'} = 0$ if i is not connected to i'. Network connections are known as *edges* (Boccaletti et al., 2006). If, for example, we define a connection

as a friendship, then $y_{ii'} = 1$ if student i and student i' are friends. In this framework, connections are dichotomous; two characters are either connected or they are not. In a more general scenario, connections may be weighted to allow for the incorporation of the strength of relationships *a priori* (Barrat, Barthélemy, Pastor-Satorras, and Vespignani, 2004). For example, if we know that student 1 and student 2 are best friends, then $y_{12} = 2$. Student 1 and student 3 may be acquaintances, so in this case $y_{13} = 1$. Suppose student 1 and student 4 do not interact, then $y_{14} = 0$. Finally, if student 1 and student 5 are enemies, we could let $y_{15} = -1$. While this general framework does have strengths in allowing for weighted connections, we limit our investigation to that of binary outcomes, indicating the presence or absence of a connection between a pair of characters. By convention, we assume that characters are not connected to themselves, or $y_{ii} = 0$. Connections are the known source of data that we model in our network analyses.

1.1.2 Direction of Connections

For any pair of characters, the connection that may exist between them can be *directed* or *undirected* (Yuan and Wang, 2007). Consider our classmates example. Assume student 1 has a love interest in student 2. Thus, student 1 feels connected to student 2, so $y_{12} = 1$. Student 2, though, may try to avoid student 1 and therefore does not reciprocate the connection. This would lead to $y_{21} = 0$. In this example, we have directed connections. When a network is directed, $y_{ii'} = 1$ indicates that a connection originates with character i and is directed to character i'. Oftentimes, directed networks have instances of reciprocity in which connections exist in both directions between characters i and i' (Hoff et al., 2002).

If we now define a connection as existing between two students if they have taken

a math course together, then the connection is undirected. It is not possible for character i to be connected to character i' but the reverse not to be true. Thus, there is no directional element to the connection. In general, a directed network can be reduced to an undirected network if the direction of the connections between pairs of characters is ignored (Newman, 2010).

1.1.3 Connections in Social Space

A connection between two characters can be dependent, in some situations, on how similar the characters are. For example, if two students are on a sports team together and associate with the same people, they are more likely to have a friendship connection than a music-focused theater student and a business-focused football player who do not have any mutual friends. We must incorporate this dependency between characters into our formulation of social space (McFarland and Brown, 1973). The two-dimensional social space thus contains latent characteristics that influence connections between characters. By placing a probability measure on these latent characteristics, the chance of a connection between two characters depends on all other connections in the network (Hoff et al., 2002). Therefore, if two students share many of the same friends, then these two students are also likely to be connected.

1.2 Latent Modeling Approach

For a network of N characters, we have an $N \times N$ adjacency matrix, Y, with each element $y_{ii'}$ indicating whether a connection is present or not between characters i and i'. Y is necessarily symmetric if connections are undirected and is potentially asymmetric if connections are directed. In general, we also allow for covariate information, H, with each element, $h_{ii'}$, encoding the covariate information shared by the pair of

characters (i, i') (e.g. $h_{ii'}$ can be an indicator for whether student i and student i' are on a sports team together). The probability of a connection existing between a pair of characters is modeled as independent of all other connections in the network, conditioned on the unobserved positions of the pairs of characters in social space (Hoff et al., 2002). Thus, once we know where character i and character i' are in social space, and since that social space is based on all connections in the network, we can model the probability of a connection between these two characters as being dependent on the distance that separates them in social space, along with any relevant covariate information. We will denote the position of character i in social space as $\mathbf{z_i}$, which is a two-dimensional coordinate to be estimated.

1.2.1 Model Specification

With this setup, we parameterize the conditional probability of a connection between characters i and i' with a logistic regression model that depends on the distance between the characters' locations in social space, along with any observed covariate information (for simplicity, we present one covariate between characters i and i'):

$$P(y_{ii'} = 1|\mathbf{z_i}, \mathbf{z_{i'}}, h_{ii'}, a, b) = \frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1 + e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||}}. \tag{1.1}$$

Note that a and b are regression coefficients. From this logistic model, we can define the likelihood for the conditional independence model as follows:

$$P(Y|\mathbf{z}, H, a, b) = \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1 + e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||}}. \tag{1.2}$$

In general, the log-likelihood, $log(P(Y|\mathbf{z}, H, a, b))$, is not concave in $\{a, b, \mathbf{z}\}$. This lack of concavity can be attributed to $a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||$ not being affine. Ultimately, this leads to the direct maximization of parameter estimates not being straightforward;

instead, to maximize the parameter estimates, we first obtain a measure of dissimilarity between each pair of characters. As a most basic measure, we use the distance between characters' rows in the adjacency matrix (i.e. if two characters have similar connections in their rows of the adjacency matrix, then the distance between them will be small, whereas two characters with vastly different connections in their rows of the adjacency matrix will have a large distance separating them in social space). Multidimensional scaling is then used to determine the approximate character positions and distances between characters (Oh and Raftery, 2001).

Since our focus is on the distances between a set of points in Euclidean space, rotations, reflections, and translations all lead to equivalent representations of the space (Handcock and Raftery, 2007). Thus, there is an infinite number of latent positions, $\mathbf{z}$, that result in the same log-likelihood. If we collect all these equivalent representations of $\mathbf{z}$ under rotations, reflections, and translations, then for this entire set, there is only one set of distances between characters. This set of equivalent representations of $\mathbf{z}$ is a *configuration.* Inferential procedures are based on particular elements that are comparable across configurations. In particular, for a given configuration, we perform inference on the element $\mathbf{z}^* = \arg\min_{R^*\mathbf{z}} tr(\mathbf{z}_0 - R^*\mathbf{z})^T(\mathbf{z}_0 - R^*\mathbf{z})$, where $\mathbf{z}_0$ is taken to be the approximate maximum likelihood estimator (MLE) of $\mathbf{z}$ obtained from multidimensional scaling, and R^* is all possible rotations, reflections, and translations. $\mathbf{z}^*$, under this specification, is thus a *procrustes transformation* of $\mathbf{z}$ since it is the element of the configuration most similar to $\mathbf{z}_0$ in terms of sum of squared positional deviations. This procrustes transformation is unique so long as $\mathbf{z}_0\mathbf{z}^T$ is nonsingular (Sibson, 1979). In its traditional form, a procrustes transformation includes dilations. Throughout this thesis, though, we only consider a procrustes transformation as including translations, reflections, and rotations, in order to preserve the distances between characters in a network.

1.2.2 Inference

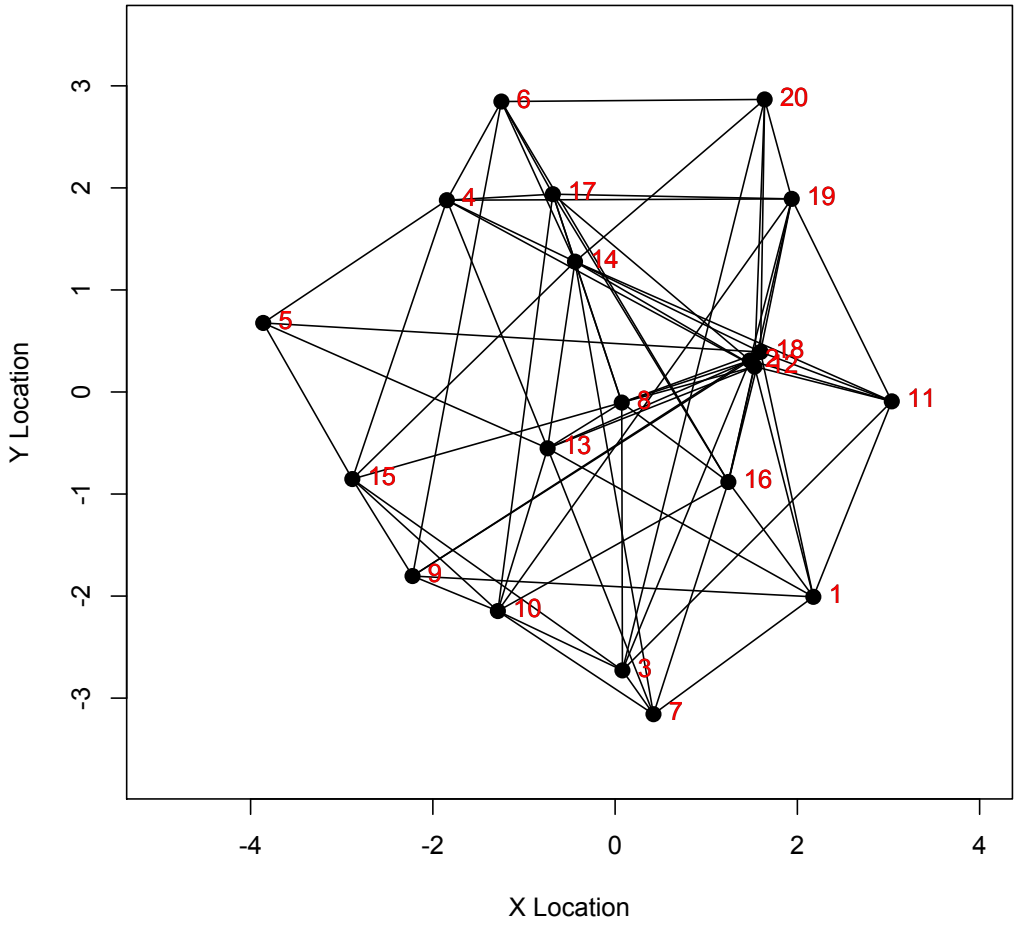
With this specification of the latent positions, $\mathbf{z}$, we can perform posterior inference on a, b, and $\mathbf{z}$ once prior distributions for the parameters are specified. This setup leads to the following Markov chain Monte Carlo (MCMC) posterior inferential procedure:

1. Identify $\mathbf{z}_0$, the zero-centered MLE of $\mathbf{z}$, via the multidimensional scaling method.
2. Let $\mathbf{z}_0$ be the starting value. Construct a Markov chain over the model parameters based on the following procedure for iteration $j+1$:
   (a) Sample a proposed $\tilde{\mathbf{z}}$ from $Q(\mathbf{z}|\mathbf{z}^{(j)})$, where Q is a symmetric proposal distribution.
   (b) Accept $\tilde{\mathbf{z}}$ for $\mathbf{z}^{(j+1)}$ with probability $\frac{P(Y|\tilde{\mathbf{z}},a^{(j)},b^{(j)},H)}{P(Y|\mathbf{z}^{(j)},a^{(j)},b^{(j)},H)}\frac{P(\tilde{\mathbf{z}})}{P(\mathbf{z}^{(j)})}$. Otherwise, update $\mathbf{z}^{(j+1)}$ with $\mathbf{z}^{(j)}$.
   (c) Store $\tilde{\mathbf{z}}^{(j+1)} = \underset{R^*\mathbf{z}^{(j+1)}}{\arg\min}\, tr(\mathbf{z}_0 - R^*\mathbf{z}^{(j+1)})^T(\mathbf{z}_0 - R^*\mathbf{z}^{(j+1)})$.
3. Update a and b with a Metropolis-Hastings algorithm.
4. Iterate between steps 2 and 3 until convergence.

This posterior inferential procedure thus results in a representation of character locations in social space, with distances between characters indicating the strength of their relationships. Social space in this way provides a convenient visualization of a network and the strength of connections between characters.

Consider our example of students in the same class year. Figure 1.1 represents an example social space of the network formed by the undirected connections of $N = 20$ students. While character 19 is connected to both character 11 and character 20, the connection between characters 19 and 20 is much stronger than that of characters 11 and 19, given by the shorter distance between characters 19 and 20 in social space.

Figure 1.1: Simulated social space of $N = 20$ characters



Due to the non-unique representation of a network in social space, posterior estimates of a and b are not of direct interest. In estimating a network, it is difficult to recover the true scale for which the network should be represented on. Connections imply closeness of characters in social space, so when connections do exist, they tend to drive characters closer together in social space, thus causing the scale to differ from the truth. Because of this, the estimates of a and b will depend on the scale of a given space. If the space is estimated on a large scale, then the magnitude of a and b will be larger than if the social network were estimated to be on a more constricted space. With social network analyses, we thus do not focus on the actual

estimates of a and b, since they often do not have meaningful contextual relevance and will depend on the particular social space being modeled. Instead, the main goal of social network analyses is to determine the location of characters in social space. Throughout this thesis, we do not directly focus on a and b. Inferential procedures for estimating these parameters are presented in the chapters below, and consideration is given as to whether convergence to the posterior distributions is met, but we limit the presentation of results of a and b from simulation studies and data applications.

1.3 Outline of Thesis

This social space based on distances between characters presented by Hoff et al. (2002) remains the setting in which our work will stay, but we build upon this network representation to incorporate additional information and extend its scope to topic modeling and spatial analyses. In Chapter 2, we situate topic models in social space, allowing for character attributes to be incorporated into the social network representation. Also, the social network provides information into topic modeling procedures, creating a feedback loop between network and topic models. In particular, we focus on text documents authored by characters in the social network. Ultimately, we combine Blei, Ng, and Jordan's (2003) Latent Dirichlet Allocation for topic modeling with Hoff et al.'s (2002) network approach to provide for a social network topic model.

In Chapter 3, we keep our focus on social network topic models, discussing extensions to the method introduced in Chapter 2. With these extensions, we allow for flexibility in our model to accommodate text documents that have multiple authors, as well as generalizing it to a greater class of mixed-membership models that can be applied outside the setting of text documents. Chapter 3 also includes a discussion of a variational expectation-maximization procedure for performing posterior inference on

model parameters. This inferential method aims to reduce computational complexity, allowing us to feasibly implement the social network topic model and its variants.

Chapter 4 shifts attention from topic models and instead focuses on spatial models. We introduce a method of traditional hierarchical spatial analysis for network data. In particular, since distances between characters in social space are indicative of the strength of their relationships, we investigate the manner in which such a space mirrors that of physical geographic distances used in traditional spatial models. With the goal of spatial models being to describe how a characteristic of interest varies over a geographic space, we aim to investigate how character attributes vary over social space. Ultimately, we introduce a method that incorporates spatial models into the framework of social network analyses. Our social network spatial model allows for character information to be included in the estimation of social space positions, as well as having the network inform our understanding of how an attribute of interest varies over the space. We examine model performance through a simulation study, and apply our method to a network of emergency response organizations in Texas.

Throughout all the models introduced in this thesis, our primary goal is to visualize social space, providing for a graphical representation of a network. We choose to represent social space in two-dimensional Euclidean space. While such a representation has its limitations, it does meet our objective of conveniently visualizing a network. We acknowledge that with the choice of two-dimensional space, we are unable to model four or more individuals as being equidistant in social space. In social network analyses with no additional covariate information, this limitation is only relevant if four or more characters are not connected to each other but are otherwise indistinguishable in their connections. As more information is added to the network (from either the topic or spatial models presented in this thesis), this potential issue becomes less worrisome, especially when compared to the benefits of visualization.

Chapter 2

Social Network Topic Model

2.1 Introduction

The analysis of text documents has become increasingly important and feasible within the growing contexts of social media and big data (Najafabadi et al., 2015). Determining applicable articles to return in a Google search, finding fellow Twitter users who discuss similar topics, and exploring email exchanges to understand a company's values all present relevant contexts for the necessity of analyzing text. Topic models allow us to cluster documents into similar groupings such that documents with shared interests are easily identifiable. By examining the words that exist within a document, we can determine how often various topics are discussed. While other topic models exist, our focus will be on that of *Latent Dirichlet Allocation* (LDA), an unsupervised Bayesian model for soft clustering text documents (Blei, Ng, and Jordan, 2003).

Methods of topic modeling fall within a larger class of hierarchical Bayesian mixed-membership models (HBMMMs). HBMMMs are applicable to any collection of discrete data and rely upon soft-clustering methods in which an object has partial inclusion in all clusters (Erosheva, Fienberg, and Lafferty, 2004; Erosheva and Fienberg,

2005). Recently, HBMMMs have grown in popularity due to their flexibility and vast applicability in the analysis of images (Barnard et al., 2003), text documents (Blei et al., 2003; Erosheva et al., 2004; Griffiths and Steyvers, 2005), disability profiles (Erosheva, 2002; Erosheva, 2003), and population genetics (Pritchard, Stephens, and Donnelly, 2000), among other areas.

In this chapter, we introduce LDA and its methodological approach to hierarchically clustering text documents. We then extend the LDA approach to include related network data, providing an additional source of information used in clustering documents. We describe a method for incorporating Hoff, Raftery, and Handcock's (2002) latent approach for modeling social networks into the context of LDA, providing for a comprehensive model that couples social networks and text documents. The chapter concludes with simulation results and a discussion of model selection.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic model for collections of discrete data, specifically that of text corpora. Documents are represented as random mixtures over latent topics, and each topic is characterized by a distribution over its words (Blei et al., 2003). Under this specification, a single text document has partial membership in each of the latent topics, which are given descriptions based on the words that have the highest probability under that topic. For example, consider an article that appears in *Nature*. Assume that there are potentially three topics to be discussed in the article, which are determined to be ecology, evolution, and genetics. These three latent topics—topic 1, topic 2, and topic 3—are given their labels—"ecology," "evolution," and "genetics"—based on the words most highly associated with them. Perhaps, for example, "plants," "green," and "ecosystem" are most com-

monly used when discussing topic 1, leading to a natural label of "ecology" for this topic. It should be noted, though, that these words can also belong to the "evolution" topic, but the word distributions would be different for the two topics. Based on the words used in the article, and since words are used to express the three topics, it can be determined that 90% of this document discusses ecology, with the remaining 10% of the article being equally split between evolution and genetics. Thus, this article has 90%, 5%, and 5% membership into each of the ecology, evolution, and genetics topics, respectively. Each word used in the article contributes to the article's membership into each of the three topics. This process of soft clustering can be extended to an entire corpus of documents and is the primary application of LDA.

2.2.1 LDA Characterization

2.2.2 Assumptions

LDA is an HBMMM, which is characterized based on its specification at four levels of assumptions: population, subject, latent variable, and sampling scheme (Airoldi, Blei, Erosheva, and Fienberg, 2014). Here, we present the assumptions as they apply to the LDA model.

Assumption 1: Population—

There are K finite sub-populations (i.e. topics) present in a text corpus, and for each of those K sub-populations, there is a probability distribution associated with each of M documents, denoted as $f(x_m|\beta_k)$, where x_m represents the words used in the m^{th} document and β_k are the relevant parameters for the k^{th} sub-population. At this level, it is assumed that within each sub-population, the observed words are independent across documents.

Assumption 2: Subject—

Each of M subjects (i.e. documents) is a mixture of the various sub-populations specified in Assumption 1. The mixture for document m, known as the mixed-membership vector, is denoted by $\theta_m = (\theta_{m[1]}, ..., \theta_{m[K]})$. Given the mixed-membership vector for a document, the distribution of the words in that document is specified as $P(x_m|\theta_m, \beta) = \sum_{k=1}^{K} \theta_{m[k]} f(x_m|\beta_k)$. Here, we assume that, conditional on the mixed-membership vector, the words are independent of one another and across documents.

Assumption 3: Latent Variable—

The mixed-membership vector for each of M documents, θ_m, is a realization of a latent variable that has distribution D_α. With this specification, the probability of observing the words in a document, given the associated parameters, is denoted $P(x_m|\alpha, \beta) = \int \left(\sum_{k=1}^{K} \theta_{m[k]} f(x_m|\beta_k) \right) D_\alpha(d\theta)$.

Assumption 4: Sampling Scheme Level—

In each of M documents, there are R_m word positions. The R_m replicates (i.e. word positions) in a document are taken to be independent of one another. Under this assumption, the probability of observing a set of words in a document, $x_m^{1:R_m}$, conditional on the associated parameters, is $P(x_m^{1:R_m}|\alpha, \beta) = \int \left(\prod_{r=1}^{R_m} \sum_{k=1}^{K} \theta_{m[k]} f(x_m^r|\beta_k) \right) D_\alpha(d\theta)$.

The last important consideration is that of the exchangeability condition to be placed on the positions in each document. Exchangeability tells us that the joint distribution of the topics for each position is invariant to permutation (Aldous, 1983). Furthermore, we assume that the topics are infinitely exchangeable—meaning that every finite subsequence is exchangeable—within a document, and thus the words, which are realizations of a topic, can be placed in any order (Blei, 2012).

2.2.3 Parameter Specification

In the context of the LDA model, the response variable represents which word in the vocabulary is used to fill a position in the text. Note that standard stop words (e.g. "the," "and," and "that") are removed during preprocessing, leaving only contextual words in the documents to be used for LDA modeling. We allow text documents to have different lengths, meaning that the number of replicates can differ across documents. It is assumed that each position in a document is filled with a word that specifies a single topic (i.e. when a word is used at a given point in a document, it is only expressing one topic), and since a document is a collection of these words, it can express possibly different topics (Blei et al., 2003). This classifies LDA as a soft-clustering model in which we do not require each document to only be part of one group; instead, it is a mixture of the various groups.

Using the LDA model specification, we let the mixed-membership vector θ_m be distributed according to a Dirichlet distribution with hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ (i.e. $D_{\boldsymbol{\alpha}} = \text{Dirichlet}(\alpha_1, ..., \alpha_K)$). The Dirichlet distribution is a multidimensional extension of the beta distribution, with probability density

$$P(\theta_m|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_{m[1]}^{\alpha_1 - 1} ... \theta_{m[K]}^{\alpha_K - 1}, \tag{2.1}$$

where $\theta_{m[k]} \geq 0$ for $k = 1, ..., K$ and $\sum_{k=1}^{K} \theta_{m[k]} = 1$. Recall that θ_m is a vector for the m^{th} document that encodes the topic proportions that inform the choice of words in that document, leading to a Dirichlet distribution as being a natural choice for θ_m.

We now introduce u_m^r, a latent topic indicator variable that specifies the topic to be expressed by the word in the r^{th} position of document m. Since we want to draw one of K categories based on the mixture of topics that are discussed in document m, which is expressed by θ_m, the topic indicator variable is specified as following a multinomial

distribution based on the mixed-membership vector (i.e. $u_m^r|\theta_m \sim \text{Multinomial}(\theta_m, 1)$) Note that u_m^r is a vector of length K in which only one element equals 1 and all the rest are 0 (i.e. we indicate which of the K topics is being utilized in the r^{th} position of document m). We can use this topic indicator in our specification of $f(x_m^r|\beta_k)$ as multinomial. As discussed above, each word expresses a specific topic, so we can say that $f(x_m^r|\beta_k) = P(x_m^r|u_{m[k]}^r = 1, \beta) = \text{Multinomial}(\beta^T u_m^r, 1)$, where β_k is a probability row vector the size of the vocabulary, which we will denote as V, and β is a $K \times V$ matrix, made up of the K random row vectors β_k. Note that $\sum_{v=1}^{V} \beta_{k[v]} = 1$. In general, a text corpus will have a large vocabulary size, but most documents will only use a relatively small subset of the words in the vocabulary. For this hierarchical model, we consider the matrix β and let each row be an independent draw from an exchangeable Dirichlet distribution with hyper-parameters $\boldsymbol{\phi} = (\phi_1, ..., \phi_V)$ (i.e. $D_{\boldsymbol{\phi}} \sim \text{Dirichlet}(\phi_1, ..., \phi_V)$). Typically, we let $\boldsymbol{\phi}$ be symmetric in that $\phi_1 = ... = \phi_V = \phi$. There is little information *a priori* for us to know which topics each vocabulary word has the highest probability of appearing in (Wallach, Mimno, and McCallum, 2009); instead, we *a priori* let each word have equal probability of being used to express each of the K topics. We will assume $\boldsymbol{\phi} = \phi$, and is thus a symmetric hyper-parameter, for the remainder of this chapter.

2.2.4 LDA Model

Given this setup, LDA can now be translated into a finite mixture model. It is to be assumed that the number of latent topics, K, is unknown but at some finite value. To be described below, we must determine the optimal number of topics. For a fixed number of topics, K, the finite mixture model has the following generative process:

1. For each of K topics, sample the length V vocabulary probability vector $\beta_k \sim$

Dirichlet(ϕ).

2. For each of M documents:

   (a) Sample the mixed-membership vector $\theta_m \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$.

   (b) For each of R_m words in document m:

      i. Sample a latent topic indicator $u^r_m|\theta_m \sim \text{Multinomial}(\theta_m, 1)$.

      ii. Sample a word $x^r_m|u^r_m, \beta \sim \text{Multinomial}(\beta^T u^r_m, 1)$

It should be noted that we set the hyper-parameters for the Dirichlet distribution on the topics to be symmetric (i.e. $\alpha_1 = ... = \alpha_K$). This symmetry can be interpreted to mean that each of the K topics are equally likely *a priori*, and thus we do not expect any topics to dominate over the others (Airoldi, Fienberg, Joutard, and Love, 2006). This is a reasonable assumption to make here, since there should be a diversity of topics without any information *a priori* of one being more dominant than the others. Additionally, we can think of the $(k, v)^{th}$ element of β as being the probability of the v^{th} vocabulary word occurring when we are discussing the k^{th} topic. This can probabilistically be viewed as $\beta_{k[v]} = P(x^r_{m[v]} = 1|u^r_{m[k]} = 1)$. Since we are using the latent topic indicator here, and that indicator is specific to each position in the text, separate instances of the v^{th} vocabulary word in the same document can be generated from different topics (Blei et al., 2003).

2.3 Extending LDA to Include Social Network Information

Oftentimes, documents will naturally have a social network to accompany the text itself. For example, consider the social media website Twitter. Individuals will Tweet

text discussing some topic of interest to them. If we take each Tweet to be a text document, we can then form a corpus of Tweets for a defined group of users. Furthermore, users are linked through the Twitter network, and we are able to define which users have social connections and which do not. In this way, we have that each Tweet is linked to a specific user, as well as who that user is socially connected to (Alvarez-Melis and Saveski, 2016). Additional examples of text documents that are linked with social networks include email exchanges (Hardin, Sarkis, and URC, 2015), political correspondences (Johnson and Orbach, 2002), blog posts and replies (Liang, 2015), and works of literature in which characters interact with each other (Kydros and Anastasiadis, 2015).

