UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MODELING AND PREDICTION OF TIME SERIES OF DIRECTED BINARY NETWORKS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Brenda Betancourt

September 2015

The Dissertation of Brenda Betancourt
is approved:

Professor Abel Rodríguez, Chair

Professor Athanasios Kottas

Professor Raquel Prado

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Modeling and Prediction of Time Series of Directed Binary Networks

by

Brenda Betancourt

Over the last few years modeling of temporal evolution of network data has become a relevant problem for different applications. In this work, we develop novel statistical methods for modeling time series of directed binary networks. The main interests are identifying events associated with structural changes over time and perform short-term link prediction of the network in future periods.

First, we introduce a Bayesian hidden Markov model that uses a stochastic block-model to describe the community structure of the network during each period. This model allow us to monitor structural changes in the network and also perform accurate short-term predictions of future links. As an alternative for link prediction, we propose two multinomial logistic regression models using different lasso type penalties including an extension of the autologistic model for network data. In this setting, we are able to provide fast computational algorithms for estimation and prediction using optimization and full Bayesian inference. The performance of the models is illustrated using both simulated and real data from a financial trading network in the NYMEX natural gas futures market.

Chapter 1

Introduction

1.1 Background

Network data, in which observations correspond to the interactions among a group of nodes, has become pervasive in disciplines as diverse as social, physical and biological sciences. Accordingly, there has been a growing interest in developing tools for the analysis of network data, particularly from a model-based perspective (for excellent reviews see Newman, 2003, Goldenberg et al., 2009 and Snijders, 2011). The focus on this work is on models for time series of binary directed networks that involve the same set of subjects at each time point. A primary goal in the analysis of this type of dynamic network data is link prediction at future times, going as far as predicting the structure of the whole network. An additional goal is to provide models to understand the mechanistic effects that drive the evolution of the network such as change-points in the network dynamic, community structure, reciprocity and inter-temporal transitivity. In the sequel, we consider a sequence of binary directed networks $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, each one observed over a common set of n nodes. The adjacency matrix of the network at time t is therefore an $n \times n$ binary matrix $\mathbf{Y}_t = [y_{i,j,t}]$, where $y_{i,j,t} = 1$ if there is a link directed from node i to node j at time t, and $y_{i,j,t} = 0$ otherwise. We adopt the convention $y_{i,i,t} \equiv 0$ so that there are no loops within the network.

In the following sections, we present a summary of basic descriptive measures for network analysis, and literature reviews for stochastic models for static and dynamic network data.

1.1.1 Descriptive measures for networks

Properties of networks and its elements are useful to explore the overall structure of a system. Some characteristics and measures of interest for vertices and networks are the following:

- **Degree:** is the number of edges connected to a node, for directed networks a node has in-degree (in-coming edges) and out-degree (out-going edges). For a network with n nodes, the node degree can take values from 0 (isolated node) to $n-1$. The degree of a node can be considered as a measure of centrality of the node in the network.

- **Betweenness:** is another measure of centrality for the nodes in the network, it is most often defined as
$$C_B(v) = \sum_{i \neq j \neq v} \frac{\mathfrak{g}(i,j|v)}{\mathfrak{g}(i,j)},$$
where $\mathfrak{g}(i,j) = \sum_v \mathfrak{g}(i,j|v)$, $\mathfrak{g}(i,j|v)$ is the total number of geodesic paths going through the node v, where a geodesic path is the shortest path between a pairs of nodes.

- **Clustering Coefficient:** The clustering coefficient measures the *transitivity* of the network as the probability with which two neighbors of the same node will be neighbors too. It is calculated as

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes}},$$

where a connected triple is a subgraph of three nodes connected by two edges, and a triangle is a set of three nodes connected to each other by three edges.

Other relevant properties of a network are its resilience, community structure and mixing patterns. Centrality measures such as the degree and betweenness reflect the importance of a node in the topology of the network, allowing us to identify subjects that can affect the network resilience by being removed causing changes in the network connectivity and global structure. In addition to computing centrality measures for individual nodes, it is often of interest to identify clusterings of nodes with a large number of edges between them and a low connectivity with other groups. This behavior is commonly observed in social networks reflecting a community structure which is natural in this type of data. More generally, it is common to observe selective linking between actors leading to mixing patterns in the network. For example, assortative mixing or homophily occurs when actors with similar attributes (e.g. degree, race, age) are more likely to relate. This phenomenon of assortativity is one possible source of community structure and consequently of transitivity in the network even though a network can have assortative mixing patterns and no community structure.

- **Assortativity Coefficient:** Assortative mixing according to a discrete node feature is characterized by a quantity e_{ij} which defines the fraction of edges in the network that connect a node of type i to one of type j.

The assortativity coefficient is $r = \dfrac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$, where $\sum_{ij} e_{ij} = 1$, and a_i and b_i are the fraction of each type of end of an link that is attached to nodes of type i. A value of $r = 1$ indicates perfect assortative mixing, $r = 0$ when there is no assortative mixing, and if the network is perfectly disassortative then r is negative and has the value $r_{\min} = \dfrac{\sum_i a_i b_i}{1 - \sum_i a_i b_i}$.

1.1.2 Models for static networks

In this work we move beyond descriptive network summaries to focus on stochastic models for array-valued data that place a probability distribution on the full network. The statistical literature on stochastic models for individual networks is well developed. The simplest such model is the class Erdös-Rényi model (Erdös & Rényi, 1959), which assumes that interactions among any two traders occur independently and with constant probability that is independent of the identity of the nodes. This class of models, although well studied from a theoretical perspective, is too simplistic to accommodate most realistic networks. As an alternative, Frank & Strauss (1986) proposed the class of exponentially weighted random graphs, also called p^* models. These models formalize the use of summary measures by including them as sufficient statistics in exponential-family models. On the other hand, Holland & Leinhardt (1981) proposed the class of p_1 models, which extend generalized linear models to array-valued data. The model assumes independence between dyads and focuses on modeling the pairs $(y_{i,j}, y_{j,i})$ jointly for $i < j$, $j = 1, \dots, n$ as follows

$$p\left(y_{i,j}, y_{j,i}\right) \propto \exp\left\{\alpha_{i,j} y_{i,j} + \beta_{i,j} y_{j,i} + \gamma_{i,j} y_{i,j} y_{j,i}\right\}. \tag{1.1}$$

A related approach was introduced in Hoff et al. (2002) using the concept of latent social space models in which the probability of a link between nodes increases as they occupy closer positions in latent social space.

Wang & Wong (1987) introduced the concept of stochastic blockmodels. Stochastic blockmodels rely on the concept of structural equivalence to identify groups of nodes with similar interaction patterns. Indeed, the partitions induced by stochastic blockmodels are driven not only by the internal relations within a particular group of nodes, but also by the interactions among these groups (Wasserman & Faust, 1994). Model-based stochastic blockmodels have been developed as array-valued extensions of traditional mixture models. For example, Nowicki & Snijders (2001) proposed a simple Bayesian model that uses a finite mixture model and a Dirichlet prior for the probabilities of the latent classes. An extension of this model that relies on infinite mixture models based on the Dirichlet process have been proposed by Kemp et al. (2006) and Xu et al. (2006). More recently Airoldi et al. (2008) introduced the idea of mixed membership stochastic blockmodels for binary networks wherein the actors can belong to more than one latent class to explore subjects with multiple roles in the network.

1.1.3 Models for dynamic networks

The class of p_1 models and other basic static models have naturally been extended to a dynamic setting in the past (e.g., see Banks & Carley, 1996; Goldenberg et al., 2009; Kolacyzk, 2009). Other relevant approaches include the dynamic versions of the latent space model of Hoff et al. (2002) presented in Sarkar & Moore (2005) and Sewell & Chen (2015), and the work of Xing et al. (2010) presenting the temporal extension of the stochastic blockmodel for community identification in social networks. On the other hand, Hanneke et al. (2010) presents a temporal version of the Exponential random graph model (ERGM). This temporal model can be used to infer links but its prediction ability is poor unless node attributes or dyadic covariates are included in the model in addition to traditional static network statistics (e.g. reciprocity, transitivity and popularity statistics). Cranmer & Desmarais (2011) present a more general temporal ERGM that includes node and dyad-level covariates with applications to political science (see also Snijders et al., 2010). In this extension, the square root of the indegree and outdegree are added as node attributes at every time point, and functions of past networks can be utilized as a dyadic covariates. For example, the inclusion of a lagged network in a similar fashion to an autoregressive term in regression analysis, or the delayed reciprocation of edges are common choices.

Moreover, approaches for link prediction in dynamic networks with a large number of nodes have been recently explored using traditional spatial and time series models. Sarkar et al. (2012) presents a nonparametric link prediction algorithm for sequences of directed binary networks where each observation in time is modeled using a moving window, and the function is estimated through kernel regression. They also incorporate pair specific features, and a spatial dimension using local neighborhoods for each node. Huang & Lin (2009) present an autoregressive integrated moving average model (ARIMA), and combine it with link occurrence scores based on similarity indices of network topology measures for link prediction in temporal weighted networks (see also da Silva & Bastos, 2012). More recently, Bliss et al. (2014) proposed a method based on similarity indices and node attributes joined with a covariance matrix adaptation evolution strategy also for link prediction.

1.2 Financial Trading Networks

Part of the motivation for this work was the unexplored area of financial trading networks using stochastic models and time series analysis. Financial trading networks are directed graphs in which nodes correspond to traders participating in a financial market, and edges represent pairwise buy-sell transactions among them that occur within a period of time. Financial trading networks contain important information about patterns of order execution in order-driven markets; hence, they provide insights into aspects of market microstructure such as market frictions, trading strategies, and systemic risks.

1.2.1 Market friction and systemic risk

Consider first the role of financial trading networks in understanding the effect of market frictions on market microstructure. In the absence of market frictions, we could expect orders from different traders to be matched randomly. However, real trading networks often exhibit features such as elevated transitivity or preferential attachment among certain groups of actors (Adamic et al., 2010), which are inconsistent with random matching. In the case of open-outcry markets, these features can be partially explained by sociological factors (for example, see Zaloom, 2004). Alternative explanations include the effect of different market roles (e.g., liquidity providers/takers) or trading strategies (e.g., long vs. short strategies), see Ozsoylev et al., 2010 or Hatfield et al., 2012.

Financial trading networks also provide information that is key in the assessment of systemic risks. Analysis of the evolution of financial trading networks can aid in tests of financial market stability (or fragility as it may be) by financial regulators to ensure that events such as a large trader failures do not serve to destabilize financial markets. In the event of a large trader failure, an understanding of their network will help guide regulators through the process of unwinding their positions and may dictate whether those positions are unwound in the open market or through a transfer to a suitable counterparty (Boyd et al., 2011). Financial trading networks can also be used to identify important traders that play a critical role in the market (for example, by acting as de facto market makers or liquidity providers). In addition, they can also help us identify frequent counterparties of specific traders which may aid in regulatory oversight by federal agencies and market exchanges alike; price distortion and manipulation may be more likely between frequent counterparties than by one agent acting in isolation (Harris et al., 1994).

1.2.2 Models for financial trading networks

The literature on the mathematical modeling of financial trading networks is limited. Theoretical approaches that explain the structure of a financial network as the outcome of a game have recently been developed (e.g., see Ozsoylev et al., 2010 and Hatfield et al., 2012), but they are of limited practical applicability. Most of the empirical work on trading networks has focused on the use of summary statistics such as degree distributions, average betweenness and clustering coefficients (Newman, 2003; Adamic et al., 2010). These type of approaches provide some interesting insights into market microstructure, but suffer from two main drawbacks. First, the summary statistics to be monitored need to be carefully chosen to ensure that relevant features of the market are captured. Although some of the game-theoretic work mentioned before might provide some insights into which network summaries should be monitored, the choice is typically difficult and the selection is often incomplete. Second, and more importantly, approaches of this type are not helpful in predicting future interactions among traders.

1.3 NYMEX Natural Gas Futures Market Data

To motivate the structure of our models, consider a time series of $T = 201$ weekly financial trading networks constructed from proprietary trades in the natural gas futures market on the New York Mercantile Exchange (NYMEX) between January 2005 and December 2008. Directed binary networks were constructed by setting $y_{i,j,t} = 1$ if there was at least one transaction in which trader i sold an option to trader j during week t. One particularity of this market is that futures were traded on the New York Mercantile Exchange (NYMEX) only through traditional open-outcry trades until September 5, 2006, and as a hybrid market that included electronic trading conducted via the CME Globex platform after that date.

Note that treating the network as binary ignores information about the transactions such as the number, maturities, and prices of the contracts. We proceed in this way for two reasons. First, in some markets (i.e., black pools) the prices and number of contracts might not be disclosed, making it impossible to apply more general models. Second, even if available, this extra data provides limited additional information about the identity of counterparties subject to contagion risks. Nonetheless, the modeling frameworks we describe in this work can be easily extended to more general types of weighted networks.

1.3.1 Descriptive analysis

A total of 970 unique traders participate in proprietary transactions at least once over the four years to December 2008. However, this list includes traders that either abandoned proprietary trading or went bankrupt during the period under study, as well as traders that entered the market after January 2005. Indeed, only between 240 and 340 traders participated in trades each week. Since we have no information about the exact times at which different traders entered or left the market, our analysis focuses on 71 traders we identified as being present in the market (although not necessarily active) during the whole period. Traders were anonymized and are identified in the plots using numbers.

Figure 1.1 presents a matrix representation for the trading network associated with the week of February 22, 2005 (traders have been reordered to make the graph easier to read). The graph suggests the existence of groups of traders that are structurally equivalent, including a large group of inactive traders that participate in very few or no transactions during this particular week, as well as small group of traders with a high number of intra-group and a relatively low number of inter-group transactions. This suggests that a stochastic block model might be a reasonable model for individual trading networks. In addition, the trading network is sparse with an average over time of 826 links (out of 4970 possible ties).

Figure 1.2 presents time series plots of the clustering coefficients and assortativity coefficients associated with each of the 201 networks. From these plots it is clear that there is a change in market microstructure on September 2006 (which corresponds the date of introduction of an electronic trading system in this market). In addition, there is some

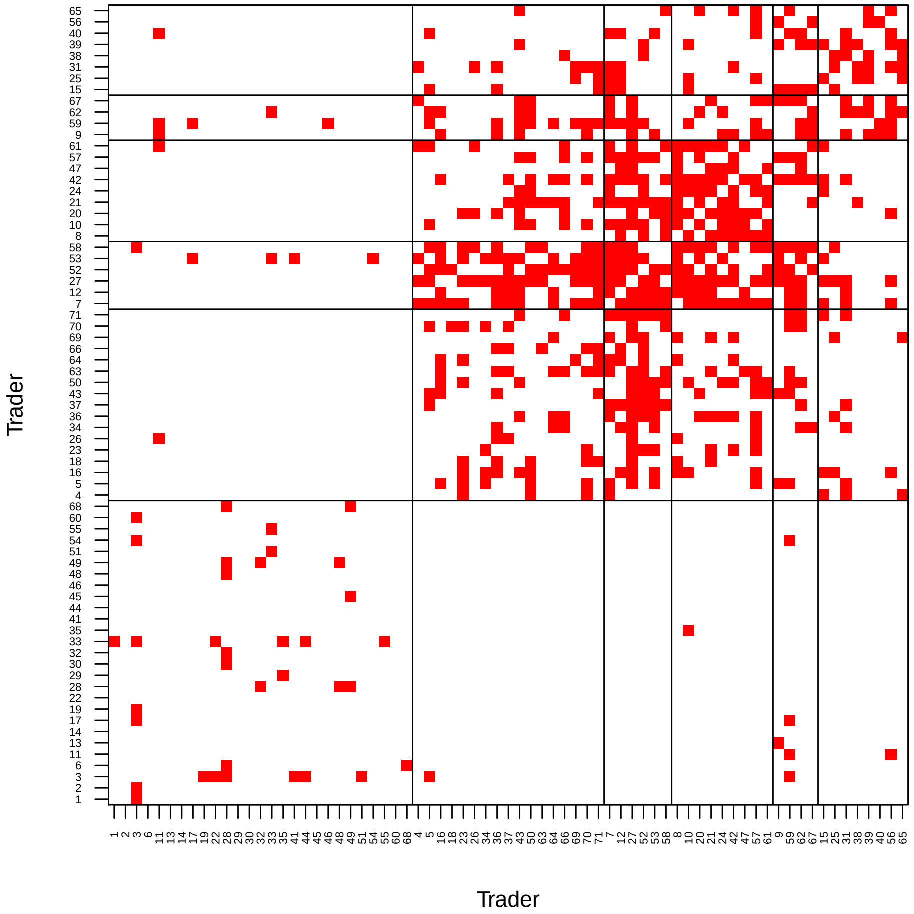

Figure 1.1: Incidence matrix for the trading network associated with the week of February 22, 2005. The solid lines suggest one possible partition of the traders into groups of structurally equivalent nodes.

evidence of two additional structural changes around January 2007 and June 2008. This suggests that a Markov switching model in which the state of the system is assumed to be constant over short periods of time could be a reasonable model for this type of data.

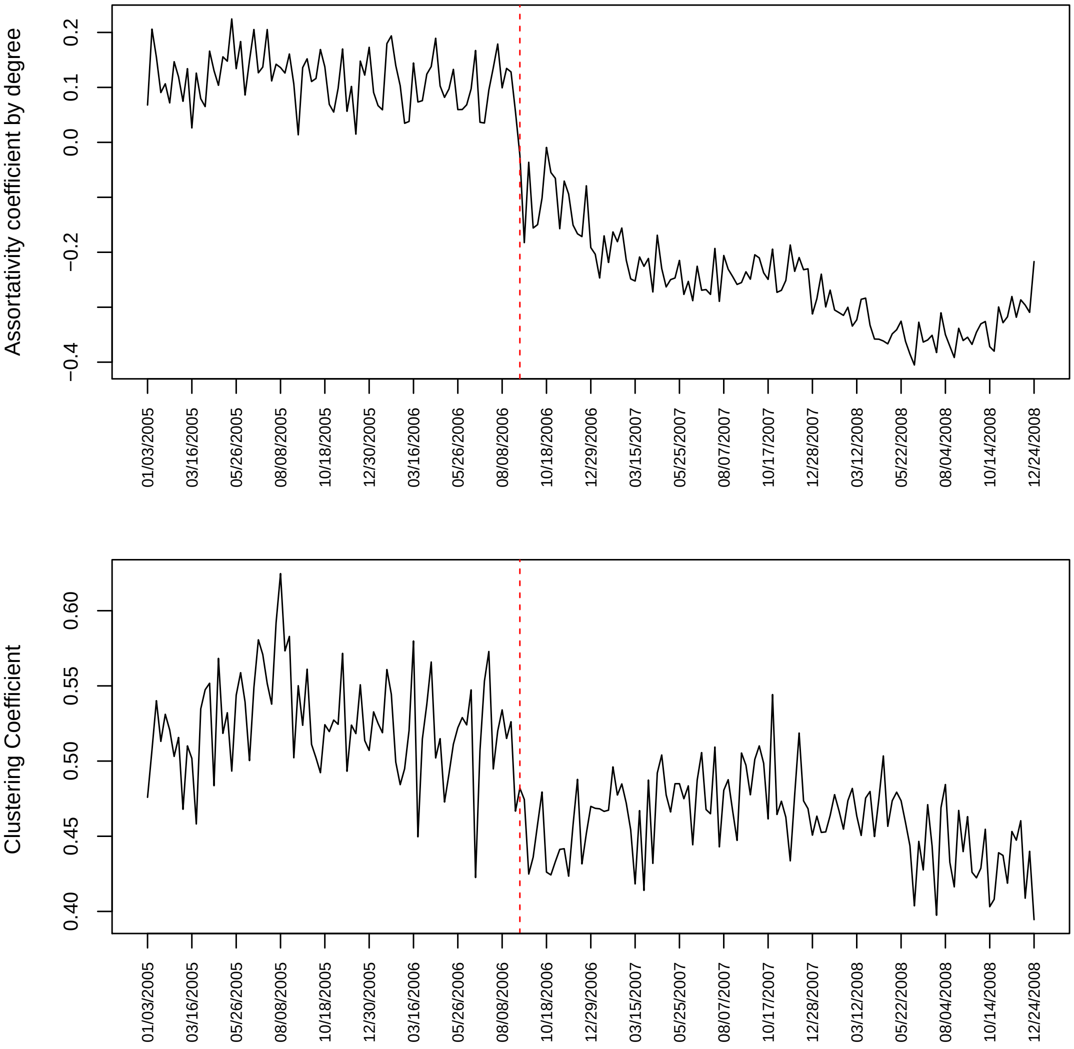

Figure 1.2: Assortativity and clustering coefficients for weekly trading networks in the NYMEX natural gas futures market between January 2005 and December 2008. The vertical dashed line indicates the date in which electronic trading was introduced in this market.

