Title
Estimation and Inference for Self-Exciting Point Processes with Applications to Social Networks and Earthquake Seismology

Author
Fox, Eric Warren

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Estimation and Inference for Self-Exciting Point Processes with Applications to Social Networks and Earthquake Seismology

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Eric Warren Fox

2015

Abstract of the Dissertation

Estimation and Inference for Self-Exciting Point Processes with Applications to Social Networks and Earthquake Seismology

by

Eric Warren Fox

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2015

Professor Frederic Paik Schoenberg, Chair

Self-exciting point processes describe random sequences of events where the occurrence of an event increases the likelihood that subsequent events occur nearby in time and space. Models for self-exciting point processes have many important applications to diverse topics such as earthquake and crime forecasting, epidemiology, invasive species, and social networks.

The first part of this dissertation discusses a new application of self-exciting point processes to modeling the times when e-mails are sent by individuals in a social network. The proposed models are fit to datasets from West Point Military Academy and the Enron Corporation, and the resulting parameter estimates characterize communication behaviors and leadership roles for users in each network. We argue that the self-exciting models adequately capture major temporal clustering features in the data and perform better than traditional stationary Poisson models.

The second part of this dissertation discusses the nonparametric method of Marsan and Lengliné (2008) for estimating space-time Hawkes point process models of earthquake occurrences. Their method provides an estimate of a station-

ary background rate for mainshocks, and a histogram estimate of the triggering function for the rate of aftershocks following an earthquake. At each step of the procedure the model estimates rely on computing the probability each earthquake is a mainshock or aftershock of a previous event. We focus on improving Marsan and Lengliné's method by proposing novel ways to incorporate a non-stationary background rate, and adding error bars to the histogram estimates which capture the sampling variability and bias in the estimation of the underlying seismic process. A simulation study is designed to validate and assess new methodology. An application to earthquake data from the Tohoku District in Japan is also discussed, and the results are compared to a well established parametric model of seismicity for this region.

The dissertation of Eric Warren Fox is approved.

Qing Zhou

Ying Nian Wu

Andrea L. Bertozzi

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2015

To my parents . . .

for their unconditional love and support

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my most sincere gratitude to my advisor Frederic P. Schoenberg for giving me the opportunity to pursue this doctorate, and for being a constant source of encouragement, guidance, and motivation. I have greatly appreciated his patience and willingness to share his experience and provide advice.

I am grateful to Andrea L. Bertozzi for supporting me as a Graduate Student Researcher and providing me with invaluable research experience. It has been a pleasure working with her research group the past several years.

I also want to thank Ying Nian Wu and Qing Zhou for being inspirational professors and for providing many helpful suggestions and comments about this work.

Finally, I wish to thank my family and friends for supporting my academic and extracurricular pursuits, and for always having my best interest in mind.

Chapter 2 is a version of Fox, E.W., Short, M.B., Schoenberg, F.P., Coronges, K.D., and Bertozzi, A.L. (2014): "Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes" revised for the Applications and Case Studies of *JASA*. I would like to thank the co-authors for their following contributions: Martin B. Short (Georgia Tech., Mathematics) for proposing the Hawkes process model of e-mail communication; Frederic P. Schoenberg (UCLA, Statistics) for providing substantial guidance; Kathryn D. Coronges (USMA, Behavioral Sciences) for providing the IkeNet e-mail dataset and survey results; and Andrea L. Bertozzi (UCLA, Mathematics) for directing and initiating this project.

Chapter 3 is a version of Fox, E.W., Schoenberg, F.P., and Gordon, J.S. (2015): "A Note on Nonparametric Estimates of Space-time Hawkes Point Process Models for Earthquake Occurrences" submitted to the *Annals of Applied Statistics*. I would like to thank the co-authors for their following contributions: Frederic P.

Schoenberg for proposing this project and providing substantial guidance; and Joshua S. Gordon (UCLA, Statistics) for helping with the data processing and simulation study.

The research in Chapter 2 was supported by ARO grant W911NF1010472, AFOSR MURI grant W911NF-11-1-0332, NSF grant DMS-0968309, and NSF grant DMS-1045536.

Vita

2009	B.A in Mathematics, Occidental College
2010–present	Private Math and Statistics Tutor
2011–2012	Teaching Assistant/Associate, UCLA Department of Statistics
2012-2013	Teaching Assistant Coordinator, UCLA Department of Statistics
2012–2014	Graduate Student Researcher, UCLA Department of Applied Mathematics
2014–present	Dissertation Year Fellowship, UCLA Graduate Division

Publications

Fox, E.W., Short, M.B., Schoenberg, F.P., Coronges, K.D., and Bertozzi, A.L. Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes. *JASA*, revised 11/14.

Fox, E.W., Schoenberg, F.P., Gordon, J.S. A Note on Nonparametric Estimates of Space-time Hawkes Point Process Models for Earthquake Occurrences. *Annals of Applied Statistics*, submitted 5/15.

CHAPTER 1

Introduction

1.1 Background

A spatial-temporal point process is a random process where any realization consists of a collection of points $\{(t_i, \mathbf{s}_i) : i = 1, \cdots, N\}$ representing the times and locations of events. The number of points N that occur in the process is not known in advance and is random. Examples of events include earthquakes, incidents of diseases, volcanic eruptions, or burglaries. Typically the spatial locations are observed in two or three spatial coordinates. However, in this dissertation, only two spatial coordinates $\mathbf{s}_i = (x_i, y_i)$ are considered, often representing the longitude and latitude of an event. Any additional information associated with a point is called a mark. Examples of marks include the earthquake moment magnitude, type of crime (theft, assault, homicide), or volcanic explosivity index.

The conditional intensity for a spatial-temporal point process is defined as the as the infinitesimal expected rate at which events occur around a time and location (t, x, y) given the history of the process:

$$\begin{aligned}&\lambda(t, x, y|H_t)\\ &= \lim_{\Delta t, \Delta x, \Delta y \downarrow 0} \frac{E[N\{(t, t+\Delta t) \times (x, x+\Delta x) \times (y, y+\Delta y)\}|H_t]}{\Delta t \Delta x \Delta y}.\end{aligned} \tag{1.1}$$

Here the history $H_t = \{(t_i, x_i, y_i, m_i) : t_i < t\}$ denotes the times, locations, and marks of all events occurring before time t. Conditional intensities are a natural way to model point processes as all finite-dimensional distributions of a simple