Recall that ultimately the goal of topic modeling is to determine how often each document discusses each of K possible topics. From this, we are able to group documents based on the topics they most commonly discuss. Using traditional LDA, this mixed-membership clustering is based on the probability with which each word in a document discusses a certain topic. Our goal is to extend this approach to include additional information from the social network that naturally accompanies many text documents in order to more accurately learn of the topic membership of each document. In particular, if we know that two individuals are socially closely related, then they are likely to share interests and discuss similar topics. This added information should help to distinguish the true topic distribution of documents.

2.3.1 Incorporating Social Networks into Topic Models

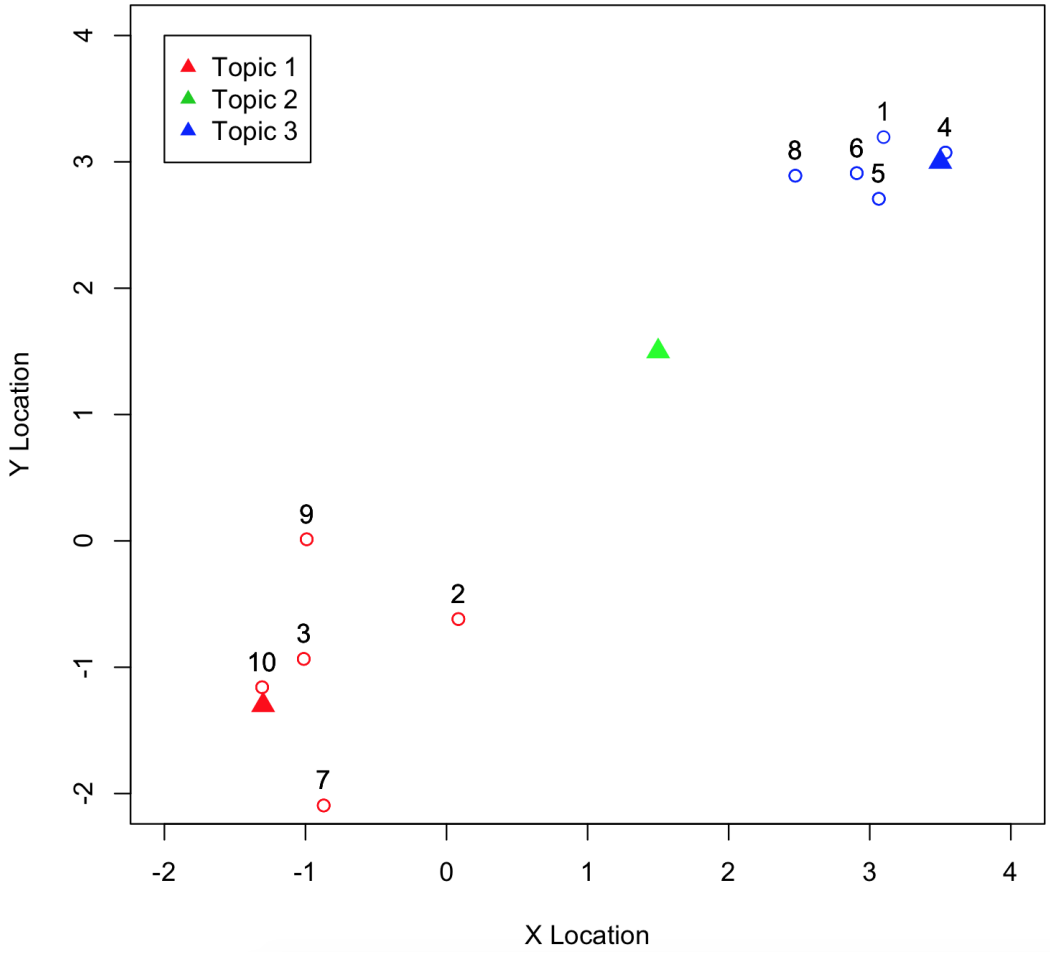
In the preceding chapter, we discussed Hoff et al.'s latent approach for representing a social network, with distances between characters indicating the strength of their relationships (2002). For example, two characters who are close together in the two-

dimensional social space have a stronger relationship than a pair of characters who are far away from each other. The goal of the latent network method is to determine the locations of characters in social space based on their connections to other characters.

If a network does exist within the framework of a text corpus, there are two issues to be considered. First, there is that of topic modeling for the text. Second, we must also determine the spatial representation of the network. As introduced above, it seems reasonable to include information from the social network in the determination of the topic membership for text documents, but this would suppose, in its current presentation, that the social network is fixed and known. This, though, is not the case; instead, we must learn of the representation of the social space. Because of this, our goal is to use information from the text documents to better learn of the degree to which characters are socially related. Two characters who often discuss similar topics are likely to be closer in social space, since this would mean that they share similarities. Thus, we can rely upon the information from the topic model to inform our social network representation. Together, the social network and text documents inform the modeling of the other, providing a loop of information that allows for both the social network and topic membership to be better represented.

To approach this task of combining the latent procedure for modeling social networks with hierarchical topic modeling via LDA, we directly place the text topics in the two-dimensional social space of the network. Based on this representation, characters who are close to a given topic are more likely to often discuss that topic in their respective text documents. Alternatively, if a character is far away from a topic, that character will not often discuss that topic. A simulated social space of $N = 10$ characters and $K = 3$ topics is displayed in Figure 2.1. With this combining of both the text and the social spaces together, and by having distances between characters and topics indicate how likely a character is to discuss a topic in a text document, we

Figure 2.1: Simulated social space of $N = 10$ characters and $K = 3$ topics



can represent θ_i, our mixed-membership vector for character i, as a function of the distance between character i and each of the K topics. It should be noted that each character, regardless of the number of documents they have written, is assumed to always be discussing the same mixture of topics. Since positions in social space are assumed to be fixed and unchanging, every document that a character authors must have the same mixed-membership representation. Thus, we are no longer concerned with individual documents. Instead, we combine all of one character's documents together and focus on the mixed-membership vector for each character in the network.

Since the mixed-membership vector for each character is a function of the distance

between that character and each topic in social space, we must specify the exact function form of this relationship. We use an exponentially decaying relationship between characters and topics based on the distance separating them. In particular, as a character has an increased distance from a topic, the mixed-membership proportion will be exponentially smaller than a topic that the character is closer to.

2.3.2 Model Specification

To formally specify the relationship between characters and topics, we must introduce some notation. Let $\mathbf{z_i}$ be the Cartesian coordinates of character i in social space. Additionally, let $\mathbf{z_{t_k}}$ be the location of the k^{th} topic in social space. Then the mixed-membership proportion for character i discussing topic k is given as $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$, where $||\mathbf{z_{t_k}} - \mathbf{z_i}||$ represents the Euclidean distance between topic k and character i.

In specifying the mixed-membership vector as being a function of both the location of the characters and the topics, our two previously independent models (i.e. LDA and the latent social network model) now share information and are inter-related. Considering a network of N characters who each solely author a document, we can now specify our combined generative social network topic model as follows:

1. Sample N character locations, $\mathbf{z} = (\mathbf{z_1}, ..., \mathbf{z_N})$, in social space from $P(\mathbf{z})$.

2. Sample K topic locations, $\mathbf{z_t} = (\mathbf{z_{t_1}}, ..., \mathbf{z_{t_K}})$, in social space from $P(\mathbf{z_t})$.

3. For each of K topics, sample the length V vocabulary probability vector $\beta_k \sim$ Dirichlet(ϕ).

4. For each of N characters:

(a) Calculate the topic membership vector, θ_i, based on the character's distance from each topic: $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$.

(b) For each of the R_i words in character i's document:

i. Sample a latent topic indicator $u_i^r|\theta_i \sim \text{Multinomial}(\theta_i, 1)$.

ii. Sample a word $x_i^r|u_i^r, \beta \sim \text{Multinomial}(\beta^T u_i^r, 1)$.

5. Sample social network parameters a and b from $P(a)$ and $P(b)$, respectively.

6. Sample the edges $Y \sim \text{Bernoulli}(\mathbf{p_{adj}})$, where $p_{adj_{ii'}} = \frac{e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||}}{1+e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||}}$ and $h_{ii'} = \frac{1}{2}\sum_{v=1}^{V}\left(\theta_i\beta_{[.v]}log\left(\frac{\theta_i\beta_{[.v]}}{\theta_{i'}\beta_{[.v]}}\right) + \theta_{i'}\beta_{[.v]}log\left(\frac{\theta_{i'}\beta_{[.v]}}{\theta_i\beta_{[.v]}}\right)\right)$.

Steps 1 and 2 of this model position both the characters and topics in social space based on their respective distributions, $P(\mathbf{z})$ and $P(\mathbf{z_t})$. The topic model is present in steps 3 and 4. The novelty in the approach described above is that each θ_i is no longer sampled from a Dirichlet distribution but instead is deterministically calculated based on the distances between character i and the topics. Finally, the social network aspect of the model is in steps 5 and 6. Note that a and b are scalars, with a relating to the base probability of a connection and b giving information regarding the impact of covariate information, after accounting for distances between characters. The unique element of this social network step is the inclusion of the text-dependent covariate, the $h_{ii'}$ term, which is the average Kullback-Leibler divergence between the word distributions used by character i and character i'. This term draws upon the text portion of the model and incorporates information from the documents directly into the determination of characters' social relationships.

Being a generative model, we assume that the documents and social connections are created based on the specified process. We use this generative process to take our data, which consists of the words in text documents, $\mathbf{x}$, and the social network

connections, Y, and draw posterior inference regarding the parameters in the model. These parameters determine the locations of characters in social space, the locations of topics in social space, the probability of words being used to express the various topics, and the topic membership vector for each character.

2.3.3 Inference

The key inferential interest is in determining the joint posterior distribution of the parameters $\mathbf{z_t}$, $\mathbf{z}$, β, $\mathbf{u}$, a, and b. Recall from Chapter 1 that inference regarding a and b will depend on the particular social space representation, so although consideration is given to these social network parameters, we place a higher interest on $\mathbf{z_t}$, $\mathbf{z}$, β, and $\mathbf{u}$. The joint posterior distribution of model parameters is specified as:

$$P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, b|\mathbf{x}, Y, \boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta) = \frac{P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, b, \mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)}{P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)}. \tag{2.2}$$

The difficulty here, though, is that $P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)$ does not have a closed-form solution, with the likelihood being expressed as:

$$\begin{aligned} &P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta) \propto \\ &\int\int\int\int\int \sum_{\mathbf{u}} \left(\prod_{i=1}^{N} \prod_{r=1}^{R_i} \prod_{k=1}^{K} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \prod_{v=1}^{V} \beta_{k[v]}^{x_{i[v]}^{r}} \right)^{u_{i[k]}^{r}} \right) \\ &\times \left(\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}}{1 + e^{a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||}} \right) \\ &\times D_{\boldsymbol{\eta}}(d\mathbf{z_t}) D_{\boldsymbol{\nu}}(d\mathbf{z}) D_{\phi}(d\beta) D_{\omega}(da) D_{\delta}(db). \end{aligned} \tag{2.3}$$

The coupling of $\mathbf{z_t}$, $\mathbf{z}$, and β in the summation of latent topics renders this likelihood intractable for analytic inference (Airoldi et al., 2006; Dickey, 1983). Because of this, approximate inferential methods are used to determine the joint posterior distribu-

tion. Common methods used for approximate inference include Markov chain Monte Carlo (MCMC) sampling, variational approximation, and the Laplace approximation (Airoldi et al., 2006; Mukherjee and Blei, 2008; Porteous et al., 2008; Teh, Newman, and Welling, 2007; Sontag and Roy, 2011). We focus here on an MCMC method to estimate the joint posterior distribution.

For each of the parameters of interest, we determine its conditional posterior distribution. Gibbs sampling is rendered inappropriate, as conjugacy is not present. The conditional posterior distributions of the parameters of interest are as follows:

$$P(\beta|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \mathbf{u}, a, b, \phi) \propto \tag{2.4}$$
$$\left[\prod_{v=1}^{V}\prod_{k=1}^{K}\beta_{k[v]}^{\phi+\left(\sum_{i=1}^{N}\sum_{r=1}^{R_i}u_{i[k]}^{r}x_{i[v]}^{r}\right)-1}\right]\left[\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}\right]$$

$$P(u_i^r|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta) \propto \prod_{k=1}^{K}\left[\left(\frac{e^{-||\mathbf{z_{t_k}}-\mathbf{z_i}||}}{\sum_{c=1}^{K}e^{-||\mathbf{z_{t_c}}-\mathbf{z_i}||}}\right)\left(\prod_{v=1}^{V}\beta_{k[v]}^{x_{i[v]}^{r}}\right)\right]^{u_{i[k]}^{r}} \tag{2.5}$$

$$P(\mathbf{z_t}|\mathbf{x}, Y, \mathbf{z}, \beta, \mathbf{u}, a, b, \boldsymbol{\eta}) \propto \tag{2.6}$$
$$P(\mathbf{z_t}|\boldsymbol{\eta})\left[\prod_{i=1}^{N}\prod_{r=1}^{R_i}\prod_{k=1}^{K}\left(\frac{e^{-||\mathbf{z_{t_k}}-\mathbf{z_i}||}}{\sum_{c=1}^{K}e^{-||\mathbf{z_{t_c}}-\mathbf{z_i}||}}\right)^{u_{i[k]}^{r}}\right]\left[\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}\right]$$

$$P(\mathbf{z}|\mathbf{x}, Y, \mathbf{z_t}, \beta, \mathbf{u}, a, b, \boldsymbol{\nu}) \propto \tag{2.7}$$
$$P(\mathbf{z}|\boldsymbol{\nu})\left[\prod_{i=1}^{N}\prod_{r=1}^{R_i}\prod_{k=1}^{K}\left(\frac{e^{-||\mathbf{z_{t_k}}-\mathbf{z_i}||}}{\sum_{c=1}^{K}e^{-||\mathbf{z_{t_c}}-\mathbf{z_i}||}}\right)^{u_{i[k]}^{r}}\right]\left[\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}\right]$$

$$P(a|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta, b, \omega) \propto P(a|\omega)\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}} \tag{2.8}$$

$$P(b|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta, a, \delta) \propto P(b|\delta)\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}. \tag{2.9}$$

Note that $h_{ii'}$ is a function of β, $\mathbf{z_t}$, and $\mathbf{z}$ in the above distributions. We recommend setting bivariate normal prior distributions for each $\mathbf{z_{t_k}}$ and $\mathbf{z_i}$. β has a symmetric Dirichlet prior distribution with parameter ϕ. The prior specifications for a and b are

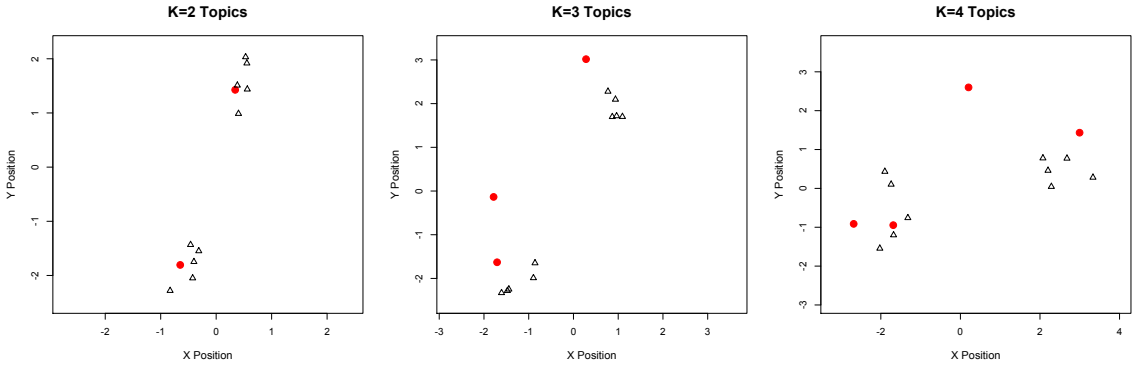
recommended to be fairly uninformative uniform distributions. When implementing MCMC inference, we set hyper-parameters $\boldsymbol{\eta}$, $\boldsymbol{\nu}$, ϕ, ω, and δ to fixed values, as opposed to directly estimating them. We use Just Another Gibbs Sampler (JAGS) to implement slice sampling from each of these conditional posterior distributions (Plummer, 2013).

2.3.4 Model Selection

In the preceding model specification, we assumed that the number of topics, K, was fixed and known. In practice, we must decide the appropriate number of topics to use in describing our data. Likelihood-based methods of deviance information criterion and Bayesian information criterion do not provide reasonable mechanisms for correctly choosing the number of topics. To handle the model selection issue, we therefore recommend relying upon visual inspection. In particular, we recommend fitting the social network topic model over a range of reasonable values of K. The resulting social spaces should be plotted. When looking at the social spaces, if adding an additional topic causes two topic locations to be relatively close, we recommend using the more parsimonious model that provides a reasonable representation of the space. For example, in Figure 2.2, we present the social space representations resulting from social network topic models fit with $K = 2$, $K = 3$, and $K = 4$ topics to data simulated with $K = 3$ topics. Visual inspection indicates that the addition of the third topic is reasonable, as the third topic is placed relatively far away from the other two topics. With addition of a fourth topic, though, the resulting social space has two topics placed in close proximity, indicating that not both are necessary. Thus, visual inspection correctly indicates that $K = 3$ topics is appropriate. We also note that consideration should be given to the interpretability of topics, providing an

additional aid to the visual inspection in selecting an appropriate K.

Figure 2.2: Social network topic model fit with $K = 2$, $K = 3$, and $K = 4$ topics



Although visual examination of the fitted models provides a mechanism for selecting the number of topics, it is imperfect. As with all mixed-membership clustering models, the choice of K is an important consideration for which work still remains to be done. The limitation of likelihood-based methods performing poorly for the social network topic model adds an additional challenge, but visual inspection, combined with interpretability considerations, renders the prescribed method as a basic mechanism for choosing a reasonable number of topics. In future work, we plan to further investigate the issues of model selection.

2.4 Simulations

To validate this model and evaluate its performance, we conduct a series of simulation studies. In general, we select each of $N = 10$ character locations from a mixture distribution of bivariate normal random variables, $P(\mathbf{z_i}) = \pi N(\boldsymbol{\mu_1}, \Sigma_1) + (1-\pi)N(\boldsymbol{\mu_2}, \Sigma_2)$. Locations for $K = 3$ topics are placed at fixed points, $\mathbf{z_{t_1}}$, $\mathbf{z_{t_2}}$, and $\mathbf{z_{t_3}}$. A total vocabulary size of $V = 60$ words is used. Three word distributions, β_1, β_2, and β_3, are drawn

from asymmetric Dirichlet distributions such that $v_1, ..., v_{20}$ have the greatest probability of being used to describe topic 1, $v_{21}, ..., v_{40}$ have the greatest probability of being used to describe topic 2, and $v_{41}, ..., v_{60}$ have the greatest probability of being used to describe topic 3. Documents for each character are then created based on the character's location-dependent mixed-membership vector, θ_i, and the vocabulary probabilities, β. The social network parameters are set at $a = 3$ and $b = -10$. We simulate network connections, $y_{ii'}$, from a Bernoulli distribution dependent on the locations and word distributions in order to create an adjacency matrix. A total of forty-five undirected connections are possible. The simulated documents and adjacency matrix are the necessary data inputs to perform our posterior inference. We implement slice sampling in `R` using the `rjags` library, which relies upon the use of JAGS. From this, we are able to draw posterior samples for each simulation in order to estimate the model parameters (Plummer, 2013). In our model specification, we place prior distributions on model parameters as $\beta_k \sim \text{Dirichlet}(0.01, ..., 0.01)$, $\mathbf{z_i} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$, $\mathbf{z_{t_k}} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$, $a \sim \text{Unif}(0.1, 5)$, and $b \sim \text{Unif}(-20, 20)$.

Our goal with the simulations is to compare the social network topic model (known hereafter as the dependence model) with the independent models of LDA and the latent social network approach (known hereafter as the independence model). In particular, for the independence model, we consider the documents by themselves without any social network influence and calculate the mixed-membership vectors $\theta_1, ..., \theta_N$ and vocabulary probability matrix β based on the original LDA method that was previously introduced. We rely upon the variational approximation method for estimating the parameters, as originally described by Blei et al. (2003). Additionally, we

consider the adjacency matrix of social connections by itself without the influence of the text documents and implement Hoff et al.'s latent approach (2002) for modeling the positions of characters, $\mathbf{z}$, in social space. Using this independence model as the current standard for analyzing text and network data, we evaluate the performance of the dependence model in estimating the parameters θ, β, and $\mathbf{z}$. Recall from Chapter 1 that the estimation of a and b is not of direct interest, and thus we do not present results for such parameters here. In each simulation, we calculate the sum of squared deviations of each parameter estimate from the truth, with the goal of minimizing this error, as follows:

$$\sum_{i=1}^{10}\sum_{k=1}^{3}\left(\hat{\theta}_{i[k]}-\theta_{i[k]}\right)^2$$
$$\sum_{k=1}^{3}\sum_{v=1}^{60}\left(\hat{\beta}_{k[v]}-\beta_{k[v]}\right)^2$$
$$\sum_{i=1}^{10}\left(\left(\hat{z}_{i[1]}-z_{i[1]}\right)^2+\left(\hat{z}_{i[2]}-z_{i[2]}\right)^2\right).$$

Since each $\mathbf{z_i}$ is a two-dimensional coordinate, we let $z_{i[1]}$ be character i's coordinate in the first dimension and $z_{i[2]}$ be character i's coordinate in the second dimension.

With both traditional LDA and our social network topic model, the labels of the fitted topics may be switched compared to the labels of the originally simulated topics. The assignment of each topic is arbitrary and the labeling of the posterior distribution may differ from that of the simulated topics. For example, originally simulated topic 1 may be labeled as topic 3 during posterior inference. To account for this potential discrepancy in labels, we permute all possible label classifications in the posterior distribution of θ and consider the specification that minimizes the sum of squared difference between the estimates of θ and β and the true parameter values.

Hoff et al.'s latent approach does not have a unique representation of the network

(2002). Translations, reflections, and rotations all lead to equivalent social spaces. Because of this, a procrustes transformation allowing for any combination of translations, reflections, and rotations is used to orient the inferred network as closely to the simulated points as possible, based on the sum of squared deviations (Sibson, 1979).

2.4.1 Assessing Convergence: A Multivariate Approach

In each simulation, the MCMC sampling method for posterior inference is run until convergence of the posterior distributions is met based on the effective sample size of a multivariate Markov chain for θ, implemented through the `mcmcse` package in `R` (Flegal, Hughes, and Vats, 2016). The effective sample size of the multivariate Markov chain, referred to as the multivariate effective sample size (or multi-ESS), accounts for correlation between parameters. In our case of $K = 3$ topics, once we know $\theta_{i[1]}$ and $\theta_{i[2]}$, we can deterministically calculate $\theta_{i[3]}$ as $\theta_{i[3]} = 1 - (\theta_{i[1]} + \theta_{i[2]})$. From this, we can see that, in general, $\theta_{i[3]}$ is perfectly correlated with $\theta_{i[1]}$ and $\theta_{i[2]}$, and $\theta_{i[1]}$ and $\theta_{i[2]}$ are also correlated. Thus, using the multi-ESS method that considers the variables as correlated allows for a better assessment of convergence of our Markov chain. The multi-ESS is calculated as follows:

$$\text{multi-ESS} = n\frac{|\Lambda|^{\frac{1}{p}}}{|\Sigma|^{\frac{1}{p}}}, \tag{2.10}$$

where Λ is the sample covariance matrix for the mean of the target density, Σ is an estimate of the Monte Carlo standard error of the mean of the target density, and p is the number of parameters we are considering (Flegal et al., 2016). Since, as discussed above, one entry in a character's mixed-membership vector is deterministic, we focus only on $K-1$ entries for each character. In particular, when $K = 3$, we only consider the first two topic membership proportions for each character. Additionally,

since locations of topics are dependent on the configuration of characters in social space, θ_i and $\theta_{i'}$ for any pair of characters are also correlated. Thus, we have $N = 10$ characters for which we are considering $K - 1 = 2$ parameters each, leading to a total of $p = 20$ model parameters. Once a multi-ESS of 400 has been reached, we consider our MCMC samples to have reasonably converged to the true posterior distributions. Additionally, a burn-in period of 500 samples is taken.

2.4.2 Simulations Set 1

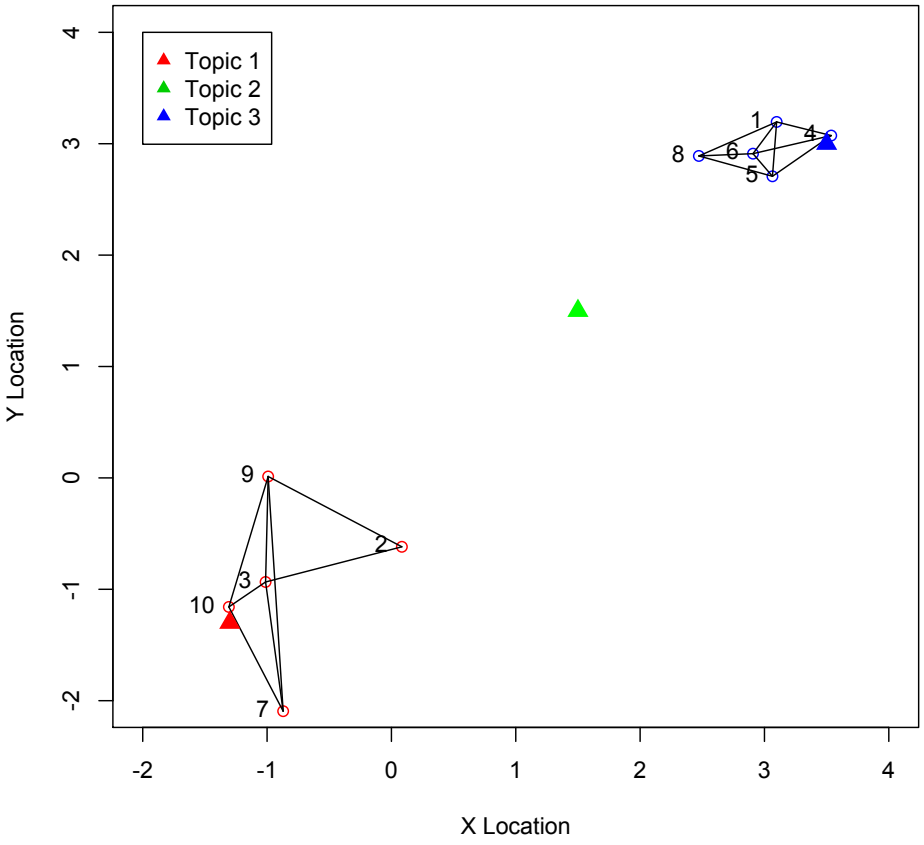
We first run 100 simulations with the $N = 10$ character locations coming from the mixture distribution with parameters

$$\pi = 0.5,\ \boldsymbol{\mu_1} = \begin{pmatrix} -1 \\ -1 \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix},\ \boldsymbol{\mu_2} = \begin{pmatrix} 3 \\ 3 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}.$$

Locations for $K = 3$ topics are placed at fixed points of $\mathbf{z_{t_1}} = (-1.3, -1.3)$, $\mathbf{z_{t_2}} = (1.5, 1.5)$, and $\mathbf{z_{t_3}} = (3.5, 3.0)$. Each character is close to one topic and relatively far from the others, and there is distinct separation between topics. One topic lies between the two clusters of characters and thus is the bridging topic that characters from both clusters are likely to discuss some of the time. While characters are well connected within their cluster, few, if any, connections exist between characters who are farther apart. A representation of one simulated social space under these conditions is presented in Figure 2.3.