Chapter 2

Bayesian Nonparametric Model for Change-Point Identification

In this chapter, we propose modeling the dynamics of networks using an extension of the Bayesian infinite-dimensional model of Kemp et al. (2006). The model we propose accounts for dependence of the network structure over time and incorporates more general hierarchical priors on the interaction probabilities as well as the partition structure. To account for changes in the network structure over time, the blockmodels associated with different time periods are linked through a hidden Markov model. The resulting model is similar in spirit to the one described in Rodriguez (2011), but allows for additional flexibility and enhanced interpretability.

2.1 Stochastic Blockmodels

A stochastic blockmodel for a network $\mathbf{Y}$ at a single point in time assumes that its entries are conditionally independent given two sets of parameters: a vector of discrete indicators $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$, where $\xi_i = k$ if and only if node i belongs to community $k = 1, \dots K$, and a $K \times K$ matrix $\boldsymbol{\Theta} = [\theta_{k,l}]$ such that $\theta_{k,l}$ represents the probability of an outgoing link from a member of community k to a member of community l. Therefore,

$$y_{i,j} \mid \boldsymbol{\xi}, \boldsymbol{\Theta} \sim \mathsf{Ber}(\theta_{\xi_i,\xi_j}).$$

Note that K represents the maximum potential number of trading communities allowed a priori. A posteriori, the effective number of trading communities K^* present in the sample could potentially be smaller than K.

A Bayesian formulation for the stochastic blockmodel is completed by eliciting prior distributions for K, $\boldsymbol{\xi}$, and $\boldsymbol{\Theta}$. In the sequel we set $K = \infty$ and let the indicators be independent a priori where $\mathsf{Pr}(\xi_i = k \mid \mathbf{w}) = w_k$ for $k = 1, 2, \dots$, and the vector of weights $\mathbf{w} = (w_1, w_2, \dots)$ be constructed so that

$$w_k = v_k \prod_{s<k} \{1 - v_s\}, \qquad\qquad v_k \sim \mathsf{Beta}(1 - \alpha, \beta + \alpha k), \tag{2.1}$$

for $0 \leq \alpha < 1$ and $\beta > -\alpha$. Note that, by setting $K = \infty$, the model allows for the effective number of components K^* to be as large as the number of traders n, for any n.

The formulation in (2.1) is equivalent to the constructive definition of the Poisson-Dirichlet process (Pitman, 1995; Pitman & Yor, 1997), with $\alpha = 0$ leading to the Dirichlet process. Hence, the implied prior on the effective number of trading communities K^* and the size of those communities, $m_1, \dots, m_{K^*}$, is given by

$$\frac{\Gamma(\beta+1)}{(\beta+\alpha K^*)\Gamma(\beta+n)} \prod_{k=1}^{K^*} (\beta+\alpha k) \frac{\Gamma(m_k-\alpha)}{\Gamma(1-\alpha)}.$$

Note that larger values of α or β favor a priori a larger effective number of K^*. Setting $\alpha = 0$ leads to the prior expected number of communities to grow logarithmically with n, while for $\alpha > 0$ the expected number components grows as a power of the number of traders.

Consider now building a prior on the matrix of interaction probabilities $\mathbf{\Theta}$. In this case we let

$$\theta_{k,l} \mid a_O, b_O, a_D, b_D \sim \begin{cases} \mathsf{Beta}(a_O, b_O) & k \neq l \\ \\ \mathsf{Beta}(a_D, b_D) & k = l. \end{cases}$$

This prior is more general than those typically used in stochastic blockmodels, as it allows the distribution of the diagonal and off-diagonal elements of $\mathbf{\Theta}$ to have different hyperparameters. This ensures additional flexibility in terms of the implied degree distribution of the network, while still ensures that both $p(\mathbf{Y})$ and $p(\mathbf{\Theta})$ are jointly exchangeable, i.e., that the distributions are invariant to the order in which traders or communities are labeled (Aldous, 1981).

2.1.1 Model-based assortativity and transitivity

In this section, we present two alternative model-based indexes to the assortativity by degree and the clustering coefficients discussed in Figure 1.2 (Rodriguez & Reyes, 2013). The prior specification for the matrix of interaction probabilities allows us to define an assortative index for the network as

$$\begin{aligned}\Upsilon &= \log\left\{\mathsf{E}(\theta_{k,k} \mid a_D, b_D)\right\} - \log\left\{\mathsf{E}(\theta_{k,l} \mid a_O, b_O)\right\} \\ &= \log\left\{\frac{a_D}{a_D + b_D}\right\} - \log\left\{\frac{a_O}{a_O + b_O}\right\},\end{aligned}$$

and a cycle-type transitivity index

$$\chi = \mathsf{Pr}(y_{i,j} = 1 \mid y_{j,k} = 1, y_{k,i} = 1, a_O, b_O, a_D, b_D, \alpha, \beta) = \frac{\chi_N}{\chi_D},$$

where

$$\begin{aligned}\chi_N = \frac{(1-\alpha)(2-\alpha)}{(\beta+1)(\beta+2)} &\frac{(a_D+2)(a_D+1)a_D}{(a_D+b_D+2)(a_D+b_D+1)(a_D+b_D)} \\ &+ 3\frac{(1-\alpha)(\beta+\alpha)}{(\beta+1)(\beta+2)}\frac{a_D}{(a_D+b_D)}\left(\frac{a_O}{a_O+b_O}\right)^2 \\ &\qquad + \frac{(\beta+\alpha)(\beta+2\alpha)}{(\beta+1)(\beta+2)}\left(\frac{a_O}{a_O+b_O}\right)^3,\end{aligned}$$

and

$$\begin{aligned}\chi_D = \frac{(1-\alpha)(2-\alpha)}{(\beta+1)(\beta+2)} &\frac{(a_D+1)a_D}{(a_D+b_D+1)(a_D+b_D)} \\ &+ 2\frac{(1-\alpha)(\beta+\alpha)}{(\beta+1)(\beta+2)}\frac{a_D}{(a_D+b_D)}\frac{a_O}{(a_O+b_O)} \\ &\qquad + \frac{(\beta+\alpha)(\beta+\alpha+1)}{(\beta+1)(\beta+2)}\left(\frac{a_O}{a_O+b_O}\right)^2.\end{aligned}$$

These indexes allow us to asses assortativity mixing and transitivity patterns in the model based on specific prior knowledge about the community structure and mixing patterns of the network.

2.2 Hidden Markov Model for Time Series of Networks

Hidden Markov models are widely used in financial (e.g., see Ryden et al., 1998 and references therein) and biological (e.g., Yau et al., 2011 and references therein) applications where there is interest in identifying structural changes in the system under study. Here, we extend the hierarchical blockmodel described in Section 2.1 to model a time series of networks. The extension is built with the goals of identifying events associated with structural changes in the network, and making short-term predictions about the structure of the network in future periods. For these reasons, we focus our attention on the use of hidden Markov models for network data.

2.2.1 Model formulation

Consider now a sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ of binary trading networks observed over T consecutive time intervals, where all networks are associated with a common set of n traders. In addition, let $\zeta_1, \ldots, \zeta_T$ be a sequence of unobserved state variables such that $\zeta_t = s$ indicates that the system is in state $s \in \{1, 2, \ldots, S\}$ during period $t = \{1, 2, \ldots, T\}$. Each state has associated with it a vector of community indicators $\boldsymbol{\xi}_s = (\xi_{1,s}, \ldots, \xi_{n,s})$ with $\xi_{i,s} \in \{1, 2, \ldots, K\}$ and a matrix of interaction probabilities $\mathbf{\Theta}_s = [\theta_{k,l,s}]$ representing, respectively, the grouping of nodes into communities and the probabilities of links occurring between communities when the system is in state s.

Analogously to our previous discussion, S and K represent the maximum number of states and the maximum number of trading communities allowed by the model a priori. A posteriori, the effective number of states S^* and the effective number of communities on each state $K_1^*, \ldots, K_S^*$ is potentially smaller than S and K, respectively.

Conditionally on the state parameters, observations are assumed to be independent, i.e.,

$$y_{i,j,t} \mid \zeta_t, \{\boldsymbol{\xi}_s\}, \{\boldsymbol{\Theta}_s\} \sim \mathsf{Ber}(y_{i,j,t} \mid \theta_{\xi_{i,\zeta_t},\xi_{j,\zeta_t},\zeta_t}).$$

Hence, the joint likelihood for the data can be written as

$$\begin{aligned} p\left(\{\mathbf{Y}_t\} \mid \{\boldsymbol{\zeta}_t\}, \{\boldsymbol{\xi}_s\}, \{\boldsymbol{\Theta}_s\}\right) &= \prod_{t}^{T} \prod_{i=1}^{n} \prod_{\substack{j=1 \\ j \neq i}}^{n} \theta_{\xi_{i,\zeta_t},\xi_{j,\zeta_t},\zeta_t}^{y_{i,j,t}} \left(1 - \theta_{\xi_{i,\zeta_t},\xi_{j,\zeta_t},\zeta_t}\right)^{1-y_{i,j,t}} \\ &= \prod_{s=1}^{S} \prod_{k=1}^{K} \prod_{l=1}^{K} \prod_{(i,j,t) \in A_{k,l,s}} \theta_{k,l,s}^{y_{i,j,t}} \left(1 - \theta_{k,l,s}\right)^{1-y_{i,j,t}}, \end{aligned}$$

where $A_{k,l,s} = \{(i,j,t) : i \neq j, \zeta_t = s, \xi_{i,\zeta_t} = k, \xi_{j,\zeta_t} = l\}$ is the set of observations associated with the interactions between communities k and l in state s.

2.2.2 Prior specification

In order to account for the persistence in network structure illustrated in Figure 1.2, we assume that the evolution of the system indicators follows a first-order Markov process with transition probabilities

$$p(\zeta_t = s \mid \zeta_{t-1} = r, \{\boldsymbol{\pi}_r\}) = \pi_{r,s},$$

where $\boldsymbol{\pi}_r = (\pi_{r,1}, \ldots, \pi_{r,S})$, the r-th row of the transition matrix $\boldsymbol{\Pi} = [\pi_{r,s}]$, must satisfy $\sum_{s=1}^{S} \pi_{r,s} = 1$. A natural prior for $\boldsymbol{\pi}_r$ is a symmetric Dirichlet distribution,

$$\boldsymbol{\pi}_r \mid \gamma \sim \mathsf{Dir}\left(\frac{\gamma}{S}, \frac{\gamma}{S}, \ldots, \frac{\gamma}{S}\right).$$

Note that, as $S \to \infty$, the induced distribution of transitions over states is equivalent to that generated by a Dirichlet process prior with concentration parameter γ (for example, see Green & Richardson, 2001). Since γ plays an important role in controlling the number of effective states S^*, its value is estimated from the data by assigning an exponential hyperprior.

The specification of the model is completed by eliciting hierarchical priors on the state-specific parameters $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_S$ and $\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_S$. Following the discussion in Section 2.1, we let $\mathsf{Pr}(\xi_{i,s} = k \mid \mathbf{w}_s) = w_{k,s}$ for $k = 1, 2, \cdots$, where $w_{k,s} = v_{k,s} \prod_{h<k} \{1 - v_{h,s}\}$ are weights constructed from a sequence $v_{1,s}, v_{2,s}, \dots$ where $v_{k,s} \sim \mathsf{Beta}(1 - \alpha_s, \beta_s + k\alpha_s)$. Again, since the hyperparameters α_s and β_s play a critical role in controlling the number of expected trading communities, they are assigned independent hyperpriors $\alpha_s \sim p(\alpha_s)$ and $\beta_s \sim p(\beta_s)$. A natural choice is to assign α_s a uniform prior on the unit interval and β_s an exponential prior, while carrying out a sensitivity analysis that involves priors that favor small values of α_s as well as priors that favor both lower and higher values for β_s.

Similarly, the interaction probabilities are assigned priors

$$\theta_{k,l,s} \mid a_{s,O}, b_{s,O}, a_{s,D}, b_{s,D} \sim \begin{cases} \mathsf{Beta}(a_{s,O}, b_{s,O}) & k \neq l \\ \mathsf{Beta}(a_{s,D}, b_{s,D}) & k = l. \end{cases}$$

where $\{a_{s,O}\}$, $\{b_{s,O}\}$, $\{a_{s,D}\}$, and $\{b_{s,D}\}$ are independent and gamma distributed with shape parameter c and unknown rates d_O, e_O, d_D and e_D, which are in turn assigned exponential priors with means λ_d and λ_e.

Figure 2.1 presents a schematic representation of the hidden Markov model with state-specific community structure associated with the generalized Chinese restaurant process, and state-specific matrix of interaction probabilities between communities.

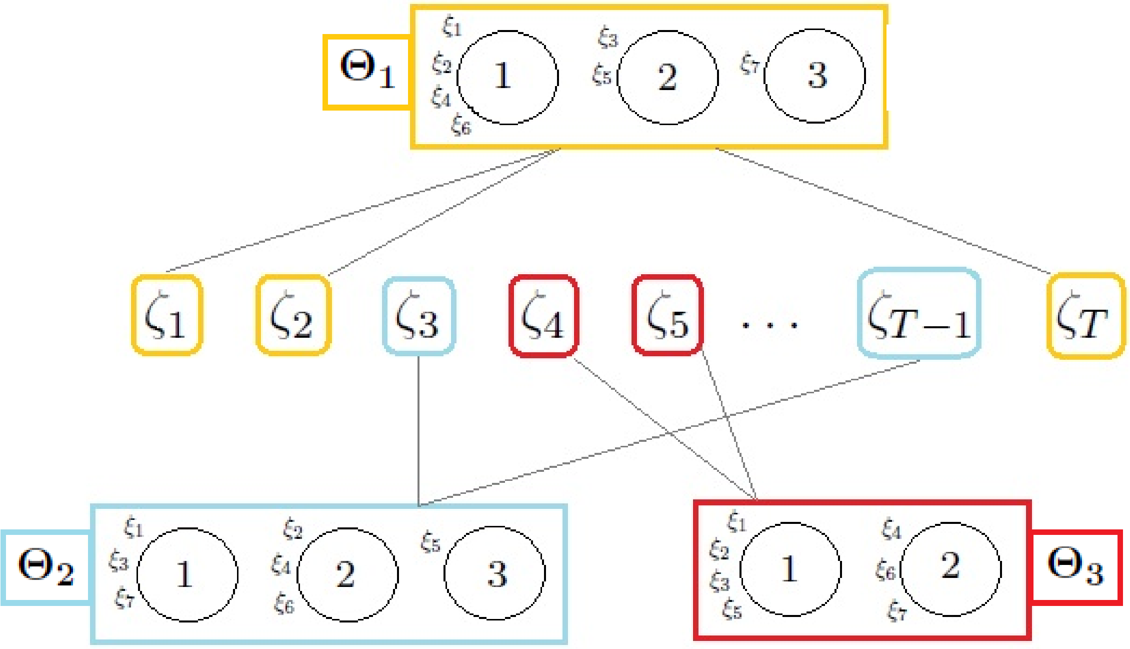

Figure 2.1: Schematic representation of hidden Markov model for dynamic networks. In this scheme, the colors represent three states visited by the system across T network emissions over time on a fixed number of seven nodes. In this particular example, the first observed state (yellow) has three communities with four, two and one nodes (respectively), and associated interaction probability matrix $\mathbf{\Theta}_1$.

2.2.3 Posterior inference and prediction

The joint posterior distribution for the model is proportional to

$$\begin{aligned} &p\left(\{\mathbf{Y}_t\} \mid \{\zeta_t\}, \{\boldsymbol{\xi}_s\}, \{\boldsymbol{\Theta}_s\}\right) p\left(\{\boldsymbol{\Theta}_s\} \mid \{a_{s,O}\}, \{b_{s.O}\}, \{a_{s,D}\}, \{b_{s.D}\}\right) \\ &\qquad p\left(\{a_{s,O}\} \mid d_O\right) p\left(\{b_{s,O}\} \mid e_O\right) p\left(\{a_{s,D}\} \mid d_D\right) p\left(\{b_{s,D}\} \mid e_D\right) \\ &\qquad p(d_O)\, p(e_O)\, p(d_D)\, p(e_D)\, p\left(\{\zeta_t\} \mid \boldsymbol{\Pi}\right) p\left(\boldsymbol{\Pi} \mid \gamma\right) p\left(\gamma\right) \\ &\qquad p\left(\{\boldsymbol{\xi}_s\} \mid \{\alpha_s\}, \{\beta_s\}\right) p\left(\{\alpha_s\}\right) p\left(\{\beta_s\}\right). \end{aligned} \tag{2.2}$$

This posterior distribution is not analytically tractable. Therefore, we implemented a Markov chain Monte Carlo algorithm (Robert & Casella, 2005) that simulates a dependent sequence of random draws from the target distribution.