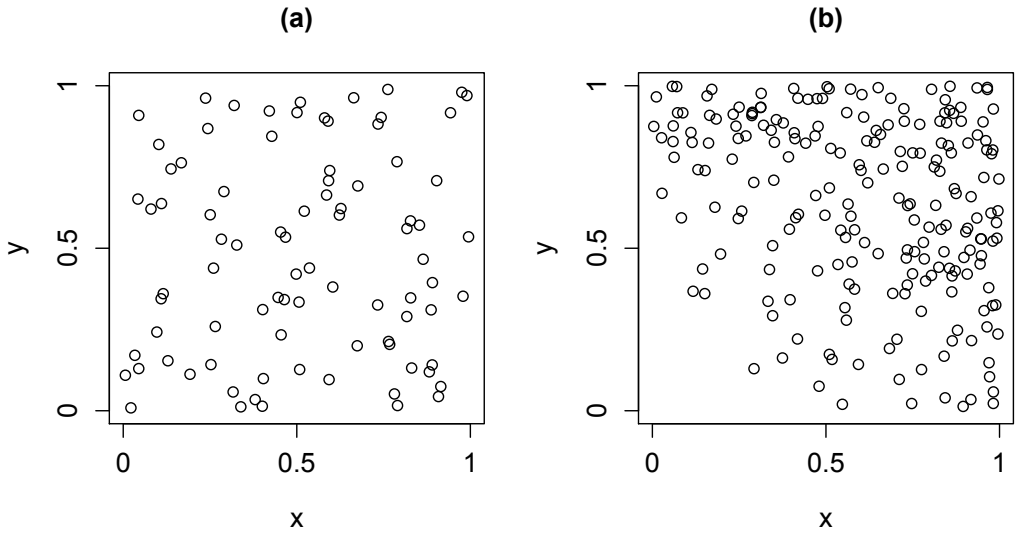

Figure 1.1. Simulated realization of the Poisson process with (a) stationary rate $\lambda(x, y) = 75$, and (b) non-stationary rate $\lambda(x, y) = 300(x^2 + y^2)$.

point process are uniquely determined by its condition intensity (Daley and Vere-Jones, 2003). Note, a point process is simple if all points occur at unique locations with probability 1.

For the Poisson process the conditional intensity function does not depend on the history, i.e. $\lambda(t, x, y|H_t) = \lambda(t, x, y)$. A Poisson process is stationary if the intensity function is constant at all times and locations: $\lambda(t, x, y) = c$, where c is a constant representing the expected number of points per unit time per unit area. For example, Figure 1.1(a) shows a realization a purely spatial stationary Poisson process with $\lambda(x, y) = 75$. A Poisson process is non-stationary if the intensity function depends on (t, x, y) in some way. For example, Figure 1.1(b) shows a realization of a purely spatial Poisson process with $\lambda(x, y) = 300(x^2 + y^2)$. For a Poisson process, the expected number of points occurring during the time interval $[0, T]$ and in region A is given by:

$$\int_0^T \int \int_A \lambda(t, x, y) dx dy dt$$

Self-exciting point processes describe random sequences of events where the occurrence of one event increases the likelihood that subsequent events occur nearby in time and space. Earthquakes are an example since the occurrence of an earthquake may trigger nearby aftershocks. As opposed to the Poisson process, the conditional intensity of a self-exciting point process depends on the past history H_t.

Many parametric models for the conditional intensity of a self-exciting point process have been proposed in the literature. The Hawkes process (Hawkes, 1971) is an important temporal model:

$$\lambda(t|H_t) = \mu(t) + \sum_{t_i<t} g(t - t_i). \tag{1.2}$$

This model classifies events into two types: background and triggered. The rate of background events at time t is modeled by the Poisson process $\mu(t)$. The rate at which an event at time t_i triggers additional events at time t is modeled by the triggering function $g(t - t_i)$, which is often assumed exponential: $g(t - t_i) = \alpha e^{\omega(t-t_i)}$. The summation term gives the contribution of all previous events to the overall intensity at time t. Note, if $t_i > t$ then $g(t - t_i) = 0$.

The temporal Hawkes process can be easily extended to the marked spatial-temporal case:

$$\lambda(t, x, y|H_t) = \mu(t, x, y) + \sum_{\{i:t_i<t\}} g(t - t_i, x - x_i, y - y_i; m_i). \tag{1.3}$$

In an application to seismology, Ogata (1998) considered many parametric forms of this model, collectively referred to as Epidemic Type Aftershock Sequences (ETAS) models. For example, one such parametrization is given by:

$$\lambda(t, x, y|H_t) = \mu + \sum_{\{i:t_i<t\}} \frac{K}{(t - t_i + c)^p} \cdot \frac{e^{\alpha(m-m_c)}}{((x - x_i)^2 + (y - y_i)^2 + d)^q},$$

where $(\mu, K, \alpha, p, c, q, d)$ are parameters to be estimated, and m_c is the fixed magnitude cut-off for the earthquake catalogue.

The parameters of model (1.3) can be estimated by maximizing the log-likelihood function (Ogata, 1998) with respect to the parameters of the model:

$$log(L) = \sum_{i=1}^{N} \log(\lambda(t_i, x_i, y_i|H_t)) - \int_0^T \int \int_S \lambda(t, x, y|H_t) dxdydt, \tag{1.4}$$

where $S \times [0, T]$ is the space-time observation window over which the process is observed. The first term of the log-likelihood is easily computed. The integral in the second term often does not have a closed form solution and must be numerically approximated (Schoenberg, 2013). Standard errors for the parameter estimates can be derived using asymptotic properties of the maximum likelihood estimators (Rathbun and Cressie, 1994). More recently, nonparametric techniques have been considered for estimating model (1.3) (see Marsan and Lengliné (2008); Mohler et al. (2011)); such methods will be described in great detail in Chapter 3.

1.2 Motivation

There is extensive application of self-exciting point processes to modeling and forecasting earthquake occurrences. Current models of the triggering function perform remarkably well at estimating and predicting properties of aftershock sequences, such such as the shape and spatial-temporal decay rate (Ogata, 1988, 1998). Estimates of the background rate and overall intensity are also useful for identifying regions with a high incidence of large seismic events, and may help guide construction regulation.

Self-exciting models have gained recent popularity due to the multitude of new and important applications to diverse areas such as criminology, epidemiology, plant ecology, finance, and social networks. The self-exciting crime models, in particular, fit the data well and have been used to improve crime forecasting with hotspot maps (Mohler et al., 2011), and infer unknown gang affiliation for acts of retaliatory violence in a gang network (Stomakhin et al., 2011). One proposed

explanation for the successful implementation of these models is that crimes spur local revenge crimes much as earthquakes trigger aftershocks.

In Chapter 2, we discuss a new application of self-exciting point processes to modeling e-mail traffic on a social network. The parameter estimates from the proposed models may be used to characterize important e-mail communication behaviors such as the baseline sending rates, average reply rates, and average response times. We also investigate the problem of using these features to infer the underlying leadership status of users in a social network. In Chapter 3, we assess and suggest ways to improve the nonparametric method of Marsan and Lengliné (2008) for estimating a space-time Hawkes process model (1.3) for earthquake occurrences. The advantage of this approach over traditional parametric models is that the shape of the triggering function does not need to specified a-priori, and a data-driven estimate is provided instead. The methods in this chapter may lead to improvements in earthquake forecasting techniques and model diagnostics.