For each of the 100 simulations, the sum of squared deviations for the estimates of θ, β, and $\mathbf{z}$ based on the independence model are calculated, as well as for the estimates of the same parameters based on our social network topic model. Under the conditions of this simulated scenario, the dependence model does relatively well in

Figure 2.3: Set 1: Simulated social space of $N = 10$ characters and $K = 3$ topics, with network connections



estimating the mixed-membership vectors and the location of characters, as compared to the independence model. In particular, we outperform the independence model 72% of the time in estimating θ and 100% of the time in estimating $\mathbf{z}$. For the word probability matrix, β, we perform better than the independence model in 13% of the simulations, with the independence model outperforming the dependence model in the remaining 87% of the simulations. Upon closer examination, we see that the sum of squared deviations are relatively small and quite close for both the independence and dependence models in terms of estimating β, indicating that the parameter is not

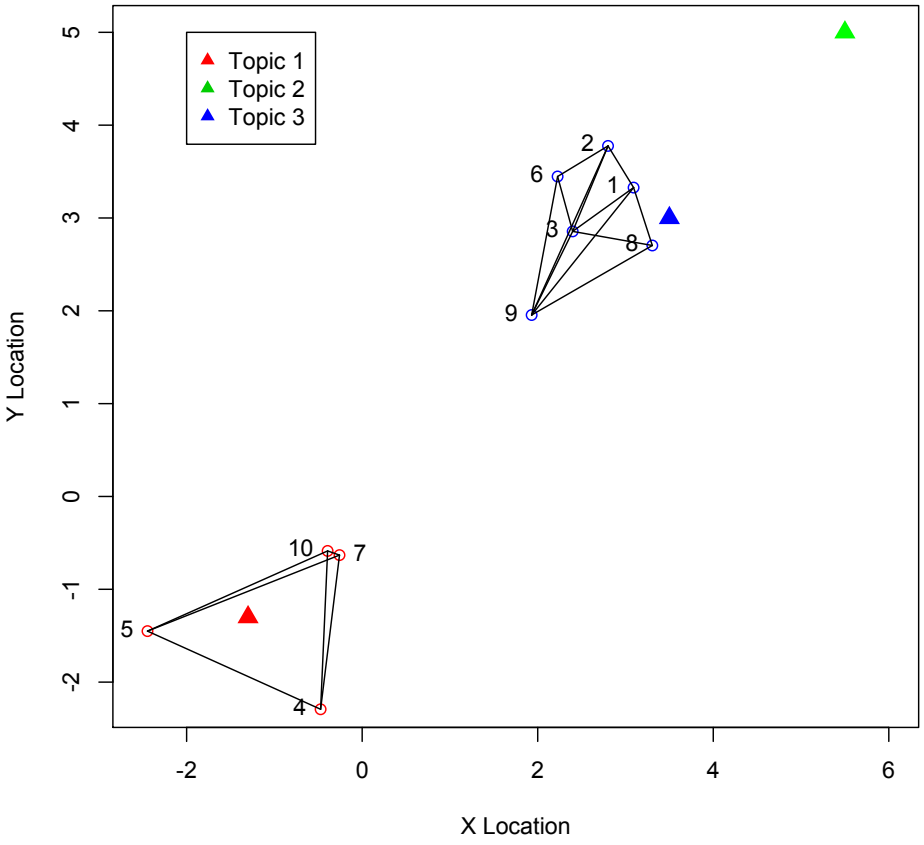
too poorly estimated in either case. With the maximal sum of squared deviations for β from the dependence model being 0.26, our performance in estimating β is fairly good. Overall, the dependence model outperforms the independence model in this situation, especially in terms of estimating θ and $\mathbf{z}$, which are the parameters we are most directly interested in.

2.4.3 Simulations Set 2

The social space is slightly modified to create a second set of simulation conditions for which we can assess model performance. In particular, the mixture distribution for the $N = 10$ simulated character locations remains unchanged from the original simulations, but the $K = 3$ topic locations are changed to $\mathbf{z_{t_1}} = (-1.3, -1.3)$, $\mathbf{z_{t_2}} = (5.5, 5.0)$, and $\mathbf{z_{t_3}} = (3.5, 3.0)$. A simulated social space under these conditions is presented in Figure 2.4. The topic that had originally been between the two clusters of characters is now moved in social space farther away from one cluster of characters and placed past the second cluster of characters. Thus, the second cluster of characters will have relatively high membership into this third topic, whereas the first cluster of characters is now quite far from the topic and will have low membership in it. Without being between the cluster of characters, estimating the location of this topic may be more difficult since less information is known about its position.

Under these conditions, the dependence model again performs quite well compared to the independence model. The dependence model outperforms the independence model in estimating θ in 83% of the simulations and $\mathbf{z}$ in 100% of the simulations. The performance of β is inferior in the dependence model, outperforming the independence model in only 8% of the simulations. The maximal sum of squared deviations in estimating β with the dependence model is 0.38, which again indicates that we are not

Figure 2.4: Set 2: Simulated social space of $N = 10$ characters and $K = 3$ topics, with network connections



doing too poorly in estimating the truth. Ultimately, the social network topic model again performs well in this scenario of increasing space between topics, particularly in estimating θ and $\mathbf{z}$.

2.4.4 Simulations Set 3

As a third set of simulated conditions, we contract the space of the original simulations, but still allow for the same variability in character locations. The mixture distribution for the $N = 10$ characters for this set of simulations has the following

parameter specifications:

$$\pi = 0.5,\ \boldsymbol{\mu_1} = \begin{pmatrix} -1 \\ -1 \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix},\ \boldsymbol{\mu_2} = \begin{pmatrix} 2 \\ 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}.$$

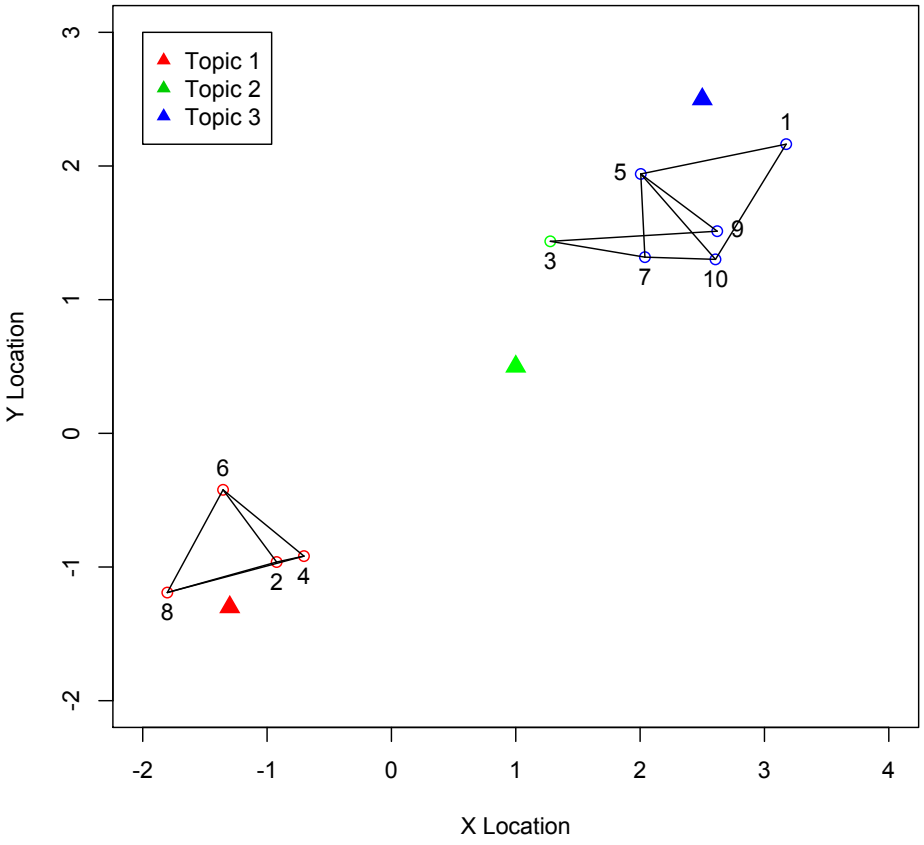
The change in specification of $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ brings characters closer together in space, and the variance components are kept unchanged from the original simulation conditions, leading to all characters being relatively closer together than in the previous simulations. Additionally, the locations of the $K = 3$ topics are moved closer together to $\mathbf{z_{t_1}} = (-1.3, -1.3)$, $\mathbf{z_{t_2}} = (1.0, 0.5)$, and $\mathbf{z_{t_3}} = (2.5, 2.5)$, so characters are likely to have higher membership in multiple topics. A simulated representation of this space is presented in Figure 2.5.

Under these simulated conditions, the dependence model generally outperforms the independence model. θ is better estimated based on the dependence model in 69% of the simulations, and $\mathbf{z}$ is better estimated 100% of the time through the dependence model compared to the independence model. Consistent with the previous two sets of simulation conditions, there is some difficulty in recovering better estimates of β with the dependence model. Under the simulation conditions specified here, the dependence model outperforms the independence model in 15% of the simulations, with the maximal sum of squared deviations for β being 0.18. These results for the specified conditions support the use of the dependence model as generally superior to the independence model in estimating the social space.

2.4.5 Simulations Under Independent Conditions

The final simulation condition is that of independence between the text topics and social network. In this case, the topics are unrelated to the social network and are

Figure 2.5: Set 3: Simulated social space of $N = 10$ characters and $K = 3$ topics, with network connections



not simulated to be in the same social space. Instead, the text documents are simulated according to the generative process described above for traditional LDA, and the social network is independently simulated using the latent approach introduced in Chapter 1. The dependence model should not perform too well, as it is assuming a relationship between the text and network that does not actually exist. When text topics are unrelated to the social network, the dependence model is outperformed by the independence model in estimating θ in 85% and $\mathbf{z}$ in 92% of the simulations. Thus, the dependence model does a relatively poor job at estimating the sample

space since it assumes relationships that are not present. In terms of estimating β, our performance is not much different than in the previous simulation conditions. The dependence model outperforms the independence model in only 10% of the simulations, and our maximal sum of squared deviations from the true value of β is 0.286 under the dependence model. Again, even under this scenario, the social network topic model is not estimating β very poorly.

2.4.6 Model Selection: Appropriateness of Social Network Topic Model

The above simulations demonstrate the ability of the social network topic model to properly estimate model parameters when there is in fact a relationship between the social network and text topics. In the case of independence between text topics and character relationships, the social network topic model should not be applied. Thus, a model selection technique must be in place to properly determine when the social network topic model is appropriate.

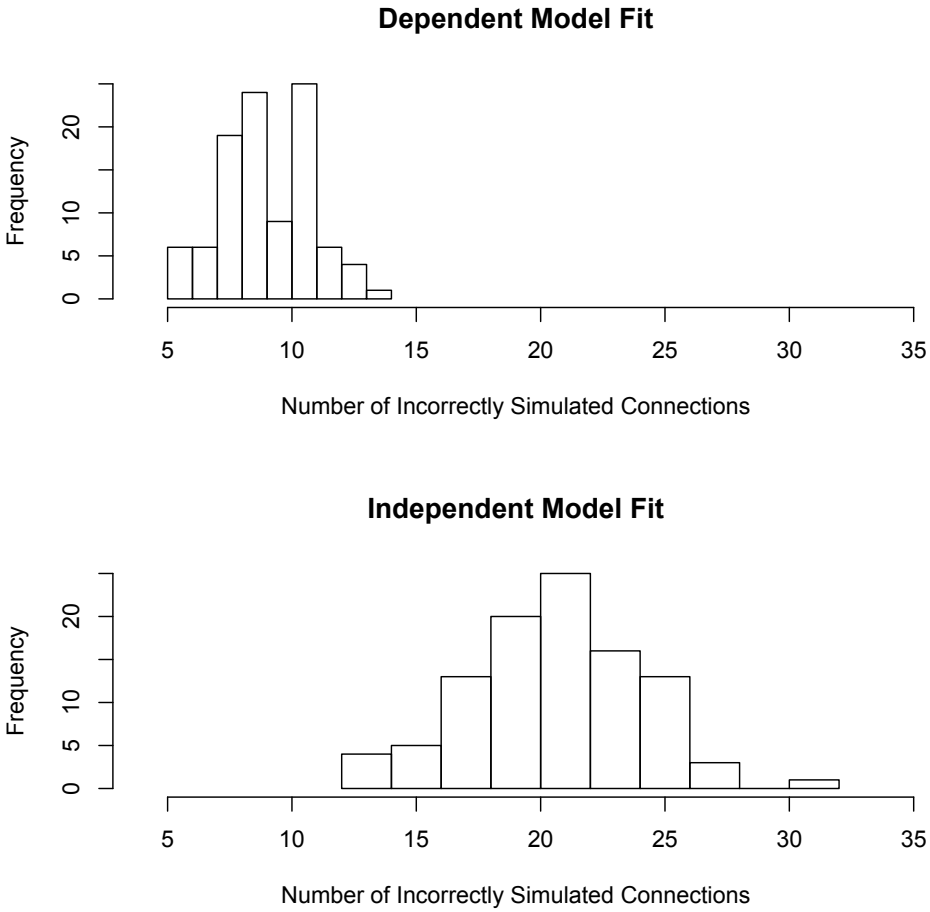
To determine whether the social network topic model is the proper model to apply to a given dataset, we rely upon a predictive distribution criterion. In particular, for a given set of data, we fit both a social network topic model and independent topic and social network models. For each of these dependence and independence models, we use the posterior means of the model parameters to simulate a large number of adjacency matrices (we recommend simulating 100 such adjacency matrices). Then, we determine how similar these simulated adjacency matrices are to the true adjacency matrix used in fitting the models. For assessing similarity, we recommend taking the sum of the absolute differences in the adjacency matrices and dividing by two in order to determine the number of incorrectly simulated connections (recall that the

adjacency matrices discussed here are symmetric, since connections are taken to be undirected, and thus we need to divide the total number of discrepancies in the adjacency matrices by two in order to get the total number of wrongly predicted connections).

If the data are truly arising from a dependent process where the social network is linked to the text topics, then the social network topic model should have fewer incorrectly simulated connections. As an example, see Figure 2.6, in which both the social network topic model and independent topic and social network models are fit to data that were originally simulated under a dependence model, and the predictive error is calculated for 100 simulations. Here we can see that there is almost no overlap between the distributions of the number of incorrectly simulated connections for the dependence model and the independence model, with the former distribution being on a lower range than the latter distribution. In particular, the 95^{th} percentile for the dependence distribution of incorrectly simulated connections is below the 5^{th} percentile for independence distribution. In this case, we determine that the social network topic model is appropriate.

Alternatively, if the data are truly arising from an independent process in which the social network is not related to the text topics, then the social network topic model is not expected to have fewer incorrectly simulated connections. To examine this situation, both the social network topic model and independent topic and social network models are fit to data originally simulated under an independence model, and the predictive error is estimated based on 100 simulations. The resulting predictive error distributions are presented in Figure 2.7. There is significant overlap in the distributions of the number of incorrectly simulated connections for the two candidate models. In particular, the 95^{th} percentile for the dependence distribution of incorrectly simulated connections is above the 5^{th} percentile for independence distribution. When

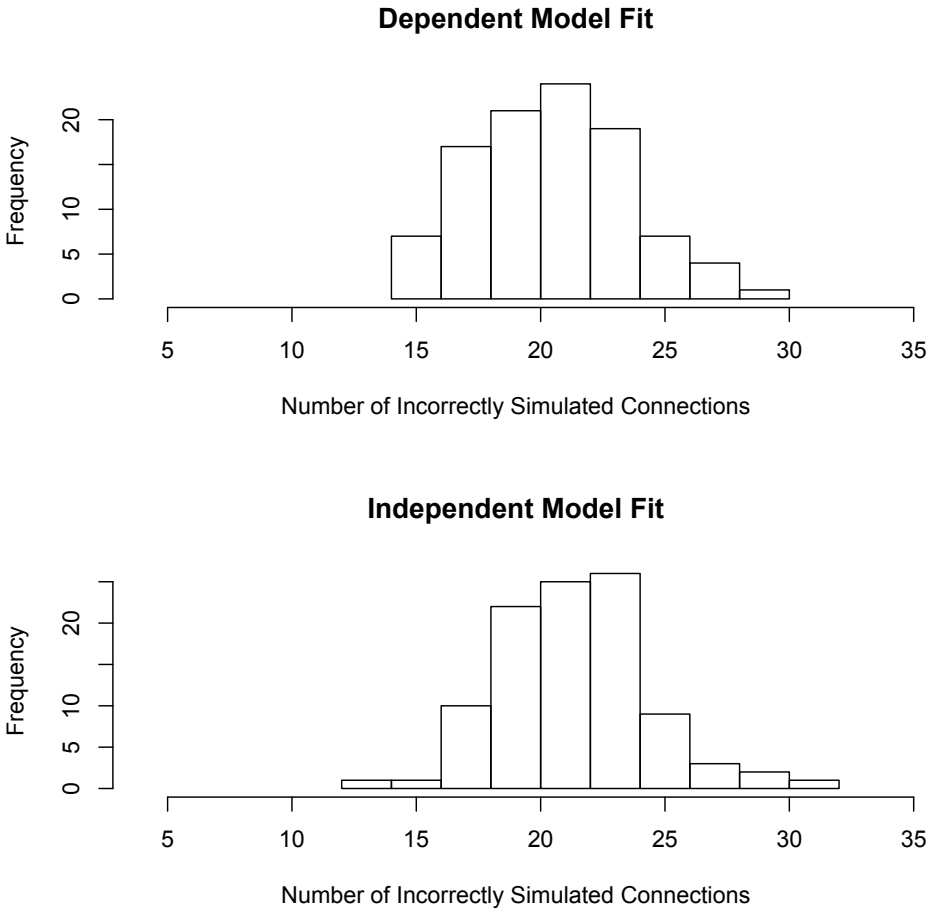
Figure 2.6: Model Selection: Data simulated under dependence model



there is significant overlap between these distributions, the simpler independence model is more appropriate to use over the social network topic model.

We performed a series of simulations to evaluate the performance of this model selection criterion. In particular, we simulated 100 datasets arising from the dependence model and 100 datasets under the independence model. For each set of simulations, we fit both the dependence and independence models, calculating our predictive distribution criterion based on the number of incorrect connections for each simulation. We examined the 95^{th} percentile of the distribution of incorrect social connections when fitting a dependence model and the 5^{th} percentile of the distribution of incorrect social

Figure 2.7: Model Selection: Data simulated under independence model



connections when fitting an independence model. Out of the 100 simulations arising from a dependence model, our percentile criterion selected the dependence model 93% of the time, indicating that it does well at properly choosing the dependence model. Similarly, out of the 100 simulations arising from an independence model, our criterion correctly selected the independence model 94% of the time. Overall, this posterior predictive criterion provides a useful mechanism for assessing model fit.

2.5 Discussion

In this chapter, we introduced a method for social networks within the framework of topic models. Latent Dirichlet Allocation was extended to include social network data, offering a unified model that incorporates both text data and social network interactions. Ultimately, text topics and characters were placed together in the same social space, which provided for a spatial representation of how characters are related to one another and to the different topics discussed in a text corpus.

The social network topic model performs well when a relationship does exist between characters and text topics. In particular, when we have situations of characters with strong social connections discussing similar topics in the documents they author, we are able to better estimate the strength of their relationships and represent their resulting network, as well as get a better understanding of the topics that each character discusses in a text document.

While this model marks the first comprehensive unification of topic modeling and social network analyses, there remains work for improvement. Currently the implementation of the social network topic model is quite computationally intensive. While JAGS is a useful statistical software for the model, we aim to make this method more computationally accessible moving forward by introducing a variational Bayes approach for posterior inference. Additionally, there is an issue of data collection (Masala, Servetti, Basso, and De Martin, 2014). Social network data has historically been difficult to obtain, and placing the restriction that the social network also be linked to text documents further limits the data available for use. While social media companies, for example, have a wealth of such data, gaining access to it is often difficult.

Additionally, in our current model specification, β is a matrix of random row

vectors. For each row of β, we do not incorporate any information *a priori* of how words are related to each other. There exists much information, though, regarding word relationships that can potentially be incorporated into this model. From a basic understanding of the English language, it is known that certain words often appear together and that they are used when discussing certain topics. If such information is incorporated, estimation of β will be better informed. In practice, though, adding such selective information to the model is challenging.

Many extensions to the social network topic model are possible. In particular, we can extend the model to scenarios of multiple characters authoring a document, as well as generalizing it to other HBMMMs that are linked to social network information. Extensions and generalizations to the social network topic model are the subject of the next chapter.

Chapter 3

Social Network Topic Model Extensions and Generalizations

3.1 Introduction

In the preceding chapter, we introduced a topic modeling approach that incorporates social network data. Our method unifies social network and text analyses into a cohesive model, allowing for a sharing of information between the network and text. Thus, we are able to better cluster documents using social network data, as well as use topic membership to inform the modeling of the social network. With this approach, we position characters and text topics in a single social space. The distances between a character and the topics in this social space are indicative of the degree of membership of that character's text documents in each of the topics.

Chapter 2 introduced Latent Dirichlet Allocation (LDA), which was used as the basis of the topic modeling aspect of our social network topic model (Blei, Ng, and Jordan, 2003). We assumed that N inter-related characters formed a social network. Each of these $i = 1, ..., N$ characters authored d_i documents that expressed a mixture

of K topics. Every character was assumed to always be discussing the same mixture of topics over their d_i documents and to be the sole author of their set of documents. Documents for each character were combined, as we considered all text written by an author to be a single document.

The Markov chain Monte Carlo (MCMC) inferential procedure introduced in Chapter 2 for the social network topic model provides one method for approximating the posterior distributions of model parameters. As previously discussed when introducing inference for the social network topic model, the MCMC approach is computationally intensive, rendering the method infeasible as the dataset size grows larger. In this chapter, we thus turn our attention to an alternative method for posterior inference that relieves computational burden.

Additionally, this chapter focuses on extensions and generalizations to the social network topic model. In its original form, the social network topic model assumes that there is one author per text document. Such a setup is realistic for many types of network text data. For example, on social media websites such as Twitter, each user solely authors Tweets, which can be considered to discuss the author's mixture of topics. In many applications, though, we do not want to restrict characters to solely authoring documents; instead, we want to allow for the flexibility of having characters co-author documents together. Consider, as an example, a network of academics who have social connections if they worked in the same department at some point in their careers. Each character authors journal articles, some of which they solely author and others of which they jointly author with other characters. In this chapter, we consider this situation of multiple characters being linked to each document.

We begin this chapter by discussing a variational expectation-maximization algorithm as a less computationally-intensive alternative to the MCMC approach of Chapter 2 for performing posterior inference. Then, we extend the social network

topic model introduced in Chapter 2 to include multiple authors of documents. Finally, we generalize our social network topic model to the case of discrete data in the form of survey responses. Focusing on the Grade of Membership model as an alternative mixed-membership specification, we incorporate social network information into the scenario of membership profiles based on binary response data.

3.2 Variational Inference for the Social Network Topic Model

When considering the social network topic model, the likelihood $P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)$ is intractable for exact inference. As discussed in Chapter 2, MCMC approaches to performing inference are quite computationally intensive. To alleviate the computational burden of the MCMC approach, we introduce here a variational Bayes approach for approximate inference. This variational expectation-maximization (EM) algorithm is based on Jensen's inequality to obtain an adjustable lower bound for the log-likelihood (Blei et al., 2003). An approximate log-likelihood function, which is characterized by *variational parameters*, is used in place of $P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)$ and is often taken to be of simpler form (e.g. independence between parameters). The resulting variational joint posterior distribution is also of simpler form, leading to more convenient posterior inference through this approximate distribution.

Our goal is to make the variational posterior distribution similar to the true posterior distribution while also reducing its complexity. In the variational EM algorithm, we iterate between updating the expected variational parameter values—achieved through a mean-field approximation where variational parameters are optimized to minimize the Kullback-Leibler divergence between the true and variational posterior

distributions—and maximizing the true model hyper-parameters with respect to the lower bound on the log-likelihood. This process renders the variational posterior distribution to be a close approximation to the true posterior distribution. As a result of the variational EM algorithm, we estimate values of the model hyper-parameters $(\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)$ and get a joint posterior distribution for $(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, b)$, based on the variational parameters.

To further discuss the variational EM method, we now introduce some model choices that will be continued throughout this section. We begin by modifying step 6 of the social network topic model presented on pages 21-22 of Chapter 2. To make the model more manageable, we remove the $h_{ii'}$ term and re-express the probability of a connection between two characters as $p_{adj_{ii'}} = \frac{e^{a(1-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{a(1-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}$, matching the model without additional covariate information presented by Hoff, Raftery, and Handcock (2002). We assume bivariate normal prior distributions for each $\mathbf{z_{t_k}}$ and $\mathbf{z_i}$ as $N(\boldsymbol{\eta_1}, \eta_2^2 \times I_2)$ and $N(\boldsymbol{\nu_1}, \nu_2^2 \times I_2)$, respectively, a symmetric Dirichlet prior distribution for β as Dirichlet(ϕ), and a uniform distribution on a over the range $[-\omega, \omega]$. Note that both $\boldsymbol{\eta_1}$ and $\boldsymbol{\nu_1}$ are two-dimensional parameters in order to match the dimension of the social space, I_2 is the 2×2 identity matrix, and ϕ is a scalar to be repeated V times. Thus, our set of model hyper-parameters is $(\boldsymbol{\eta_1}, \eta_2, \boldsymbol{\nu_1}, \nu_2, \phi, \omega)$.