To derive the algorithm, we rely on the fact that the joint posterior distribution in (2.2) can be factorized as

$$\begin{aligned} &p\left(\{\boldsymbol{\Theta}_s\} \mid \{\boldsymbol{\xi}_s\}, \{\zeta_t\}, \{a_{s,O}\}, \{b_{s,O}\}, \{a_{s,D}\}, \{b_{s,D}\}, \{\mathbf{Y}_t\}\right) \times \\ &\qquad p\left(\{\boldsymbol{\xi}_s\}, \{\zeta_t\}, \{a_{s,O}\}, \{b_{s,O}\}, \{a_{s,D}\}, \{b_{s,D}\},\right. \\ &\qquad\qquad \left. d_O, e_O, d_D, e_D, \{\alpha_s\}, \{\beta_s\}, \gamma \mid \{\mathbf{Y}_t\}\right) \end{aligned} \tag{2.3}$$

Since the values of $\theta_{k,l,s}$ are conditionally independent a posteriori given the observations, the indicators $\{\zeta_t\}$ and $\{\xi_{i,s}\}$, and the prior parameters $\{a_{s,O}\}$, $\{b_{s,O}\}$, $\{a_{s,D}\}$ and $\{b_{s,D}\}$, the first term in (2.3) is easy to sample from. Furthermore, conditionally on the other parameters in the model, the state indicators $\zeta_1, \ldots, \zeta_T$ are sampled jointly using a forward-backward algorithm (Rabiner, 1986), while the full conditional distribution for each collection of indicators $\xi_{1,s}, \ldots, \xi_{n,s}$ is sampled using a collapsed (marginal) Gibbs sampler (Neal, 2000). Details of the algorithm are discussed in Appendix A.

Given a sample from the Markov chain Monte Carlo algorithm,

$$\Big(\{\mathbf{\Theta}_s^{(b)}\}, \{\boldsymbol{\xi}_s^{(b)}\}, \{\zeta_t^{(b)}\}, \{a_{s,O}^{(b)}\}, \{b_{s,O}^{(b)}\}, \{a_{s,D}^{(b)}\}, \\ \{b_{s,D}^{(b)}\}, d_O^{(b)}, e_O^{(b)}, d_D^{(b)}, e_D^{(b)}, \{\alpha_s^{(b)}\}, \{\beta_s^{(b)}\}, \gamma^{(b)}\Big), \qquad b = 1, \ldots, B,$$

obtained after an appropriate burn-in period, point and interval estimates for model parameters can be easily obtained by computing the empirical mean and/or the empirical quantiles of the posterior distribution. The label switching associated with the state and community indicators produces identifiability problems. However, we focus our inference on identifiable functions of the parameters. For example, the posterior co-clustering probabilities, $\omega_{t,t'} = \mathsf{Pr}(\zeta_t = \zeta_{t'} \mid \{\mathbf{Y}_t\})$ are identifiable and can be estimated as

$$\hat{\omega}_{t,t'} = \mathsf{Pr}(\zeta_t = \zeta_{t'} \mid \{\mathbf{Y}_t\}) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\zeta_t^{(b)} = \zeta_{t'}^{(b)}),$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The estimates can be arranged into a co-clustering matrix $[\hat{\omega}_{t,t'}]$, which can in turn be used to identify the state of the system at each time period through a decision-theoretic approach (e.g., see Lau & Green, 2007). A similar procedure can be used to identify communities on each period.

The samples from the posterior distribution can also be used as the basis for prediction. For this purpose, note that the probability of a link from node i to node j in the unobserved period $T+1$ can be estimated by

$$\mathsf{E}\left(y_{i,j,T+1} \mid \{\mathbf{Y}_t\}\right) \approx \frac{1}{B}\sum_{b=1}^{B} \pi_{\zeta_T^{(b)},s}^{(b)}\, \theta_{\xi_{i,s}^{(b)},\xi_{j,s}^{(b)}}.$$

Using a simple 0/1 utility function, a future link from node i to node j is predicted as $\hat{y}_{i,j,T+1} = \mathbb{I}(\mathsf{E}\{y_{i,j,T+1} \mid \{\mathbf{Y}_t\}\} > f)$, for some threshold f that reflects the relative cost associated false positive and false negative links.

2.3 Application to the NYMEX natural gas futures market

In this section we analyze the sequence of $T = 201$ weekly financial trading networks described in detail in Section 1.3. The results presented in this section are based on 100,000 iterations collected after a burn-in period of 10,000 iterations. Convergence of the algorithm was diagnosed using the single-chain approach discussed in Geweke (1992) and by a visual evaluation of trace plots. We monitored the log-likelihood function, as well as the number of active states S^* and the mean and variance at a few selected times of the assortativity and transitivity indexes $\{\Upsilon_t\}$ and $\{\chi_t\}$. In terms of hyperparameters, the maximum number of states is set to $S = 30$, the prior means for γ and $\{\beta_s\}$ are assigned exponential priors with unit mean, and the priors for d_O, e_O, d_D and e_D are exponential distributions with mean 2. This specification implies that, a priori, $\mathsf{E}(\Upsilon_t) = 0$ for all $t = 1, \ldots, T$, so that we favor neither assortative nor dissasortive trading communities a priori.

2.3.1 Identifying changes in market microstructure

Figure 2.2 presents the posterior estimate of the co-clustering matrix for the latent states $\zeta_1, \ldots, \zeta_T$, along with a point estimator for the grouping of networks into states (recall Section 2.2.3). This point estimator suggests that the structure of the trading networks alternates between four highly persistent states. The first state runs between early January 2005 and early September 2006, when the electronic market is introduced. The presence of a change point at this date is not surprising in light of the descriptive analysis presented in Section 1.3. The second state runs between early September 2006 and early May 2007, when the system transitions to a new state for a short period of 3 months. After that, the system

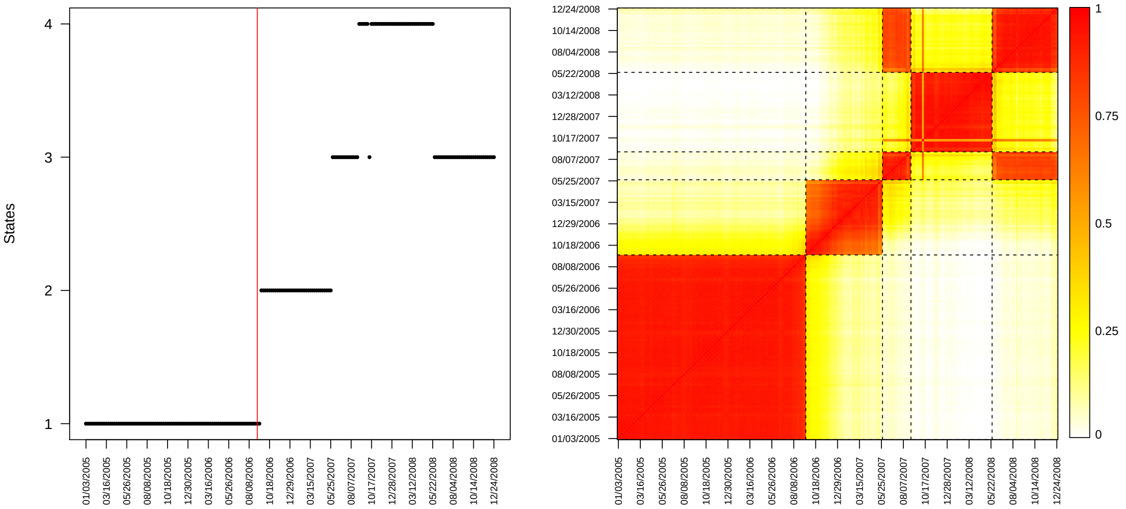

Figure 2.2: In the left panel, point estimate of the states for the 201 weeks observed for the trading network with the vertical line indicating the introduction of the electronic platform on week 85. On the right, mean posterior pairwise incidence matrix for the weekly networks, illustrating the uncertainty associated with this point estimate.

seems to transition to a fourth state in early August 2007 (interestingly, the beginning of the recent financial crises), where it stays for 37 weeks before returning to the third state in early June 2008 (which coincides with some of the largest drops in the S&P500 energy sector index over the last 13 years). Also, it is clear from the heatmap that, although there is some uncertainty associated with this point estimate of the system states (mostly in time of the transitions between states three and four), this uncertainty is relatively low.

Figure 2.3 shows estimates of the community structure associated with two different weeks, that of October 11, 2005 ($t = 40$) and that of November 14, 2007 ($t = 145$). We selected these dates because they are representative of states 1 and 4. Note that, although there are some similarities, the overall structure of the communities is quite different. State 1 is characterized by a large group of 25 mostly inactive traders, while all other traders

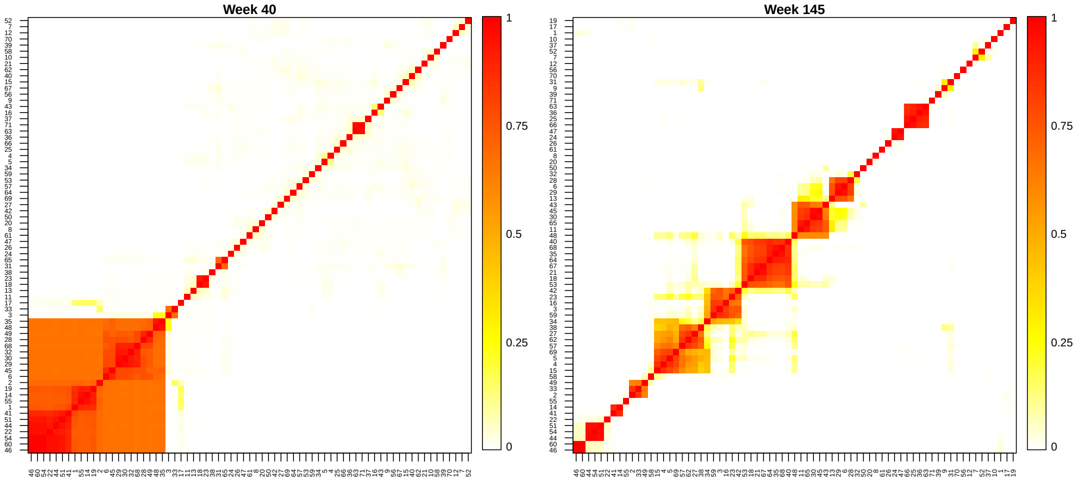

Figure 2.3: Mean posterior pairwise incidence matrices of traders for $t = 40$ from state 1 and $t = 145$ from state 4.

tend to fall, for the most part, into singleton clusters. On the other hand, while state 4 also exhibits a number of singleton clusters, it also shows a number of small communities comprising between 5 and 10 traders each. Because of the fact that communities are driven by trading strategies, this suggest granularity before the introduction of electronic trading and heard behavior after the hybrid trading exchange.

Figure 2.4 shows time series plots for the estimates of the assortativity and transitivity indexes $\Upsilon_1, \ldots, \Upsilon_T$ and $\chi_1, \ldots, \chi_T$. Recall that these quantities are model-based alternatives to the assortativity by degree and the clustering coefficient presented in Figure 1.2. Both sets of plots share some common features, revealing mild assortativity and higher transitivity before September 2006 and highly disassortative networks with lower transitivity afterwards. This makes sense because we would expect that the introduction of an electronic market would limit the effect of social connections among traders (which

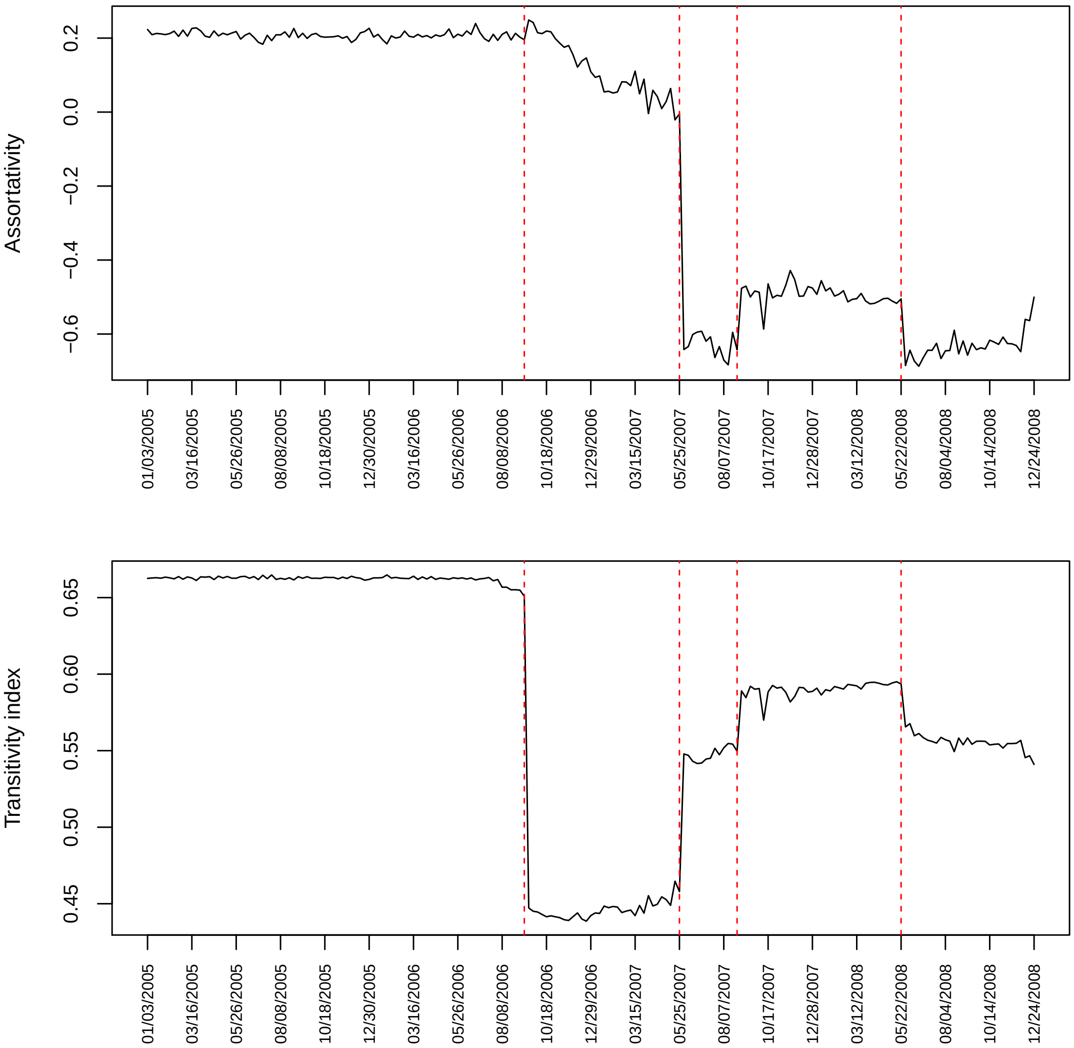

Figure 2.4: Time series plot for assortativity and transitivity indexes. The vertical line represents the transitions across states identified from Figure 2.2.

Table 2.1: Prior and posterior point estimates and credibility intervals for some model hyperparameters.

Parameter	Posterior	Posterior 95%	Prior	Prior 95%
	mean	credible interval	mean	credible interval
α_{40}	0.748	$(0.198, 0.942)$	0.500	$(0.025, 0.975)$
α_{100}	0.524	$(0.240, 0.851)$	0.500	$(0.025, 0.975)$
α_{145}	0.587	$(0.264, 0.785)$	0.500	$(0.025, 0.975)$
β_{40}	1.225	$(0.034, 4.526)$	1.000	(0.025, 3.689)
β_{100}	2.518	$(0.107, 7.563)$	1.000	(0.025, 3.689)
β_{145}	1.223	$(0.034, 4.510)$	1.000	(0.025, 3.689)
γ	0.344	$(0.090, 0.814)$	1.000	(0.025, 3.689)

tend to be assortative and transitive) and favor connections based on differential trending strategies (which tend to be disassortative).

Finally, Table 2.1 shows point estimates and credible intervals associated with some hyperparameters in the model, both a priori and a posteriori. In all cases, the posterior estimates appear to be more concentrated and be centered around different values than the prior.

2.3.2 Link prediction

As we discussed in the introduction, besides identifying points of change in the microstructure of the market, one of our goals is to predict future trading partnerships. To assess the predictive capabilities of the model we ran an out-of-sample crossvalidation exercise where we held out the last ten weeks in the dataset and made one-step-ahead predictions for the structure of the held-out networks. More specifically, for each $t = 191, 192, \ldots, 200$ we use the information contained in $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$ to estimate the model parameters and obtain predictions for $\hat{\mathbf{Y}}_{t+1}$ for different values of the threshold f. Each of these predictions is compared against the observed network $\mathbf{Y}_{t+1}$, the number of false

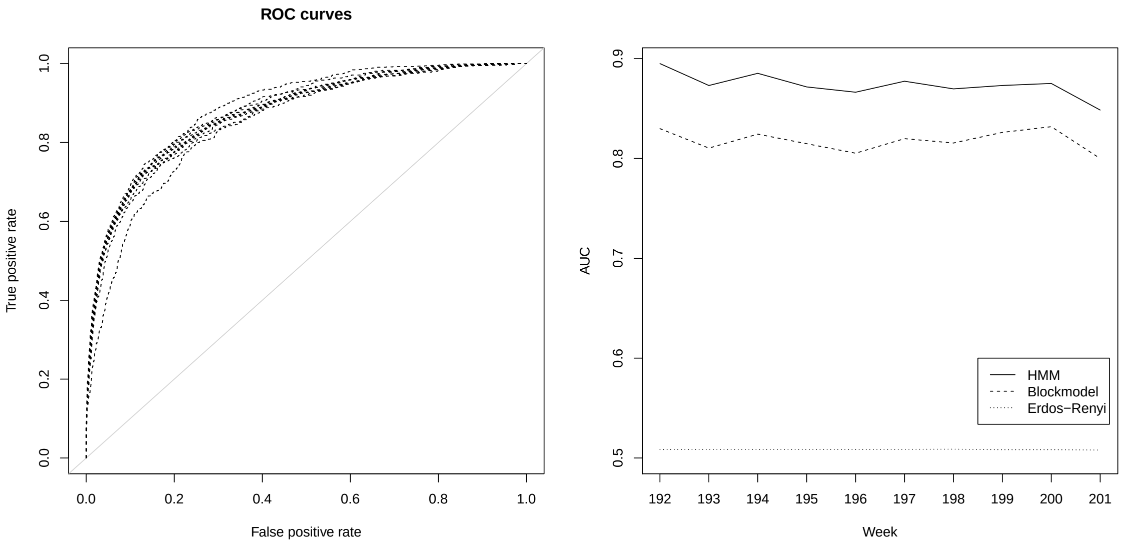

Figure 2.5: The left panel shows ten operating characteristic curves associated with one-step-ahead out of sample predictions from our hidden Markov model. The right panel shows a time series plot of the area under the receiver operating characteristic curves for three different models.

and true positives is computed, and a receiver operating characteristic (ROC) curve is constructed. For comparison purposes, the same exercise was repeated first by fitting a single blockmodel (which corresponds to taking $S = 1$) and then by fitting a single Erdös-Renyi model (which corresponds to $S = 1$ and $K_1 = 1$) to the whole collection of networks. Figure 2.5 shows the ten operating characteristic curves associated with one-step-ahead out of sample predictions from our hidden Markov model, along with estimates of the area under the receiver operating characteristic curves (AUC) for all three models. The proposed hidden Markov model shows superior predictive capabilities, with AUCs between 0.85 and 0.9.

2.3.3 Sensitivity analysis

To assess the effect of our prior choice on posterior inference we conducted a sensitivity analysis where the model was fitted with somewhat different priors. In particular, we used independent Beta priors with mean 1/10 and variance 9/1100 for each α_s, as well as exponential priors with means 1/3 and 3 for each β_s. On the other hand, exponential priors with mean 2 were also used for d_O, e_O, d_D and e_D. Although inferences on the community structure were somewhat affected by prior choices, inferences on the state parameters as well as the assortativity and transitivity indexes and the predictive performance were essentially unchanged.