CHAPTER 2

Point Process Analysis of E-mail Networks

Several studies on e-mail communication have shown that the times when individuals send e-mails deviate from a stationary Poisson process (Barabási, 2005; Malmgren et al., 2008). Two important properties of the stationary Poisson process are that the mean number of events per unit time is constant, and the time intervals between consecutive events (inter-event or waiting times) follows an exponential distribution. Barabási (2005) provided empirical evidence showing that the inter-event times for e-mails are better approximated by a heavy-tailed power law distribution. Essentially, this means the sending times for a typical e-mail user are highly clustered: short periods with lots of activity are separated by long periods when no messages are sent.

To account for the clustering and uneven waiting times observed in e-mail traffic Barabási (2005) proposed a priority queue model, in which high priority e-mails are responded to more quickly than low priority e-mails. We take a different approach by considering self-exciting point process models for e-mail traffic. In general, self-exciting point processes describe random collections of events where the occurrence of one event increases the likelihood that another event occurs shortly thereafter. E-mail traffic may be viewed as a self-exciting point process since each e-mail received by an individual increases the likelihood that reply e-mails are sent shortly thereafter. In other words, sending an e-mail can trigger a chain of messages sent between individuals in rapid succession.

The application of self-exciting point processes to modeling and characterizing

social networks is a relatively new research topic. Some recent work includes self-exciting models for retaliatory acts of violence in a Los Angeles gang networks (Stomakhin et al., 2011; Hegemann et al., 2012) and face-to-face conversation sequences in a company (Masuda et al., 2012). As in these previous works, we model event times (e-mails) on a social network as a multivariate Hawkes process (Hawkes, 1971; Hawkes and Oakes, 1974) with an exponential triggering function.

This work is primarily focused on describing, modeling, and analyzing two interesting e-mail network datasets: the IkeNet dataset collected from the log files of e-mail transactions between 22 officers attending West Point Military Academy over a one-year period, and the Enron dataset collected from 151 employees over a three-year period before the company's demise. The IkeNet dataset offers a unique opportunity to study e-mail communication on a small and relatively flat social network, in which all officers in the network are enrolled in the same academic program. The Enron dataset, on the other hand, is much larger and users in this network exhibit a complex and rich corporate hierarchy. Moreover, it is perhaps the only corporate e-mail corpus freely available to the public for research. Using these datasets we seek to address the following questions:

(a) Do the estimated self-exciting models perform significantly better than stationary Poisson models and account for the observed temporal clustering in e-mail network traffic?

(b) Does the incorporation of diurnal and weekly trends into the baseline (background) rate at which e-mail conversations are initiated provide an overall better fit to the observed network data?

(c) How can the estimated parameters be used to characterize important communication behaviors, such as the average reply rate and response time, for individuals in the network and the network as a whole?

(d) How can various features of e-mail communication, estimated from the self-exciting models, be used to predict and rank leaders within a social network?

The prediction of network leadership from communication patterns is an important question. Many methods have been proposed in the literature to address this issue (Shetty and Adibi, 2005; Tyler et al., 2005; Creamer et al., 2009). Our contribution is to show that a point process analysis provides additional insight into the leadership roles and hierarchy underlying a communication network. A distinctive aspect of both the IkeNet and Enron datasets is that ground-truth about the actual leadership status of individuals in these networks is readily available, and provides a means to evaluate and validate our proposed covariates for inferring leadership.

This chapter is organized as follows: In Section 2.1 we provide some descriptive statistics for the IkeNet dataset. In Section 2.2 we propose various self-exciting models for e-mail communication networks and fit these to the IkeNet data using an EM-type procedure. In Section 2.3 we describe how to use our parameter estimates to characterize communication behaviors and predict leadership for the IkeNet social network. In Section 2.3 we also discuss model comparisons and diagnostics. In Section 2.4 we compare the models fit to the Enron and IkeNet datasets and use parameter estimates for the Enron e-mail network to describe and discriminate leadership roles within the corporate hierarchy. In the Discussion Section we summarize and speculate about our results and suggest possible future directions for this research. In Appendix A we spell out the simulation algorithm we use to generate realizations of the IkeNet e-mail network from the fitted self-exciting models.

2.1 IkeNet Dataset and Descriptive Statistics

The IkeNet dataset contains the sender, receiver, timestamp, and identification for each message sent between 22 officers in a closed network over a one-year period beginning in May 2010. E-mails were sent with Blackberries, which were given to the officers as incentive for their participation in the study. The officers were anonymized in the data for privacy, therefore we will refer to them by number (1–22) instead of name. Only 3.3% of e-mails sent in the IkeNet dataset have more than one recipient; thus for simplicity we treat each sender-recipient pair as an e-mail (e.g. one e-mail sent to three recipients is coded as three separate e-mails). After removing duplicates and instances when officers sent messages to themselves, we are left with a total of approximately 8400 e-mails.

Each officer was asked in a questionnaire to list the officers, within the network, whom they considered strong team and military leaders. This supplementary survey data, provided with the IkeNet e-mail data, allows for a particularly unique opportunity to make connections between e-mail communication behaviors and leadership attributes. Many previous studies of e-mail activity have only focused on describing and modeling temporal communication patterns (e.g. Barabási (2005); Malmgren et al. (2008)), and have not looked at the relationships between those communication patterns and the attributes and perceptions of users in the network. Questions such as how one might predict perceived leadership status using only observations of network communication are addressed in Section 2.3.

Descriptive statistics for the IkeNet dataset reveal daily, weekly and seasonal trends in e-mail traffic. Figure 2.1 is a histogram of the number of e-mails sent in the network each hour of the day, over the yearlong observation window. This plot reveals a clear diurnal rhythm: e-mails were most frequently sent mid-day and activity diminished during the night. Decreased activity during lunch and dinner is also visible, around noon and seven p.m. Figure 2.2 is a bar plot of the

number of e-mails sent each day of the week. The e-mail activity among these officers was evidently substantially greater during weekdays (Mon.–Fri.) than on the weekend.

Figure 2.3 is a time series plot of the number of e-mails sent in the network each day. The smoother curve helps reveal monthly trends. For instance, there was a drop in network activity in January; this was probably due to the holidays and officers being out of town. The time series plot exposes two days with an unusually high amount of e-mail traffic. The first peak occurred on 02 February 2011 (162 e-mails sent) and coincided with escalating violence in the Egyptian revolution. The second peak occurred on 02 May 2011 (166 e-mails sent) and coincided with the assassination of Osama bin Laden. These outliers are also present in Figure 2.4, a right skewed histogram which shows that on a typical day, fewer than thirty e-mails are sent within the network.