Recall that we are ultimately attempting to determine the joint posterior distribution $P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \mathbf{x}, Y, \boldsymbol{\eta_1}, \eta_2, \boldsymbol{\nu_1}, \nu_2, \phi, \omega)$, which we approximate by a simpler variational distribution. In formulating the variational likelihood, we assume that model parameters $(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a)$ are independent, thus alleviating the coupling of parameters in the likelihood that causes slow mixing, as discussed in Chapter 2. Based on this

information, we formulate our variational joint posterior distribution as follows:

$$
\begin{aligned}
&q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a|\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}) \qquad (3.1)\\
&\quad = q(a|\tilde{\omega}) \left(\prod_{i=1}^{N} \left(q(\mathbf{z_i}|\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}, \tilde{\nu}_{2i}) \prod_{r=1}^{R_i} q(u_i^r|\tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK}) \right) \right)\\
&\qquad \times \left(\prod_{k=1}^{K} q(\mathbf{z_{t_k}}|\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}, \tilde{\eta}_{2k}) q(\beta_k|\tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV}) \right).
\end{aligned}
$$

In this variational posterior distribution, we take $q(a|\tilde{\omega})$ to be a uniform distribution over the range $[-\tilde{\omega}, \tilde{\omega}]$, $q(\mathbf{z_i}|\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}, \tilde{\nu}_{2i})$ to be a bivariate normal distribution centered at $\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}$ with variance $\tilde{\nu}_{2i}^2 \times I_2$, $q(u_i^r|\tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK})$ to be a multinomial distribution with parameters $\tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK}$, $q(\mathbf{z_{t_k}}|\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}, \tilde{\eta}_{2k})$ to be a bivariate normal distribution centered at $\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}$ with variance $\tilde{\eta}_{2k}^2 \times I_2$, and $q(\beta_k|\tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV})$ to be a Dirichlet distribution with parameters $\tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV}$.

Given this variational joint posterior distribution, we must optimize the variational parameters $(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega})$ to minimize the Kullback-Leibler divergence between the true and variational posterior distributions:

$$
\begin{aligned}
&(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}{}^*, \tilde{\eta}_2{}^*, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}{}^*, \tilde{\nu}_2{}^*, \tilde{\phi}^*, \tilde{\tau}^*, \tilde{\omega}^*) \qquad (3.2)\\
&= \underset{(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega})}{\arg\min} D\left(q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a|\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}) || P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a|\mathbf{x}, Y, \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega) \right).
\end{aligned}
$$

The E-step of the variational EM algorithm consists of finding the resulting estimates of the variational parameters. To do this, we determine an expression for the Kullback-Leibler divergence, take its derivative with respect to each variational parameters, and set each of these derivatives equal to zero. We then have update equations for the

variational parameters, given as follows (see the Appendix for derivations):

$$\text{Solve for } \tilde{\boldsymbol{\eta}}_{\mathbf{1k}}: \quad \frac{2(\boldsymbol{\eta}_{\mathbf{1}}-\tilde{\boldsymbol{\eta}}_{\mathbf{1k}})}{\eta_2^2}-\frac{1}{2}\sum_{i=1}^{N}\sum_{r=1}^{R_i}\Bigg[\tilde{\tau}_{irk}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(\frac{\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}}{||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||}\right)\left(1-\frac{e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2)\right)^{\frac{1}{2}}}}{\sum_{c=1}^{K}e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}}\right)\Bigg]=0$$

$$\text{Solve for } \tilde{\eta}_{2k}: \quad \frac{2}{\tilde{\eta}_{2k}}-\frac{2\tilde{\eta}_{2k}}{\eta_2^2}-2\sum_{i=1}^{N}\sum_{r=1}^{R_i}\Bigg[\tilde{\tau}_{irk}\tilde{\eta}_{2k}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(1-\frac{e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}}{\sum_{c=1}^{K}e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}}}\right)\Bigg]=0$$

$$\text{Solve for } \tilde{\boldsymbol{\nu}}_{\mathbf{1i}}: \quad \frac{2\left(\boldsymbol{\nu}_{\mathbf{1}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}\right)}{\nu_2^2}-\frac{1}{2}\sum_{r=1}^{R_i}\sum_{k=1}^{K}\Bigg[\tilde{\tau}_{irk}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(\frac{\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}}{||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||}\right)-\frac{\tilde{\tau}_{irk}}{\sum_{c=1}^{K}e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}}$$
$$\times\sum_{c=1}^{K}\Bigg(e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(\frac{\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}}{||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||}\right)\Bigg)\Bigg]-\frac{1}{2}\sum_{i'\neq i}\Bigg[\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(\frac{\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}}{||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||}\right)\left(y_{ii'}\tilde{\omega}-\frac{e^{\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}\right)}}{1+e^{\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}\right)}}\right)\Bigg]=0$$

$$\text{Solve for } \tilde{\nu}_{2i}: \quad \frac{2}{\tilde{\nu}_{2i}}-\frac{2\tilde{\nu}_{2i}}{\nu_2^2}-2\sum_{r=1}^{R_i}\sum_{k=1}^{K}\Bigg[\tilde{\tau}_{irk}\tilde{\nu}_{2i}\Bigg(\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}$$
$$-\frac{\sum_{c=1}^{K}\left(e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}\right)}{\sum_{c=1}^{K}e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{-\frac{1}{2}}}}\Bigg)\Bigg]$$
$$-2\sum_{i'\neq i}\Bigg[\tilde{\nu}_{2i}\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{-\frac{1}{2}}$$
$$\times\left(y_{ii'}\tilde{\omega}+\frac{e^{\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{-\frac{1}{2}}\right)}}{1+e^{\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1i'}}||+2\left(\tilde{\nu}_{2i}^2+\tilde{\nu}_{2i'}^2\right)\right)^{-\frac{1}{2}}\right)}}\right)\Bigg]=0.$$

$$\text{Solve for } \tilde{\omega}: \quad \frac{1}{\tilde{\omega}} + \sum_{i=1}^{N}\sum_{i'=i+1}^{N}\left[y_{ii'}\left(1 - \left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}} - \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i'}}|| + 2\left(\tilde{\nu}_{2i}^2 + \tilde{\nu}_{2i'}^2\right)\right)^{-\frac{1}{2}}\right) - \frac{1 - \left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}} - \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i'}}|| + 2\left(\tilde{\nu}_{2i}^2 + \tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}}{1 + e^{\tilde{\omega}\left(1 - \left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}} - \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i'}}|| + 2\left(\tilde{\nu}_{2i}^2 + \tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}\right)}}\right] = 0$$

$$\tilde{\phi}_{kv} = \phi + \sum_{i=1}^{N}\sum_{r=1}^{R_i}\tilde{\tau}_{irk}x_{i[v]}^{r}$$

$$\tilde{\tau}_{irk} = e^{-1-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}} - \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}|| + 2\left(\tilde{\eta}_{2k}^2 + \tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}} + \sum_{v=1}^{V} x_{i[v]}^{r}\left(\Psi(\tilde{\phi}_{kv}) - \Psi(\sum_{j=1}^{V}\tilde{\phi}_{kj})\right)} - \sum_{c=1}^{K} e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{c}} - \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}|| + 2\left(\tilde{\eta}_{2c}^2 + \tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}$$

In this specification, $\Psi(.)$ is the digamma function. When optimizing each variational parameter, we use the most recent estimates of the model parameters and other variational parameters.

The variational posterior distribution is conditional on the observed data, $(\mathbf{x}, Y)$, so we can explicitly express the optimized variational parameters as being functions of the data: $q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a|\tilde{\boldsymbol{\eta}}_{\mathbf{1}}{}^{*}(\mathbf{x}), \tilde{\eta}_2{}^{*}(\mathbf{x}), \tilde{\boldsymbol{\nu}}_{\mathbf{1}}{}^{*}(\mathbf{x}, Y), \tilde{\nu_2}{}^{*}(\mathbf{x}, Y), \tilde{\phi}^{*}(\mathbf{x}), \tilde{\tau}^{*}(\mathbf{x}), \tilde{\omega}^{*}(Y))$. Equation (3.2) is optimized based on the given data, and thus the resulting optimized variational parameters $(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}{}^{*}, \tilde{\eta}_2{}^{*}, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}{}^{*}, \tilde{\nu_2}{}^{*}, \tilde{\phi}^{*}, \tilde{\tau}^{*}, \tilde{\omega}^{*})$ are functions of $\mathbf{x}$ and Y. Because of this conditioning on the observed data, the optimization of variational parameters is character specific, depending both on the words used and social connections.

The procedure outlined thus far represents the E-step of the variational EM algorithm. In this step, we updated the variational parameters using a mean-field approximation of the posterior distribution—that is, using the simpler variational posterior distribution and optimizing the variational parameters to minimize the Kullback-Leibler divergence—given the most recent estimates of the model hyper-parameters. We now turn our attention to the M-step, in which we maximize the lower bound

for the log-likelihood over the model hyper-parameters, based on the most recent estimates of variational parameters $(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}{}^*, \tilde{\eta_2}^*, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}{}^*, \tilde{\nu_2}^*, \tilde{\phi}^*, \tilde{\tau}^*, \tilde{\omega}^*)$, and thus, in turn, update $(\boldsymbol{\eta_1}, \eta_2, \boldsymbol{\nu_1}, \nu_2, \phi, \omega)$. In doing so, we are finding the pseudo maximum likelihood estimates based on the expected sufficient statistics of the variational posterior distribution, computed in the E-step.

As shown in the Appendix, the pseudo maximum likelihood estimates for the model parameters are given as follows:

$$\boldsymbol{\eta_1} = \frac{1}{K}\sum_{k=1}^{K} \tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}$$

$$\eta_2 = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\tilde{\eta}_{2k}^2 + \left(\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}} - \boldsymbol{\eta_1}\right)^T\left(\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}} - \boldsymbol{\eta_1}\right)\right)}$$

$$\boldsymbol{\nu_1} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}$$

$$\nu_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\nu}_{2i}^2 + \left(\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}} - \boldsymbol{\nu_1}\right)^T\left(\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}} - \boldsymbol{\nu_1}\right)\right)}$$

$$\omega = \infty$$

$$\text{Solve for } \phi: \quad KV\left(\Psi(V\phi) - \Psi(\phi)\right) + \sum_{k=1}^{K}\sum_{v=1}^{V}\left(\Psi(\tilde{\phi}_{kv}) - \Psi\left(\sum_{j=1}^{V}\tilde{\phi}_{kj}\right)\right) = 0.$$

In determining the values of each of these pseudo maximum likelihood estimates, we use the most recent values of the variational parameters and other model hyper-parameters. Note that the estimate of $\omega = \infty$ implies that the prior distribution for a is estimated to be uninformative. Since ω is not used to update any variational parameters, this estimate has no direct effect on the variational posterior distributions.

Unlike the optimized variational parameters, the estimates of the model hyper-parameters are not character specific. Iterating between optimizing the variational parameters and maximizing the model hyper-parameters leads to the inferential re-

sults that are of interest. In particular, this variational EM algorithm gives us point estimates of the model hyper-parameters and also provides an approximate joint posterior distribution for $(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a)$, which is the primary inferential goal for the social network topic model. Since the variational posterior distributions are being conditioned on point estimates of the hyper-parameters, the variability in the posterior distributions will be understated as compared to a fully Bayesian approach. This variational EM algorithm has been introduced for a simplified version of the social network topic model, but it can be extended to the case of allowing for covariate information $h_{ii'}$ in modeling social network connections. Additionally, this inferential procedure can be modified to fit the extensions presented in the rest of this chapter (i.e. the co-authorship social network topic model and the social network Grade of Membership model).

3.3 Model Extensions to Allow for Multiple Authors of a Document

Extending the social network topic model to allow for multiple authors of each document is natural, given that many text documents do in fact have multiple authors. The setup of the modified social network topic model remains similar to that of Chapter 2. We have a network of N characters and K topics situated together in social space. Character i authors d_i documents that express a mixture of the K topics. We define θ_i as the membership of character i into each of the K topics. This membership is determined based on the distance between characters and topics in social space, given specifically as $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$, where $||\mathbf{z_{t_k}} - \mathbf{z_i}||$ represents the Euclidean distance between topic k and character i. We extend the social network topic model by

now allowing for documents to have multiple authors. Of character i's d_i documents, some (if not all) of them may be co-authored with other characters. A character can still be the sole author of some documents, but in general we allow for situations of co-authorship. In this scenario, we have a total of M documents in our corpus.

3.3.1 Model Specification

The co-authorship model shares similarities to the original social network topic model. In particular, steps 1, 2, 3, and 4 of the co-authorship model can all be found in the original social network topic model presented on pages 21-22. Variations from the original social network topic model are **bolded** in the model specification below. Consider the network described above of N characters who each author d_i documents **for a total of M documents. The generative process for this co-authorship social network topic model is given as follows:**

1. Sample N character locations, $\mathbf{z} = (\mathbf{z_1}, ..., \mathbf{z_N})$, in social space from $P(\mathbf{z})$.

2. Sample K topic locations, $\mathbf{z_t} = (\mathbf{z_{t_1}}, ..., \mathbf{z_{t_K}})$, in social space from $P(\mathbf{z_t})$.

3. For each of K topics, sample the length V vocabulary probability vector $\beta_k \sim$ Dirichlet($\boldsymbol{\phi}$).

4. For each of N characters, calculate the topic membership vector, θ_i, based on the character's distance from each topic: $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$.

5. **For each of the M documents and R_m words in the document:**

   (a) **Sample a latent topic indicator $u_m^r|\theta_m^* \sim$ Multinomial($\theta_m^*, 1$), where $\theta_m^* = \frac{1}{n_m} \sum_{i \in A_m} \theta_i$, A_m is the set of authors, and n_m is the number of authors for document m.**

(b) Sample a word $x_m^r|u_m^r, \beta \sim \text{Multinomial}(\beta^T u_m^r, 1)$.

6. **Sample social network parameters a, b, and c from $P(a)$, $P(b)$, and $P(c)$, respectively.**

7. Sample the edges $Y \sim \text{Bernoulli}(\mathbf{p_{adj}})$, where $p_{adj_{ii'}} = \frac{e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'}}}{1+e^{a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'}}}$, $h_{ii'} = \frac{1}{2}\sum_{v=1}^{V}\left(\theta_i\beta_{[.v]}log\left(\frac{\theta_i\beta_{[.v]}}{\theta_{i'}\beta_{[.v]}}\right) + \theta_{i'}\beta_{[.v]}log\left(\frac{\theta_{i'}\beta_{[.v]}}{\theta_i\beta_{[.v]}}\right)\right)$, and $s_{ii'}$ **is an indicator for whether the pair jointly author a document.**

In this co-authorship model, each of the M documents is individually considered instead of being combined for each character as in the single authorship model. Furthermore, in the original social network topic model, we sample a latent topic indicator for each document based on a single character's mixed-membership vector, whereas here we instead use an average mixed-membership vector for all characters who contribute to that document. For example, if three characters co-author a document, we take the mean of their mixed-membership vectors to get the overall topic distribution for that document. Under this specification, if a character is an author on a document, then that document will discuss some topic that is of interest to the character. Thus, by averaging over all co-authoring characters' mixed-membership vectors, we get the topic membership for what is likely to be discussed in the document. Taking a simple average supposes that each character contributes equally to the document's topic distribution. Additionally, this model differs from the original social network topic model in that probabilities of network connections can now depend on whether two characters co-author a document. In many real networks, if two characters co-author a document, they have a higher probability of a network connection. The co-authorship social network topic model is thus distinguished from the original social network topic model in that there is an additional source of observed data: the co-authorship connections, denoted as S.

3.3.2 Inference

Inferential methods for this model remain comparable to those of Chapter 2. We let $\mathbf{x}$ be the collection of words used in the documents, Y be an adjacency matrix of social network connections, and S be a co-authorship matrix, with entry $s_{ii'}$ indicating whether character i and character i' co-author a document. In this co-authorship model, the likelihood can then be specified as follows:

$$P(\mathbf{x}, Y, S|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta, \zeta) \propto \tag{3.3}$$
$$\int\int\int\int\int\int \sum_{\mathbf{u}} \left(\prod_{m=1}^{M} \prod_{r=1}^{R_m} \prod_{k=1}^{K} \left(\frac{1}{n_m} \sum_{i \in A_m} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \right) \prod_{v=1}^{V} \beta_{k[v]}^{x^r_{m[v]}} \right)^{u^r_{m[k]}} \right)$$
$$\times \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}} D_{\boldsymbol{\eta}}(d\mathbf{z_t}) D_{\boldsymbol{\nu}}(d\mathbf{z}) D_\phi(d\beta) D_\omega(da) D_\delta(db) D_\zeta(dc).$$

Note that this likelihood is also intractable for analytic inference due to the coupling of $\mathbf{z_t}$, $\mathbf{z}$, and β. The full conditional posterior distributions for this co-authorship social network topic model are:

$$P(\beta|\mathbf{x}, Y, S, \mathbf{z}, \mathbf{z_t}, \mathbf{u}, a, b, c, \phi) \propto$$
$$\left[\prod_{v=1}^{V} \prod_{k=1}^{K} \beta_{k[v]}^{\phi + (\sum_{m=1}^{M} \sum_{r=1}^{R_m} u^r_{m[k]} x^r_{m[v]}) - 1} \right] \left[\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}} \right]$$

$$P(u^r_m|\mathbf{x}, Y, S, \mathbf{z}, \mathbf{z_t}, \beta, a, b, c) \propto \prod_{k=1}^{K} \left[\left(\frac{1}{n_m} \sum_{i \in A_m} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \right) \right) \left(\prod_{v=1}^{V} \beta_{k[v]}^{x^r_{m[v]}} \right) \right]^{u^r_{m[k]}}$$

$$P(\mathbf{z_t}|\mathbf{x}, Y, S, \mathbf{z}, \beta, \mathbf{u}, a, b, c, \boldsymbol{\eta}) \propto$$
$$P(\mathbf{z_t}|\boldsymbol{\eta}) \left[\prod_{m=1}^{M} \prod_{r=1}^{R_m} \prod_{k=1}^{K} \left(\frac{1}{n_m} \sum_{i \in A_m} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \right) \right)^{u^r_{m[k]}} \right]$$
$$\times \left[\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||+cs_{ii'})}} \right]$$

$$P(\mathbf{z}|\mathbf{x}, Y, S, \mathbf{z_t}, \beta, \mathbf{u}, a, b, c, \boldsymbol{\nu}) \propto$$
$$P(\mathbf{z}|\boldsymbol{\nu}) \left[\prod_{m=1}^{M} \prod_{r=1}^{R_m} \prod_{k=1}^{K} \left(\frac{1}{n_m} \sum_{i \in A_m} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \right) \right)^{u^r_{m[k]}} \right]$$
$$\times \left[\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}} \right]$$

$$P(a|\mathbf{x}, Y, S, \mathbf{z}, \mathbf{z_t}, \beta, b, c, \omega) \propto P(a|\omega) \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}$$

$$P(b|\mathbf{x}, Y, S, \mathbf{z}, \mathbf{z_t}, \beta, a, c, \delta) \propto P(b|\delta) \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}$$

$$P(c|\mathbf{x}, Y, S, \mathbf{z}, \mathbf{z_t}, \beta, a, b, \zeta) \propto P(c|\zeta) \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}|| + cs_{ii'})}}.$$

Note that $h_{ii'}$ is a function of β, $\mathbf{z_t}$, and $\mathbf{z}$ in the above distributions. Following the original social network topic model, we set a symmetric Dirichlet prior distribution for β with parameter ϕ. Additionally, we recommend setting a bivariate normal prior distribution for each $\mathbf{z_{t_k}}$ and $\mathbf{z_i}$ and fairly uninformative uniform prior distributions for a, b, and c. Slice sampling can be implemented to sample from these conditional posterior distributions via Just Another Gibbs Sampler (JAGS) (Plummer, 2013). The MCMC sampler should draw from the posterior distributions until convergence is met based on the same multi-ESS criterion described in Chapter 2 (Flegal, Hughes, and Vats, 2016). This approach is quite computationally intensive, and fitting medium-to-large sized datasets often becomes infeasible via the introduced MCMC methods. The variational inference approach introduced in Section 3.2 can be extended to the co-authorship social network topic model, providing an alternative method for inference.

3.4 Extension to the Grade of Membership Model

Recall that in Chapter 2 we introduced LDA as a topic model that meets the specifications of a hierarchical Bayesian mixed-membership model (HBMMM). Although we focused our attention to text documents and LDA, our model can be generalized to consider any HBMMM that is linked to social network data. By altering Assumptions 1-4 presented on pages 13-14 of Chapter 2, we can specify different HBMMMs that will be appropriate for different types of data (Airoldi, Fienberg, Joutard, and Love, 2006). The general framework of our model can be extended to any HBMMM so long as a social network fits within its framework.

We introduce here the Grade of Membership (GoM) model as another specification of an HBMMM. The GoM model has frequent applicability, as it is appropriate for discrete response data. Erosheva (2002; 2007) considered the GoM model as an HBMMM for binary response data over J characteristics, and we follow along in our development of the GoM model using a similar characterization. In particular, θ_i is the mixed-membership vector that describes how character i mixes between different *profiles*. We use the disability data from the National Long Term Care Survey (NLTCS) as our motivating example to introduce the social network GoM model. Individuals who participated in the NLTCS were classified as either being disabled or healthy for $J = 16$ activities, some of which are physical and others cognitive (Erosheva, Fienberg, and Joutard, 2007).

In this context, individuals were measured for a binary response (healthy/disabled) on whether they could complete each of J activities. Using this information, we can determine a number of K disability profiles. All individuals are then mixtures of these different profiles. Suppose there are $K = 4$ profiles. Based on the healthy/disabled responses for individuals with high membership in each profile, we may perhaps label

profile 1 as "physically and cognitively healthy" (all responses are "healthy"), profile 2 as "physical healthy, cognitively impaired" (all responses to physical questions are "healthy," but all cognitive responses are "disabled"), profile 3 as "physically disabled, cognitively healthy" (all responses to physical questions are "disabled," but all cognitive responses are "healthy") and profile 4 as "physically and cognitively disabled" (all responses are "disabled"). Individuals with high membership in profile 3, for example, are likely to have full cognitive functionality but are unable to independently complete physical activities, such as lifting heavy laundry baskets.

If we also have information relating to the social connections between individuals included in the survey, we could potentially situate the relationships of characters within the context of disability profiles. Thus, individuals who are socially close together would be likely, by assumption, to have similar characteristics. Intuitively, this is a reasonable assumption. For example, we expect people who are physically fit to be socially close since they may often play golf together. Similarly, for individuals with physical disabilities who have no cognitive impairments, we would expect them to be socially close since they are likely to stay inside and play board games, for example, since this requires little physical ability. With this setup, we generalize the social network topic model to allow for network data within the GoM framework.

3.4.1 Model Specification

To formally specify the relationship between characters and profiles, we adapt the notation from the social network topic model. Since the social network topic model and the social network GoM model share many similarities, we **bold** the differences between the two models. Let $\mathbf{z_i}$ be the Cartesian coordinates of character i in social space. Additionally, let $\mathbf{z_{t_k}}$ be the location of the k^{th} **profile in social space.** Then

the mixed-membership proportion for character i expressing **profile** k is given as $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$, where $||\mathbf{z_{t_k}} - \mathbf{z_i}||$ represents the Euclidean distance between **profile** k **and character** i**.**

In specifying the mixed-membership vector as being a function of both the location of the characters and the profiles, two previously independent models (i.e. GoM and the latent social network models) now share information and are inter-related. Considering a network of N **characters, we specify our combined generative social network GoM model as follows:**

1. Sample N character locations, $\mathbf{z} = (\mathbf{z_1}, ..., \mathbf{z_N})$, in social space from $P(\mathbf{z})$.

2. Sample K **profile locations**, $\mathbf{z_t} = (\mathbf{z_{t_1}}, ..., \mathbf{z_{t_K}})$, in social space from $P(\mathbf{z_t})$.

3. **For each of** K **profiles and J responses, sample the "success" probability** $\beta_{[kj]} \sim$ **Beta**(ϕ, ϕ)**.**

4. For each of N characters:

   (a) Calculate the **profile membership vector,** θ_i, based on the character's distance from each **profile**: $\theta_{i[k]} = \frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}}$.

   (b) For each of the J responses:

      i. Sample a **latent profile indicator** $u_{ij}|\theta_i \sim$ Multinomial$(\theta_i, 1)$.

      ii. Sample a **response** $x_{ij}|u_{ij}, \beta \sim$ Bernoulli$(\beta_{[.j]}^T u_{ij})$.

5. Sample social network parameters a and b from $P(a)$ and $P(b)$, respectively.

6. Sample the edges $Y \sim$ Bernoulli$(\mathbf{p_{adj}})$, where $p_{adj_{ii'}} = \frac{e^{a+bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||}}{1+e^{a+bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||}}$ and $h_{ii'} = \frac{1}{2}\sum_{\mathbf{j=1}}^{\mathbf{J}} \left(\theta_i \beta_{[.j]} log\left(\frac{\theta_i \beta_{[.j]}}{\theta_{i'} \beta_{[.j]}}\right) + \theta_{i'} \beta_{[.j]} log\left(\frac{\theta_{i'} \beta_{[.j]}}{\theta_i \beta_{[.j]}}\right)\right)$.

As in the social network topic model, steps 1 and 2 of this social network GoM model position the characters and the profiles together in social space based on their

respective distributions $P(\mathbf{z})$ and $P(\mathbf{z_t})$. The GoM portion of the model is present in steps 3 and 4. Similar to the case in the social network topic model, we deviate here from a traditional GoM model in that we specify θ as being deterministically calculated as a function of the distance between characters and profiles in social space, as opposed to the traditional setup of having θ be a realization from a Dirichlet distribution. Thus, characters who have high membership in a profile will be close to that profile in social space. The social network portion of the model is evident in steps 5 and 6, with a and b being scalars coming from distributions $P(a)$ and $P(b)$, respectively. In the GoM context, $h_{ii'}$ is now the average Kullback-Leibler divergence between response distributions for character i and character i', such that characters who have similar response patterns will be closer together in social space. The $h_{ii'}$ term draws upon the GoM portion of the model and incorporates such information into the determination of whether two characters are connected.