2.4 Discussion

We have presented a class of hidden Markov models that, in the context of financial trading networks, have clear potential for market regulatory oversight. Key application of these models include identifying specific events (such as large trader failures or specific changes in market rules) that affect market stability, as well as identifying frequent trading counterparties that might be likely collusion partners or particularly at risk in case of bankruptcies. Here, we have focused on models for binary networks where only the presence/absence of transactions over a week is recorded. However, when information about volumes is available, the model can be easily extended to incorporate this information.

Changes in the network structure may cause some traders to fail while new traders can enter the market and become prominent. This model could also be extended to incorporate information about trader survivability using continuation ratio models. However, in this particular application, we have no information about the exact times at which the traders entered or failed in the market.

Although the use of a hidden Markov model allows us to account for time dependence and is useful for identifying structural changes in the system, the computational complexity might be too restrictive for predictive purposes in large networks. In the following chapters we explore extensions of p_1 models as well as autologistic models that might allow for improved and more efficient predictions.

Chapter 3

Bayesian fused lasso regression for link prediction

In this chapter, we propose modeling time series of networks using an extension of the class of p_1 models presented in Holland & Leinhardt (1981) (see section 1.1.2). In the modeling approach discussed here, the model parameters are set to be time dependent to add flexibility and account for alterations in the network structure over time. One challenging feature of the network data we are working with is high-dimensionality. In particular, the number of parameters in our proposed model will always be larger than the number of available observations. To deal with this issue we resort to fused lasso regression by imposing an L^1 penalty on consecutive differences of the model parameters. In a Bayesian setting, this is equivalent to assuming double exponential priors on the differences of the coefficients in contiguous time points. Here, we explore two different computational approaches for our model. First, full Bayesian inference is presented and implemented using two different sampling schemes. As an alternative, we also carry out maximum a posteriori (MAP) estimation utilizing an optimization approach to overcome the computational burden of a full Bayesian analysis as the number of nodes in the network increases.

3.1 Model Formulation

We consider an extension of (1.1) in which the pairs $\{(y_{i,j,t}, y_{j,i,t}) : i < j\}$ are modeled using a logistic model of the form

$$p\left(y_{i,j,t}, y_{j,i,t}\right) \propto \exp\left\{\alpha_{i,j,t} y_{i,j,t} + \beta_{i,j,t} y_{j,i,t} + \gamma_{i,j,t} y_{i,j,t} y_{j,i,t}\right\}, \tag{3.1}$$

where $\alpha_{i,j,t}$ and $\beta_{i,j,t}$ represent the baseline probabilities of a directed link between nodes i and j, and $\gamma_{i,j,t}$ controls the level of dependence between $y_{i,j,t}$ and $y_{j,i,t}$. For example, $\gamma_{i,j,t} = 0$ implies that $y_{i,j,t}$ and $y_{j,i,t}$ are conditionally independent with $\mathsf{Pr}(y_{i,j,t} = 1) = \exp\left\{\alpha_{i,j,t}\right\} / \left(1 + \exp\left\{\alpha_{i,j,t}\right\}\right)$ and $\mathsf{Pr}(y_{j,i,t} = 1) = \exp\left\{\beta_{i,j,t}\right\} / \left(1 + \exp\left\{\beta_{i,j,t}\right\}\right)$. On the other hand, $\gamma_{i,j,t} > 0$ favors outcomes in which $y_{i,j,t} = y_{j,i,t}$ (a phenomenon often called positive reciprocity), while $\gamma_{i,j,t} < 0$ favors situations in which $y_{i,j,t} \neq y_{j,i,t}$ (often called negative reciprocity). Hence, by allowing the values of $y_{i,j,t}$ and $y_{j,i,t}$ to be potentially correlated the model can accommodate reciprocity in the network.

Model (3.1) can be conveniently reformulated as a multinomial logistic regression.

Using the following specification:

$$z_{i,j,t} = \begin{cases} 1 \quad \text{iff } (y_{i,j,t}, y_{j,i,t}) = (1,0) \\ 2 \quad \text{iff } (y_{i,j,t}, y_{j,i,t}) = (0,1) \\ 3 \quad \text{iff } (y_{i,j,t}, y_{j,i,t}) = (1,1) \\ 4 \quad \text{iff } (y_{i,j,t}, y_{j,i,t}) = (0,0) \end{cases},$$

the likelihood can be rewritten as

$$p\left(z_{i,j,t} = r | \boldsymbol{\theta}_{i,j,t}\right) = \frac{\exp\left(\theta_{i,j,r,t}\right)}{\sum_{s=1}^{4} \exp\left(\theta_{i,j,s,t}\right)}, \tag{3.2}$$

where $\theta_{i,j,1,t} = \alpha_{i,j,t}$, $\theta_{i,j,2,t} = \beta_{i,j,t}$, $\theta_{i,j,3,t} = \alpha_{i,j,t} + \beta_{i,j,t} + \gamma_{i,j,t}$, and $\theta_{i,j,4,t} = 0$. Note that the total number of parameters associated with each pair (i,j) is $p = 3T$.

3.1.1 Prior specification

The parameters in the multinomial logistic model we just described are time dependent. Hence, it is natural and useful to take into account the information about their temporal correlation structure in the estimation process. In particular, we are interested in a random walk model with double exponential priors of the form:

$$\theta_{i,j,r,t} = \theta_{i,j,r,t-1} + \epsilon_{i,j,r,t}, \qquad \epsilon_{i,j,r,t} \sim \mathsf{DE}(0, 1/\lambda),$$

where DE represents the double exponential distribution, and $\lambda > 0$ is the parameter that controls the shrinkage level in the differences of the coefficients. A dynamic model of this type on the parameters leads to the joint prior

$$p(\boldsymbol{\theta}_{i,j,r} | \lambda) \propto \exp\left\{ -\lambda \sum_{t=1}^{T} |\theta_{i,j,r,t} - \theta_{i,j,r,t-1}| \right\},$$

where $\theta_{i,j,r,0} = \hat{\theta}_{r,0}$ is assumed known and elicited using an empirical Bayes procedure. Since we assume that $\theta_{i,j,r,0}$ is known, this pairwise difference prior is proper and belongs to the class of Markov random fields and can be associated with the class of conditionally autoregressive (CAR) priors frequently used in spatial statistics and image processing. In addition, assuming double exponential priors is equivalent to imposing L_1 penalty functions on the differences of the parameters in contiguous time points. This penalty type is commonly referred to as the fused lasso with tuning parameter λ. An extensive review of the fused lasso and its theoretical properties is presented in Rinaldo (2009).

Let $\mathbf{\Theta}_{i,j,r} = (\theta_{i,j,r,0}, \theta_{i,j,r,1}, \theta_{i,j,r,2}, \ldots, \theta_{i,j,r,T})$ be the vector of parameters over time for class r, and $\mathbf{\Theta}_{i,j} = \{\mathbf{\Theta}_{i,j,1}, \mathbf{\Theta}_{i,j,2}, \mathbf{\Theta}_{i,j,3}\}$ the vector of all nonzero parameters for the pair of nodes (i,j). The posterior distribution of the parameters is given by

$$\sum_{i<j} \left\{ V_{i,j}(\mathbf{\Theta}_{i,j}) - \lambda \sum_{r=1}^{3} \|\mathbf{L}\mathbf{\Theta}_{i,j,r}\|_1 \right\} \tag{3.3}$$

where

$$V_{i,j}(\mathbf{\Theta}_{i,j}) = \sum_{t=1}^{T} \left\{ \theta_{i,j,z_{i,j,t},t} - \log \sum_{s=1}^{4} \exp\left(\theta_{i,j,s,t}\right) \right\}$$

is the (unpenalized) log-likelihood, $\|\cdot\|_1$ denotes the L_1-norm, and $\mathbf{L}$ is a matrix of dimension $T \times (T+1)$ of the form

$$\mathbf{L} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Note that, given λ, (3.3) can be broken down into $n(n-1)/2$ problems, each one corresponding to fitting a multinomial regression for each pair of nodes in the network. In the sequel we focus on algorithms that can be used to solve each of these independent problems, which are then implemented in a parallel environment. Hereinafter, we drop the subindex (i, j) to simplify notation.

3.2 Full Bayesian Inference

In order to perform Bayesian inference with a multinomial likelihood, we exploit the data-augmentation method based on Pólya-Gamma latent variables proposed by Polson et al. (2013). Using this approach, the multinomial likelihood can be represented as a mixture of normals with Pólya-Gamma mixing distribution. This approach allows for a full conjugate hierarchical representation of the model and simple posterior inference through Gibbs sampling.

For the Bernoulli case, note that

$$\frac{[\exp(\psi_t)]^{y_t}}{1+\exp(\psi_t)} \propto \exp(\kappa_t \psi_t) \int_0^\infty \exp\{-\omega_t \psi_t^2/2\} p(\omega_t) d\omega_t$$

where ψ_t is the log odds of $y_t = 1$, $\kappa_t = y_t - 1/2$ and $p(\omega_t)$ is the Pólya-Gamma density with parameters $(1, 0)$. Hence, by augmenting the model with the latent variable ω_t, conditional Gaussianity for the multinomial likelihood can be achieved.

Similarly for the multinomial case, conditional on $\{\omega_{r,t}\}$ for $r = 1, 2, 3$, the contribution of observation t to the likelihood function in our model can be written as

$$L(\theta_{r,t}) \propto \exp\left\{-\frac{\omega_{r,t}}{2}(\theta_{r,t} - C_{r,t})^2 + \kappa_{t,r}(\theta_{r,t} - C_{r,t})\right\}$$

with $\kappa_{r,t} = \mathbb{I}(z_t = r) - 1/2$, $C_{r,t} = \log \sum_{s \neq r} \exp(\theta_{r,t})$ and

$$(\omega_{r,t}|\mathbf{\Theta}) \sim \mathsf{PG}(1, \theta_{r,t} - C_{r,t})$$

where PG denotes a Pólya-Gamma distribution. This results in a Gaussian likelihood with observations $y^*_{r,t} \sim N(\theta_{r,t}, \omega^{-1}_{r,t})$, where $y^*_{r,t} = \kappa_{r,t}/\omega_{r,t} + C_{r,t}$.

The following sections describe computational algorithms that can be used for posterior estimation in our proposed model. First, we describe two different sampling algorithms for full Bayesian inference. Estimation results with these sampling schemes are identical but we are interested in comparing their efficiency (see section 3.6).

3.2.1 Latent variables approach

Using the fact that the double exponential distribution can be expressed as a scale mixture of normals with exponential mixing density (Park & Casella, 2008) :

$$\frac{a}{2}\exp(-a|x|) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau}} \exp\left(\frac{x^2}{2\tau}\right) \frac{a^2}{2} \exp\left(-\frac{a^2\tau}{2}\right) d\tau,$$

the proposed model can be expressed as a simple dynamic linear model for $1 \leq t \leq T$ as:

$$\begin{aligned} y^*_{r,t} &= \theta_{r,t} + \epsilon_{r,t}, & \epsilon_{r,t} &\sim N(0, \omega^{-1}_{r,t}), \\ \theta_{r,t} &= \theta_{r,t-1} + \varepsilon_{r,t}, & \varepsilon_{r,t} &\sim N(0, \tau^2_{r,t-1}), \end{aligned}$$

where $\tau_{r,t-1}$ are the latent parameters associated with the mixture updated as

$$\left(1/\tau^2_{r,k} | \boldsymbol{\Theta}_r, \lambda\right) \sim \mathsf{IGau}\left(\sqrt{\frac{\lambda^2}{(\theta_{r,k+1} - \theta_{r,k})^2}}, \lambda^2\right),$$

for $k = 0, \ldots, T-1$, where IGau denotes the Inverse Gaussian distribution (Kyung et al., 2010). Note that this hierarchical representation of the model, in spite of involving an improper prior, leads to a proper posterior distribution. We rely on the dynamic linear model representation to update the parameters in a component-wise fashion using a forward filtering backward sampling (FFBS) algorithm (Frühwirth-Schnatter, 1994; Carter & Kohn, 1994). Details of the algorithm are presented in Appendix B.

3.2.2 Direct sampling

Note that the full conditional prior on $\theta_{r,t}$ only involves the two nearest neighbors such that for $1 \leq t \leq T-1$:

$$\boldsymbol{\pi}(\theta_{r,t}|\theta_{r,t-1}, \theta_{r,t+1}) \propto \exp\left\{-\lambda(|\theta_{r,t} - \theta_{r,t-1}| + |\theta_{r,t+1} - \theta_{r,t}|)\right\}.$$

Hence, the conditional posterior distribution of $\theta_{r,t}$ is a mixture of truncated normal distributions with three components:

$$\begin{aligned}(\theta_{r,t}|y^*_{r,t}, \theta_{r,t-1}, \theta_{r,t+1}, \omega_{r,t}) \sim w_1 \mathsf{TN}(\mu^{(1)}_{r,t}, \sigma_{r,t}; \theta_{r,t} < \xi_{r,t}) + \\ w_2 \mathsf{TN}(\mu^{(2)}_{r,t}, \sigma_{r,t}; \theta_{r,t} > \zeta_{r,t}) + w_3 \mathsf{TN}(\mu^{(3)}_{r,t}, \sigma_{r,t}; \xi_{r,t} < \theta_{r,t} < \zeta_{r,t})\end{aligned}$$

where $\sigma_{r,t} = 1/\sqrt{\omega_{r,t}}$, $\xi_{r,t} = \min\{\theta_{r,t-1}, \theta_{r,t+1}\}$, $\zeta_{r,t} = \max\{\theta_{r,t-1}, \theta_{r,t+1}\}$, and the means of the truncated normal distributions are given by the following expressions:

$$\mu^{(1)}_{r,t} = y^*_{r,t} + \frac{2\lambda}{\omega_{r,t}}, \;\; \mu^{(2)}_{r,t} = y^*_{r,t} - \frac{2\lambda}{\omega_{r,t}}, \;\; \mu^{(3)}_{r,t} = y^*_{r,t}.$$

In addition, the conditional posterior probabilities of the components of the mixture are given by

$$\begin{aligned}
w_1 &= \exp\left\{\frac{\omega_{r,t}}{2}\mu^{(1)}_{r,t} - \lambda(\xi_{r,t} + \zeta_{r,t})\right\} \Phi\left(\frac{\xi_{r,t} - \mu^{(1)}_{r,t}}{\sigma_{r,t}}\right) \\
w_2 &= \exp\left\{\frac{\omega_{r,t}}{2}\mu^{(2)}_{r,t} + \lambda(\xi_{r,t} + \zeta_{r,t})\right\} \Phi\left(\frac{-\zeta_{r,t} - \mu^{(2)}_{r,t}}{\sigma_{r,t}}\right) \\
w_3 &= \exp\left\{\frac{\omega_{r,t}}{2}\mu^{(3)}_{r,t} - \lambda(\zeta_{r,t} - \xi_{r,t})\right\} \times \left[\Phi\left(\frac{\zeta_{r,t} - \mu^{(3)}_{r,t}}{\sigma_{r,t}}\right) - \Phi\left(\frac{\xi_{r,t} - \mu^{(3)}_{r,t}}{\sigma_{r,t}}\right)\right],
\end{aligned}$$

where Φ represents the Gaussian cumulative distribution function.

In principle, the efficiency of this algorithm is limited by the use of the full conditional distributions for posterior sampling. However, this approach avoids the introduction of the latent variables $\{\tau_{r,k}\}$ discussed in section 3.2.1. The direct characterization of the posterior distribution for our model is similar to the work of Hans (2009) on Bayesian lasso regression with Gaussian likelihoods.

3.2.3 Link prediction

In this case, Monte Carlo posterior samples of the parameters at time $T+1$ can be obtain as:

$$\theta_{i,j,r,T+1}^{(b)} \sim \mathsf{DE}(\theta_{i,j,r,T}^{(b)}, 1/\lambda), \quad b = 1, \ldots, B.$$

Hence, we can estimate (for $i < j$) the probability of a directed link from node i to node j at time $T+1$ as

$$\hat{p}\left(y_{i,j,T+1} = 1 \mid \mathbf{Y}_T\right) = \frac{1}{B} \sum_{b=1}^{B} \left\{ p\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,0) \mid \boldsymbol{\theta}_{i,j,T+1}^{(b)} \right] \right.$$
$$\left. + p\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,1) \mid \boldsymbol{\theta}_{i,j,T+1}^{(b)} \right] \right\},$$

with a similar expression being valid for $\hat{p}\left(y_{j,i,T+1} = 1 \mid \mathbf{Y}_T\right)$.

3.3 Posterior Mode Estimation

For posterior mode estimation, we develop an extension of the Split Bregman method proposed by Ye & Xie (2011) to solve general optimization problems with convex loss functions and L^1 penalized parameters (see also Goldstein & Osher, 2009). The iterative algorithm involves the reformulation of (4.5) as a constrained problem with the linear restriction $\mathbf{L\Theta} = \mathbf{b}$, and the introduction of a vector of dual variables $\mathbf{v}$ used to split the optimization problem into more tractable steps. Furthermore, we also rely on a second-order Taylor approximation to the multinomial likelihood for the implementation.

The proposed algorithm consists on repeating the following steps until convergence for each vector of parameters $\mathbf{\Theta}_r$:

$$\text{(i)}\quad \mathbf{\Theta}_r^{(m+1)} = \arg\max_{\mathbf{\Theta}} \quad V(\mathbf{\Theta}^{(m)}) - \langle \mathbf{v}_r^{(m)}, \mathbf{L\Theta}_r^{(m)} - \mathbf{b}_r^{(m)} \rangle - \tfrac{\mu}{2} \| \mathbf{L\Theta}_r^{(m)} - \mathbf{b}_r^{(m)} \|_2^2$$

$$\text{(ii)}\quad \mathbf{b}_r^{(m+1)} = \mathfrak{T}_{\lambda_2 \mu^{-1}} \left(\mathbf{L\Theta}_r^{(m+1)} + \mu^{-1} \mathbf{v}_r^{(m)} \right)$$

$$\text{(iii)}\quad \mathbf{v}_r^{(m+1)} = \mathbf{v}_r^{(m)} + \delta \left(\mathbf{L\Theta}_r^{(m+1)} - \mathbf{b}_r^{(m+1)} \right)$$

where $\mathbf{v}_r$ is a vector of dual variables, and $\mathfrak{T}_\lambda(\mathbf{w}) = [t_\lambda(w_1), t_\lambda(w_2), \ldots]'$ is a thresholding operator with $t_\lambda(w_i) = \operatorname{sgn}(w_i) \max\{0, |w_i| - \lambda\}$, and $0 < \delta \leq \mu$. Convergence of the algorithm is guaranteed for any chosen value of μ, we follow previous literature and set $\delta = \mu$ for our implementation.