Also of interest are descriptive statistics for the number of e-mails sent between officers in the network. Figure 2.5 is a graphical representation of a matrix whose entries are the number of e-mails sent from officer i (column) to j (row). Notice that this matrix is not symmetric, since the number of e-mails sent from officer i to j may be different from the number of e-mails sent from j to i. The e-mail network itself is shown in Figure 2.6 with node sizes proportional to the number of e-mails sent by each officer, and edge widths proportional to the number of messages sent between officers. Officers 9, 18, and 13 stand out in this plot for sending the highest number of e-mails in the network. The matrix and network plots reveal pairs of officers that communicate frequently with each other, as well as those officers that communicate infrequently with the network as a whole. For instance, officer pair (9,18) particularly stands out as being most prolific, as these officers sent a total of 1042 e-mails to each other. In contrast, officers 1 and 21 are distant from the network and have very few e-mail interactions. Both plots also illustrate the overall sparsity in e-mail communication on this closed network.

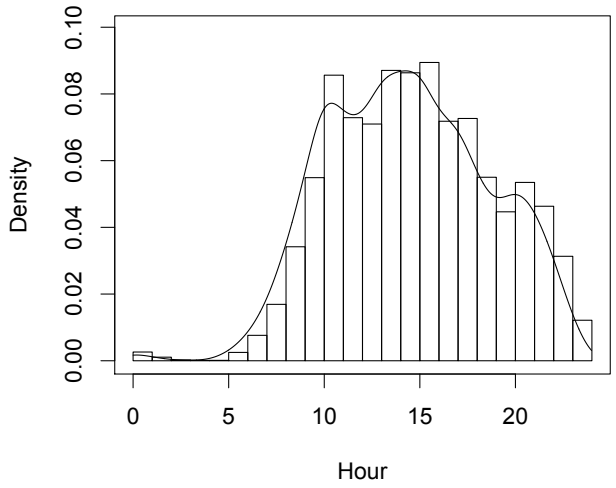

Figure 2.1. Histogram density of the number of e-mails sent each hour of the day over the one-year observation window. The smoother curve was formed using kernel density estimation with a fixed bandwidth (Scott, 1992).

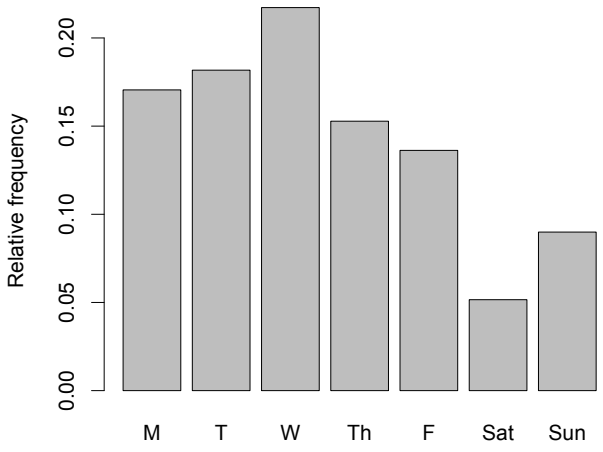

Figure 2.2. Proportion of e-mails sent each day of the week over the one-year observation window.

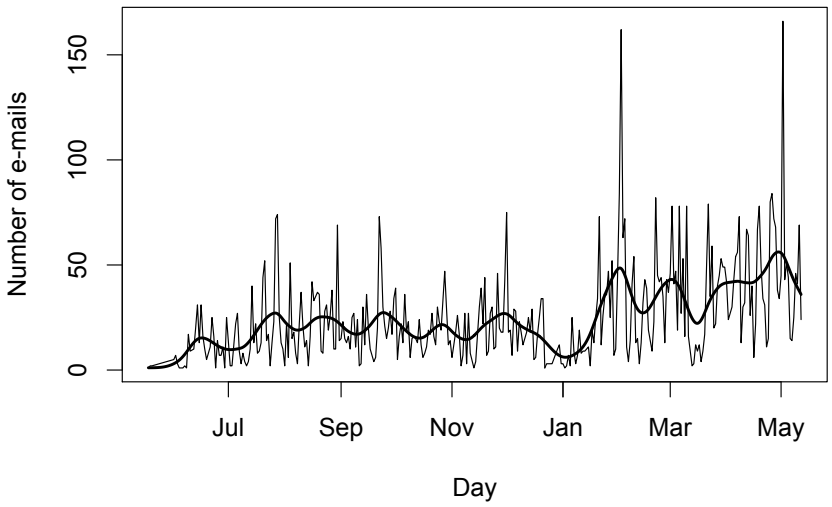

Figure 2.3. Time series plot of number of e-mails sent by date.

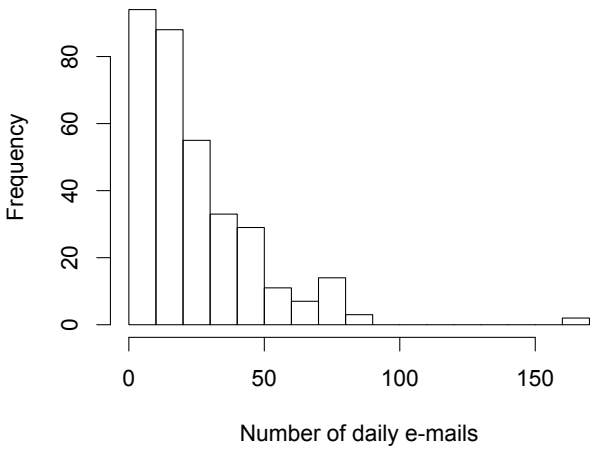

Figure 2.4. Histogram of the number of daily e-mails.

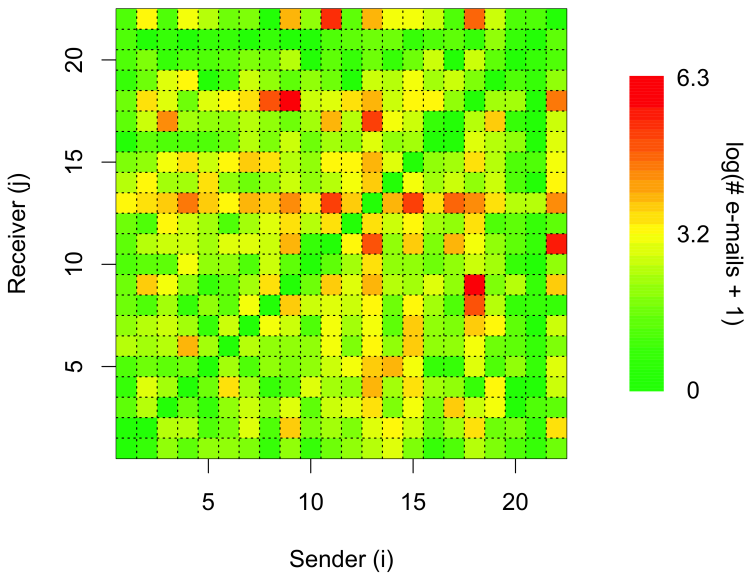

Figure 2.5. Matrix plot of the logarithm of the number of e-mails sent from officer i (column) to j (row) for the IkeNet dataset. The red and orange cells indicate pairs of officers that communicate frequently through e-mail. Likewise, the yellow and green cells indicates moderate to low communication between officer pairs.