This generative process describes how social connections and character responses are created. Using our data—consisting of social connections and character responses to the J survey questions—and this generative process, it thus remains for posterior inference to be drawn regarding model parameters. These parameters determine the locations of characters in social space, the locations of profiles in social space, the probability of a "healthy" response to each question, and the disability profile membership for each character.

3.4.2 Inference

Similar to the social network topic model, our primary interest is in determining the joint posterior distribution of $\mathbf{z_t}$, $\mathbf{z}$, β, $\mathbf{u}$, a, and b, specified as:

$$P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, b|\mathbf{x}, Y, \boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta) = \frac{P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, b, \mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)}{P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)}.$$

For this social network GoM model setup, the likelihood $P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta)$ also does not have a closed form solution. The likelihood can be expressed as:

$$P(\mathbf{x}, Y|\boldsymbol{\eta}, \boldsymbol{\nu}, \phi, \omega, \delta) \propto \tag{3.4}$$
$$\int\int\int\int\int \sum_{\mathbf{u}} \left(\prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \beta_{[kj]}^{x_{ij}} \left(1 - \beta_{[kj]}\right)^{1-x_{ij}} \right)^{u_{ij[k]}} \right)$$
$$\times \prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}}{1 + e^{a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||}} D_{\boldsymbol{\eta}}(d\mathbf{z_t}) D_{\boldsymbol{\nu}}(d\mathbf{z}) D_{\phi}(d\beta) D_{\omega}(da) D_{\delta}(db).$$

In this likelihood, $\mathbf{z_t}$, $\mathbf{z}$, and β are again coupled in the summation of latent profiles, leading to the likelihood being intractable for analytic inference. Approximate inference can be performed in a similar manner to the social network topic model through MCMC sampling or a variational EM algorithm.

Although posterior distributions are not conjugate, we present here the conditional posterior distributions for the model parameters, which share similarities to their corresponding parameters in the social network topic model:

$$P(\beta|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \mathbf{u}, a, b, \phi) \propto$$
$$\left[\prod_{j=1}^{J} \prod_{k=1}^{K} \beta_{[kj]}^{\phi + \left(\sum_{i=1}^{N} u_{ij[k]} x_{ij}\right)} \left(1 - \beta_{[kj]}\right)^{\phi + \left(\sum_{i=1}^{N} u_{ij[k]} (1 - x_{ij})\right)} \right]$$
$$\times \left[\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}} \right]$$
$$P(u_{ij}|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta) \propto \prod_{k=1}^{K} \left[\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \beta_{[kj]}^{x_{ij}} \left(1 - \beta_{[kj]}\right)^{1-x_{ij}} \right]^{u_{ij[k]}}$$
$$P(\mathbf{z_t}|\mathbf{x}, Y, \mathbf{z}, \beta, \mathbf{u}, a, b, \boldsymbol{\eta}) \propto$$
$$P(\mathbf{z_t}|\boldsymbol{\eta}) \left[\prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{e^{-||\mathbf{z_{t_k}} - \mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}} \right)^{u_{ij[k]}} \right] \left[\prod_{i=1}^{N} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}}{1 + e^{(a + bh_{ii'} - ||\mathbf{z_i} - \mathbf{z_{i'}}||)}} \right]$$

$$P(\mathbf{z}|\mathbf{x}, Y, \mathbf{z_t}, \beta, \mathbf{u}, a, b, \boldsymbol{\nu}) \propto$$
$$P(\mathbf{z}|\boldsymbol{\nu}) \left[\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K}\left(\frac{e^{-||\mathbf{z_{t_k}}-\mathbf{z_i}||}}{\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}}-\mathbf{z_i}||}}\right)^{u_{ij[k]}}\right]\left[\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}\right]$$
$$P(a|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta, b, \omega) \propto P(a|\omega)\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}$$
$$P(b|\mathbf{x}, Y, \mathbf{z}, \mathbf{z_t}, \beta, a, \delta) \propto P(b|\delta)\prod_{i=1}^{N}\prod_{i'=i+1}^{N}\frac{e^{y_{ii'}(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{(a+bh_{ii'}-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}.$$

In the above representation, $h_{ii'}$ is a function of β, $\mathbf{z_t}$ and $\mathbf{z}$. Similar to the social network topic model, we recommend setting a bivariate normal prior distribution for each $\mathbf{z_{t_i}}$ and $\mathbf{z_i}$. $\beta_{[kj]}$ has a symmetric beta prior distribution with parameter ϕ. The prior specifications for a and b are recommended to be fairly uninformative uniform distributions. MCMC sampling methods can be implemented via JAGS until convergence is met based on the multi-ESS criterion (Flegal et al., 2016). Alternatively, the variational inference approach introduced in Section 3.2 can be modified for the social network GoM model.

3.5 Discussion

In this chapter, we discussed extensions to the social network topic model that was introduced in Chapter 2. In particular, we focused on two main extensions: the extended co-authorship social network topic model and the generalized social network Grade of Membership model. The co-authorship model allowed for multiple authors of each text document, extending the original social network topic model to a broader case of how many documents are in fact authored. Future work for the co-authorship social network topic model should focus on the weighting scheme for θ_m^*. In the co-authorship model, recall that θ_m^* is currently the average of mixed-membership vectors for all au-

thors of document m. Perhaps a better combination of authors' mixed-membership vectors would be more appropriate. For example, a weighted average of each author's θ_i could be calculated, with weights being proportional to the number of connections that the character has in the network. Under this weighting scheme, more popular characters would have greater influence over the topic specification of a document.

The Grade of Membership generalization of the social network topic model keeps the networking aspect of the original model but generalizes the mixed-membership portion from that of topic modeling to the GoM model. In this case, binary responses to multiple survey questions replace words as the discrete data used for the mixed-membership modeling. This generalization extends the social network topic model into another framework of discrete data soft clustering that naturally arises in many applications, particularly those of survey responses.

The introduced extensions and generalizations to the social network topic model demonstrate the flexibility with which the model can be applied. Hierarchical Bayesian mixed-membership models that are linked to social network data can be modeled, allowing for social networks to inform mixed-membership modeling, and discrete data attributes to inform the social network representation.

In the social network topic models and the social network GoM model, there is an issue of computational intensity that characterizes each method, particularly when inference is performed via MCMC methods. The computational limitation of the introduced models poses a serious drawback to the methods. As a resolution to this computational issue, a variational approximation for posterior inference was introduced, providing an alternative to MCMC sampling.

Chapter 4

Social Networks in the Context of Spatial Models

4.1 Introduction

On November 22, 2016, the temperature at my house in Henrietta, New York, was 32 degrees Fahrenheit. At the same time, the temperature at the University of Rochester campus was 33 degrees. Intuitively, we expect that nearby locations share similar temperatures, among other characteristics, and are thus highly correlated. The temperature at a given time at a house in Henrietta should be almost exactly the same as that outside a dorm at the University of Rochester, given their close proximity. Now, consider the temperature at my house on this day and at a house in Richmond, Virginia, where it was 54 degrees. Being so far away from one another, the temperatures at these two locations are less correlated. Thus, as I move farther away from my home, the temperature at my current location is increasingly less associated with the temperature at my house. Eventually, at a far enough distance away, the temperatures will be completely unrelated. But how far is far enough?

Data such as these temperature readings at specific locations form a class of spatial data. In particular, these spatial data are point-referenced. Generally, point-referenced spatial data consists of a set of fixed sites or locations and measurements of a characteristic of interest at each one of these sites (Finley and Banerjee, 2013). While other forms of spatial data (e.g. areal and point-pattern) exist, we focus only on point-referenced spatial data. Our interest is in modeling how a characteristic of interest, which we call the *response*, changes continuously over the entire space, even though we have a fixed number of sites for which measurements are taken. For example, even though I am only measuring the temperature at three locations—my house in Henrietta, a dorm at the University of Rochester, and a house in Richmond, Virginia—temperature continuously exists over the entire space, which perhaps is the eastern United States. Thus, we want to create a surface representing a characteristic of interest on a continuous scale for an entire space (Banerjee and Finley, 2009).

In order to model, for example, temperature over the eastern United States, we have to know how temperature varies with location. Intuitively, it seems reasonable that the temperatures at nearby locations will be quite similar whereas temperatures at far away locations will differ and be unrelated. This reasoning provides the foundation of spatial modeling. A correlation structure must be put in place to describe how a response is related for two sites that are a given distance apart (Banerjee, Carlin, and Gelfand, 2015). Sites that are close together will be highly correlated whereas sites farther apart have less of a relationship. Different rates of decay in association will be appropriate for different situations and will be discussed further below.

4.2 Point-Referenced Data Specifications

Point-referenced spatial data models are characterized by three inter-related features: stationarity, variograms, and isotropy (Banerjee et al., 2015). Changing the specifications for each of these features will create varying models that are appropriate for different spatial relationships. To further introduce these topics, some notation is necessary. Let $\mathbf{z_i}$ be the location of site i for which we are interested in modeling a response, $s(\mathbf{z_i})$. In total, assume there are N sites. All these sites are contained within the fixed space of interest, which we will denote as D. Typically, we take D to be a subset of $\mathbb{R}^2$. When the space of locations is at least two dimensional, we refer to the response's variation over this space as a *spatial process*. This process will explain the entire space, allowing us to make predictions at locations for which no sites are present. We will restrict ourselves to focusing on the two-dimensional case of a spatial process. Thus, continuing with our example, we are interested in modeling the change in temperature over latitude and longitude.

4.2.1 Stationarity, Variograms, and Isotropy

We assume that a spatial process has an average response at each site i, given as $\mu(\mathbf{z_i}) = E(s(\mathbf{z_i}))$. Also, we let the variance of $s(\mathbf{z_i})$ exist for all sites in D. With these assumptions, $s(\mathbf{z})$, for all $\mathbf{z} \in D$, is a *Gaussian process* if, for any set of $N \geq 1$ sites, $\mathbf{s} = (s(\mathbf{z_1}), ..., s(\mathbf{z_N}))^T$ has a multivariate normal distribution. Altering the specifications of the mean and variance will lead to different *stationarity* properties.

Typically, we require spatial processes to be at least *intrinsically stationary* (Myers, 1989). For intrinsic stationarity, we are interested in looking at how the response changes as we move from site $\mathbf{z_i}$ to $\mathbf{z_i} + \mathbf{g}$. Note that $\mathbf{g}$ is a *lag*, or a change in location, defined by both magnitude and direction. The mean response at these two

locations should remain unchanged, or $E\left[s(\mathbf{z_i}+\mathbf{g})-s(\mathbf{z_i})\right]=0$. When this is not true, we can subtract out the mean process. Additionally, for intrinsic stationarity, the variance of the difference in characteristics between the two locations should depend only on the lag, $\mathbf{g}$, not the particular locations of the sites themselves. More formally, $var\left[s(\mathbf{z_i}+\mathbf{g})-s(\mathbf{z_i})\right]=2\gamma(\mathbf{g})$. The function of the variance of the difference, denoted $2\gamma(\mathbf{g})$, is called the *variogram*, with $\gamma(\mathbf{g})$ being the *semi-variogram*. Changing the specification of the variogram leads to different covariance structures between the responses, which will be discussed in detail below.

Upon closer examination, it is apparent that the variogram for an intrinsically stationary process is a function of the covariance between locations, depending only on the lag between sites:

$$\begin{aligned}
2\gamma(\mathbf{g}) &= var\left[s(\mathbf{z_i}+\mathbf{g})-s(\mathbf{z_i})\right] \\
&= var\left[s(\mathbf{z_i}+\mathbf{g})\right]+var[s(\mathbf{z_i})]-2cov[s(\mathbf{z_i}+\mathbf{g}),s(\mathbf{z_i})] \\
&= C(0)+C(0)-2C(\mathbf{g}) \\
&= 2[C(0)-C(\mathbf{g})].
\end{aligned}$$

Note that, in general, $C(\mathbf{g})=cov[s(\mathbf{z_i}+\mathbf{g}),s(\mathbf{z_i})]$, expressing the correlation between the responses at two sites separated by a lag of $\mathbf{g}$.

Recall that we typically want locations far away from each other to be less correlated than locations close together. Spatial processes that incorporate this principle are *ergodic* (Cressie, 2003). In particular, ergodic spatial processes are characterized by $C(\mathbf{g})\to 0$ as $||\mathbf{g}||\to\infty$. Note that $||\mathbf{g}||$ is the Euclidean distance between sites and is a positive scalar, whereas the lag $\mathbf{g}$ is a D-dimensional vector of both magnitude and direction.

It is usually reasonable and desirable to assume that the variogram is a function

only of the distance between two sites. Under this assumption, any two sites that are the same distance away will have the same correlation. In other words, we are focusing only on the distance $||\mathbf{g}||$ and not the lag $\mathbf{g}$ itself. Spatial processes meeting this assumption of $\gamma(\mathbf{g}) = \gamma(||\mathbf{g}||)$ are called *isotropic*. *Anisotropic* covariance functions are ones where $\gamma(\mathbf{g}) \neq \gamma(||\mathbf{g}||)$. We focus here only on isotropic spatial correlations. Note that isotropic spatial processes are invariant under rotations. For simplicity, we will hereafter denote the distance $||\mathbf{g}||$ by d.

If a spatial process is isotropic and additionally is intrinsically stationary, then it is a *homogeneous spatial process*. In this case, we have that the mean responses at two nearby locations contained on our space are expected to be the same, or $E[s(\mathbf{z_i} + \mathbf{g}) - s(\mathbf{z_i})] = 0$, and the variance of the difference in responses is a function only of the distance between the two locations, or $var[s(\mathbf{z_i} + \mathbf{g}) - s(\mathbf{z_i})] = 2\gamma(d)$. For homogenous processes, the direction of the lag does not matter; we are only interested in the distance between sites (Banerjee et al., 2015). We assume homogeneity of our spatial processes for the remainder of this chapter.

4.2.2 Variogram Specification

As mentioned above, to fully specify a spatial process, we must determine what form the variogram will take. Stationarity and isotropy provide information relating to the properties of the space, but the choice of variogram will decide the shape of the spatial correlation. Many variogram specifications exist for modeling the covariance between responses at two locations. All variograms, though, share three common aspects—the *nugget*, *partial sill*, and *range*—and differ only in terms of their functional form (Banerjee and Finley, 2009).

The semi-variogram is a function of the nugget, partial sill, and range (Banerjee

and Finley, 2009). The functional form chosen for the semi-variogram will model how the covariance between the responses at two locations dissipates with increasing distance. Common semi-variogram forms include linear, exponential, spherical, powered exponential, Gaussian, wave, power law, and Matérn. We focus our exploration on the exponential semi-variogram, which is given as follows:

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - e^{-\phi d}) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}, \tag{4.1}$$

where τ^2 denotes the nugget, σ^2 the partial sill, and $\frac{1}{\phi}$ the range.

Generally, the nugget is the variability of the non-spatial random error that exists in our model, providing a description of the amount of variability in measurements from their mean response. In particular, the covariance between a location and itself is 0, so the semi-variogram at a distance of 0 will equal 0, or $\gamma(0) = 0$. As we approach 0 from any positive distance, though, we have that $\lim_{d \to 0^+} \gamma(d) = \tau^2$, since there is some natural variability in observed responses about their mean (Banerjee et al., 2015). Thus, the nugget represents the variability at a given site. Alternatively, σ^2 is the spatial-specific variability. In other words, the partial sill tells us how much variability there is in responses due to being at different locations across the space. Together, the nugget and partial sill form the *sill*, given by $\tau^2 + \sigma^2$ (Cressie, 2003).

The sill represents the maximum variability between two sites. Recall that the semi-variogram tells us how the correlation changes as we look at locations farther away from each other. Depending on the functional form, there is a certain distance between sites at which we reach maximum variability. In this case, the two locations are uncorrelated. When we have reached this point, we are at the sill. More formally, $\lim_{d \to \infty} \gamma(d) = \tau^2 + \sigma^2$. The distance for which we first reach the sill is the given by the range, $\frac{1}{\phi}$ (Banerjee and Finley, 2009). For an exponential semi-variogram, the

sill is only reached asymptotically, so we typically focus on the *effective range*. The effective range is the distance at which the correlation between two sites is negligible, usually taken to be 0.05.

Recall that the variogram describes the relationship between the responses observed at two sites that are distance d apart. In particular, we have that $\gamma(d) = C(0) - C(d)$. If we take the limit of this variogram as the distance between sites gets infinitely large, we have that $\lim_{u\to\infty} \gamma(u) = C(0) - \lim_{u\to\infty} C(u) = C(0)$, since sites infinitely far away are assumed to be completely unrelated. Using this conclusion and rearranging the expression for $\gamma(d)$, we can express the covariance in terms of the variogram. Specifically, we can write $C(d) = \lim_{u\to\infty} \gamma(u) - \gamma(d)$. Under this specification, we can directly model how the responses at two sites that are distance d away from each other will covary, based on the choice of functional form for the variogram. The covariance representation of the exponential variogram is:

$$C(d) = \begin{cases} \sigma^2 e^{-\phi d} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}. \tag{4.2}$$

In practice, it may not necessarily be obvious what functional form best models the covariance structure of a given space. For example, I may know that temperatures at nearby locations are similar and those far away are not related, but does the intensity of the relationship decrease exponentially? Along a Gaussian curve? Or perhaps in some other manner? To determine a potential covariance structure for the model, we use the observed data at the finite locations for which we measure a response to formulate the empirical semi-variogram, which is a nonparametric estimate of the semi-variogram (Bohling, 2005). We can compare this empirical estimate to the shapes of the theoretical semi-variograms, ultimately choosing the functional form that seems most reasonable for our data.

If we let B_d be the set of all pairs of locations that are distance d apart from each other and n_d be the number of pairs that are distance d apart, then we can write the empirical variogram as follows:

$$\hat{\gamma}(d) = \frac{1}{n_d} \sum_{(\mathbf{z_i}, \mathbf{z_{i'}}) \in B_d} [s(\mathbf{z_i}) - s(\mathbf{z_{i'}})]^2. \tag{4.3}$$

Essentially, we look at the average squared deviations in responses between all sites that are a given distance apart and use this as our nonparametric estimate of the semi-variogram. It should be acknowledged that this empirical estimate is imperfect. A noticeable issue is that of requiring a set of points be exactly distance d apart. Generally, each pair of characters will be a unique distance apart, which leads to an unstable estimate of the empirical semi-variogram as it is currently described. Because of this, we allow for B_d to be the set of points that fall within a neighborhood of d. In particular, we split d into intervals, determine which pairs fall into each interval, and use these sets to update our definition of B_d (Banerjee et al., 2015). Perhaps, for example, we look at $[0, 1)$, $[1, 2)$, $[2, 3)$, and $[3, 4)$, where the interval $[0, 1)$ contains all pairs of characters that are within 1 unit from each other. Although the selection of appropriate intervals must be decided, this empirical semi-variogram acts as the best estimator for benchmarking which theoretical semi-variogram seems most reasonable.

4.3 Spatial Model Specification

Once the stationarity and isotropy assumptions have been met and the dependence structure specified through a variogram, we have the essential aspects of the spatial process covariance structure in place and can formally create a model for describing a space. We now turn our attention to fully specifying a Bayesian spatial model for point-referenced data.

4.3.1 Basic Model

At each of N sites, we measure some response, which is what we ultimately want to be able to model over the entire space. We specify a basic model for the response measured at sites i as

$$s(\mathbf{z_i}) = \mu(\mathbf{z_i}) + w(\mathbf{z_i}) + \epsilon(\mathbf{z_i}), \quad i = 1, ..., N. \tag{4.4}$$

Note that the response is modeled as having three components, all of which can be site dependent: a mean structure, $\mu(\mathbf{z_i})$, a spatial effect, $w(\mathbf{z_i})$, and a random error term, $\epsilon(\mathbf{z_i})$ (Finley, Banerjee, Cook, and Bradford, 2013). The spatial effect and random error term together form the residual portion of the model.

The mean structure, $\mu(\mathbf{z_i})$, can be expressed in terms of linear predictors, $\mathbf{x}^T(\mathbf{z_i})\beta$. Continuing with our motivating example, the average temperature at location $\mathbf{z_i}$ can be explained by altitude, proximity to water, and other covariates of interest that may impact temperature. The mean structure is not of direct focus in our research, but covariates can be included in the model if deemed appropriate. Otherwise, an intercept-only model may be used, in which case we assume an overall, site-independent mean response (Banerjee et al., 2015). We adopt an intercept-only spatial model for the remainder of this chapter.

We previously introduced two terms in the residual portion of our model, $w(\mathbf{z_i})$ and $\epsilon(\mathbf{z_i})$. Under this partition, the $w(\mathbf{z_i})$ term captures the residual spatial association and is a realization from a zero-centered stationary Gaussian spatial process (Finley et al., 2013). The $w(\mathbf{z_i})$ portion of the residual thus depends on the partial sill, σ^2, and the range, $\frac{1}{\phi}$, as these are the spatial influences on the modeled response. With the spatial influences fully captured by $w(\mathbf{z_i})$, $\epsilon(\mathbf{z_i})$ remains as the uncorrelated pure error, or the non-spatial residuals (Finley et al., 2013). Capturing the random deviations of

the responses that are not attributable to the spatial process, $\epsilon(\mathbf{z_i})$ depends on the nugget, τ^2. With this specification of the model for $s(\mathbf{z_i})$, the variability depends only on τ^2, σ^2, and ϕ, whereas the expectation is constant but could potentially depend on any covariates of interest.

Introducing the notation of $\mathbf{s}$ as an $N \times 1$ data vector, β as a scalar representing the overall mean of the space, and Σ as an $N \times N$ covariance matrix to be more carefully specified below, and under the assumption of a Gaussian spatial model, we can represent our response as follows:

$$\mathbf{s}|\beta, \Sigma \sim N(\beta, \Sigma). \tag{4.5}$$

Based on our above examination of the model $s(\mathbf{z_i}) = \mu(\mathbf{z_i}) + w(\mathbf{z_i}) + \epsilon(\mathbf{z_i})$, for $i = 1, ..., N$, we have that Σ depends on the nugget, partial sill, and range. We can thus specify $\Sigma = \sigma^2 M(\phi) + \tau^2 I_N$. Here, M is a correlation matrix with elements $m_{ii'} = \rho(\mathbf{z_i} - \mathbf{z_{i'}}, \phi)$, ρ is a valid isotropic correlation function, indexed by the range parameter ϕ, and I_N is an $N \times N$ identity matrix. For example, we may choose ρ to be an exponential correlation function. With this specification of Σ, we can update our representation of $\mathbf{s}$:

$$\mathbf{s}|\beta, \sigma^2, \tau^2, \phi \sim N(\beta, \sigma^2 M(\phi) + \tau^2 I_N). \tag{4.6}$$

In completing our Bayesian model, independent prior distributions are typically placed on the model parameters, giving us $P(\beta, \sigma^2, \tau^2, \phi) = P(\beta)P(\sigma^2)P(\tau^2)P(\phi)$. Although uninformative prior distributions are often appropriate for β, τ^2, and σ^2, the prior for ϕ should be fairly informative, as there is not much information in the data regarding the range parameter (Banerjee et al., 2015). We can place a normal prior distribution on β, with mean λ and variance ζ^2, and inverse-gamma prior distributions on σ^2 and τ^2, with hyper-parameters (α_1, α_2) and (η_1, η_2), respectively. We want to restrict the

range of ϕ to a set of reasonable values through a uniform prior distribution over the range (ν_1, ν_2).

4.3.2 Hierarchical Model

This basic model can be re-expressed as a hierarchical model where there is a distribution on the spatial effects themselves. Recall that, by assumption, the spatial effects in the basic model are realizations from a zero-centered Gaussian spatial process. It thus is reasonable to place a normal distribution on the spatial effects. Considering the spatial effects as random, we can condition the observed response $\mathbf{s}$ on β, τ^2, and the spatial random effects, $\mathbf{w}$:

$$\mathbf{s}|\beta, \tau^2, \mathbf{w} \sim N(\beta + \mathbf{w}, \tau^2 I_N) \tag{4.7}$$

$$\mathbf{w}|\sigma^2, \phi \sim N(0, \sigma^2 M(\phi)) \tag{4.8}$$

The model for $\mathbf{w}$ is known as the *process model*, as it is a process used to capture the spatial dependence present between sites. Again, we place independent priors on β, σ^2, τ^2, and ϕ.

Note that this hierarchical model reduces down to $P(\mathbf{s}|\beta, \sigma^2, \tau^2, \phi)P(\beta, \sigma^2, \tau^2, \phi)$ if we marginalize over $\mathbf{w}$ (Banerjee et al. 2015). In practice, it is often more ideal to fit the marginalized model than the hierarchical model because the variance matrix $\sigma^2 M(\phi) + \tau^2 I_N$ in the marginalized model has more stable behavior than $\sigma^2 M(\phi)$ in the hierarchical model. Recall that M is a correlation matrix where each element, $m_{ii'}$, depends on the locations of sites $\mathbf{z_i}$ and $\mathbf{z_{i'}}$. If any pair $\mathbf{z_i}$ and $\mathbf{z_{i'}}$ are near each other, then $\sigma^2 M(\phi)$ will be close to being a singular matrix, whereas with the addition of τ^2 to the diagonal, $\sigma^2 M(\phi) + \tau^2 I_N$ will not be (Banerjee et al., 2015). Additionally, the dimensionality of our model is greatly reduced when we do not have to directly

sample the spatial random effects, $\mathbf{w}$.

In practice we use the marginalized model and later recover the spatial posterior distribution. Since $P(\mathbf{w}|\mathbf{s}) = \int\int\int\int P(\mathbf{w}, \beta, \sigma^2, \tau^2, \phi, \mathbf{s})P(\beta, \sigma^2, \tau^2, \phi|\mathbf{s})d\beta d\sigma^2 d\tau^2 d\phi$, we can thus obtain posterior realizations of $\mathbf{w}$ through composition sampling based on the posterior samples of β, σ^2, τ^2, and ϕ.