Efficiency of this algorithm is mainly constrained by the maximization of $\mathbf{\Theta}_r$ in the first step. To accelerate it, we replace $V(\mathbf{\Theta})$ by its second-order Taylor expansion around the current iterate and proceed to perform component-wise optimization (e.g., see Krishnapuram & Hartemink, 2005). Using this substitution, subproblem (i) is differentiable and the estimate of a component $\theta_{r,t}$ of $\mathbf{\Theta}_r$ for $1 < t < T - 1$ is updated as:

$$\begin{aligned} \hat{\theta}_{r,t}^{(m+1)} = \left(G_{r,t}^{(m)} - 2\mu \right)^{-1} \Big[& G_{r,t}^{(m)} \hat{\theta}_{r,t}^{(m)} - g_{r,t}^{(m)} - (v_{r,t}^{(m)} - v_{r,t-1}^{(m)}) \\ & - \mu(\hat{\theta}_{r,t+1}^{(m)} + \hat{\theta}_{r,t-1}^{(m)} + b_{r,t-1}^{(m)} - b_{r,t}^{(m)}) \Big], \end{aligned}$$

where $g_{r,t}^{(m)} = \frac{\partial V}{\partial \theta_{r,t}}\Big|_{\mathbf{\Theta}_r^{(m)}}$ and $G_{r,t}^{(m)} = -\frac{\partial^2 V}{\partial \theta_{r,t}^2}\Big|_{\mathbf{\Theta}_r^{(m)}}$ are the gradient and the information in the direction of $\theta_{r,t}$ evaluated in the current iterate values. The updates for $t = T$ are obtained in a similar fashion with the respective adjustments.

Note that in the maximum a posteriori estimates obtain in this fashion the coefficient differences ($\mathbf{b} = \mathbf{L\Theta}$) can be exactly zero. This induces a block partition of the parameters that is suitable for change-point identification (Rojas & Wahlberg, 2014; Harchaoui & Levy-Leduc, 2008).

3.3.1 Link prediction

Given a point estimate $\hat{\boldsymbol{\theta}}_{i,j,T}$ based on an observed sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, the probability of a directed link from node i to node j at time $T+1$ is estimated as

$$\begin{aligned}\hat{p}\left(y_{i,j,T+1} = 1 \mid \mathbf{Y}_T\right) = p\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,0) \mid \hat{\boldsymbol{\theta}}_{i,j,T}\right] \\ + p\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,1) \mid \hat{\boldsymbol{\theta}}_{i,j,T}\right],\end{aligned}$$

with a similar expression being valid for $\hat{p}\left(y_{j,i,T+1} = 1 \mid \mathbf{Y}_T\right)$.

3.4 Literature review on computation methods

The literature on algorithms for parameter estimation for linear regression with a fused lasso penalty is extensive. This is a challenging problem because the fused lasso penalty is not a separable and smooth function, and traditional optimization methods fail under these conditions. In particular, some algorithms that provide a solution path for sequential increments of the regularization parameter have been developed for the Fused Lasso Signal Approximator (FLSA) where the design matrix is $\mathbf{X} = \mathbf{I}$ (Friedman et al., 2007; Höfling, 2010b), and for a general full rank design matrix $\mathbf{X}$ (Tibshirani & Taylor, 2011) only in the case of gaussian regression.

In this work, we are interested in fused lasso penalized multiclass logistic regression. Friedman et al. (2010) explores coordinate descent regularization paths for logistic and multinomial logistic regression by using iteratively reweighted least squares (IRLS) but only for lasso, ridge and elastic net penalties (see also Krishnapuram & Hartemink, 2005). Höfling (2010a) proposes a coordinate-wise algorithm for the fused lasso that can be extended to logistic regression using iterative reweighted least squares (IRWLS), but no path solution algorithms have been fully developed for the multinomial logistic regression model setting. Recently, Yu et al. (2013) introduced a Majorization-Minimization (MM) algorithm for fused lasso penalized generalized linear models that benefits from parallel processing. They also present a good comparison with other existing algorithms including regularization path and first-order methods. For a fixed set of penalization parameters, several optimization algorithms have been proposed for fused lasso problems with general smooth and convex loss functions but not for the specific case of multinomial logistic regression. Liu et al. (2010) proposes an Efficient Fused Lasso Algorithm (EFLA) which solves a FLSA subproblem via a Subgradient Finding Algorithm. Goldstein & Osher (2009) use the split Bregman iteration method to deal with a set of image processing problems that can be treated as general L_1 penalized problems. Motivated by this idea, Ye & Xie (2011) developed the split Bregman based algorithm for the generalized fused lasso.

From a Bayesian perspective, a general hierarchical model for penalized linear regression that includes the fused lasso penalty is presented in Kyung et al. (2010) for the

Gaussian case (Park & Casella, 2008; Hans, 2009). In addition, Scott & Pillow (2012) used the data augmentation approach with Pólya-Gamma latent variables similar to the one we use here for full Bayesian inference of neural spike data counts observed over time by proposing a dynamic negative-binomial factor model with an autoregressive structure.

3.5 Penalty Parameter Selection

The value of the penalty parameter λ has a direct impact on the quality of the estimates and predictions generated by the model. Our default approach is to use time series cross-validation. However, selection of the optimal tuning parameter through cross-validation can be computationally expensive for large data (Tibshirani et al., 2005). A popular alternative method is the use of some model selection criteria (e.g AIC, BIC). In this section we describe the two methods for tuning parameter selection.

3.5.1 Time series cross-validation

We perform time series cross-validation by training the model on an observed sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$, and performing a one-step-ahead prediction for $\mathbf{Y}_{t+1}$ for a grid of values of λ. We repeat this procedure to obtain a set of predicted networks $\hat{\mathbf{Y}}_{t+1}, \ldots, \hat{\mathbf{Y}}_{t+m}$ for $t + m \leq T$, each of these predictions is compared against the respective observed networks, the number of false and true positives is computed, and a receiver operating characteristic (ROC) curve is constructed. The optimal penalty parameter is chosen as the value of λ in the grid that provides the highest area under the curve (AUC) average over the m predicted networks in the testing dataset.

3.5.2 Bayesian information criterion

Our approach is to select the penalty λ from among a pre-specified grid of values by maximizing the following Bayesian Information Criterion (BIC)

$$BIC_{\lambda} = \sum_{i<j} \left[2V_{i,j}(\hat{\mathbf{\Theta}}_{i,j}) - \mathcal{K}_{i,j}(\lambda) \log(T-1) \right],$$

where $\mathcal{K}_{i,j}(\lambda)$ is an estimate of the number of degrees of freedom when the penalty parameter λ is used to compute the MAP estimate. In the case of the fused lasso, Tibshirani & Taylor (2011) showed that the number of non-zero blocks of coefficients in $\hat{\mathbf{\Theta}}_{i,j}$ is rough unbiased estimate of the degrees of freedom.

3.6 Illustrations

The purpose of this section is to evaluate the performance of our model and compare it with the temporal Exponential Random Graph (tERGM) in terms of their link prediction ability. We used the **xergm** package in R to estimate the tERGM (Leifeld et al., 2014). More specifically, the tERGM is estimated with the btergm function, which implements the bootstrapped pseudolikelihood procedure presented in Desmarais & Cranmer (2012). The model includes all the typical ERGM terms, the square root of in and out-degrees as node covariates, and the lagged network and the delayed reciprocity to model cross-temporal dependencies. The results presented here are based on 1,000 MCMC simulations with other function parameters left as the default values (see btergm documentation for more details).

We evaluate the predictive capabilities of the models on both simulated and the real trading network data by performing one-step-ahead predictions for the structure of the last ten weeks in a similar fashion to the out-of-sample cross-validation exercise discussed in section 2.3.2. Before assessing prediction accuracy, we evaluate the performance of the two sampling schemes for Bayesian inference and also compare their efficiency with the optimization method for posterior mode estimation.

3.6.1 MCMC Performance

In order to asses and compare the performance of the latent variable FFBS and the direct sampling MCMC algorithms, we simulated data from our model. The parameters across all the pairs of nodes were randomly drawn from double exponential distributions as $\theta_{r,t} \sim \mathsf{DE}(\theta_{r,t-1}, 1/\lambda)$ with a true concentration parameter value of $\lambda = 3$. As a measure of efficiency, we use the effective sample size (ESS) computed as:

$$ESS = \frac{B}{1 + 2\sum_{k=1}^{K} \rho(k)}$$

where B is the number of post burn-in samples, $\rho(k)$ is the autocorrelation at lag k, and K is the cutoff lag point according to the initial monotone sequence estimator (Geyer, 1992).

We computed the effective sample size and the CPU run time in seconds for each pair of nodes based on 20,000 iterations after a burn-in period of 2,000 iterations. Table 3.1 shows the results obtained by averaging over 5 runs for each sampling scheme, including the relative efficiency of the algorithms standardizing for CPU run time. From these results it is clear the latent parameters scheme for the fused lasso is much more efficient than the direct sampler that uses the mixture of truncated normals. Based on these results, in the following sections we perform time series cross-validation and prediction for the Bayesian approach using the latent variable FFBS algorithm.

Table 3.1: Average ESS and CPU times per pair of nodes for MCMC algorithms for fused lasso model

Scheme	**ESS**	**CPU(s)**	**Rel.ESS**
Direct	113	9111.69	0.012
FFBS	4504	3403.79	1.32

It is useful to contrast the execution time of the MCMC algorithms with that of the optimization method, which is only 13.48 seconds in average for each pair of nodes using a stopping criteria of 10^{-5} for the relative error with the norm.

3.6.2 Simulation study

In this example we simulated a dataset from our model. The parameters across all the pairs of nodes were randomly drawn from double exponential distributions as $\theta_{r,t} \sim \mathsf{DE}(\theta_{r,t-1}, 1/\lambda)$ with a true penalty parameter value of $\lambda = 12$ using initial values $\theta_{r,0} = 0$. The resulting network is relatively dense with an average number of links of 2682 each week (out of 4970 possible ties), and it shows low reciprocity and high transitivity.

First, we evaluate the predictive ability using a setup similar to calibration cross-validation (CCV) by partitioning the data into three sets. The first set is used for modeling and consists of the first 181 weeks of data. Selection of the optimal penalization parameter was performed on the calibration set corresponding to weeks $t = 182, \ldots, 191$, by searching the value of λ that maximizes the mean AUC over the predictions of these ten weeks. Finally, we report out-of-sample prediction accuracy on the validation set consisting of the last ten weeks of data. The search of λ was conducted over a grid of 28 values between 0.5 and 14.

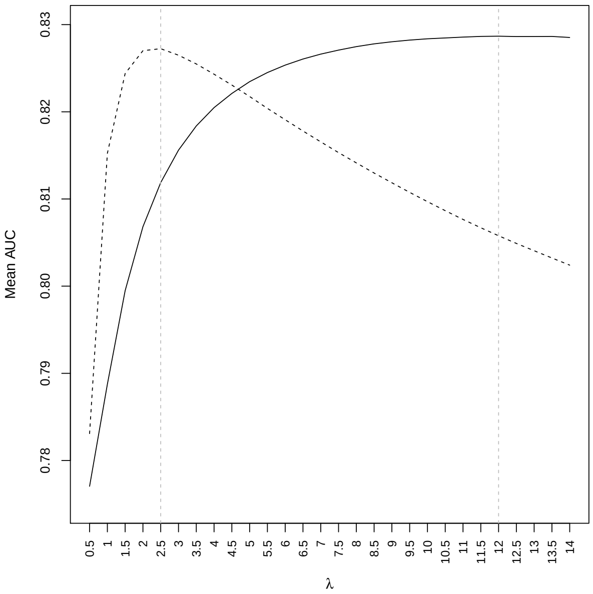

Figure 3.1: Mean AUC over weeks $t = 182, 183, \ldots, 191$ for simulated data over a grid of values of λ. The solid line corresponds to the Bayesian algorithm, and the dotted line to the optimization method.

Figure 3.1 shows that the cross-validation optimal values of λ are 2.5 using the optimization method, and 12 for the Bayesian scheme. Note that the Bayesian algorithm is approximately recovering the true penalty parameter value. In this particular case, the optimal tuning parameter obtained with BIC for the Bregman optimization method coincides with the cross-validation value of $\lambda = 2.5$. For this reason, only one set of prediction results is presented for the Bregman method.

Figure 3.2 shows the ten operating characteristic curves associated with the out of sample predictions for the last ten weeks using our model with FFBS algorithm and $\lambda = 12$, and the estimates of the area under the receiver operating characteristic curves (AUC) for the tERGM and our proposed model for the two different estimation methods. The prediction accuracy for the Bregman optimization and FFBS Bayesian algorithms is almost

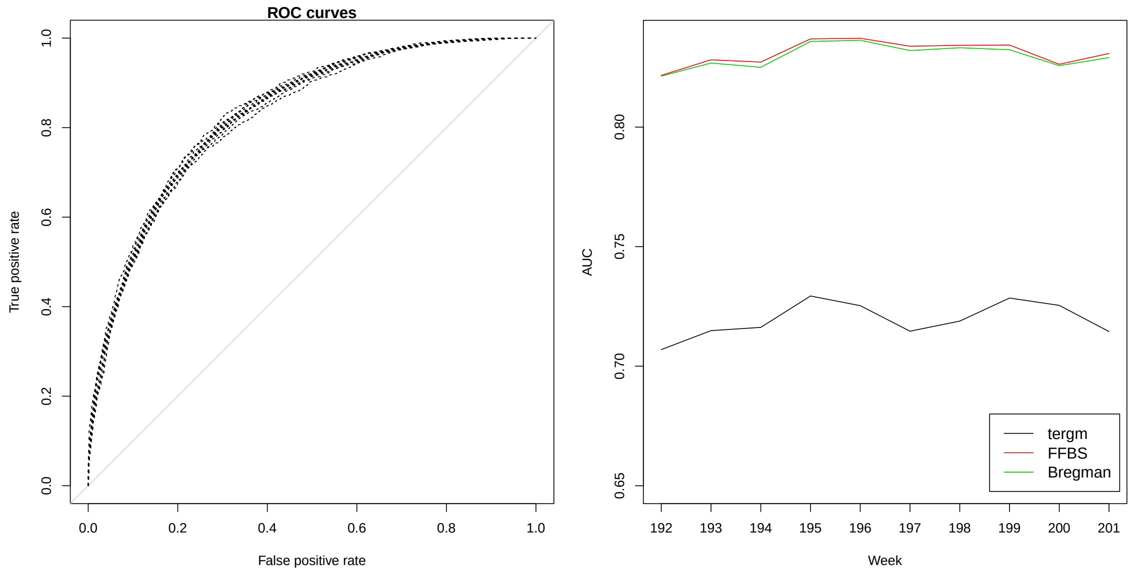

Figure 3.2: Plots of the ten operating characteristic curves associated with one-step-ahead out of sample predictions from the fused lasso model with FFBS algorithm (leftt panel). Area under the curves (AUC) for the temporal ERGM, and the fused lasso model with FFBS algorithm and Bregman optimization for simulated data.

identical, with the Bayesian scheme showing a very small advantage. In this scenario, our model outperforms the tERGM by at least 10% in the AUC values. The prediction accuracy of the fused lasso model is good with an average AUC value of 83% over the 10 weeks, while the performance of the temporal ERGM is only fair with an average AUC value of 72% .

3.6.3 NYMEX financial trading network

In this section we analyze the sequence of real weekly financial trading networks. Exploration of this data showed that this trading network is sparse, and consistently shows very high reciprocity, moderate transitivity, mixing patterns and community structure (see section 1.3).

Analogously to the previous section, we use the CCV method and select the op-

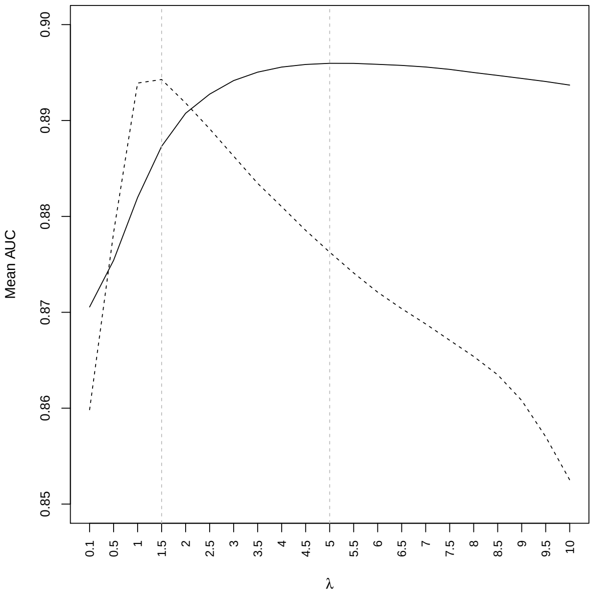

Figure 3.3: Mean AUC over weeks $t = 182, 183, \ldots, 191$ for real trading network data over a grid of values of λ. The solid line corresponds to the Bayesian algorithm, and the dotted line to the optimization method.

timal penalization parameter by searching the value of λ that maximizes the mean AUC over the predictions of ten weeks $t = 182, \ldots, 191$. The search was conducted over a grid of 21 values between 0.01 and 10. Figure 3.3 shows that the optimal values of λ are 1.5 using the optimization algorithm, and 5 for the Bayesian scheme. In order to select the penalty parameter for the Bregman optimization method utilizing BIC, we search the optimal values over a grid of 43 values between 0.1 and 20. The resulting optimal parameter value is $\lambda = 3$.

Figure 3.4 shows the ten operating characteristic curves associated with the out-of-sample predictions for the last ten weeks for our model estimated with FFBS algorithm and $\lambda = 5$, and the estimates of the area under the receiver operating characteristic curves (AUC) for the tERGM, the hidden Markov model (HMM) introduced in chapter one, and the

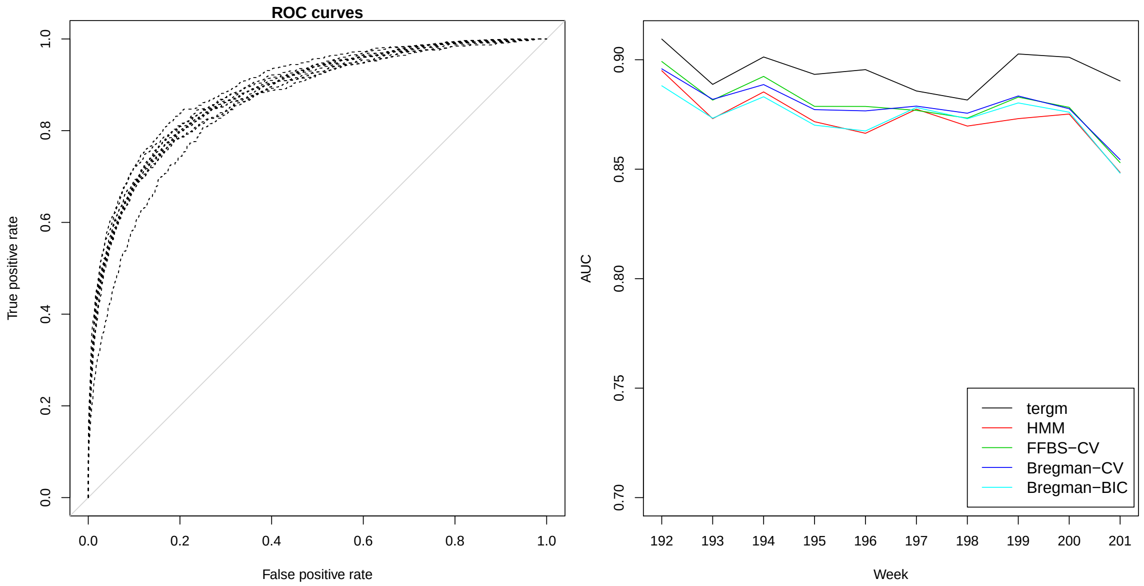

Figure 3.4: Plots of the ten operating characteristic curves associated with one-step-ahead out of sample predictions from the fused lasso model with FFBS algorithm (leftt panel). Area under the curves (AUC) for the tERGM, the HMM from chapter one, and the fused lasso model with FFBS algorithm and Bregman optimization for the trading network. CV (cross-validation) and BIC represent the two methods for tuning parameter selection.

fused lasso model for the two different estimation schemes and tuning parameter selection methods. The prediction accuracy of the HMM model and the Bregman-BIC optimization algorithm are comparable, and show the lowest AUC values across methods. On the other hand, the performances of the optimization and FFBS Bayesian algorithms using cross-validation are very similar. The prediction accuracy of our proposed model is good with an average AUC value of 88% over the 10 weeks. However, in this particular case the tERGM slightly but consistently outperforms all versions of our model.