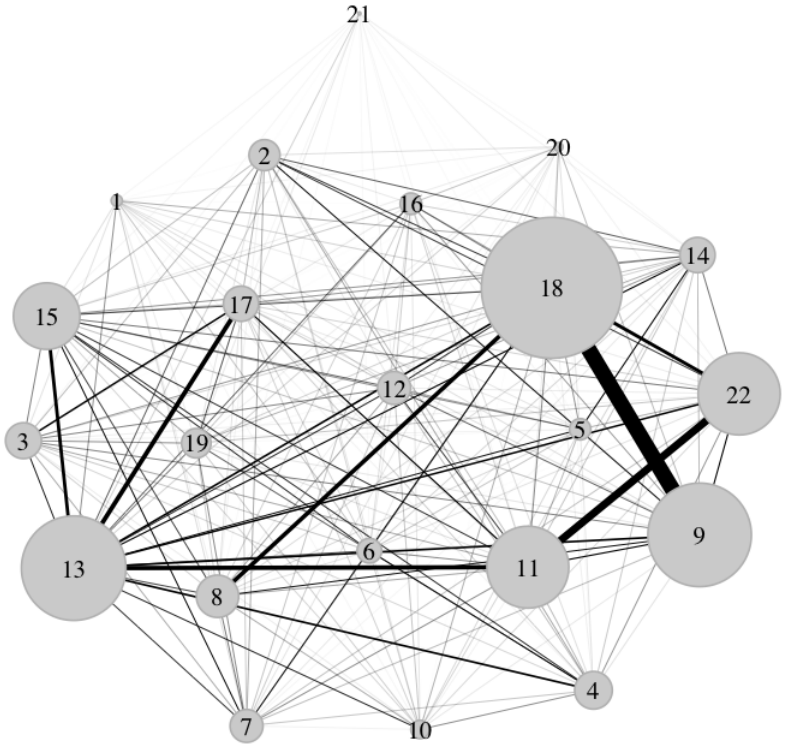

Figure 2.6. Plot of the IkeNet e-mail network with node sizes proportional to the number of e-mails sent by each officer, and edge widths proportional to the number of e-mails sent between officers.

2.2 Self-Exciting Models for IkeNet E-mail Activity

In this section we extend the temporal Hawkes process (1.2) to model e-mail activity on a social network, and fit these models to the IkeNet dataset. Like earthquakes, e-mail communications may be viewed as a branching processes. The 'mainshocks' are the times when an officer initiates e-mail conversations; the 'aftershocks' are the reply e-mails, which are sent in response to e-mails received from other officers in the network. Our approach is similar to that of Halpin and De Boeck (2013), though we model e-mail traffic on a network, not just between two people, and propose ways to account for circadian and weekly trends.

We primarily consider models of e-mail activity from an egocentric point of view, with the self-exciting point processes placed on the nodes (users) of the network to model the rate of sending e-mails. Other relational views as considered in Perry and Wolfe (2013) and Zipkin et al. (2015) include, for instance, the modeling of dyadic interactions whereby the point processes are placed on the edges of the network to measure the rate of e-mail communication between pairs of users. The dyadic models of Zipkin et al. (2015) are fit to the IkeNet dataset and applied to the problem of filling in missing communication data on this social network.

For a thorough introduction to point processes, conditional intensities, and closely related constructs, see Daley and Vere-Jones (2003). Here we briefly review a few necessary preliminaries.

A point process is a random collection of points, with each point falling in some observed metric space, S. Here, as in many applications, the observed space is a portion of the real time line, $[0, T]$, and our observations of the e-mail network may be considered a sequence of 22 point patterns, or equivalently a single multivariate point pattern. Point processes are typically modeled by specifying their associated conditional intensity processes, as all finite-dimensional

distributions of a simple point process are uniquely characterized by its conditional intensity process, assuming it exists. For a temporal point process on a closed time interval $[0, T]$, the conditional intensity may be defined as the infinitesimal expected rate at which points occur around time t, given the entire history, H_t, of the point process up to time t:

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{E[N(t, t + \Delta t)|H_t]}{\Delta t}.$$

The Hawkes process given by (1.2) is an important conditional intensity model for a self-exciting point process. It may readily be extended to model the rate at which each IkeNet officer i sends e-mails at time t (hours) given all messages received by i at times $r_k^i < t$:

$$\begin{aligned} \lambda_i(t) &= \mu_i + \sum_{r_k^i < t} g_i(t - r_k^i) \\ &= \mu_i + \theta_i \sum_{r_k^i < t} \omega_i e^{-\omega_i (t - r_k^i)}. \end{aligned} \tag{2.1}$$

In the context of e-mails, the background rate μ_i can be interpreted as that rate at which officer i sends e-mails that are not replies to e-mails received from other officers. In other words, μ_i is the baseline rate at which i initiates new e-mail threads. Each message received by officer i at time r_k^i elevates the overall rate of sending e-mails at time $t > r_k^i$, through the triggering function $g_i(t - r_k^i)$, which is assumed to be exponential. Time t is expressed continuously as hours since midnight on the day when the first e-mail was sent in the network.

In model (2.1), the background rate μ_i is assumed to be constant over the observation window $[0, T]$. This is unrealistic in light of the diurnal and weekly non-stationarities suggested in Figures 2.1 and 2.2. Non-stationary forms for the background rate will be discussed subsequently in Section 2.2.1.

The exponential triggering function is perhaps not unreasonable. For instance, Figure 2.7 shows that the survival function of the inter-event times for the observed e-mails sent by each officer in the network falls reasonably close to the 95%

confidence envelope formed from 100 simulated realizations of the IkeNet e-mail network (Appendix A) using estimated model (2.1). This plot indicates that the inter-event time distribution for the estimated model closely resembles that of the observed data.

As an illustration of model (2.1), the top panel in Figure 2.8 shows the estimated conditional intensity for officer 13, $\hat{\lambda}_{13}(t)$, over a three-day time period. The clustering in the times when e-mails are sent and received are easily discerned in this plot, and are characteristic of Hawkes point processes.