4.4 Social Networks as Spatial Data

Let us now change the scene. Instead of modeling the behavior of temperature over a set of locations, let us suppose we are interested in how an attribute varies over a social network. For example, consider your group of closest acquaintances. The people whom you are most strongly connected to probably share many of the same characteristics as you. Similarities exist between you and these friends, resulting in you having a strong friendship with them. Perhaps you and your best friend often attend exercise classes together, which thus leads to the two of you having similar body mass indexes (BMIs). Alternatively, there are people who you are not as close with. These are people who probably do not share as many of the same interests as you and are thus socially farther away from you. Continuing with our example, perhaps these people are acquaintances who do not share your enthusiasm for exercise and have higher BMIs. Within the framework of this example, the spatial locations have been exchanged for a social network, but otherwise we are still interested in observing how a response varies over a space.

An added complexity must be considered with social networks. Whereas locations on a traditional space are fixed and exactly positioned, a social network does not have a fixed orientation. At minimum, we know which characters in the network are connected to other characters, but we are left to create a visual representation of what the

social space actually looks like. We draw upon Hoff, Raftery, and Handcock's (2002) latent approach for modeling a network, where the adjacency matrix of connections between characters determines the representation of the network in social space. If we take this network as being fixed and have a response for each character, which is known in the network context as a *node attribute*, our situation closely resembles that of the traditional spatial setting. An added complexity arises, though, since the network is not in fact fixed. We introduce here a method for simultaneously estimating the social space and a spatial process over that space.

While previous work has been done to integrate social network information into spatial models (Zheng, Salganik, and Gelman, 2006; Radil, 2011), none have replaced the spatial aspect of the model with a social network. Instead, the social influence has previously been considered as an additional term in the random-effects model presented above and traditional spatial locations are known (Emch et al., 2012). In this case, there would be a mean response, a spatial effect, a random error term, and a network effect based on covariate information. We instead focus on using the latent social network representation introduced in Chapter 1 to replace the spatial locations used in traditional spatial models.

Consider a group of $N = 100$ people who all have at least one friend in the group. We record the BMI for each person in this group. It seems reasonable that close friends will probably have similarities in their BMIs. Friends are likely to be more or less active together, so this social support should lead to friends having similar BMIs (Estabrooks, Bradshaw, Dzewaltowski, and Smith-Ray, 2008; Shelton et al., 2011; Yu et al., 2011). Can we use the theory of traditional spatial models to describe how BMI, for example, changes over a social network? We turn now to answering questions similar to this one through the formal specification of a social network spatial model that allows us to describe the variability of an attribute over a social network while

also jointly modeling the social space.

4.4.1 Model Specification

For this social network spatial model, the spatial process is dependent on the social network representation, along with the position of characters in social space being influenced by their spatial effects. For each of the N characters in the social network—who have unknown social location $\mathbf{z_i}$ and known social connections $y_{ii'}$—we thus have a measured attribute, $s(\mathbf{z_i})$, coming from a spatial process with unknown parameters ϕ, σ^2, and τ^2. With this setup and a, b, and β as scalars, our social network spatial model can be presented as follows:

$$\mathbf{s}|\beta, \sigma^2, \tau^2, \phi, \mathbf{z} \sim N(\beta, \sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N) \tag{4.9}$$

$$Y|\mathbf{s}, a, b, \mathbf{z} \sim \text{Bernoulli}(\mathbf{p_{adj}}), \text{where } p_{adj_{ii'}} = \frac{e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}. \tag{4.10}$$

Analogous to the traditional spatial models above, $M(\phi, \mathbf{z})$ is a correlation matrix with elements $m_{ii'} = \rho(\mathbf{z_i} - \mathbf{z_{i'}}, \phi)$ and ρ being a valid isotropic correlation function. To determine an appropriate form of the correlation function, we recommend examining the empirical semi-variogram based on a social network spatial model fit through two-step inference. In particular, a social network model should be fit independent of the spatial data, following the methods described in Chapter 1. Then, these locations should be considered fixed and the traditional empirical semi-variogram calculated over this social space. Visual examination of the semi-variogram provides a foundation for an appropriate correlation function. Also note that $||.||$ indicates the absolute value for scalar quantities (e.g. $s(\mathbf{z_i})$) and the Euclidean distance for two-dimensional variables (e.g. $\mathbf{z_i}$).

Independent prior distributions are placed on the model parameters, leading to

a prior specification of $P(\beta, \sigma^2, \tau^2, \phi, \mathbf{z}, a, b) = P(\beta)P(\sigma^2)P(\tau^2)P(\phi)P(\mathbf{z})P(a)P(b)$. Natural choices for prior distribution specifications are $\beta \sim N(\lambda, \zeta^2)$, $\sigma^2 \sim IG(\alpha_1, \alpha_1)$, $\tau^2 \sim IG(\eta_1, \eta_2)$, $\phi \sim \text{Unif}(\nu_1, \nu_2)$, $\mathbf{z_i} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Lambda}\right)$, $a \sim \text{Unif}(\kappa_1, \kappa_2)$, $b \sim \text{Unif}(\psi_1, \psi_2)$. Following traditional spatial analysis methods, the hyper-parameters for each of these prior distributions are set to fixed values, as opposed to being directly estimated. Note that in the prior distribution for β, ζ^2 is often taken to be large such that this prior is uninformative. When an uninformative prior distribution is desired for β, an uninformative uniform distribution may also be an appropriate choice. As noted above for traditional spatial models, the prior distribution for ϕ should be fairly informative, and thus ν_1 and ν_2 should be reasonably chosen to restrict the range of ϕ. Information relating to hyper-parameter specification can be obtained from the two-step empirical semi-variogram described above.

Similar to traditional spatial models, we can analogously represent the social network spatial model in hierarchical form. Placing a distribution on the spatial random effects, $\mathbf{w}$, we can condition the observed spatial response $\mathbf{s}$ on β, τ^2, and $\mathbf{w}$:

$$\mathbf{s}|\beta, \tau^2, \mathbf{w} \sim N(\beta + \mathbf{w}, \tau^2 I_N) \tag{4.11}$$

$$\mathbf{w}|\sigma^2, \phi, \mathbf{z} \sim N(0, \sigma^2 M(\phi, \mathbf{z})) \tag{4.12}$$

$$Y|a, b, \mathbf{z}, \mathbf{s} \sim \text{Bernoulli}(\mathbf{p_{adj}}), \text{where } p_{adj_{ii'}} = \frac{e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}. \tag{4.13}$$

We again place independent prior distributions on β, σ^2, τ^2, ϕ, $\mathbf{z}$, a, and b. With this hierarchical model, it is often better to marginalize over $\mathbf{w}$ and use the basic model representation, as the variance matrix $\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N$ of this social network spatial model has more stable behavior than the variance $\sigma^2 M(\phi, \mathbf{z})$ of the hierarchical model. As is the case of traditional spatial models, we then recover the spatial distribution

$\mathbf{w}|\mathbf{s}, Y$ through composition sampling based on the posterior samples of β, σ^2, τ^2, ϕ, $\mathbf{z}$, a, and b.

4.4.2 Inference

We present here inference for the social network spatial model, marginalized over the random spatial effects. The joint posterior distribution of the model parameters is represented as follows:

$$
\begin{aligned}
&P(\phi, \tau^2, \sigma^2, \beta, \mathbf{z}, a, b|\mathbf{s}, Y) \qquad (4.14)\\
&\quad \propto P(\phi)P(\tau^2)P(\sigma^2)P(\beta)P(\mathbf{z})P(a)P(b)P(\mathbf{s}|\beta, \tau^2, \sigma^2, \phi, \mathbf{z})P(Y|a, b, \mathbf{z}, \mathbf{s})\\
&\quad \propto (\sigma^2)^{-\alpha_1-1}(\tau^2)^{-\eta_1-1}\left(\prod_{i=1}^{N} e^{-\frac{1}{2}\mathbf{z_i}^T\Lambda^{-1}\mathbf{z_i}} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}\right)\\
&\qquad \times |\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N|^{-\frac{1}{2}} e^{-\frac{\alpha_2}{\sigma^2}-\frac{\eta_2}{\tau^2}-\frac{1}{2}\left(\left(\frac{\beta-\lambda}{\zeta}\right)^2+(\mathbf{s}-\beta)^T(\sigma^2 M(\phi,\mathbf{z})+\tau^2 I_N)^{-1}(\mathbf{s}-\beta)\right)}\\
&\qquad \times \mathbb{1}_{(\nu_1,\nu_2)}(\phi)\mathbb{1}_{(\kappa_1,\kappa_2)}(a)\mathbb{1}_{(\psi_1,\psi_2)}(b).
\end{aligned}
$$

From here, we derive the full conditional posterior distribution of β in closed form as:

$$P(\beta|\mathbf{s}, Y, \sigma^2, \tau^2, \phi, \mathbf{z}) \propto e^{-\frac{1}{2}\frac{(\beta-\hat{\beta})^2}{V_\beta}}. \qquad (4.15)$$

Note that:

$$
\begin{aligned}
\hat{\beta} &= \left(\frac{1}{\zeta^2} + J_N^T(\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N)^{-1} J_N\right)^{-1}\left(\frac{\lambda}{\zeta^2} + J_N^T(\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N)^{-1}\mathbf{s}\right)\\
V_\beta &= \left(\frac{1}{\zeta^2} + J_N^T(\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N)^{-1} J_N\right)^{-1}.
\end{aligned}
$$

Additionally, J_N is a length N column vector with all entries equal to $\frac{1}{N}$. Thus $\beta \sim N(\hat{\beta}, V_\beta)$. If the prior distribution for β is specified as uninformative, the mean of the posterior distribution is a weighted average of the data, with each character's attribute being weighted by the average variability.

For the remaining parameters in this marginalized model, we have the following full conditional posterior distributions:

$$P(\sigma^2|\mathbf{s}, Y, \tau^2, \phi, \beta, \mathbf{z}) \propto$$
$$(\sigma^2)^{-\alpha_1-1}|\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{s}-\beta)^T(\sigma^2 M(\phi,\mathbf{z})+\tau^2 I_N)^{-1}(\mathbf{s}-\beta)+\frac{\alpha_2}{\sigma^2}}$$
$$P(\tau^2|\mathbf{s}, Y, \sigma^2, \phi, \beta, \mathbf{z}) \propto$$
$$(\tau^2)^{-\eta_1-1}|\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{s}-\beta)^T(\sigma^2 M(\phi,\mathbf{z})+\tau^2 I_N)^{-1}(\mathbf{s}-\beta)+\frac{\eta_2}{\tau^2}}$$
$$P(\phi|\mathbf{s}, Y, \sigma^2, \tau^2, \beta, \mathbf{z}) \propto |\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{s}-\beta)^T(\sigma^2 M(\phi,\mathbf{z})+\tau^2 I_N)^{-1}(\mathbf{s}-\beta)} \mathbb{1}_{(\nu_1,\nu_2)}(\phi)$$
$$P(a|\mathbf{s}, Y, \mathbf{z}, b) \propto \prod_{i=1}^{N}\prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}} \mathbb{1}_{(\kappa_1,\kappa_2)}(a)$$
$$P(b|\mathbf{s}, Y, \mathbf{z}, a) \propto \prod_{i=1}^{N}\prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}} \mathbb{1}_{(\psi_1,\psi_2)}(b)$$
$$P(\mathbf{z}|\mathbf{s}, Y, \sigma^2, \tau^2, \phi, \beta, a, b) \propto |\sigma^2 M(\phi, \mathbf{z}) + \tau^2 I_N|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{s}-\beta)^T(\sigma^2 M(\phi,\mathbf{z})+\tau^2 I_N)^{-1}(\mathbf{s}-\beta)}$$
$$\times \prod_{i=1}^{N} e^{-\frac{1}{2}\mathbf{z_i}^T\Lambda^{-1}\mathbf{z_i}} \prod_{i'=i+1}^{N} \frac{e^{y_{ii'}(a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||)}}{1+e^{a+b||s(\mathbf{z_i})-s(\mathbf{z_{i'}})||-||\mathbf{z_i}-\mathbf{z_{i'}}||}}.$$

Noting that, aside from β, the full conditional posterior distributions for the parameters of this marginalized model do not have convenient forms, inference is performed through the use of Just Another Gibbs Sampling (JAGS), where the posterior distribution for β is sampled from via a Gibbs sampler and the remaining non-conjugate posterior distributions are sampled from using slicer samplers (Plummer, 2013).

After calculating the posterior distributions of model parameters, we then perform composition sampling to recover the posterior distribution of the random spatial effects, $\mathbf{w}$. The full conditional posterior distribution of the random spatial effects is

$\mathbf{w}|\mathbf{s}, Y, \sigma^2, \tau^2, \phi, \mathbf{z} \sim N(\mu_w, \Sigma_w)$, where the mean and variance are given as follows:

$$\mu_w = ((\sigma^2 M(\phi, \mathbf{z}))^{-1} + n(\tau^2 I_N)^{-1})^{-1}(n(\tau^2 I_N)^{-1}\bar{s})$$
$$\Sigma_w = ((\sigma^2 M(\phi, \mathbf{z}))^{-1} + n(\tau^2 I_N)^{-1})^{-1}.$$

In this notation, $\bar{s}$ is the overall observed mean response. At the j^{th} iteration of the composition sampler, the mean, μ_w, and variance, Σ_w, are evaluated at the j^{th} values of the sampled marginal posterior distributions of σ^2, τ^2, ϕ, and $\mathbf{z}$. The samples drawn from this posterior distribution collectively form the marginal posterior distribution of $\mathbf{w}$.

4.5 Simulations

To evaluate the social network spatial model presented above, we perform a series of simulations. In particular, for each simulation, we generate a social network of $N = 100$ characters. For each character in the network, we simulate an attribute according to a spatial model, using the positions of the characters in social space as the spatial locations. We use an exponential variogram with parameters $\sigma^2 = 0.1$, $\tau^2 = 0.02$, $\phi = 0.5$, and $\beta = 1$ to simulate character attributes, $\mathbf{s} = (s(\mathbf{z_1}), ..., s(\mathbf{z_N}))$. Our goal is to use the simulated adjacency matrix and attributes as data for the social network spatial model to determine how well we are able to recover spatial variations and character locations in social space. As noted in Chapter 1, we are not directly interested in the values of a and b, as their magnitudes will depend on the particular representation of the social space, but we do give attention to whether convergence to their posterior distributions has been achieved.

Since location of characters in social space are estimated, the values of spatial parameters σ^2, τ^2, and ϕ will also be dependent on the scaling of the social space.

In particular, if social space is estimated to be on a large scale, the estimated spatial parameters will also be larger than if the space were on a smaller range. Because of this, we are interested in the proportion of the variability attributable to the spatial process, or $\frac{\sigma^2}{\sigma^2+\tau^2}$. Regardless of the range on which the social space is estimated to be on, the proportion of variability attributable to the spatial effect is comparable. Higher values of this proportion indicate that the spatial process is explaining more of the observed variability in the response. In the simulations below, we therefore assess the ability of the social network spatial model to properly recover $\frac{\sigma^2}{\sigma^2+\tau^2}$. The range, $\frac{1}{\phi}$, is not of direct inferential interest in these simulations, but it has a reasonable interpretation in the context of any given social space representation. The overall mean effect should not be affected by the social space representation, and thus β remains of direct inferential interest in the simulations.

In performing inference using the social network spatial model, we jointly model the social network and spatial correlations using the marginalized model where the spatial effects, $\mathbf{w}$, are recovered via composition sampling. We simulate 100 such spaces and repeat the inferential process for each, determining the average estimates of the spatial parameters σ^2, τ^2, ϕ, and β, along with the social network parameters $\mathbf{z}$, a, and b. For the estimation of the network, we specify the prior distributions for the logistic parameters as $a \sim \text{Unif}(-10, 10)$ and $b \sim \text{Unif}(-20, 20)$ and the two-dimensional network locations as $\mathbf{z_i} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$. A flat prior distribution is used for β, along with $\sigma^2 \sim IG(2, 0.5)$, $\tau^2 \sim IG(2, 0.02)$, and $\phi \sim \text{Unif}(0.01, 1)$.

In drawing samples from our posterior distributions via JAGS, we run the sampler until satisfactory convergence is reached based on the multivariate effective sample size criterion presented in Chapter 2, applied to the parameters σ^2, τ^2, ϕ, β, a, and b (Flegal, Hughes, and Vats, 2016). Summary results for the parameter estimates are

presented in Table 4.1.

Table 4.1: Set 1: Simulation summary statistics

	Truth	Mean	Coverage
$\frac{\sigma^2}{\sigma^2+\tau^2}$	0.83	0.802	90%
β	1.00	1.007	99%

Overall, the social network spatial model performs relatively well in recovering the true values of the model parameters. For the proportion of variability attributable to the spatial effect, the coverage is reasonably high at 90%, indicating that the posterior intervals frequently cover the truth and are thus well estimating the ratio. A slight negative bias is noticed, which suggests that the social network spatial model is attributing a bit more of the variability to random error. Since the social locations are estimated, there will be additional variability captured by τ^2, leading to the observed result. The overall spatial effect, β, is generally well estimated with high coverage and a posterior mean that closely matches the simulated value.

Although we are not directly interested in the values of a and b, we do pay attention to the sign of their estimates, particularly that of b. Consider a situation of three characters who are equidistant in social space. If b is negative, the two characters with the most similar attributes will be the most likely to have a connection. Alternatively, if b is positive, the two characters with the most dissimilar attributes will have the highest probability of a connection. To further investigate this, we simulated data and fit the social network spatial model with two sets of prior distribution specifications. First, we considered the fairly uninformative prior distributions for $a \sim \text{Unif}(-10, 10)$ and $b \sim \text{Unif}(-20, 20)$, as stated above. In this case, the correct sign is determined for each parameter (a is estimated as positive and b is estimated as negative). Additionally, we considered a more informative set of priors distributions for $a \sim \text{Unif}(0, 10)$ and $b \sim \text{Unif}(-20, 0)$. Under both prior distribution specifications,

similar results were reached in estimating all model parameters, indicating that either specification of the prior distributions can be appropriate, if the correct knowledge is used to inform the sign of a and b.

For the social network aspect of the model, we look at the sum of squared deviations of the estimated locations from the true character locations. After drawing all the samples from the posterior distribution of the character locations, we allow for a procrustes transformation of each sample, using the truth as our reference locations. Additionally, we fit an independent social network model that does not include any spatial information, similarly applying a procrustes transformation, and compare the sum of squared deviations to those for the social network spatial model. From this comparison, we see that the social network spatial model performs better than the independently fit social network model without spatial information in 80% of the simulations. The social network spatial model generally gains information by incorporating the spatial attributes of each character, providing a well-estimated social space representation.

In addition to the simulation conditions specified above, we also seek to investigate the sensitivity of results to the choice of prior distributions, specifically on σ^2 and τ^2. Until now, we have considered informative inverse-gamma prior distributions for both σ^2 and τ^2, setting the mean of each distribution at the true parameter value. Instead, we now investigate simulation results when uninformative inverse-gamma prior distributions, specifically $IG(0.0001, 0.0001)$, are used for both σ^2 and τ^2. Informative priors distributions are used for a and b as Unif$(0, 10)$ and Unif$(-20, 0)$, respectively. Summary results are presented in Table 4.2. We note that it took 50,000 iterations to reach convergence compared to the 10,000 iterations used in the informative simulations above.

In Table 4.2, we see that the overall mean effect, β, is still well estimated, regard-

Table 4.2: Set 2: Simulation summary statistics

	Truth	Mean	Coverage
$\frac{\sigma^2}{\sigma^2+\tau^2}$	0.83	0.778	83%
β	1.00	1.008	100%

less of the choice of prior distributions. We also note that the estimates of a and b, along with character locations, $\mathbf{z}$, are not noticeably affected by the choice of prior distributions on σ^2 and τ^2. In this case of uninformative prior distributions for σ^2 and τ^2, the proportion of variability explained by the spatial process is not too well estimated. With the uninformative prior distribution, larger values of σ^2 and τ^2 were sampled. Additionally, τ^2 is more positively biased than σ^2, leading to less of the variability being attributed to the spatial process. Since the true value of τ^2 is small, the influence of sampling larger values when using this uninformative prior distribution has a sizable impact on its estimation.

Based on these simulation results, we thus recommend that fairly informative prior distributions be chosen for σ^2 and τ^2. In particular, inverse-gamma prior distributions should be used with means equal to plausible values for σ^2 and τ^2. To determine plausible values for the mean of each prior distribution, we recommend an empirical Bayesian approach. We examine the empirical semi-variogram fit through two-step inference and use the resulting estimates of σ^2 and τ^2 as the prior means. This will lead to informative prior distributions, in which case the posterior means are generally well estimated.

4.5.1 Model Selection

The above simulations demonstrate the ability of the social network spatial model to effectively recover both the spatial and the social network parameters that were simu-

lated under this dependence model. In practice, though, we must determine whether a spatial effect does in fact exist over a social network. If the social network is unrelated to an observed attribute, and thus the observed attribute is being independently generated under some regression model (Banerjee et al., 2015), the social network spatial model (known hereafter as the dependence model) is not appropriate and the independent social network and intercept-only regression models (known hereafter as the independence model) should be fit to describe the space. To determine whether the dependence model is more appropriate than the independence model, we rely upon the deviance information criterion (DIC).

DIC is a generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for hierarchical models (Spiegelhalter, Best, Carlin, and van der Linde, 2002). When MCMC posterior inference is used to approximate a posterior distribution, DIC is particularly useful. In its general specification, deviance is defined as $D(\theta) = -2log(P(Y|\theta)) + C$, where Y is the data, θ are the unknown parameters of interest, and C is a constant. Taking the expectation of the deviance, we get $\bar{D} = E_\theta [D(\theta)]$. The larger this quantity, the worse the model fit. With likelihood-based model selection criteria, it is ideal to penalize for the number of parameters in the model. We can determine the effective number of parameters in our model as $p_D = \frac{1}{2}\hat{var}(D(\theta))$. Combining these two pieces, we calculate the DIC as $DIC = p_D + \bar{D}$ (Gelman, Carlin, Stern, and Rubin, 2004).

We evaluate the effectiveness of using DIC to properly choose when the dependence model is appropriate through a series of simulations. First, we simulate data based on the dependence model, in which case there is a relationship between the social network and the spatial effect. For this set of dependent data, we then fit both the dependence model and the independence model. The DIC is calculated for each model fit, with the expectation that the social network spatial model will yield the smaller value of DIC

and thus be chosen as the optimal model. 100 datasets are simulated from a dependent process and evaluated using both the dependence model and the independence model. In this particular setup, we simulate networks of $N = 50$ characters, with social network parameters $a = 1.5$ and $b = -5$. For the spatial portion of the model, we generate character attributes, $\mathbf{s}$, based on a spatial model parameters $\sigma^2 = 0.1$, $\tau^2 = 0.02$, $\phi = 0.5$ and $\beta = 1$ over the social space.

In addition to the simulations generated under the dependence model, we simulate a scenario where the social network is independent of the character attributes. We then fit both the dependence model and the independence model. The DIC is calculated in each case, and in this scenario, we expect that the dependence model will not yield a value of DIC higher than the independence model, and thus the independence model will be preferred. 100 datasets are simulated and fit under such conditions. For these simulations, we again generate networks of $N = 50$ characters and set the social network parameter for the independence model at $a = 1.5$. Note that no covariate information aside from the distances between characters is used in this independence model, and thus no b term is necessary. Node attributes in this intercept-only model are generated by a normal distribution with mean 0 and standard deviation 0.1, or $N(0, 0.1^2)$.

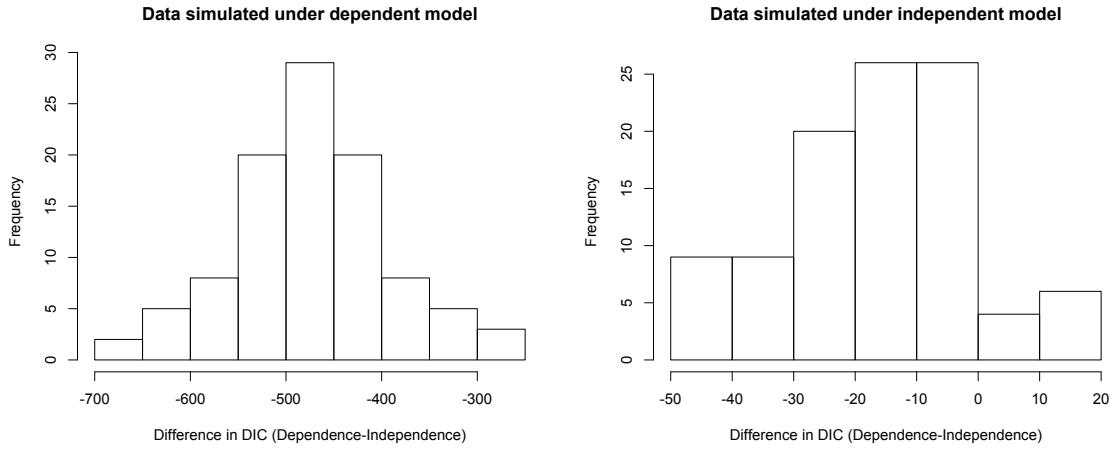
Regardless of which model we are using to simulate the data (dependence or independence), we use the following prior distribution specifications: $\sigma^2 \sim IG(2, 0.1)$, $\tau^2 \sim IG(2, 0.02)$, $\phi \sim \text{Unif}(0.01, 1)$, $\beta \sim N(0, 10000)$, $\mathbf{z_i} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$, $a \sim \text{Unif}(-10, 10)$, and $b \sim \text{Unif}(-20, 20)$. The results for the simulations are presented in Table 4.3. When the data are simulated from the dependence model, the fitted dependence model yields the smaller mean DIC. In fact, in all 100 simulations, the dependence model is chosen as better compared to the independence model. When

the independence model is used for simulating data, both the dependence and independence models perform fairly similarly in terms of DIC. Since the independence model is more parsimonious, the independence model should thus be chosen as optimal in this case. Figure 4.1 demonstrates that for any given dataset that is truly arising from a dependent process, the difference in DIC for the dependence and independence models is large in magnitude, with the dependence model fit always yielding the smaller DIC in our simulations. Alternatively, for any given dataset that arises from the independence setting, the difference in DIC for the dependence and independence models is not too pronounced and is generally characterized by a magnitude close to 0.