Change-point identification

As we mentioned in section 3.3, the maximum a posteriori estimates of the parameters in the fused lasso regression model can be used to identify changes in the network

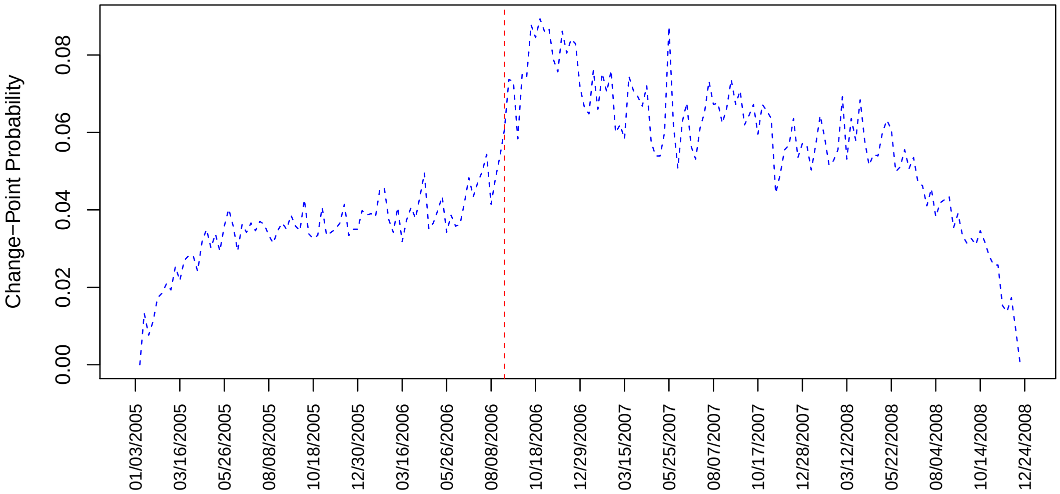

Figure 3.5: Time series of the estimated change-point probability for the trading network. The vertical line represents the introduction of electronic trading in the market at week 85.

structure over time. In this case, we use an indicator variable that assigns a value of 1 if at least one of the three parameters for pair (i, j) change from time t to time $t+1$, and 0 otherwise. The average fraction over all pairs of nodes provides a rough estimation of the change-point probability over time.

Figure 3.5 presents the time series of the estimated change-point probability for the trading network obtained using the Bregman optimization algorithm with $\lambda = 1.5$. The vertical line represents the introduction of electronic trading in the market, and right after this event we observe the maximum change probability over the 201 weeks. This result is consistent with previous exploration of this data.

3.7 Discussion

Even though the tERGM seems to have a slightly better performance than our fused lasso model, the proposed model shows competitive prediction accuracy results in this setting and superior results in the simulated data scenario. The model can be easily extended to weighted networks by replacing the multinomial likelihood with an appropriate member of the exponential family. Similarly, a variation of the model can be devised for undirected networks.

In terms of the two different estimation approaches, the Bayesian scheme is useful when the main interest is to perform full inference of the model parameters. The results on simulated and real data showed that the prediction ability of the optimization approach is competitive compared to the Bayesian method, and the algorithm is far more computationally efficient. In addition, the optimization approach also provides a way of exploring change points in the network dynamics. On the other hand, the cross-validation approach for tuning parameter selection provides slightly better results than the BIC method but the computational cost is considerably higher. This can be a limitation as the number of nodes in the network increases.

Chapter 4

Sparse autologistic model for link prediction

In this chapter, we extend the notion of an autologistic model to perform link prediction in directed binary networks. In the much simpler case of a binary time series $y_1, y_2, \ldots, y_T$ with $y_t \in \{0, 1\}$, a first order autologistic model with parameters α and ξ assumes that

$$\operatorname{logit} \mathsf{Pr}(y_t = 1 \mid y_{t-1}, \ldots, y_1) = \eta_t = \alpha + \xi y_{t-1} \tag{4.1}$$

where $\operatorname{logit} x = \log\{x/(1-x)\}$ and the unknown parameters α and ξ control the structure of the temporal dependence (in particular, note that $\xi = 0$ implies that the observations are independent and identically distributed). This implies that

$$p(y_2, \ldots, y_T \mid y_1, \alpha, \xi) = \prod_{t=2}^{T} \frac{\exp\{y_t(\alpha + \xi y_{t-1})\}}{1 + \exp\{\alpha + \xi y_{t-1}\}}.$$

Autologistic models for spatio-temporal binary data have been discussed in Zhu et al. (2008) and Zheng & Zhu (2008). However, these models for spatio-temporal data cannot be directly applied to the type of time series data discussed in this work because they are not designed to account for common features of directed network data such as reciprocity and transitivity. In contrast, the class of models discussed here are specifically designed to account for these features, and its parameters have a direct interpretation in terms of the properties of the networks. Furthermore, unless the system is observed for a very long time, the number of parameters in our auto logistic model will typically be much larger than the number of available observations. In this case, to control overfitting we impose an L^1 penalty on the model parameters.

4.1 Model Formulation

We consider an extension of (4.1) in which the pairs $\{(y_{i,j,t}, y_{j,i,t}) : i < j\}$ are assumed conditionally independent given the history of the network, and each pair $(y_{i,j,t}, y_{j,i,t})$ is modeled using a logistic model of the form

$$p\left(y_{i,j,t}, y_{j,i,t} \mid \mathbf{Y}_{t-1}\right) \propto \exp\left\{\eta_{i,j,t,1} y_{i,j,t} + \eta_{i,j,t,2} y_{j,i,t} + \eta_{i,j,t,3} y_{i,j,t} y_{j,i,t}\right\}, \qquad (4.2)$$

where the natural parameters $\eta_{i,j,t,1} = f_{i,j,1}(\mathbf{Y}_{t-1})$, $\eta_{i,j,t,2} = f_{i,j,2}(\mathbf{Y}_{t-1})$ and $\eta_{i,j,t,3} = f_{i,j,3}(\mathbf{Y}_{t-1})$ depend on time only through $\mathbf{Y}_{t-1}$.

A full specification of the model requires that we specify the form of the log-linear predictors $f_{i,j,1}$, $f_{i,j,2}$ and $f_{i,j,3}$. A tempting option is to make these predictors dependent of all entries of $\mathbf{Y}_{t-1}$, including all high order interactions. However, such an approach leads to models with an extremely high number of parameters that is computationally unmanageable even for networks with a relatively small number of nodes. On the other hand, while focusing only on first order effects associated with the entries of $\mathbf{Y}_{t-1}$ can substantially reduce the number of parameters, the resulting model ignores interactions that could be expected to be important. We take a middle ground approach and include in the specification of the linear predictors a subset of the first and second order effects that are associated with the interactions of nodes i and j among themselves and with other nodes during the previous period. In particular, we set

$$\begin{aligned} f_{i,j,l}(\mathbf{Y}_{t-1}) = \alpha_{i,j,l} + \beta_{i,j,l} y_{i,j,t-1} + \gamma_{i,j,l} y_{j,i,t-1} + \sum_{k \neq i,j} \delta_{i,j,k,l} y_{i,k,t-1} + \sum_{k \neq i,j} \phi_{i,j,k,l} y_{k,j,t-1} \\ + \sum_{k \neq i,j} \psi_{i,j,k,l} y_{j,k,t-1} + \sum_{k \neq i,j} \omega_{i,j,k,l} y_{k,i,t-1} + \rho_{i,j,l} y_{i,j,t-1} y_{j,i,t-1} \\ + \sum_{k \neq i,j} \xi_{i,j,k,l} y_{i,k,t-1} y_{k,j,t-1} + \sum_{k \neq i,j} \zeta_{i,j,k,l} y_{j,k,t-1} y_{k,i,t-1} \end{aligned} \tag{4.3}$$

for $l = 1, 2, 3$. To better motivate this specification for the log-linear predictors, consider for example the structure of $f_{i,j,1}(\mathbf{Y}_{t-1})$ in (4.3). As we showed before in section 3.1, we can roughly interpret $f_{i,j,1}(\mathbf{Y}_{t-1})$ as controlling the probability of a directed link between i and j. Hence, $\alpha_{i,j,1}$ can be interpreted as the baseline probability of a link between nodes i and j, the coefficient $\beta_{i,j,1}$ can be interpreted as the persistence in the relationship (e.g., if $\beta_{i,j,1} > 0$ then once trader i starts selling to trader j, they tend to keep selling in future periods), the coefficients $\{\delta_{i,j,k,1} : k \neq i, j\}$ and $\{\phi_{i,j,k,1} : k \neq i, j\}$ capture substitution/diversification effects (e.g., if $\delta_{i,j,k,1} > 0$ then it is more likely that i will sell to j if it sold to k in the previous term), and $\xi_{i,j,k,1}$ capture transitivity effects such as disintermediation (e.g., if $\xi_{i,j,k,1} > 0$ then trader i is more likely to sell to j if in the previous period i sold to k and k sold to j, so that i and j tend to cut k as middleman).

The conditional independence assumption in our model is key because it dramatically simplifies computation. However, its main drawback is that it cannot capture intra-temporal transitivity (i.e., an increase/decrease in the probability of a link between nodes i and j at time t if they both link to a third node k at time t). One reason why this is not a major concern for us is that including the second order interactions in (4.3) do allow us to capture inter-temporal transitivity (i.e., an increase/decrease in the probability of a link between nodes i and j at time t if they both linked to a third node k at time $t-1$), which is often a more interesting and realistic type of transitivity effect.

In the sequel it is convenient to reformulate (4.2) as a multinomial logistic regression (see section 3.1) such that the likelihood can be rewritten as:

$$p\left(z_{i,j,t} = r | \mathbf{x}_{i,j,t}, \boldsymbol{\alpha}^*_{i,j}, \boldsymbol{\Theta}_{i,j}\right) = \frac{\exp\left(\alpha^*_{i,j,r} + \mathbf{x}'_{i,j,t}\boldsymbol{\theta}_{i,j,r}\right)}{\sum_{s=1}^{4} \exp\left(\alpha^*_{i,j,s} + \mathbf{x}'_{i,j,t}\boldsymbol{\theta}_{i,j,s}\right)}, \tag{4.4}$$

where $\boldsymbol{\alpha}^*_{i,j} = (\alpha_{i,j,1}, \alpha_{i,j,2}, \alpha_{i,j,1} + \alpha_{i,j,2} + \alpha_{i,j,3}, 0)$, $\boldsymbol{\Theta}_{i,j} = (\boldsymbol{\theta}_{i,j,1}, \boldsymbol{\theta}_{i,j,2}, \boldsymbol{\theta}_{i,j,3}, \mathbf{0})$, for $r = 1, 2$ the vector of parameters is:

$$\begin{aligned}\boldsymbol{\theta}_{i,j,r} = (&\beta_{i,j,r}, \gamma_{i,j,r}, \delta_{i,j,1,r}, \ldots, \delta_{i,j,n,r}, \phi_{i,j,1,r}, \ldots, \phi_{i,j,n,r}, \psi_{i,j,1,r}, \ldots, \psi_{i,j,n,r},\\ &\omega_{i,j,1,r}, \ldots, \omega_{i,j,n,r}, \rho_{i,j,r}, \xi_{i,j,1,r}, \ldots, \xi_{i,j,n,r}, \zeta_{i,j,1,r}, \ldots, \zeta_{i,j,n,r})',\end{aligned}$$

$\boldsymbol{\theta}_{i,j,3}$ is the corresponding vector with the sum of the parameters over the three log-linear predictors, and

$$\begin{aligned}\mathbf{x}_{i,j,t} = (&y_{i,j,t-1}, y_{j,i,t-1}, y_{i,1,t-1}, \ldots, y_{i,n,t-1}, y_{1,j,t-1}, \ldots, y_{n,j,t-1}, y_{j,1,t-1}, \ldots, y_{j,n,t-1},\\ &y_{1,i,t-1}, \ldots, y_{n,i,t-1}, y_{i,j,t-1}y_{j,i,t-1}, y_{i,1,t-1}y_{1,j,t-1}, \ldots, y_{i,n,t-1}y_{n,j,t-1},\\ &y_{j,1,t-1}y_{1,i,t-1}, \ldots, y_{j,n,t-1}y_{n,i,t-1})'.\end{aligned}$$

Note that each of the vectors $\boldsymbol{\theta}_{i,j,r}$ has dimension $d = 3 + 6(n-2)$, so that the total number of parameters associated with the pair (i, j) is $p = 9 + 18(n-2)$. This grows linearly with the number of nodes, but it is constant as a function of time.

4.1.1 Prior specification

The dimensionality of the autologistic model we just described will typically be quite large. In fact, the number of parameters in the model will often be larger than the number of observations available to estimate them. Hence, we impose double exponential priors on the coefficients in order to address overfitting. This approach is equivalent to using a regularized likelihood approach based on L^1 penalty functions commonly called lasso regression.

More specifically, the posterior distribution of the autologistic model is:

$$\sum_{i<j} \left\{ V_{i,j}(\boldsymbol{\alpha}^*_{i,j}, \boldsymbol{\Theta}_{i,j}) - \lambda \|\boldsymbol{\Theta}_{i,j}\|_1 \right\} \tag{4.5}$$

where

$$V_{i,j}(\boldsymbol{\alpha}^*_{i,j}, \boldsymbol{\Theta}_{i,j}) = \sum_{t=2}^{T} \left\{ \alpha^*_{i,j,z_{i,j,t}} + \mathbf{x}'_{i,j,t} \boldsymbol{\theta}_{i,j,z_{i,j,t}} \right\} - \log \left(\sum_{s=1}^{4} \exp \left\{ \alpha^*_{i,j,s} + \mathbf{x}'_{i,j,t} \boldsymbol{\theta}_{i,j,s} \right\} \right)$$

is the (unpenalized) log-likelihood, $\| \cdot \|_1$ denotes the L_1-norm, and $\lambda > 0$ is the penalty parameter that controls the shrinkage level of the coefficients towards zero. Note that the intercept parameters $\left\{ \alpha^*_{i,j,r} \right\}$ receive a flat prior as they remain unpenalized. As a consequence of the conditional independence assumption, (4.5) can be broken down into $n(n-1)/2$ estimation problems, each one corresponding to fitting a separate lasso multinomial regression for each pair of nodes in the network.

4.1.2 Model properties

Because our estimation procedure reduces to fitting independent L^1 regularized multinomial logistic regressions for each pair of nodes, the procedure shares all the positive (and negative) properties of this type of approaches. The book of Bühlmann & van de Geer (2011) presents a complete study of L^1-penalization based statistical methods for high-dimensional data including an extensive literature review on its properties (see also Vidaurre et al., 2013). For example, Fan & Li (2001) establishes results of consistency and variable selection consistency of penalized likelihood estimators for linear and generalized linear models in the case $p \ll T$. These results are valid under regularity conditions and guarantee the oracle property that the limiting distribution of the non-zero coefficients is the same as if only the relevant features had been included in the model.

A more challenging problem is the high-dimensional case where $p \gg T$. In this scenario the Lasso solution is not unique and it may not recover the true model but the solution is still sparse and recovers the relevant features. Zhao & Yu (2006) proved that for linear regression, the irrepresentable condition is a sufficient and almost necessary condition for least squares Lasso model selection consistency for both the classical fixed p scenario and p growing with T (see also Greenshtein, 2006). Meinshausen & Bühlmann (2006) independently found a similar result for lasso consistency in random Gaussian designs applied to graphical models. Wainwright (2009) consolidated the previous results by providing bounds on the minimum sample size for correct model selection for the Gaussian case in terms of p and the number of relevant features s. On the other hand, Kakade et al. (2010) explores consistency and sparsity of L^1 regularized estimators under the strong convexity property of the loss function for general exponential families. Using the Restricted Eigenvalue (RE) condition (Bickel et al., 2009), they show that, when p grows with T, the convergence rate for L^1 regularized GLM's is also $O(\frac{s \log p}{T})$. This result was previously provided in Zhao & Yu (2006) for linear regression.

The previous results do not directly apply to the regularization problem (4.5) because in our case p grows with the number of nodes and not the number of observations in time. However, Lee et al. (2014) established consistency and model selection consistency of M-estimators with geometrically decomposable penalties for quadratic loss functions and generic exponential families. The results rely on the conditions of restricted strong convexity and irrepresentability on the Fisher information matrix (see also Negahban et al. (2012)). This is a very general setting that encompasses the L^1 regularized multinomial logistic problem of interest, extending these desirable properties to our modeling approach.

4.1.3 Literature review on computation methods

Here, we are interested in implementing a Lasso penalized multiclass logistic regression. Some of the available algorithms to approach this problem include the stochastic gradient descent (SGD) algorithm for L1-penalized loss minimization introduced by Shalev-Shwartz & Tewari (2011) (see also Carpenter, 2008; Tsuruoka et al., 2009). Even though these algorithms are a feasible option, they are particularly useful when the training set is large. Goldstein & Osher (2009) proposes a coordinate descent method combined with a split Bregman iteration to deal with a set of image processing problems that can be easily adapted to general L_1 penalized problems and applied in this context.

In addition, Friedman et al. (2010) presents coordinate descent regularization paths for logistic and multinomial logistic regression by using iteratively reweighted least squares (IRLS) for lasso, ridge and elastic net penalties. From a Bayesian perspective, Cawley et al. (2007) proposes a refined cross-validation technique for lasso multinomial logistic regression by marginalizing over the penalty parameter. Implementation of coordinate descent methods combined with cross-validation or a Bayesian approach is computationally inefficient due to the high-dimensionality and the large number of optimization subproblems based on the number of nodes in the network. More recently, Tutz et al. (2015) introduce the Categorically Structured (CATS) lasso for variable selection in multinomial logit models. This approach is similar in spirit to a group Lasso and it is designed to distinguish between mandatory and optional predictors. This can be an interesting direction for future work in our modeling setting.