The parameters of model (2.1) characterize general e-mail communication habits of each officer. For instance, θ_i can be interpreted as the reply rate for officer i, since it is the expected number of reply e-mails[1] sent by officer i per e-mail received from another officer in the network, as

$$\lim_{T\to\infty} \int_{r_i^k}^{T} \theta_i \omega_i e^{-\omega_i (t - r_k^i)} dt = \lim_{T\to\infty} \theta_i (1 - e^{-\omega_i (T - r_k^i)}) = \theta_i.$$

The integrated triggering function over a finite time period will be slightly less than θ_i, but for the IkeNet data, where $T = 8640$ hours and $\omega^{-1} << T$ (see Table 2.1), θ_i will be extremely close to the expected number of replies per e-mail received for officer i. The speed at which officer i replies to e-mails is governed by the parameter ω_i, with larger values of ω_i indicating faster response times for officer i. Indeed, ω_i^{-1} is the expected number of hours it takes for officer i to reply to a typical e-mail.

[1]Note, in this work, a 'reply e-mail' is directed towards the network, and is not necessary sent directly back to the user that sent the original e-mail which triggered the reply. The distinction between a 'reply' and 'non-reply' e-mail is that a reply e-mail is triggered by and sent in response to a previously received e-mail, while a non-reply e-mail is not provoked by a received e-mail and indicates the initiation of a discussion thread.

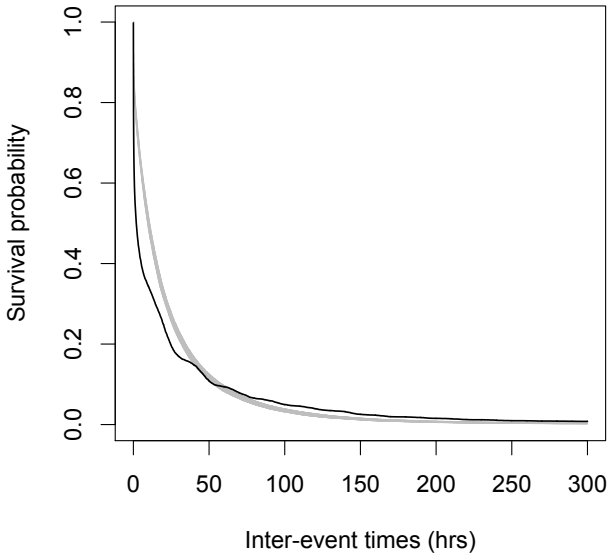

Figure 2.7. Survivor plot of the inter-event times for e-mails sent by each officer in the network (black line). A 95% confidence envelope was formed by simulating the network 100 times from the fitted model (2.1) and computing the survivor function for each realization. The pointwise 0.025 and 0.975 quantiles of the simulated survivor functions are plotted in gray.

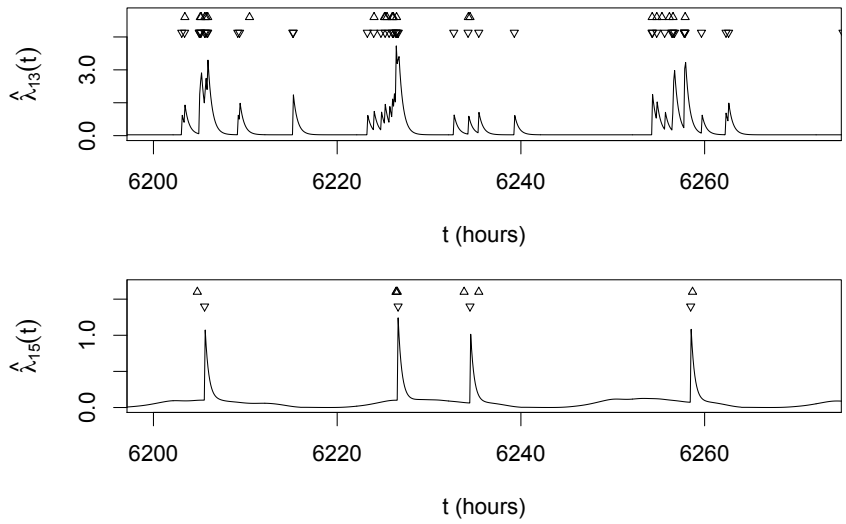

Figure 2.8. Top panel shows the estimated conditional intensity for officer 13 over a three-day period using the Hawkes model with the stationary background rate (2.1). The bottom panel shows the estimated conditional intensity for officer 15 over the same three-day period using the Hawkes model with the non-stationary background rate (2.2). The downward triangles represent the times when messages are received, while the upward triangles represent the times when messages are sent.

2.2.1 Non-stationary Background Rate

Model (2.1) makes the assumption that the background rate is a stationary Poisson process, which means in this context that the rate of creating new e-mail threads is constant at all times. This is not realistic due to the presence of circadian and weekly trends in e-mail traffic (see Figures 2.1 and 2.2). Malmgren et al. (2008) argued that the clustering and heavy-tails in the inter-event distribution of times when e-mails are sent is partially a consequence of rhythms in human activity (e.g. sleep, meals, work, etc.), and the authors explicitly modeled periodicities in e-mail communication as a non-stationary Poisson process. We take a similar approach by considering a non-stationary background rate for our Hawkes process model (2.1) of e-mail traffic:

$$\begin{aligned}\lambda_i(t) &= \nu_i \mu(t) + \sum_{r_k^i < t} g_i(t - r_k^i) \\ &= \nu_i \mu(t) + \theta_i \sum_{r_k^i < t} \omega_i e^{-\omega_i (t - r_k^i)}, \end{aligned} \tag{2.2}$$

where ν_i is a user specific parameter and $\mu(t)$ is a shared baseline density function that accounts for daily and weekly rhythms in e-mail activity. We define the integral of $\mu(t)$ to equal 1 over the observation window $[0, T]$. Our estimate of $\mu(t)$, denoted $\hat{\mu}(t)$, is found nonparametrically by a weighted kernel smoothing estimate over the e-mails sent by all officers (Figure 2.9); the details of this estimation procedure are given subsequently. Since $\int_0^T \nu_i \mu(t) dt = \nu_i$, the parameter ν_i can be interpreted as the expected number of background events, or non-reply e-mails, sent by officer i over the time interval $[0, T]$.

If we let $m \in \{0, \cdots, 59\}$ be the minute, $h \in \{0, \cdots, 23\}$ the hour, and $d \in \{0, \cdots, 6\}$ the day ($Mon = 0, \cdots, Sun = 6$) corresponding to time $t \in [0, T]$,

then our estimate of $\mu(t)$ is given by $\hat{\mu}(t) = Z \cdot \hat{f}(h + m/60)w(d)$, where

$$\hat{f}(h + m/60) = \frac{1}{\sigma} \sum_{k=1}^{N} P_k K \left(\frac{h + m/60 - h_k}{\sigma} \right)$$

$$= \frac{1}{\sigma} \sum_{k=1}^{N} P_k \frac{1}{\sqrt{2\pi}} e^{-\frac{(h+m/60-h_k)^2}{2\sigma^2}}, \tag{2.3}$$

$$w(d) = \sum_{k=1}^{N} P_k I(d_k = d), \tag{2.4}$$

and P_k is a probability weight that sums to one over $k \in \{1, \cdots, N\}$, where N is the total number of observed messages sent in the network. The notation h_k and d_k denote the hour after midnight and day of week for the k^{th} e-mail sent in the network. The constant of proportionality Z is chosen to ensure that $\hat{\mu}(t)$ integrates to 1 over $[0, T]$. An accurate approximation of Z can be found using a Riemann sum.