Table 4.3: Mean DIC for evaluating model fit

	Fitted Model	
Simulating Model	Dependence	Independence
Dependence	363.74	837.06
Independence	2497.82	2513.77

Figure 4.1: Model selection using DIC



As a general rule, the DIC can be used to determine whether the dependent social

network spatial model is appropriate to fit to a dataset. If the DIC for the fitted dependence model is much smaller than the DIC for a fitted independence model, then the dependence model should be fit. If the DIC for both the fitted dependence and independence models are similar, then the more parsimonious independence model should be chosen over the dependence model.

4.6 Data Analysis

With the simulations demonstrating the feasibility of the model and its ability to properly estimate the correlation of a response over a social network, we now turn our attention to fitting such a model on a relevant dataset. Drabek, Tamminga, Kilijanek, and Adams (1981) have compiled data on the multi-organizational networks of search and rescue activities in Texas. Organizations that interact during search and rescue efforts are considered to have a connection and otherwise are unconnected. Additionally, we have the average reverse rank in decision-making processes during responses to search and rescue initiatives, as judged by objective informants, for each organization. Organizations with higher ranks are, on average, more influential in the decision-making process. We model the natural logarithm of this rank score in order to reduce the skewness in the untransformed values. A total of twenty organizations are included in this network. Our model assumes that organizations interact with others that have similar levels of influence.

We fit an independent social network model without spatial information and present the resulting social space in Figure 4.2. This provides a preliminary representation of how the organizations are related. With the social space representation in place, we then have each organization's point in social space indicate its observed influence score (i.e. the color of each node indicates the observed spatial attribute).

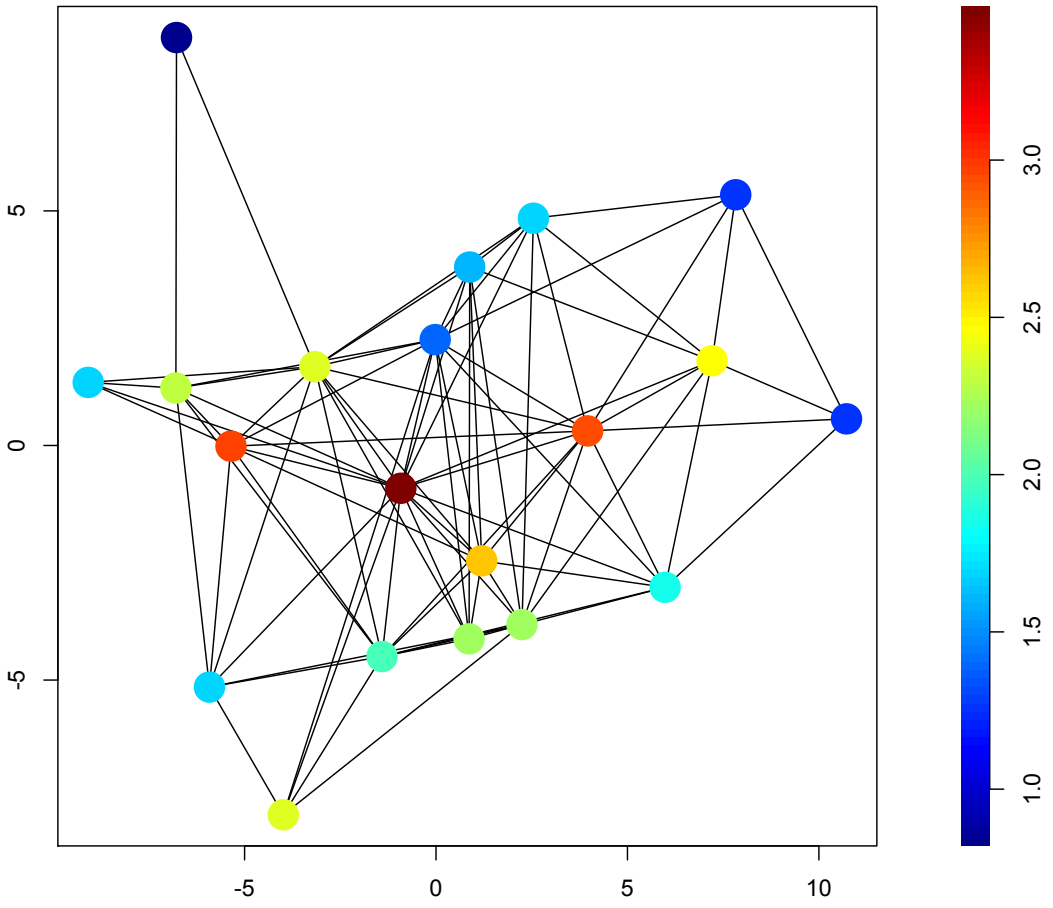
Thus, from preliminarily examining the social space, it is clear that, in general, similarly important organizations also happen to be socially close together. In the middle of the space, there are a few highly connected and highly influential organizations. Along the fringes of the social space are placed organizations that are not as highly connected, and they also tend to be less influential in the decision making process. From this exploration, a spatial effect does seem to be present over the social space.

We apply our social network spatial model to this dataset, jointly modeling the social network and the spatial effects of the emergency organizations and their log rank scores. A total of eighty-four undirected network ties exists between organizations. Based on visually examining the empirical semi-variogram for a fixed social network in Figure 4.3, we adopt an exponential covariance structure for our model, as this appears to reasonably describe the covariance of the observed response with increasing distance.

When the data are fit with a dependent social network spatial model, the DIC is calculated as 463.2. In comparison, the DIC when the data are fit with independent social network and regression models is 533.0. In the case of this emergency organizations data, the DIC indicates that the social network spatial model is appropriate to use for analysis.

We draw 100,000 samples from the posterior distributions of σ^2, τ^2, ϕ, β, $\mathbf{z}$, a, and b. Prior distributions are set as $\sigma^2 \sim IG(2, 0.5)$, $\tau^2 \sim IG(2, 0.01)$, $\phi \sim \text{Unif}(0.1, 1)$, $\beta \sim N(0, 10000)$, $\mathbf{z_i} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right)$, $a \sim \text{Unif}(-5, 5)$, and $b \sim \text{Unif}(-20, 20)$. The sill parameters are chosen such that the prior means are based on the empirical semi-variogram estimates, which in this case are $\tau^2 = 0$ and $\sigma^2 = 0.47$. The multivariate effective sample size diagnostic (Flegal et al., 2016) indicates convergence of the posterior distributions, and trace plots presented in Figure 4.4 visually sup-

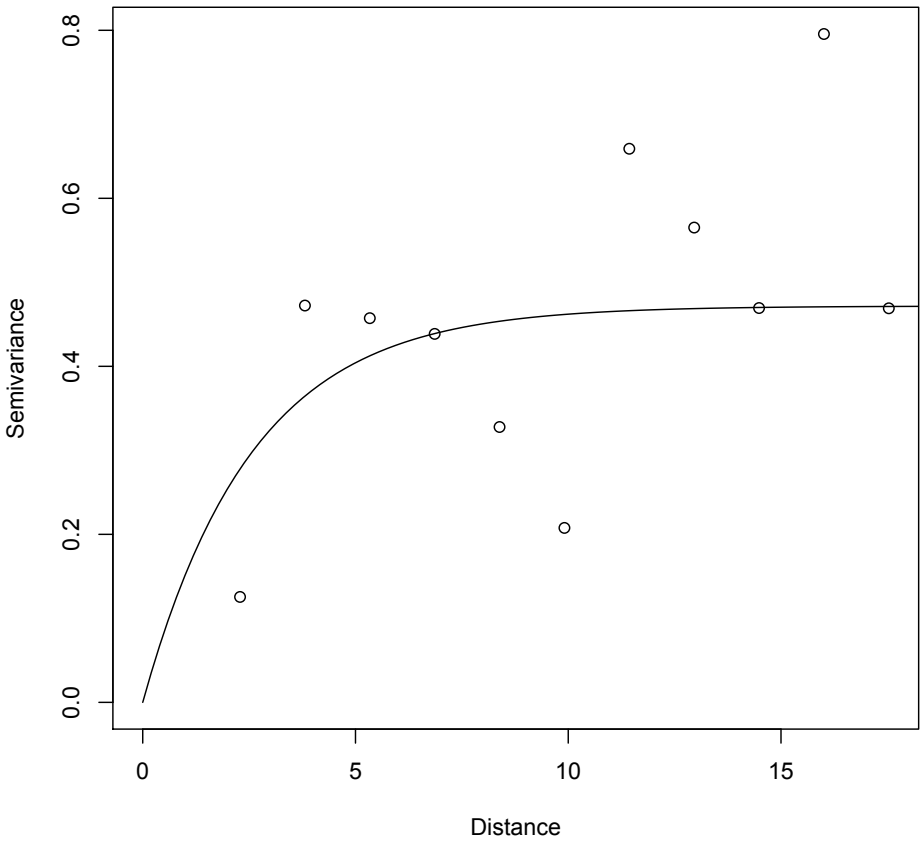
Figure 4.2: Estimated organization social space with measured spatial attribute



port convergence. Our resulting mean parameter estimates, along with 95% posterior intervals, are presented in Table 4.4.

We transform each sample of the posterior distribution of character social space locations, $\mathbf{z}$, using a procrustes transformation with an independent social network fit as the reference. In doing so, we equivalently represent the fitted posterior distribution of social locations with a more reasonable orientation and additionally avoid any issues of a rotating posterior distribution. Based on the posterior distributions of these parameters, composition sampling is used to determine the posterior distribution of the spatial effects, $\mathbf{w}$. The mean spatial effects for each of the twenty organizations

Figure 4.3: Empirical semi-variogram



are graphically displayed in Figure 4.5. This provides a visual representation of how log decision rank varies over this social space, based on our exponential covariance model. We can see that a few highly connected organizations are also influential in the decision-making process. Organizations on the fringe of the social space are less influential, and moderately influential organizations are socially close and interact with each other. Overall, important emergency response organizations are relatively similar to each other. Highly influential organizations tend to be socially close together, indicating that they often respond together. Organizations that work together are typically similar in their influence.

Figure 4.4: Trace plots of posterior samples for model parameters

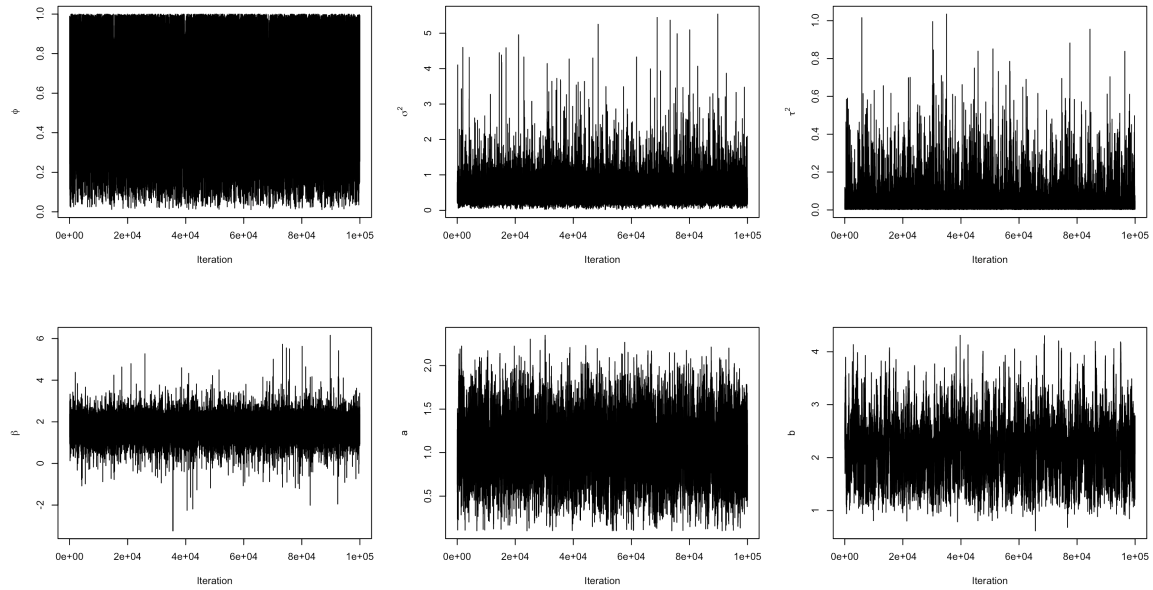
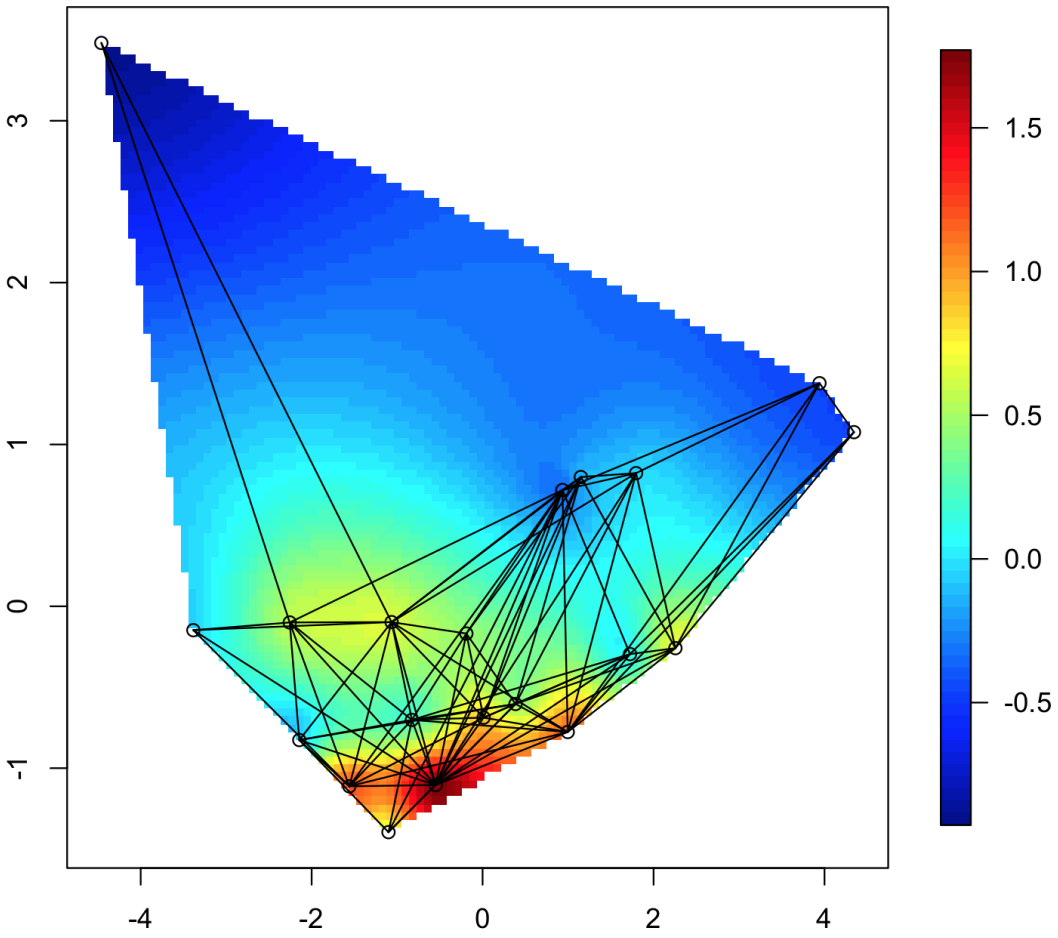


Table 4.4: Posterior summary statistics

	Mean	95% Posterior Interval
$\frac{\sigma^2}{\sigma^2+\tau^2}$	0.976	(0.874, 0.997)
ϕ	0.551	(0.130, 0.976)
β	1.746	(0.874, 2.501)
a	1.067	(0.477, 1.680)
b	2.145	(1.351, 3.173)

In examining Figure 4.5, there is a noticeable organization in the upper-left corner of the social space that is relatively socially far from all other organizations. Additionally, this organization has a low influence score. Upon further examination, this organization is found to be the Pipe Creek Volunteer Fire Department. Comparatively, the Pipe Creek Volunteer Fire Department is geographically far away from many other departments located in Texas. The geographic separation leads to the Pipe Creek Volunteer Fire Department not having as many connections, with its social connections being to the Emergency Medical Services in Kendall County (19.8

Figure 4.5: Mean posterior spatial effects



miles away) and the Bandera County Civil Defense Emergency Management (13.8 miles away). These organizations are two of the closer geographic locations to Pipe Creek. In this case, the geographic orientation of organizations directly impacts the social network connections, and since the Pipe Creek Volunteer Fire Department has few social connections and is not very influential, its estimated social location is relatively far away from all other emergency organizations.

From an inspection of our mean posterior estimates of the model parameters, we notice a few interesting conclusions. First, 97.6% of variability in the model is spatial in nature as opposed to random variability. Thus, the spatial model accounts for much

of the variability in the space, indicating that there is indeed a strong social-spatial effect. This is visually supported by the results in Figure 4.5, where a spatial effect is noticeable. For this spatial effect, we have ϕ estimated at 0.551, leading to a range estimate of $\frac{1}{\phi} = 1.815$. Inspecting Figure 4.5, this range seems reasonable for this social space representation. In general, as we move 1.815 units away from a character, the maximal variability is nearly reached. Characters with similar attributes tend to fall within a range of 1.815 units of each other, whereas those with dissimilar attributes are generally separated by more than a distance of 1.815 units.

Additionally, it should be noted that b is estimated to be positive. This positive estimate of b indicates that for organizations that are the same distance apart, those that have similar influence scores are estimated to have a lower probability of a connection than two similarly spaced organizations with dissimilar influence scores. In the context of emergency organizations, though, this perhaps is not unreasonable. For organizations that are equally socially close together, it is plausible that those with dissimilar influence scores will more often respond to calls together, as there perhaps is less of a need for multiple highly influential organizations who are socially all the same distance apart to respond to the same emergency.

4.7 Discussion

In this chapter, we jointly modeled a social network and the spatial correlations of nodal attributes over this network, providing a comprehensive social network spatial model. By simultaneously updating the social space and spatial correlations, the presented model provides a loop of information in which characters' locations in social space are used as the spatial locations in modeling spatial correlations and observed spatial attributes are used to better inform characters' locations in social

space. Our introduced method draws upon traditional spatial models in which site locations are fixed, but by incorporating the social network as the new basis for locations, additional variability is present. By jointly modeling the social network and spatial correlations, we are able to reasonably describe how an attribute varies over a social network and better model the resulting social space. We presented one such illustration of an attribute varying over a social space through the examination of a network of emergency response organizations in Texas, finding that similarly important organizations tend to be socially close together.

The introduced model marks the first comprehensive unification of spatial modeling and social network analyses where social locations are used as the basis for modeling spatial correlations. The social network spatial model performs well in describing a spatial effect over a social network. Spatial variations are generally well estimated, and the resulting social space closely matches the true orientation of characters. The slight bias in parameter estimation by the social network spatial model should be noted, though, and investigated in future work. Additionally, there is an issue of data availability. For the social network spatial model, a network with a continuous nodal attribute is needed. While such data are seemingly reasonable to collect (e.g. GPA for a class of high school students), privacy becomes a potential problem due to identifiability. As technology increasingly leads to the availability of network data, though, this limitation of data availability will become less troublesome.

With this foundation set, the next step is then to extend more sophisticated traditional spatial modeling techniques to the context of social networks. For example, spatiotemporal modeling of attributes over a social network is of interest. Both spatial attributes and social networks have the potential to change with time. By incorporating this temporal nature of social networks and spatial attributes into a comprehensive model, we could perhaps better describe changes to the space.

Chapter 5

Discussion

5.1 Remarks

Social networks have become increasingly essential to understanding relationships between individuals. The range of applications of network analyses is quite expansive, as networks can be used to more efficiently target marketing efforts, intervene with health behaviors, and introduce policy changes. Interpersonal networks have become more commonly available due to the influx of social media and technology. Models for representing such networks are powerful tools in providing an understanding of how closely related a set of individuals are.

Network data, being present in many contexts, provides an additional source of information that can be introduced into other existing models. We focused in this thesis on incorporating networks into hierarchical Bayesian mixed-membership models (HBMMMs) for the soft clustering of discrete data as well as into spatial models.

5.1.1 Mixed-Membership Models

HBMMMs were introduced as a general class of soft-clustering models, with our focus being specifically on Latent Dirichlet Allocation (LDA), a topic model for a collection of text documents introduced by Blei, Ng, and Jordan (2003). Offering a mechanism for clustering documents across a set of topics, LDA provides thematic context for documents, grouping together those that often discuss similar topics. In many instances, text documents—whether they are Tweets from an inter-related group of users or emails exchanged between colleagues—are linked to specific members of a network. We introduced a social network topic model to incorporate network data that naturally accompanies many text documents into the LDA framework.

We unified LDA with the latent approach for network modeling introduced by Hoff, Raftery, and Handcock (2002) by placing text topics and members of a network together in a single social space. Social space is characterized by distances between entities indicating the strength of their relationships. For example, two members of the network who are close together have a strong connection. By placing the text topics directly into the social space with the network of characters, we modeled how often each character discusses a given topic based on the distance between that character and topic in social space. In doing so, the social network topic model simultaneously updates the network representation and groups together text documents, allowing for information from the network to inform clustering as well as having the text documents provide information into the relationships in the social network.

Providing this feedback loop of shared information between the network and topic models, the social network topic model was shown to outperform LDA and the latent network approach when independently used to cluster documents linked to a network. With our introduced model, we are able to incorporate one of today's greatest sources

of information—that of network data—into an existing clustering framework.

Additionally, we extended the social network topic model to include multiple characters as co-authors of each document. The topic model with social network data was also generalized to allow for network relationships to be included in the context of the Grade of Membership model (Erosheva, 2002), an HBMMM for multiple discrete characteristics measured on a group of individuals. This extension introduced the flexibility by which network information can be incorporated into hierarchical clustering on a more general level. Finally, as an alternative to Markov chain Monte Carlo (MCMC) sampling, we introduced a variational expectation-maximization method for performing posterior inference for parameters in the social network topic model and its variants.

5.1.2 Spatial Models

Spatial models were introduced for describing a characteristic's variation over a given space. Using a hierarchical random-effects model, the Bayesian approach to modeling the spatial effect of an attribute of interest provided the motivation for our work. Since the latent network representation provides a social space based on Euclidean distances, we introduced a method for modeling how an attribute varies over the social network based on traditional spatial models. In particular, we developed a social network spatial model that simultaneously updates the network representation and the spatial effect. This allows for network relationships to provide information to our spatial model, as well as having the spatial model inform the network representation.

Similar to the feedback loop of the unified model introduced in the context of mixed-membership models, the social network spatial model incorporates network information directly into the modeling of a response over a group of people. Such

a model can be applied to any situation in which characteristics are known for each member of a network. The social network spatial model is thus potentially appropriate for describing how body mass index (BMI), political ideology, and moral values, for example, are related to the inter-personal relationships among a group of individuals. We provided an illustrative examination of the social network spatial model through a network of emergency response organizations that each have an influence score, thus demonstrating the ability of the model to represent both the spatial process of the influence score and the social network representation in a cohesive space.

5.2 Future Work

In this thesis, we demonstrated that the social network topic model performs reasonably well. Our next step is to apply the model in a "real data" setting. With the introduction of a variational expectation-maximization inferential method, we have derived a method to overcome the computational complexities of the MCMC inference methods. We now seek to apply this less computationally-intensive method in order to gain insight into a real-world situation of social networks linked to text documents.

Although we generalized the social network topic model to the Grade of Membership model, we would like to further generalize our work to all HBMMMs, providing a flexible framework for the soft clustering of discrete data that is linked to a social network. We believe that this will take an important step forward in introducing social network data as an additional source of information into the clustering context.

Also, additional focus should be placed on model selection within the context of these mixed-membership models with social network data. In any clustering situation, there is a question of how many clusters are appropriate for modeling a set of data. As we discussed in Chapter 2, selecting the appropriate number of clusters currently

remains a challenge. In addition to exploring other model selection criteria, we are also interested in including the number of clusters as a model parameter. Through reversible jump MCMC methods, the social network topic model can be expanded to directly include model selection. Jumping between different numbers of clusters, the inclusion of the reversible jump may offer a way for evaluating the correct number of clusters within the framework of modeling the space.

The social network spatial model has demonstrated its strength for jointly modeling network relationships and spatial effects. We now seek to explore applications of this model in order to better describe how an attribute varies over a network. In particular, we believe there are worthwhile applications in examining how a health measurement, such as BMI, varies over a social network. By understanding how BMI is related to network connections, health interventions can perhaps be introduced to a network in order to most greatly impact changes to BMI.

There is a natural extension to the social network spatial model to that of a spatiotemporal model, with attributes changing over time. Additionally, networks are often dynamic and changing; people make new connections and drop old connections that they no longer value. In future work, we seek to position dynamic networks within the framework of spatial and spatiotemporal models. In this way, we can better model changes to the network as well as improving our understanding of how characteristics vary in social space.

In thesis, our motivation is to make better use of the vast social network data that is increasingly available. Social networks naturally arise in many contexts that otherwise have well developed statistical methodology, and we aim to improve upon such models by incorporating a source of information that is often untouched. We hope the work presented in this thesis provides a foundation for better describing how inter-personal relationships impact the many aspects of our lives.

Bibliography

Airoldi, E., Blei, D., Erosheva, E., & Fienberg, S. (Eds.) (2014). *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC.

Airoldi, E., Fienberg, S., Joutard, C., & Love, T. (2006). Discovering latent patterns with hierarchical bayesian mixed-membership models. Tech. Rep. CMU-ML-06-101, Machine Learning Department, Carnegie Mellon University.

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII-1983*, (pp. 1–198). Springer, Berlin.

Alvarez-Melis, D., & Saveski, M. (2016). Topic modeling in twitter: Aggregating tweets by conversations. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, (pp. 519–522).

Banerjee, S., Carlin, B., & Gelfand, A. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press.

Banerjee, S., & Finley, A. (2009). Introduction to spatial data and models. In *ENAR Hierarchical Modeling and Analysis*.

Barnard, K., Duygulu, N., de Freitas, D., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*(1107-1135).

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Science*, *101*(11), 3747–3752.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, (pp. 993–1022).