4.2 Posterior Mode Estimation

In this section we describe a relatively simple computational algorithm similar to iterative reweighed least squares to perform posterior mode estimation in our lasso regression model. In particular, dropping the subindex (i, j) to simplify notation, we solve (4.5) by iteratively setting

$$\left(\hat{\boldsymbol{\alpha}}^{*(m+1)}, \hat{\boldsymbol{\Theta}}^{(m+1)}\right) = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\alpha}^{*}, \boldsymbol{\Theta} \mid \hat{\boldsymbol{\alpha}}^{*(m)}, \hat{\boldsymbol{\Theta}}^{(m)}\right)$$

until convergence, where $Q(\boldsymbol{\alpha}^{*}, \boldsymbol{\Theta} \mid \tilde{\boldsymbol{\alpha}}^{*}, \tilde{\boldsymbol{\Theta}})$ is a surrogate function obtained by replacing $V(\boldsymbol{\alpha}^{*}, \boldsymbol{\Theta})$ in (4.5) by its second-order Taylor expansion around the current iterate (e.g., see Krishnapuram & Hartemink, 2005). However, rather than attempting to solve the problem using block-wise updates, we proceed with component-wise steps. In particular, the estimate of a component $\theta_{r,k}$ of $\boldsymbol{\theta}_r$ is updated as

$$\hat{\theta}_{r,k}^{(m+1)} = \text{soft}\left(\hat{\theta}_{r,k}^{(m)} - \frac{g_{r,k}^{(m)}}{G_{r,k}^{(m)}}; \frac{-\lambda}{G_{r,k}^{(m)}}\right), \tag{4.6}$$

where $\text{soft}(w, \lambda) = \text{sgn}(w)\max\{0, |w| - \lambda\}$ is the soft-thresholding operator, $g_{r,k}^{(m)} = \frac{\partial V}{\partial \theta_{r,k}}$ and $G_{r,k}^{(m)} = -\frac{\partial^2 V}{\partial \theta_{r,k}^2}$ are the gradient and the information in the direction of $\theta_{r,k}$ evaluated in the current iterate values. The update equation for the intercepts is similar to a simple IRLS iteration, $\hat{\alpha}_r^{*(m+1)} = \hat{\alpha}_r^{*(m)} - \frac{g_{r,0}^{(m)}}{G_{r,0}^{(m)}}$.

Note that in the maximum a posteriori estimates obtained using this algorithm, the coefficients can be exactly zero automatically performing model variable selection and therefore addressing substantive questions about the presence of substitution and transitivity effects in the network. However, when the regression matrix is not full rank (for example, when $p >> T$), interpretation of the individual effects is difficult because of confounding. To address this issue we focus on identifying regression coefficients for which there is no evidence of significance. These are selected by identifying the effects that lie in the orthogonal complement of the column space of $\mathbf{X}_{i,j}(\mathcal{A}_\lambda)$, the submatrix that contains the columns associated with variables that have been identified as significant.

4.2.1 Selection of the penalty parameter and link prediction

Our default approach is to select the penalty λ from among a pre-specified grid of values by maximizing the Bayesian Information Criteria described in section 3.5.2, where $\mathcal{K}_{i,j}(\lambda) = \text{rank}\{\mathbf{X}_{i,j}(\mathcal{A}_\lambda)\}$ is an estimate of the number of degrees of freedom when the penalty parameter λ is used to compute $(\hat{\boldsymbol{\alpha}^*}_{i,j}, \hat{\boldsymbol{\Theta}}_{i,j})$, and $\mathbf{X}_{i,j}(\mathcal{A}_\lambda)$ is a $(T-1) \times d$ matrix whose t-th row contains a subset of elements of $\mathbf{x}_{i,j,t}$ and whose columns correspond to the covariates for which $\hat{\boldsymbol{\theta}}_{i,j,r}$ is different from zero for at least one value $r = 1, 2, 3$ (Park & Hastie, 2007; Zou et al., 2007; Tibshirani & Taylor, 2012). Note that for all values of λ, the degrees of freedom satisfy the condition $0 \leq \mathcal{K}_\lambda \leq \min\{d, T-1\}$.

Similar to the discussion in the previous chapter, given a point estimate $(\hat{\boldsymbol{\alpha}}^*_{i,j}, \hat{\boldsymbol{\Theta}}_{i,j})$ based on an observed sample $\mathbf{Y}_1, \dots, \mathbf{Y}_T$, we can estimate (for $i > j$) the probability of a directed link from node i to node j at time $T+1$ as

$$\begin{aligned}\hat{p}\left(y_{i,j,T+1} = 1 \mid \mathbf{Y}_T\right) = p&\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,0) \mid \hat{\boldsymbol{\alpha}}^*_{i,j}, \hat{\boldsymbol{\Theta}}_{i,j})\right] \\ &+ p\left[(y_{i,j,T+1}, y_{j,i,T+1}) = (1,1) \mid \hat{\boldsymbol{\alpha}}^*_{i,j}, \hat{\boldsymbol{\Theta}}_{i,j})\right],\end{aligned}$$

with a similar expression being valid for $\hat{p}\left(y_{j,i,T+1} = 1 \mid \mathbf{Y}_T\right)$.

4.3 Illustrations

In this section we assess the predictive accuracy of the sparse autologistic model on simulated data and the real financial trading network, and compare it with the tERGM by performing one-step ahead predictions of the last ten weeks in a similar fashion to previous chapters. In addition, we compare the sparse autologistic model with the Bayesian hidden Markov model and fused lasso models discussed in this work.

4.3.1 Simulation study

In this example we simulated a dataset consisting of $T = 201$ observations from our model such that, for each of the 2485 pairs of nodes, only six non-zero coefficients are present for each class, $r = 1, 2, 3$. Three of the non-zero coefficients for each class correspond to $\alpha_{i,j,r}$, $\beta_{i,j,r}$, and $\gamma_{i,j,r}$. We randomly draw these parameters from common

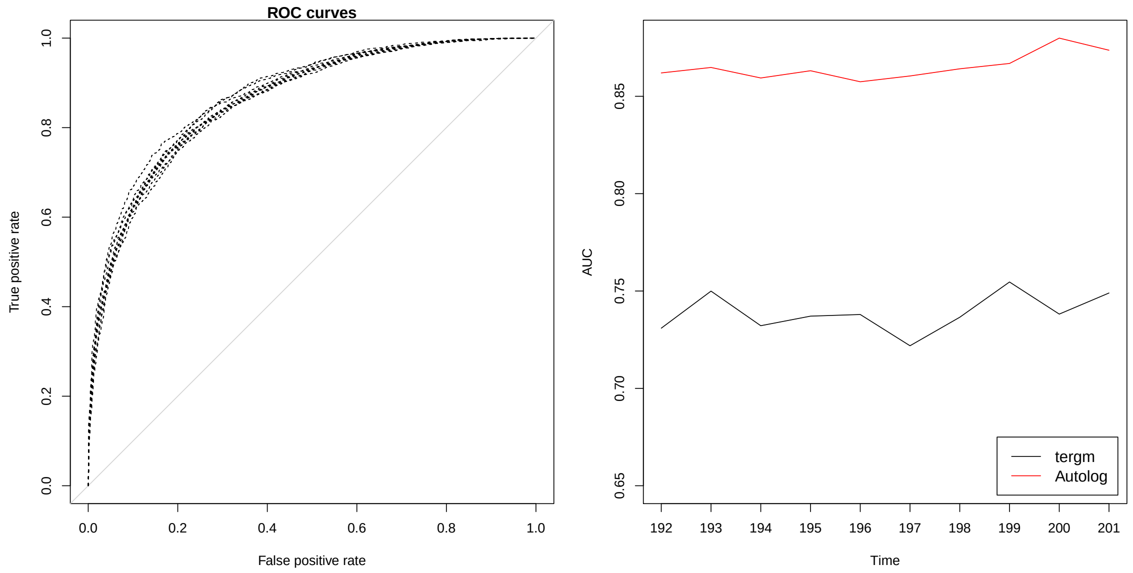

Figure 4.1: Plots of the ten operating characteristic curves associated with one-step-ahead out of sample predictions from the autologistic model, and area under the curves (AUC) for the tERGM and the autologistic model for the simulated dataset.

Gaussian distributions across pairs (e.g. $\alpha_{i,j,r} \sim N(\bar{\alpha}_r, \tau_r^2)$). The other relevant coefficients correspond to $\xi_{i,j,k,r}$ for three different values of k. Four groups of pairs of traders of similar sizes were simulated with different selections of the three values of k, and the respective parameter values of $\xi_{i,j,k,r}$ were fixed equal within each of the groups and with opposites signs to the global mean of $\alpha_{i,j,r}$ and $\beta_{i,j,r}$.

Under this simulation scheme, the persistence of the relationship between the nodes i and j and a few transitive relationships drive the network structure and dynamics over time. The resulting network is relatively dense with a average number of links of 2971 (out of 4970 possible ties), and it shows low reciprocity and high transitivity. In this case, the optimal penalty parameter was searched over a grid of 55 values between 0.01 and 35 obtaining an optimal value of $\lambda = 12$. Figure 4.1 shows the ten operating characteristic curves associated with one-step-ahead out of sample predictions from our autologistic model,

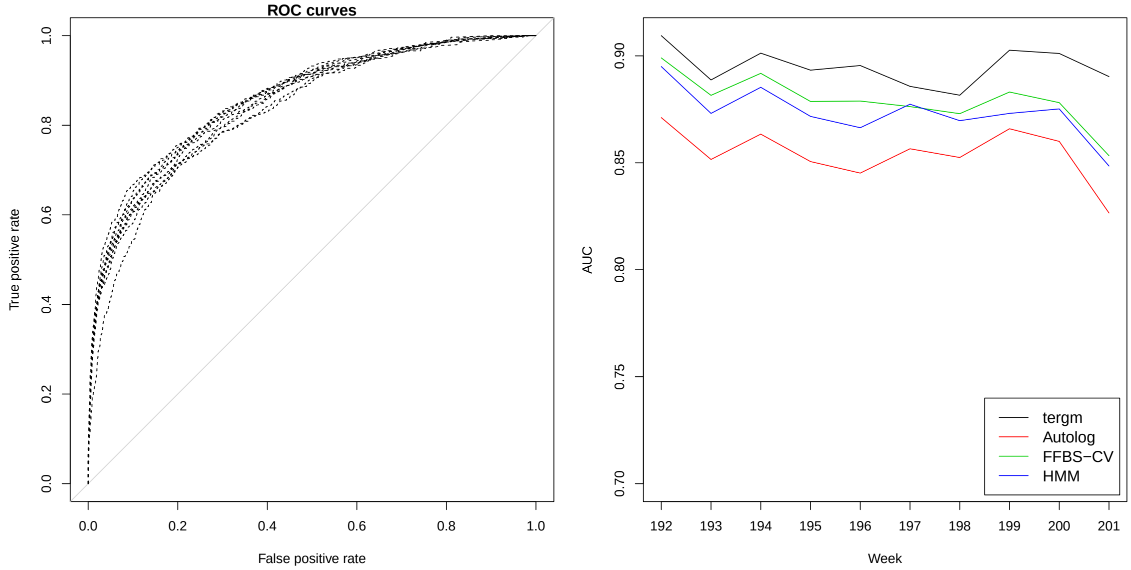

Figure 4.2: Plots of the ten operating characteristic curves associated with one-step-ahead out of sample predictions from the autologistic model, and area under the curves (AUC) for the the tERGM, autologistic model, hidden Markov model and fused lasso model for the trading network.

along with estimates of the area under the receiver operating characteristic curves (AUC) for the proposed model and the tERGM. According to these results, the predictive accuracy of the tERGM is fair as it only reaches AUC values below 75% in most cases. Our autologistic model outperforms the tERGM by at least 11% in the AUC values for the ten predicted weeks showing good prediction accuracy.

4.3.2 NYMEX financial trading network

Selection of the optimal penalization parameter in this case was performed by searching over a grid of 40 values between 0.01 and 20 obtaining an optimal value of $\lambda = 10$. Figure 4.2 shows the ten operating characteristic curves associated with the out of sample predictions from the autologistic model, and the estimates of the area under the receiver

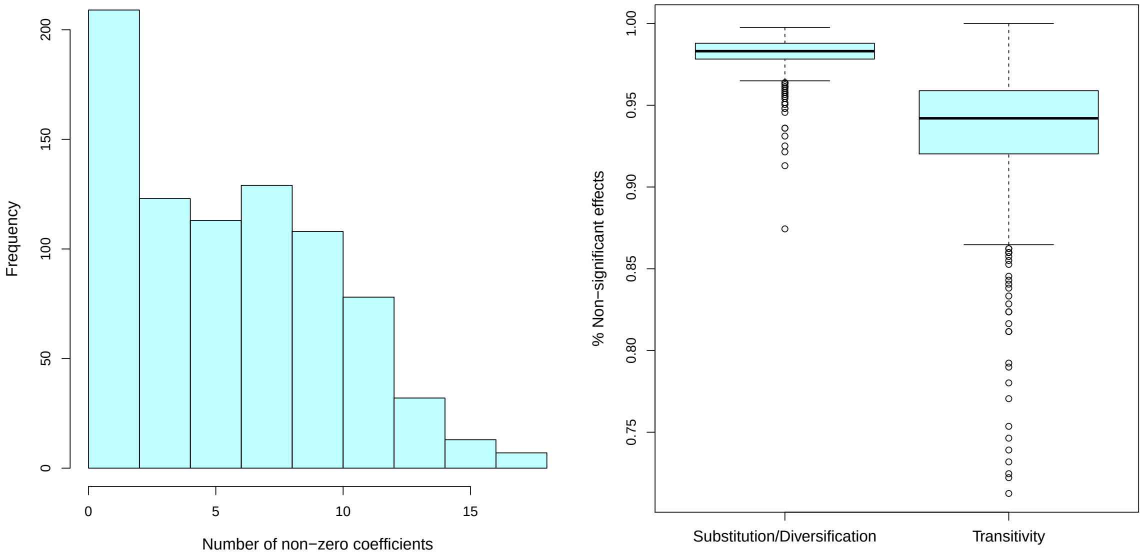

Figure 4.3: Histogram of the number non-zero coefficients and boxplots of the percentage of non-significant effects of substitution/diversification and transitivity for 812 pairs of traders.

operating characteristic curves (AUC) for the tERGM, autologistic model, hidden Markov model and the fused lasso model proposed in this work. In this case, the sparse autologistic model shows the lowest AUC values acroos models. and the temporal ERGM outperforms our proposed model by 3% to 6% in the AUC values. However, the prediction performance of the autologistic model is good with an average AUC value of 85% over the 10 weeks.

Now, we turn our attention to the interpretation of the regression coefficients. Our goal is to determine the relative importance of the different types of effects discussed in Section 4.1, in controlling the evolution of the trading network. Indeed, one advantage of this model over the ones previously presented in this dissertation is its interpretability.

Recall that the trading network has $n = 71$ traders, so that the autologistic model includes 1251 covariates (excluding the unpenalized intercept). Of these 1251 coefficients, 6 capture persistence effects, 3 capture inter-temporal reciprocity, 828 capture substitution/diversification effects, and 414 capture inter-temporal transitivity. Since the number of covariates is much larger than the number of observations, these effects are confounded with each other, complicating the interpretation of the model. To address this issue we focus on identifying effects for which there is no evidence of significance (see Section 4.2). Although the dimensionality of the problem prevents us from identifying exactly which effects do affect the evolution of the relationship between a given pair of traders, it allows us to clearly identify which effects do not.

Table 4.1: Number of effects, and non-presence percentage over 812 pairs of traders.

	Persistence	**Reciprocity**	**Subs/Div**	**Transitivity**
# of effects	6	3	828	414
% non-present	72.8	88.8	0.0	0.4

Our results suggest that the regression models for each pair of traders tend to be quite sparse. Indeed, only 812 out of 2485 pairs of traders have at least one non-zero regression coefficient aside from the intercept and the number of non-zero coefficients is very low across these pairs (see first panel Figure 4.3). Furthermore, for the 812 pairs that show at least one significant effect, a large percentage have no persistence (72.8%) or inter-temporal reciprocity (88.8%) coefficients that are significant (see Table 4.1). In contrast, each one of these 812 pairs presents at least one substitution/diversification coefficient that might be significant, and the vast majority (99.6%) present at least one potentially significant inter-temporal transitivity effect (again, see Table 4.1). This clearly suggests that second-order effects (substitution/diversification and transitivity) are much more important in this financial trading network than first order effects (persistence and reciprocity). The relatively low importance that persistence seems to have on the evolution of the trading network is particularly surprising, specially given that previous analysis of the network suggested that the structure remained stable over long periods of time (e.g., see Section 2.3).

To better understand the relative importance of the second order effects, we also compute the number of non-significant effects of each type for each pair of traders (see Figure 4.3). Note that, although all 812 pairs present at least one substitution/diversification effect, the number of these effects that might be significant on each pair is relatively small. In contrast, the number of potentially significant coefficients associated with inter-temporal transitivity effects tends to be larger, with a few pairs presenting more than 25% of potentially significant effects. These results suggests that the evolution of this trading network is driven in majority by inter-temporal transitivity effects.

Chapter 5

Conclusions

Through this work we found interesting results about the structure of financial trading networks and the matching trends of traders in the New York Mercantile Exchange (NYMEX) natural gas futures market. As we show in our illustration, by developing a dynamic fully probabilistic hidden Markov model for array-valued data we are able to monitor structural changes in the network while at the same time making accurate short-term predictions of future links. However, the complexity of this model increases rapidly with the number of nodes making efficient short term prediction of future outcomes of the system a challenge for big network data. The Bayesian fused lasso model provides a highly efficient and accurate alternative for link prediction, and the sparse autologistic model can be used to speed up prediction while still getting some insights about the presence of substitution and transitivity effects in the network.

As an alternative to increase efficiency of the algorithms for posterior mode estimation in the L_1 penalized models, implementation of stochastic gradient descent methods can be useful for big network data with a large number of observations over time. Other interesting application for the models presented in this work is the cosponsorship networks of legislation in the U.S. Senate and U.S. House of Representatives available for the years 1973 to 2004. Exploration of this data through descriptive analysis and community structure identification has been performed in the past but no dynamic models have been applied in this context (Fowler, 2006a,b; Porter et al., 2007).

Bibliography

Adamic, L., Brunetti, C., Harris, J. H. & Kirilenko, A. A. (2010). Trading networks. Technical report, MIT Sloan School of Management.

Airoldi, E., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* **11**, 581–598.

Ang, A. & Bekaert, G. (2002a). International asset allocation with regime shifts. *Review of Financial Studies* **15**, 1137–1187.

Ang, A. & Bekaert, G. (2002b). Regime switches in interest rates. *Journal of Business Economics and Statistics* **20**, 163–182.

Banks, D. & Carley, K. M. (1996). Models for network evolution. *Journal of Mathematical Sociology* **21**, 173–196.

Bartlett, P. L., Mendelson, S. & Neeman, J. (2012). l_1-regularized linear regression: persitence and oracle inequalities. *Probability Theory and Related Fields* **154**, 193–224.

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2**, 183–202.

Bertoin, J. (2006). *Random fragmentation and coagulation processes.* Cambridge studies in advanced mathematics. Cambridge University Press.