To get an initial estimate of $\hat{\mu}(t)$ we select equal probability weights $P_k = 1/N$, making (2.3) the standard kernel density estimate of the histogram of the number of e-mails sent by hour of day (Figure 2.1). For this kernel smoothing we choose a gaussian kernel $K(\cdot)$ with bandwidth σ set to the default value suggested by Scott (1992). To account for weekly trends $\hat{f}(\cdot)$ is multiplied by a weight $w(d)$, which is simply the proportion of all observed messages sent in the network on day d when $P_k = 1/N$ (Figure 2.2). Our initial estimate of the background rate density $\hat{\mu}(t)$, with equal probability weights, is plotted as the dashed curve in Figure 2.9. Note that $\hat{\mu}(t)$ is periodic, with period equal to one week (7 days / 168 hours), i.e $\hat{\mu}(t + 168) = \hat{\mu}(t)$, and one period of $\hat{\mu}(t)$ is shown in this figure. In Section 2.2.3, we will explain how to improve our estimate of $\hat{\mu}(t)$ by using the probabilities each e-mail is either a non-reply (background event) or reply (offspring event) to simultaneously estimate the model parameters and nonparametric background rate density.

To illustrate the fitted model, the lower panel of Figure 2.8 shows the estimated

conditional intensity for officer 15 under model (2.2). The troughs in the estimated conditional intensity in Figure 2.8 correspond to times when few e-mails are sent and received.

2.2.2 Alternative Model

One shortcoming of models (2.1) and (2.2) is that the reply rate θ_i for officer i does not depend on who sends an e-mail to i. According to this model, officer i sends the same expected number of reply messages to each e-mail received, regardless of the sender j. In order to incorporate some pairwise interactions between officers we consider the following alternative Hawkes process model for the rate at which officer i sends e-mails at time t:

$$\begin{aligned}\lambda_i(t) &= \nu_i \mu(t) + \sum_j \sum_{r_k^{ij} < t} g_{ij}(t - r_k^{ij}) \\ &= \nu_i \mu(t) + \sum_j \sum_{r_k^{ij} < t} \theta_{ij} \omega_i e^{-\omega_i (t - r_k^{ij})}. \end{aligned} \tag{2.5}$$

The triggering function, $g_{ij}(t - r_k^{ij})$, gives the contribution of the k^{th} message officer i receives from j at time r_k^{ij} to the conditional intensity at time t. The inner summation is over all messages officer i receives from j at times $r_k^{ij} < t$, and the outer summation is over all officers j in the network. Note that one may also model each officer pair (dyad) so that a distinct ω_{ij} and ν_{ij} is estimated for each receiver i and sender j, however with the current dataset this may not be advisable due to the sparsity in the number of e-mails sent between certain pairs of individuals (Figure 2.5) and the large number of additional parameters to estimate.

The parameters of model (2.5) help characterize e-mail communication behaviors between officers. For each officer i, there are twenty-one parameters θ_{ij}, each of which may be interpreted as the expected number of replies i sends per e-mail received from j. This additional information is gained at the expense of adding

twenty more parameters per network member than model (2.2). (Instances when officers send e-mails to themselves have been removed, so the reply rate θ_{ii} is not included in model (2.5).) A more in-depth comparison between models (2.2) and (2.5) is provided in Section 2.3.

2.2.3 Parameter Estimation

The parameters of models (2.1), (2.2), and (2.5) can be estimated by an expectation-maximization type of algorithm (Veen and Schoenberg, 2008; Marsan and Lengliné, 2008). Recall that for a self-exciting point process each event is either a background event or an offspring event (i.e. triggered by a previous event). This classification of events as background or offspring is referred to as the branching structure of the process. In most applications the branching structure is an unobserved or latent variable. For instance, it is not known whether an earthquake is an aftershock or mainshock, or in the case of IkeNet e-mail traffic, whether a message is a reply or non-reply. The EM algorithm works iteratively by first estimating the branching structure of a self-exciting point process (E-step), and then estimating model parameters (M-step) by maximizing the expected log-likelihood function, given the current estimate of the branching structure. Marsan proposed the EM algorithm as a way to estimate the conditional intensity nonparametrically, using a histogram estimator for the triggering function. Many authors have since applied the EM algorithm to parametric Hawkes process models (Lewis and Mohler, 2010; Hegemann et al., 2012), yielding closed form estimators for model parameters.

For the remainder of this section we will describe how to use an EM-type procedure to estimate the parameters of model (2.2). Models (2.1) and (2.5) can be estimated similarly. In particular, model (2.1) is just a special case of model (2.2) with $\mu(t) = 1/T$, where T is the length of the observation window in hours.

For the IkeNet dataset let s_l^i be the time when the l^{th} e-mail was sent by officer i, r_k^i be the time when the k^{th} e-mail was received by i, and N_i^{send} and N_i^{rec} be the number of messages sent and received by i. We may define the true branching structure for the e-mail network using the following random variables:

$$\psi_l^i = \begin{cases} 1 & \text{if } s_l^i \text{ is a non-reply message (background event)} \\ 0 & \text{otherwise,} \end{cases} \tag{2.6}$$

$$\chi_{kl}^i = \begin{cases} 1 & \text{if } s_l^i \text{ is a reply to message } r_k^i, \text{ where } s_l^i > r_k^i \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

The log-likelihood function (Ogata, 1978) for the conditional intensity defined in model (2.2) is given by

$$\begin{aligned} l_i(\Omega_i) = logL_i(\Omega_i) &= \sum_{k=1}^{N_i^{send}} \log(\lambda_i(s_k^i)) - \int_0^T \lambda_i(t)dt \\ &= \sum_{k=1}^{N_i^{send}} \log(\lambda_i(s_k^i)) - \left(\nu_i + \theta_i \sum_{k=1}^{N_i^{rec}} [1 - e^{-\omega_i(T - r_k^i)}] \right), \end{aligned} \tag{2.8}$$

where $\Omega_i = \{\nu_i, \theta_i, \omega_i\}$ is the parameter space for officer i. Recall that $\int_0^T \nu_i \mu(t)dt = \nu_i$ since $\mu(t)$ is a density function over $[0, T]$. In order to find the parameters $\hat{\Omega}_i$ that maximize (2.8) directly, numerical optimization techniques must be used. However, when incorporating information about the branching structure we instead work with the complete data log-likelihood function, which is more tractable for maximization, and decomposes additively into a likelihood function for the background process and a likelihood function for the triggering processes:

$$\begin{aligned} l_i^c(\Omega_i) = &\underbrace{\sum_{l=1}^{N_i^{send}} \psi_l^i log(\nu_i \mu(s_l^i)) - \int_0^T \nu_i \mu(t)dt}_{l_i^\mu} \\ &+ \underbrace{\sum_{k=1}^{N_i^{rec}} \left[\sum_{\{l: s_l^i > r_k^i\}} \chi_{kl}^i log(g_i(s_l^i - r_k^i)) - \int_{r_k^i}^T g_i(t - r_k^i)dt \right]}_{l_i^g}. \end{aligned} \tag{2.9}$$