Boccaletti, S., Latora, V., Morena, Y., Chavez, M., & Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*, 175–308.

Bohling, G. (2005). Introduction to geostatistics and variogram analysis.

Chang, J., Boyd-Graber, J., & Blei, D. (2009). Connections between the lines: Augmenting social networks with text. In *Proceedings of the 15th CVM SKIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cressie, N. (2003). *Statistics for Spatial Data, Revised Edition*. Wiley-Interscience Publication.

Dickey, J. (1987). Multiple hypergeometric function: Probablistic interpretations and statistical uses. *Journal of the American Statistical Association*, *78*, 628–637.

Drabek, T., Tamminga, H., Kilijanek, T., & Adams, C. (1981). Data from managing multiorganizational emergency responses: Emergent search and rescue network in natural disaster and remote area settings. In *Program on Technology, Environment, and Man Monograph 33*. Institute for Behavioral Science, University of Colorado.

Emch, M., Root, E., Giebultowicz, S., Ali, M., Perez-Heydrich, C., & Yunus, M. (2012). Integration of spatial and social network analysis in disease transmission studies. *Annals of the Association of American Geographers*, *105*(5), 1004–1015.

Erosheva, E. (2002). *Grade of membership and latent structure models with application to disability survey data*. Ph.D. thesis, Department of Statistics, Carnegie Mellon University.

Erosheva, E. (2003). Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, vol. 7, (pp. 501–510).

Erosheva, E., & Fienberg, S. (2005). Bayesian mixed membership models for soft clustering and classification. *Classification—The Ubiquitous Challenge*, (pp. 11–26).

Erosheva, E., Fienberg, S., & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, *1*(2), 502–537.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Science*, *97*(22), 11885–11892.

Estabrooks, P., Bradshaw, M., Dzewaltowski, D., & Smith-Ray, R. (2008). Determining the impact of walk kansas: applying a team-building approach to community physical activity promotion. *Annals of Behavioral Medicine*, *36*, 1–12.

Finley, A., & Banerjee, S. (2013). Point-referenced spatial modeling. In *The SAGE Handbook of Multilevel Modeling*. Sage Publishing.

Finley, A., Banerjee, S., & Carlin, B. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, *19*(4), 1–24.

Finley, A., Banerjee, S., Cook, B., & Bradford, J. (2013). Hierarchical bayesian spatial models for predicting multiple forest variables using waveform lidar, hyperspectral imagery, and large invetory datasets. *International Journal of Applied Earth Observation and Geoinformation*, *22*, 147–160.

Finley, A., Banerjee, S., & Gelfand, A. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, *63*(13), 1–28.

Flegal, J., Hughes, J., & Vats, D. (2016). *mcmcse: Monte Carlo Standard Errors for MCMC*.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis: Second Edition*. CRC Press.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, *101*(Suppl. 1), 5228–5235.

Handcock, M., & Raftery, A. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A*, *170*(2), 301–354.

Hardin, J., Sarkis, G., & URC, P. (2015). Network analysis with the enron email corpus. *Journal of Statistics Education*, *23*(2).

Hoff, P., Raftery, A., & Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098.

Johnson, J., & Orbach, M. (2002). Perceiving the political landscape: ego biases in cognitive political networks. *Social Network*, *24*, 291–310.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233.

Kydros, D., & Anastasiadia, A. (2015). Social newtork analysis in literature: The case of the great eastern by a. emirikos. In K. Dimadis (Ed.) *Proceedings of the 5th European Congress of Modern Greek Studies of the European Society of Modern Greek Studies*, vol. 4, (pp. 681–702).

Liang, J. (2014). *Mining political blogs with network based topic models*. Master's thesis, Department of Statistical Science and Department of Economics, Duke University.

Masala, E., Servetti, A., Basso, S., & De Martin, J. (2014). *Challenges and issues on collecting and analyzing large volumes of network data measurements*, (pp. 203–212). Springer.

McFarland, D., & Brown, D. (1973). Social distances as metric: A systematic introduction to smallest space analysis. In E. Laumann (Ed.) *Bonds of Pluralism: The Form and Substance of Urban Social Networks*. Wiley.

Mukherjee, I., & Blei, D. (2008). Relative performance guarantees for approximate inference in latent dirichlet allocation. In *Neural Information Processing System*.

Myers, D. (1989). To be or not to be...stationary? that is the question. *Mathematrical Geology*, *21*(3), 347–362.

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., N., S., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1).

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

Oh, M., & Raftery, A. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, *96*, 1031–1044.

Plummer, M. (2013). *JAGS version 3.4.0 user manual*.

Porteous, I., Newman, D., Ihler, A., Asucion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 569–577).

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Radil, S. (2011). *Spatializing social networks: Making space for theory in spatial analysis*. Ph.D. thesis, Unversity of Illinois at Urbana-Champaign.

Raftery, A., & Lewis, S. (1992). One long run with diagnostics: Implemenation strategies for markov chain monte carlo. *Statistical Science*, *7*, 493–497.

Raftery, A., & Lewis, S. (1995). The number of iterations, convergence diagnostics and generic metropolis algorithms. In W. Gilks, D. Spiegelhalter, & S. Richardson (Eds.) *Pratical Markov Chain Monte Carlo*. Chapman and Hall.

Shelton, R., McNeill, L., Puleo, E., Wolin, K., Emmons, K., & Bennett, G. (2011). The association between social factors and physical activity among low-income adults living in public housing. *American Journal of Public Health*, *101*, 2102–2110.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B*, *41*, 217–229.

Sontag, D., & D., R. (2011). Complexity of inference in latent dirichlet allocation. In *Neural Information Processing System*.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*(4), 583–639.

Teh, Y., Newman, D., & Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 19*.

Wallach, H., Mimno, D., & McAllum, A. (2009). Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems 22*, (pp. 1973–1981).

Yu, G., Renton, A., Schmidt, E., Tobi, P., Bertotti, M., Watts, P., & Lais, S. (2011). A multilevel analysis of the association between social networks and support on leisure time physical activity: Evidence from 40 disadvantaged areas in london. *Health Place*, *17*(1023-1029).

Yuan, B., & Wang, B. (2007). Growing directed networks: organization and dynamics. *New Journal of Physics*, *9*.

Zheng, T., Salganik, M., & Gelfand, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*, *101*(474), 409–423.

Appendix: Variational Inference

This appendix is dedicated to the derivation of the variational inference algorithm presented in Section 3.2 of Chapter 3. In particular, we focus here on the derivation of the lower bound of the log-likelihood, the update equations for the variational parameters, and the pseudo maximum likelihood estimates for model hyper-parameters. Recall that the variational posterior distribution is formulated as follows:

$$
\begin{aligned}
& q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_1, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_1, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}) \qquad (1) \\
& \quad = q(a|\tilde{\omega}) \left(\prod_{i=1}^{N} \left(q(\mathbf{z}_i | \tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}, \tilde{\nu}_{2i}) \prod_{r=1}^{R_i} q(u_i^r | \tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK}) \right) \right) \\
& \qquad \times \left(\prod_{k=1}^{K} q(\mathbf{z_{t_k}} | \tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}, \tilde{\eta}_{2k}) q(\beta_k | \tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV}) \right).
\end{aligned}
$$

This variational posterior distribution is used as an approximation to the true posterior distribution, $P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \mathbf{x}, Y, \boldsymbol{\eta}_\mathbf{1}, \eta_2, \boldsymbol{\nu}_\mathbf{1}, \nu_2, \phi, \omega)$ and eliminates the dependencies between parameters that makes analytic inference based on the true posterior distribution infeasible. The set of variational parameters $\left(\tilde{\boldsymbol{\eta}}_1, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_1, \tilde{\nu}_2, \tilde{\beta}, \tilde{\tau}, \tilde{\omega} \right)$ are optimized in order to yield a variational posterior distribution that is a reasonable approximation to the true posterior distribution.

To begin optimizing the variational parameters, we bound the log-likelihood using Jensen's inequality (Jordan, Ghahramani, Jaakola, and Saul, 1999). The bound on the log-likelihood is found as follows:

$$
\begin{aligned}
& log\left(P(\mathbf{x}, Y | \boldsymbol{\eta}_\mathbf{1}, \eta_2, \boldsymbol{\nu}_\mathbf{1}, \nu_2, \phi, \omega)\right) \qquad (2) \\
& = log \int\int\int\int \sum_{\mathbf{u}} P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_\mathbf{1}, \eta_2, \boldsymbol{\nu}_\mathbf{1}, \nu_2, \phi, \omega) \, d\mathbf{z_t} d\mathbf{z} d\beta da \\
& = log \int\int\int\int \sum_{\mathbf{u}} P(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_\mathbf{1}, \eta_2, \boldsymbol{\nu}_\mathbf{1}, \nu_2, \phi, \omega) \\
& \qquad \times \frac{q\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_1, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_1, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}\right)}{q\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_1, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_1, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}\right)} d\mathbf{z_t} d\mathbf{z} d\beta da
\end{aligned}
$$

$$
\begin{aligned}
\geq & \int\int\int\int \sum_{\mathbf{u}} q\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}\right) \\
& \qquad \times log\left(P\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega\right)\right) d\mathbf{z_t} d\mathbf{z} d\beta da \\
& \qquad - \int\int\int\int \sum_{\mathbf{u}} q\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}\right) \\
& \qquad log(q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega})) d\mathbf{z_t} d\mathbf{z} d\beta da \\
= & \mathbf{E}_q\left(log\left(P\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega\right)\right)\right) \\
& - \mathbf{E}_q(log(q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}))).
\end{aligned}
$$

For any variational posterior distribution $q(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega})$, we can thus obtain a lower bound on the log-likelihood. The difference between the left-hand and the right-hand sides of the expression (2) is the Kullback-Leibler divergence between the variational and true posterior distributions (Blei, Ng, and Jordan, 2003). Denoting the lower bound as $L(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}; \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega)$, we have have the following result:

$$
\begin{aligned}
& log\left(P\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega\right)\right) \qquad (3) \\
& = L(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}; \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega) \\
& + D\left(q\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}\right) || P\left(\mathbf{z_t}, \mathbf{z}, \beta, \mathbf{u}, a, \mathbf{x}, Y | \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega\right)\right).
\end{aligned}
$$

As a result of expression (3), maximizing the lower bound with respect to the variational parameters is equivalent to minimizing the Kullback-Leibler divergence between the the variational and true posterior distributions. Ultimately, we maximize the lower bound in our optimization of the variational parameters, thus minimizing the Kullback-Leibler divergence.

Factorizing both the true and variational posterior distributions, we can expand the lower bound for a more convenient representation:

$$
\begin{aligned}
& L(\tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2, \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2, \tilde{\phi}, \tilde{\tau}, \tilde{\omega}; \boldsymbol{\eta}_{\mathbf{1}}, \eta_2, \boldsymbol{\nu}_{\mathbf{1}}, \nu_2, \phi, \omega) \qquad (4) \\
& \quad = \mathbf{E}_q\left(log\left(P\left(\mathbf{z_t} | \boldsymbol{\eta}_{\mathbf{1}}, \eta_2\right)\right)\right) + \mathbf{E}_q\left(log\left(P\left(\mathbf{z} | \boldsymbol{\nu}_{\mathbf{1}}, \nu_2\right)\right)\right) + \mathbf{E}_q\left(log\left(P\left(\beta | \phi\right)\right)\right) \\
& \qquad + \mathbf{E}_q\left(log\left(P\left(\mathbf{u} | \mathbf{z_t}, \mathbf{z}\right)\right)\right) + \mathbf{E}_q\left(log\left(P\left(a | \omega\right)\right)\right) + \mathbf{E}_q\left(log\left(P\left(\mathbf{x} | \mathbf{u}, \beta\right)\right)\right) \\
& \qquad + \mathbf{E}_q\left(log\left(P\left(Y | \mathbf{z_t}, \mathbf{z}, a\right)\right)\right) - \mathbf{E}_q\left(log\left(q\left(\mathbf{z_t} | \tilde{\boldsymbol{\eta}}_{\mathbf{1}}, \tilde{\eta}_2\right)\right)\right) - \mathbf{E}_q\left(log\left(q\left(\mathbf{z} | \tilde{\boldsymbol{\nu}}_{\mathbf{1}}, \tilde{\nu}_2\right)\right)\right) \\
& \qquad - \mathbf{E}_q\left(log\left(q\left(\beta | \tilde{\phi}\right)\right)\right) - \mathbf{E}_q\left(log\left(q\left(\mathbf{u} | \tilde{\tau}\right)\right)\right) - \mathbf{E}_q\left(log\left(q\left(a | \tilde{\omega}\right)\right)\right).
\end{aligned}
$$

To calculate this lower bound, we begin by setting the distributional assumptions and writing out the distributions on the log scale. Then, we present expressions for the necessary expectations in equation (4).

Following with the specification in Chapter 3, we have the following distributional forms for the true prior distributions:

$$\begin{aligned}
\mathbf{z_{t_k}}|\boldsymbol{\eta_1}, \eta_2 &\sim N\left(\begin{pmatrix}\eta_{1[1]}\\ \eta_{1[2]}\end{pmatrix}, \eta_2^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)\\
\mathbf{z}_i|\boldsymbol{\nu_1}, \nu_2 &\sim N\left(\begin{pmatrix}\nu_{1[1]}\\ \nu_{1[2]}\end{pmatrix}, \nu_2^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)\\
\beta_k|\phi &\sim \text{Dirichlet}(\phi)\\
a|\omega &\sim \text{Unif}(-\omega, \omega).
\end{aligned}$$

Additionally, we have the following specifications for the variational distributions:

$$\begin{aligned}
\mathbf{z_{t_k}}|\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}, \tilde{\eta}_{2k} &\sim N\left(\begin{pmatrix}\tilde{\eta}_{1k[1]}\\ \tilde{\eta}_{1k[2]}\end{pmatrix}, \tilde{\eta}_{2k}^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)\\
\mathbf{z}_i\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}, \tilde{\nu}_{2i} &\sim N\left(\begin{pmatrix}\tilde{\nu}_{1i[1]}\\ \tilde{\nu}_{1i[2]}\end{pmatrix}, \tilde{\nu}_{2i}^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)\\
\beta_k|\tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV} &\sim \text{Dirichlet}(\tilde{\phi}_{k1}, ..., \tilde{\phi}_{kV})\\
u_i^r|\tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK} &\sim \text{Multinomial}(\tilde{\tau}_{ir1}, ..., \tilde{\tau}_{irK})\\
a|\tilde{\omega} &\sim \text{Unif}(-\tilde{\omega}, \tilde{\omega}).
\end{aligned}$$

As explicitly written above, note that $\boldsymbol{\eta_1}$, $\boldsymbol{\nu_1}$, $\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}$ (for $k = 1, ..., K$), and $\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}$ (for $i = 1, ..., N$) are two-dimensional parameters. Both dimensions of each of these parameters will be updated simultaneously. Additionally, ϕ is a symmetric parameter to be repeated V times.

Next, we formally specify the explicit form of each distribution, on the log scale:

$$\begin{aligned}
log\left(P\left(\mathbf{z_t}|\boldsymbol{\eta_1}, \eta_2\right)\right) &= \sum_{k=1}^{K}\left(-log(2\pi) - 2log(\eta_2) - \frac{1}{2\eta_2^2}\left(\mathbf{z_{t_k}} - \boldsymbol{\eta_1}\right)^T\left(\mathbf{z_{t_k}} - \boldsymbol{\eta_1}\right)\right)\\
log\left(P\left(\mathbf{z}|\boldsymbol{\nu_1}, \nu_2\right)\right) &= \sum_{i=1}^{N}\left(-log(2\pi) - 2log(\nu_2) - \frac{1}{2\nu_2^2}\left(\mathbf{z}_i - \boldsymbol{\nu_1}\right)^T\left(\mathbf{z}_i - \boldsymbol{\nu_1}\right)\right)\\
log\left(P\left(\beta|\phi\right)\right) &= \sum_{k=1}^{K}\left(log\Gamma(V\phi) - Vlog\Gamma(\phi) + \sum_{v=1}^{V}(\phi - 1)log(\beta_{k[v]})\right)\\
log\left(P\left(a|\omega\right)\right) &= -log(2\omega)\\
log\left(P\left(\mathbf{u}|\mathbf{z_t}, \mathbf{z}\right)\right) &= \sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K} u_{i[k]}^r\left(-||\mathbf{z_{t_k}} - \mathbf{z_i}|| - log\left(\sum_{c=1}^{K} e^{-||\mathbf{z_{t_c}} - \mathbf{z_i}||}\right)\right)\\
log\left(P\left(\mathbf{x}|\mathbf{u}, \beta\right)\right) &= \sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K}\sum_{v=1}^{V} u_{i[k]}^r x_{i[v]}^r log(\beta_{k[v]})
\end{aligned}$$

$$log\left(P\left(Y|\mathbf{z_t},\mathbf{z},a\right)\right)=\sum_{i=1}^{N}\sum_{i'=i+1}^{N}\left(y_{ii'}\left(a\left(1-||\mathbf{z}_i-\mathbf{z}_{i'}||\right)\right)\right)-log\left(1+e^{a(1-||\mathbf{z}_i-\mathbf{z}_{i'}||)}\right)$$

$$log\left(q\left(\mathbf{z_t}|\tilde{\boldsymbol{\eta}}_{\mathbf{1}},\tilde{\eta}_2\right)\right)=\sum_{k=1}^{K}\left(-log(2\pi)-2log(\tilde{\eta}_{2k})-\frac{1}{2\tilde{\eta}_{2k}}\left(\mathbf{z_{t_k}}-\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}\right)^T\left(\mathbf{z_{t_k}}-\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}\right)\right)$$

$$log\left(q\left(\mathbf{z}|\tilde{\boldsymbol{\nu}}_{\mathbf{1}},\tilde{\nu}_2\right)\right)=\sum_{i=1}^{N}\left(-log(2\pi)-2log(\tilde{\nu}_{2i})-\frac{1}{2\tilde{\nu}_{2i}}\left(\mathbf{z}_i-\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}\right)^T\left(\mathbf{z}_i-\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}\right)\right)$$

$$log\left(q\left(\beta|\tilde{\phi}\right)\right)=\sum_{k=1}^{K}log\Gamma\left(\sum_{v=1}^{V}\tilde{\phi}_{kv}\right)-\sum_{v=1}^{V}log\Gamma\left(\tilde{\phi}_{kv}\right)+\sum_{v=1}^{V}\left(\tilde{\phi}_{kv}-1\right)log\left(\beta_{k[v]}\right)$$

$$log\left(q\left(a|\tilde{\omega}\right)\right)=-2log\left(\tilde{\omega}\right)$$

$$log\left(q\left(\mathbf{u}|\tilde{\tau}\right)\right)=\sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K}u_{i[k]}^{r}log\left(\tilde{\tau}_{irk}\right).$$

Taking the expectations with respect to the variational distribution gives us the necessary components for the lower bound on the true likelihood. In this development, we use results of Salter-Townshend and Murphy (2012) and Blei et al. (2003) to compute the expectation of social network parameters and β.

$$\mathbf{E}_q\left(log\left(P\left(\mathbf{z_t}|\boldsymbol{\eta}_{\mathbf{1}},\eta_2\right)\right)\right)=-Klog(2\pi)-2Klog(\eta_2)$$
$$-\frac{1}{2\eta_2^2}\sum_{k=1}^{K}\left(\tilde{\eta}_{2k}^2+\left(\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}-\boldsymbol{\nu}_{\mathbf{1}}\right)^T\left(\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}-\boldsymbol{\eta}_{\mathbf{1}}\right)\right)$$

$$\mathbf{E}_q\left(log\left(P\left(\mathbf{z}|\boldsymbol{\nu}_{\mathbf{1}},\nu_2\right)\right)\right)=-Nlog(2\pi)-2Nlog(\nu_2)$$
$$-\frac{1}{2\nu_2^2}\sum_{i=1}^{N}\left(\tilde{\nu}_{2i}^2+\left(\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}-\boldsymbol{\nu}_{\mathbf{1}}\right)^T\left(\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}-\boldsymbol{\nu}_{\mathbf{1}}\right)\right)$$

$$\mathbf{E}_q\left(log\left(P\left(\beta|\phi\right)\right)\right)=Klog\left(V\phi\right)-KVlog\Gamma\left(\phi\right)$$
$$+\left(\phi-1\right)\sum_{k=1}^{K}\sum_{v=1}^{V}\left(\Psi\left(\tilde{\phi}_{kv}\right)-\Psi\left(\sum_{j=1}^{V}\tilde{\phi}_{kj}\right)\right)$$

$$\mathbf{E}_q\left(log\left(P\left(a|\omega\right)\right)\right)=-2log(\omega)$$

$$\mathbf{E}_q\left(log\left(P\left(\mathbf{u}|\mathbf{z_t},\mathbf{z}\right)\right)\right)=\sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K}-\tilde{\tau}_{irk}\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{k}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}||+2\left(\tilde{\eta}_{2k}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}$$
$$-\tilde{\tau}_{irk}log\left(\sum_{c=1}^{K}e^{-\left(||\tilde{\boldsymbol{\eta}}_{\mathbf{1}\boldsymbol{c}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}||+2\left(\tilde{\eta}_{2c}^2+\tilde{\nu}_{2i}^2\right)\right)^{\frac{1}{2}}}\right)$$

$$\mathbf{E}_q\left(log\left(P\left(\mathbf{x}|\mathbf{u},\beta\right)\right)\right)=\sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K}\sum_{v=1}^{V}\tilde{\tau}_{irk}x_{i[v]}^{r}\left(\Psi\left(\tilde{\phi}_{kv}\right)-\Psi\left(\sum_{j=1}^{V}\tilde{\phi}_{kj}\right)\right)$$

$$\mathbf{E}_q\left(log\left(P\left(Y|\mathbf{z_t},\mathbf{z},a\right)\right)\right) = \sum_{i=1}^{N}\sum_{i'=i+1}^{N} y_{ii'}\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1}i'}||+2\left(\tilde{\nu}_{2i}^2-\tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}\right) - log\left(1+e^{-\tilde{\omega}\left(1-\left(||\tilde{\boldsymbol{\nu}}_{\mathbf{1}\boldsymbol{i}}-\tilde{\boldsymbol{\nu}}_{\mathbf{1}i'}||+2\left(\tilde{\nu}_{2i}^2-\tilde{\nu}_{2i'}^2\right)\right)^{\frac{1}{2}}\right)}\right)$$

$$\mathbf{E}_q\left(log\left(q\left(\mathbf{z_t}|\tilde{\boldsymbol{\eta}}_{\mathbf{1}},\tilde{\eta}_2\right)\right)\right) = -Klog(2\pi) - 2\sum_{k=1}^{K} log(\tilde{\eta}_{2k}) - 1$$

$$\mathbf{E}_q\left(log\left(q\left(\mathbf{z}|\tilde{\boldsymbol{\nu}}_{\mathbf{1}},\tilde{\nu}_2\right)\right)\right) = -Nlog(2\pi) - 2\sum_{i=1}^{N} log(\tilde{\nu}_{2i}) - 1$$

$$\mathbf{E}_q\left(log\left(q\left(\beta|\tilde{\phi}\right)\right)\right) = \sum_{k=1}^{K}\left(log\Gamma\left(\sum_{v=1}^{V}\tilde{\phi}_{kv}\right) + \sum_{v=1}^{V}\left(\left(\tilde{\phi}_{kv}-1\right)\left(\Psi\left(\tilde{\phi}_{kv}\right)-\Psi\left(\sum_{j=1}^{V}\tilde{\phi}_{kj}\right)\right)-log\Gamma\left(\tilde{\phi}_{kv}\right)\right)\right)$$

$$\mathbf{E}_q\left(log\left(q\left(a|\tilde{\omega}\right)\right)\right) = -log(2\tilde{\omega})$$

$$\mathbf{E}_q\left(log\left(q\left(\mathbf{u}|\tilde{\tau}\right)\right)\right) = \sum_{i=1}^{N}\sum_{r=1}^{R_i}\sum_{k=1}^{K}\tilde{\tau}_{irk}log\left(\tilde{\tau}_{irk}\right)$$

The above expectations are combined based on equation (4) to provide an expression for the lower bound. We then determine which values of the variational parameters maximize this lower bound. The optimization of each of the variational parameters is achieved by setting the derivative of $L(\tilde{\boldsymbol{\eta}}_{\mathbf{1}},\tilde{\eta}_2,\tilde{\boldsymbol{\nu}}_{\mathbf{1}},\tilde{\nu}_2,\tilde{\phi},\tilde{\tau},\tilde{\omega};\boldsymbol{\eta}_{\mathbf{1}},\eta_2,\boldsymbol{\nu}_{\mathbf{1}},\nu_2,\phi,\omega)$ with respect to each of the variational parameters equal to zero and solving for the resulting form of the variational parameters. The optimizing equations for these parameters are dependent on the current parameter values of the true model hyper-parameters, $(\boldsymbol{\eta}_{\mathbf{1}},\eta_2,\boldsymbol{\nu}_{\mathbf{1}},\nu_2,\phi,\omega)$, as well on the current values of the other variational parameters. In the optimization routine, the variational parameters are updated sequentially. The resulting expressions for updating each of the variational parameters are presented in Section 3.2 of Chapter 3.

To complete the variational EM algorithm, it remains for us to determine pseudo maximum likelihood estimates for the model hyper-parameters, $(\boldsymbol{\eta}_{\mathbf{1}},\eta_2,\boldsymbol{\nu}_{\mathbf{1}},\nu_2,\phi,\omega)$. To estimate these hyper-parameters, we maximize the variational lower bound on the likelihood based on the current estimates of the variational parameters. Thus, we take the derivative of the lower bound with respect to the model hyper-parameters. Setting each of these derivatives equal to zero and solving for an expression for each model hyper-parameters yields the pseudo maximum likelihood estimates of interest. The resulting expressions for the pseudo maximum likelihood estimates of the model hyper-parameters are presented in Section 3.2 of Chapter 3.

Iterating between updating the variational parameters and updating the model hyper-parameters ultimately results in the necessary components for posterior inference. In particular, we are left with posterior point estimates of the model hyper-parameters, as well as having an approximate posterior distributions, complete with point estimates of the variation posterior distribution hyper-parameters. Thus, the variational EM algorithm provides us with all the information necessary regarding model parameters of interest, particularly $(\mathbf{z_t}, \mathbf{z}, \beta, a)$.