Bickel, P. J., Ritov, Y. A. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Bliss, C., Frank, M., Danforth, C. & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediciton in dynamic social networks. *J. Comput. Sci.* .

Boyd, N., Harris, J. & Nowak, A. (2011). The role of speculators during periods of financial distress. *Journal of Alternative Investments* **14**, 10–25.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data. Methods, theory and applications.* Springer Series in Statistics. Springer.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics* **2**, 1153–1194.

Carpenter, B. (2008). Lazy sparse stochastic gradient descent for regularized multinomial logistic regression. Technical report, Alias-i.

Carter, C. & Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Cawley, G., Talbot, N. & Girolami, M. (2007). Sparse multinomial logistic regression via bayesian l1 regularisation. In *Advances in Neural Information Processing Systems*, volume 19, pp. 209–216. MIT Press.

Cranmer, S. & Desmarais, B. (2011). Inferential network analysis with exponential random graph models. *Political Analysis* **19**, 66–86.

Desmarais, B. & Cranmer, S. (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications* **391**, 1865–1876.

Efron, B. & Morris, C. (1972). Limiting the risk of bayes and empirical bayes estimators–part ii: The empirical bayes case. *Journal of the American Statistical Association* **67**, 130–139.

Efron, B. & Morris, C. (1973). Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

Erdös, P. & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae* **6**, 290–297.

Escobar, M. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* **96**, 1348–1360.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Filardo, A. (1994). Business cycle phases and their transitional dynamics. *Journal of Business, Economics, and Statistics* **12**, 299–308.

Fowler, J. (2006a). Connecting the congress: A study of cosponsorship networks. *Political Analysis* **14**, 456–487.

Fowler, J. (2006b). Legislative cosponsorship networks in the us house and senate. *Social Networks* **28**, 454–465.

Frank, O. & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81**, 832–842.

Friedman, J., Hastie, T., Hoefling, H. & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **2**, 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and markov Switching Models.* New York: Springer.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, Eds. J. M. Bernardo, J. Berger, A. P. Dawid & A. F. M. Smith, pp. 169–193. Oxford: Oxford University Press.

Geyer, C. (1992). Practical markov chain monte carlo. *Statistical Science* **7**, 473–483.

Goldenberg, A., Zheng, A. X., Fienberg, S. E. & Airoldi, E. M. (2009). A survey of statitical network models. *Found. Trends Mach. Learn.* **2**, 129–233.

Goldstein, T. & Osher, S. (2009). The split bregman method for l1-regularized problems the split bregman method for l1-regularized problems the split bregman method for l1 regularized problems. *SIAM Journal on Imaging Sciences* **2**, 323–343.

Green, P. & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Greene, W. H. (2002). *Econometric Analysis 5th ed.* New York: Prentice Hall.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *The Annals of Statistics* **34**, 2637–2386.

Guidolin, M. & Timmermann, A. (2005). Economic implications of bull and bear regimes in uk stock and bond returns. *The Economic Journal* **115**, 111–143.

Handcock, M. (2000). *Progress in Statistical Modeling of Drug User and Sexual Networks.* University of Washington, Center for Statistics and the Social Sciences: unpublished manuscript.

Hanneke, S., Fu, W. & Xing, E. P. (2010). Discrete temporal models of social networks. *Electronical Journal of Statistics* **4**, 585–605.

Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845.

Harchaoui, Z. & Levy-Leduc, C. (2008). Catching change-points with lasso.

Harris, J., Christie, W. & Schulta, P. (1994). Why did NASDAQ market makers stop avoiding odd eighth quotes? *Journal of Finance* **49**, 1841–1860.

Hatfield, J. W., Kominers, S. D., Nichifor, A., Ostrovsky, M. & Westkamp, A. (2012). Stability and competitive equilibrium in trading networks. Technical report, Standford University.

Hirose, K., Tateishi, S. & Konishi, S. (2013). Tuning parameter selection in sparse regression modeling. *Computational Statistics and Data Analysis* **59**, 28–40.

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

Höfling, H. (2010a). A coordinate-wise optimization algorithm for the fused lasso. arXiv:1011.6409 [stat.CO].

Höfling, H. (2010b). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics* **19**, 984–1006.

Holland, P. & Leinhardt, K. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**, 33–65.

Huang, Z. & Lin, D. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing* **21**, 286–303.

Kakade, S., Shamir, O., Sindharan, K. & Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the Thir-*

teenth International Conference on Artificial Intelligence and Statistics (AISTATS-10), Eds. Y. W. Teh & D. M. Titterington, volume 9, pp. 381–388.

Kemp, C., Tenenbaum, J. B., Griffiths, T., Yamada, T. & Ueda, N. (2006). Learning systems of concepts with an infinite relational data. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence(UAI 2006)*. Cambridge, MA, USA.

Kim, C., Morley, J. & Nelson, C. (2001). Does an international tradeoff between risk and return explain mean reversion in stock prices? *Journal of Empirical Finance* **8**, 403–426.

Kolacyzk, E. D. (2009). *Statistical Analysis of Network Models*. Springer.

Krishnapuram, B. & Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 957–968.

Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis* **5**, 369–412.

Lau, J. W. & Green, P. (2007). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558.

Lee, J. D., Sun, Y. & Taylor, J. E. (2014). On model selection consistency of regularized m-estimators.

Leifeld, P., Cranmer, S. & Desmarais, B. (2014). ***xergm**: Extensions of Exponential Random Graph Models*. R package.

Liben-Nowell, D. & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **58**, 1019–1031.

Liu, J., Yuan, L. & Ye, J. (2010). An efficient algorithm for a class of fused lasso problems. In *The ACM SIG Knowledge Discovery and Data Mining*. Washington, DC.

Lu, Z., Savas, B., Tang, W. & Dhillon, I. (2010). Supervised link prediction using multiple sources. In *10th International Conference on Data Mining (ICDM), IEEE*, pp. 923–928.

Meinshausen, N. & Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Negahban, S. N., Ravikumar, P., Wainwright, M. & Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review* **45**, 167–256.

Nowicki, K. & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**, 1077–1087.

Ozsoylev, H. N., Walden, J. & Yavuz, R., M. D. Bildik (2010). Investor networks in the stock market. Technical report, University of California, Berkeley.

Park, J. & Newman, M. (2004). Statistical mechanics of networks. *Physical Review E* **70**.

Park, M. & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**, 659–677.

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.

Perez-Quiroz, G. & Timmermann, A. (2000). Firm size and cyclical variations in stock returns. *Journal of Finance* **55**, 1229–1262.

Perry, P. O. & Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society, Series B* **75**, 821–849.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. & Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25(2)**, 855–900.

Polson, N., Scott, J. & Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables. arXiv:1205.0310v3.

Porter, M., Mucha, P., Newman, M. & Friend, A. (2007). Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications* **386**, 414–438.

Rabiner, L. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine* pp. 4–15.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Annals of Statistics* **37**, 2922–2952.

Robert, C. & Casella, G. (2005). *Monte Carlo Statistical Methods 2nd ed.* New York: Springer.

Rodriguez, A. (2011). Modeling the dynamics of social networks using Bayesian hierarchical blockmodels. *Statistical Analysis and Data Mining* p. In press.

Rodriguez, A. & Reyes, P. E. (2013). On the statistical properties of random blockmodels for network data. Technical report, University of California - Santa Cruz.

Rojas, C. & Wahlberg, B. (2014). On change point detection using the fused lasso method. arXiv:1401.5408 [math.ST].

Ryden, T., Terasvirta, T. & Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics* **13**, 217–244.

Sarkar, P., Chakrabarti, D. & Jordan, M. (2012). Nonparametric link prediction in dynamic networks. In *Proceedings of the 29th International Conference in Machine Learning*. Edinburgh, Scotland, UK.

Sarkar, P. & Moore, A. (2005). Dynamic social network analysis using latent space models. *SIGKDD Explorations: Special Edition on Link Mining* **7**, 31–40.

Scott, J. & Pillow, J. W. (2012). Fully bayesian inference for neural models with negative-binomial spiking. In *Advances in Neural Information Processing Systems 25*.

Sewell, D. K. & Chen, Y. (2015). Analysis of the formation of the structure of social networks by using latent space models for ranked dynamic networks. *Journal of the Royal Statistical Society, Series C* .

Shalev-Shwartz, S. & Tewari, A. (2011). Stochastic methods for l_1-regularized loss minimization. *Journal of Machine Learning Research* **12**, 1865–1892.

da Silva, P. & Bastos, R. (2012). Time series based link prediction. In *IEEE World Congress on Computational Intelligence*. Brisbane, Australia.

Snijders, T. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure* **3(2)**.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology* **37**, 129–151.

Snijders, T. A. B., Steglich, C. & van de Bunt, G. (2010). Introduction to actor-based models for network dynamics. *Social Networks* **32**.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* **67**, 91–108.

TIBSHIRANI, R. & TAYLOR, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39**, 1335–1371.

TIBSHIRANI, R. & TAYLOR, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40**, 1198–1232.

TSURUOKA, Y., TSUJII, J. & ANANIADOU, S. (2009). Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp. 477–485. Suntec, Singapur.

TUTZ, G., PÖSSNECKER, W. & UHLMANN, L. (2015). Variable selection in general multinomial logits. *Computational Statistics and Data Analysis* **82**, 207–222.

VAN DE GEER, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36**, 614–645.

VIDAURRE, D., BIELZA, C. & LARRANAGA, P. (2013). A survey of l_1 regression. *International Statistical Review* **81**, 361–387.

WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming. *IEEE Transactions on Information Theory* **55**, 2183–2202.

Wang, Y. & Wong, G. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* **82**, 8–19.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications.* New York: Cambridge University Press.

Xing, E. P., Fu, W. & Song, L. (2010). A state-space mixed memebership blockmodel for dynamic network tomography. *Annals of Applied Statistics* **4**, 535–566.

Xu, Z., Tresp, V., Yu, K. & Kriegel, H. (2006). Infinite hidden relational models. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence(UAI 2006).* Cambridge, MA, USA.

Yau, C., Papaspiliopoulos, O., Roberts, G. O. & Holmes, C. C. (2011). Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society, Series B* **73**, 37–57.

Ye, G. & Xie, X. (2011). Split bregman method for large scale fused lasso. *Computational Statistics and Data Analysis* **55**, 1552–1569.

Yu, D., Won, J., Lee, T., Lim, J. & Yoon, S. (2013). High-dimensional fused lasso regression using majorization-minimization and parallel processing. arXiv:1306.1970v2 [stat.ME].

Zaloom, C. (2004). Time, space and technology in financial networks. In *The Network*

Society: A Cross-cultural Perspective, Ed. M. Castells, pp. 198–216. Edward Elgar Publishing Limited.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

Zheng, Y. & Zhu, J. (2008). Markov chain monte carlo for a spatial-temporal autologistic regression model. *Journal of Computational and Graphical Statistics* **17**, 123–137.

Zhu, J., Huang, H. & Wu, J. (2005). Modeling spatial-temporal binary data using random markov fields. *Journal of Agricultural, Biological, and Environmental Statistics* **10**, 212–225.

Zhu, J., Zheng, Y., Carroll, A. & Aukema, B. (2008). Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood. *Journal of Agricultural, Biological, and Environmental Statistics* **13**, 84–98.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *Annals of Statistics* **35**, 2173–2192.

Appendix A

Posterior sampling for HMM model

Here, we provide the details of the MCMC algorithm discussed in Section 2.2.3. The algorithm proceeds by updating the model parameters from the following full conditional distributions:

(a) For each $i = 1, \ldots, n$ and occupied states s, $\xi_{i,s} = k$ with probability

$$\mathsf{Pr}(\xi_{i,s} = k \mid \cdots, \mathbf{Y})$$

$$= \begin{cases} (m_k^{-i} - \alpha_s) \prod_{l=1}^{K^*_{s,-i}} \frac{p(\{y_{i,j,t} : (i,j,t) \in A^{i}_{k,l,s}\})}{p(\{y_{i,j,t} : (i,j,t) \in A^{-i}_{k,l,s}\})} \frac{p(\{y_{j,i,t} : (i,j,t) \in A^{i}_{k,l,s}\})}{p(\{y_{j,i,t} : (i,j,t) \in A^{-i}_{k,l,s}\})}, & k \leq K^*_{s,-i} \\ \\ (\beta_s + \alpha_s K^*_{s,-i}) \prod_{l=1}^{K^*_{s,-i}} p(\{y_{i,j,t} : (i,j,t) \in A^{-i}_{l,s}\}) & \\ \qquad\qquad p(\{y_{j,i,t} : (i,j,t) A^{-i}_{l,s}\}), & k = K^*_{s,-i} + 1, \end{cases}$$

where $K^*_{s,-i} = \max_{j\neq i}\{\xi_{j,s}\}$, $m_k^{-i} = \sum_{j\neq i}\mathbb{I}_{(\xi_{j,s}=k)}$,

$$\begin{aligned}
A^{-i}_{k,l,s} &= \{(i',j',t) : i' \neq j' \neq i, \zeta_t = s, \xi_{i',\zeta_t} = k, \xi_{j',\zeta_t} = l\},\\
A^{i}_{k,l,s} &= \{(i',j',t) : i' = i, \zeta_t = s, \xi_{j',\zeta_t} = l\} \bigcup A^{-i}_{k,l,s},\\
A^{-i}_{l,s} &= \{(j,t) : j \neq i, \zeta_t = s, \xi_{j,\zeta_t} = l\},
\end{aligned}$$

and the marginal predictive distribution, $p(\{y_{i,j,t} : (i,j,t) \in A\})$ is given by

$$\frac{\Gamma(\sum_A y_{i,j,t} + a_s)\Gamma(|A| + b_s - \sum_A y_{i,j,t})}{\Gamma(a_s + b_s + |A|)}\frac{\Gamma(a_s + b_s)}{\Gamma(a_s)\Gamma(b_s)}.$$

and $|A|$ is the number of elements in A.

(b) Since the prior for $\theta_{k,l,s}$ is conditionally conjugate, we update these parameters for $k, l \in \{1,\ldots,K^*_s\}$ by sampling from

$$\theta_{k,l,s} \mid \cdots, \mathbf{Y} \sim \text{Beta}\left(\sum_{A_{k,l,s}} y_{i,j,t} + a_s, m_{k,l,s} + b_s - \sum_{A_{k,l,s}} y_{i,j,t}\right)$$

for $A_{k,l,s} = \{(i,j,t) : i \neq j, \zeta_t = s, \xi_{i,\zeta_t} = k, \xi_{j,\zeta_t} = l\}$ and $m_{k,l,s} = |A_{k,l,s}|$.

(c) Since the prior for the transition probabilities is conditionally conjugate, the posterior full conditional for $\boldsymbol{\pi}_r$, $r = 1,\ldots,S$ is the Dirichlet distribution

$$p(\boldsymbol{\pi}_r \mid \cdots, \mathbf{Y}) = \prod_{s=1}^{S} \pi_{r,s}^{\gamma/S + n_{rs} - 1}$$

for $n_{rs} = |\{t : \zeta_{t-1} = r, \zeta_t = s\}|$.

(d) The posterior full conditional of γ is

$$p(\gamma \mid \cdots, \mathbf{Y}) \propto p(\gamma)\prod_{s=1}^{S}\frac{\Gamma(\gamma)}{\Gamma(\gamma + n_s)}\gamma^{L_s}$$

where $n_s = |\{t : \zeta_t = s\}|$ and $L_s = \sum_r \mathbb{I}_{n_{s,r}>0}$ for $n_{s,r} = |\{t : \zeta_{t-1} = s, \zeta_t = r\}|$. Since this distribution has no standard form, we update γ using a random walk Metropolis-Hastings algorithm with symmetric log-normal proposal,

$$\log\{\gamma^{(p)}\} \mid \gamma^{(c)} \sim \mathsf{Normal}\left(\log\{\gamma^{(c)}\}, \kappa_\gamma^2\right)$$

where κ_γ^2 is a tuning parameter chosen to get an average acceptance rate between 30% and 40% .

(e) The posterior full conditional of the pairs $(a_{s,O}, b_{s,O})$ and $(a_{s,D}, b_{s,D})$ has the following general form:

$$p(a_s, b_s \mid \cdots, \mathbf{Y}) \propto p(a_s \mid d)p(b_s \mid e) \prod_{k=1}^{S}\prod_{l=1}^{S} p(y_{i,j,t} \mid A_{k,l,s}, m_{k,l,s})$$

for the marginal predictive $p(y_{i,j,t} \mid A_{k,l,s}, m_{k,l,s})$ as defined in step (b), $A_{k,l,s} = \{(i,j,t) : i \neq j, \zeta_t = s, \xi_{i,\zeta_t} = k, \xi_{j,\zeta_t} = l\}$ and $m_{k,l,s} = |A_{k,l,s}|$. Since no direct sampler is available for this distribution, we update each pair using a random walk Metropolis-Hastings algorithm with bivariate log-normal proposals,

$$\left(\log\{a_s^{(p)}\}, \log\{b_s^{(p)}\}\right)^t \quad \mid \quad \left(a_s^{(c)}, b_s^{(c)}\right)^t \quad \sim \quad \mathsf{Normal}\left[\left(\log\{a_s^{(c)}\}, \log\{b_s^{(c)}\}\right)^t, \mathbf{\Sigma}_{ab}\right]$$

where $\mathbf{\Sigma}_{ab}$ is a tuning parameter matrix chosen independently for diagonal and off-diagonal pairs of parameters.

(f) The parameters of the Poisson-Dirichlet process (α_s, β_s) can be jointly updated using the algorithm described in Escobar & West (1995).

(g) The posterior full conditional distributions for the hyperparameters d_O, e_O, d_D, and e_D

correspond to gamma distributions with shape parameter $(cS^* + 1)$ and rate parameters $(\sum_{S^*} a_{s,O} + \lambda_d)$, $(\sum_{S^*} b_{s,O} + \lambda_e)$, $(\sum_{S^*} a_{s,D} + \lambda_d)$, $(\sum_{S^*} b_{s,D} + \lambda_e)$, respectively.

Appendix B

Posterior sampling for fused lasso model using FFBS

Here, we present the details of the forward filtering backward sampling algorithm for the Bayesian implementation of the fused lasso multinomial regression model discussed in Section 3.2.1.

The filtered moments for $\theta_{r,t}$ are given recursively by the following expressions:

$$M_{r,t} = \left[\omega_{r,t} + u_{r,t}^{-1}\right]^{-1}$$

$$m_{r,t} = M_{r,t}\left[(\kappa_{r,t} + \omega_{r,t}C_{r,t}) + u_{r,t}^{-1}m_{r,t-1}\right],$$

with $u_{r,t} = M_{r,t-1} + \tau^2_{r,t-1}$.

After drawing $\theta_{r,T}$ from $N(m_{r,T}, M_{r,T})$, the parameters are updated backwards as

$\theta_{r,t} \sim N(a_{r,t}, A_{r,t})$, where

$$A_{r,t}^{-1} = M_{r,t}^{-1} + \tau_{r,t}^{-2}$$

$$a_{r,t} = A_{r,t} \left[M_{r,t}^{-1} m_{r,t} + \tau_{r,t}^{-2} \theta_{r,t+1} \right],$$

resulting in a posterior sample for the block of parameters $\mathbf{\Theta}_r$.