Since the true branching structure is unobserved, we estimate model parameters by maximizing the expected complete data log-likelihood, which is found by replacing ψ_l^i and χ_{kl}^i in (2.9) with the estimated probabilities each event is either background or offspring:

$$B_l^i = \text{probability sent message } s_l^i \text{ is background} = \frac{\hat{\nu}_i \hat{\mu}(s_l^i)}{\hat{\lambda}_i(s_l^i)}, \tag{2.10}$$

$$\begin{aligned} O_{kl}^i &= \text{probability receiving message } r_k^i \text{ triggers sending message } s_l^i \\ &= \begin{cases} \dfrac{\hat{g}_i(s_l^i - r_k^i)}{\hat{\lambda}_i(s_l^i)} & s_l^i > r_k^i \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.11}$$

Moreover, these probabilities can also be used to get a more accurate estimate of the non-stationary background rate $\hat{\mu}(t)$ using weighted kernel density estimation (2.3 and 2.4). This leads to the EM-type algorithm for estimating model (2.2):

Step 1. Initialize parameters estimates $(\hat{\nu}_i^{(0)}, \hat{\theta}_i^{(0)}, \hat{\omega}_i^{(0)})$ for each officer i. Initialize the background rate density $\hat{\mu}^{(0)}(t)$ using equal probability weights $P_k^{(0)} = 1/N$ for each event $k \in \{1, \cdots, N\}$ in (2.3) and (2.4). Set the iteration index $m = 0$.

Step 2. For each officer i, find $B_l^{i(m+1)}$ and $O_{kl}^{i(m+1)}$ using the parameter estimates and background density from iteration m.

Step 3. Estimate the background rate density, $\hat{\mu}^{(m+1)}(t)$, using the weighted KDE defined in (2.3) and (2.4), setting $P_k^{(m+1)} = B_k^{(m+1)} / \sum_{k=1}^N B_k^{(m+1)}$ where B_k is the probability that e-mail $k \in \{1, \cdots, N\}$ is non-reply (background) at iteration $m+1$. The bandwidth σ is found using the estimate from Scott (1992).

Step 4. Estimate parameters by maximizing the expected complete data log-

likelihood using the probability estimates from Step 2:

$$\hat{\nu}_i^{(m+1)} = \sum_{l=1}^{N_i^{send}} B_l^{i(m+1)}$$

$$\hat{\theta}_i^{(m+1)} = \frac{\sum_{k=1}^{N_i^{rec}} \sum_{\{l:s_l^i > r_k^i\}} O_{kl}^{i(m+1)}}{N_i^{rec} - \sum_{k=1}^{N_i^{rec}} e^{-\hat{\omega}_i^{(m)} \Delta r_k^i}}$$

$$\hat{\omega}_i^{(m+1)} = \frac{\sum_{k=1}^{N_i^{rec}} \sum_{\{l:s_l^i > r_k^i\}} O_{kl}^{i(m+1)}}{\sum_{k=1}^{N_i^{rec}} \sum_{\{l:s_l^i > r_k^i\}} O_{kl}^{i(m+1)} (s_l^i - r_k^i) + \sum_{k=1}^{N_i^{rec}} \hat{\theta}_i^{(m+1)} \Delta r_k^i e^{-\hat{\omega}_i^{(m)} \Delta r_k^i}}$$

where $\Delta r_k^i = T - r_k^i$.

Step 5. Update $m \leftarrow m + 1$ and repeat Steps 2–5 until convergence when $\left| \sum_i \left[l_i(\hat{\Omega}_i^{(m+1)}) - l_i(\hat{\Omega}_i^{(m)}) \right] \right| < \epsilon$ for some small value ϵ (in practice we set $\epsilon = 10^{-3}$).

The estimators in Step 4 are found by setting the partial derivates of the expected complete data log-likelihood (2.9), with respect to each of the parameters, equal to zero. The convergence criteria in Step 5 is in terms of the log-likelihood function in (2.8). The convergence of this EM-type algorithm for the self-exciting models is apparent in Figure 2.10.

Parameter estimates, standard errors, and maximum log-likelihood values (2.8) for the Hawkes process models (2.1, 2.2, and 2.5) are given in Tables 2.1, 2.2, and 2.3. Since estimated model (2.5) contains twenty-one reply rates $\hat{\theta}_{ij}$ we instead present the average reply rate $\hat{\bar{\theta}}_i = \sum_j \hat{\theta}_{ij} \cdot N_{ij}^{rec} / N_i^{rec}$, where N_{ij}^{rec} is the number of messages officer i received from j, for each officer in Table 2.3. Notice that the parameter estimates for models (2.2) and (2.5) presented in these tables are similar. This result is consistent with model (2.2) being contained within model (2.5) (it is the case with $\theta_{ij} = \theta_i$ for each sender j and recipient i pair).

The standard errors in Tables 2.1, 2.2, and 2.3 are found by simulating each model 100 times (Appendix A) using the EM parameter estimates from the observed data. For each simulated realization of the network, the parameters are then re-estimated, resulting in 100 sets of re-estimated parameters for each model. Standard errors are computed by taking the root-mean-square deviation between the parameter re-estimates from the simulation and the parameter estimate from the observed data.

By simulating the network repeatedly, one can also form 95% confidence envelopes for the non-stationary background rate density $\hat{\mu}(t)$ (Figure 2.9). The gray error bound in this figure is formed by simulating fitted model (2.5) 100 times (Appendix A) and re-estimating the background rate for each simulated realization of the e-mail network. Note that the background rate from the observed network (solid black curve) falls reasonably within the 95% confidence bands, indicating that the estimated background rate for the model is consistent with the estimate from the observed data.

Inspection of Tables 2.1 and 2.2 reveals that model (2.2) outperforms model (2.1) since it has larger maximum log-likelihood values for every officer. This suggests that inclusion of the non-stationary background rate provides an overall better fit to the network data. The maximum log-likelihood values for model (2.5) (see Table 2.3) are greater than model (2.2) for each officer; however, due to the large number of parameters, model (2.5) does not outperform model (2.2) typically (as well as overall) by a statistically significant margin according to the Akaike Information Criterion (AIC) of Akaike (1974). Diagnostic comparisons between each model are discussed in greater detail in Section 2.3.4.

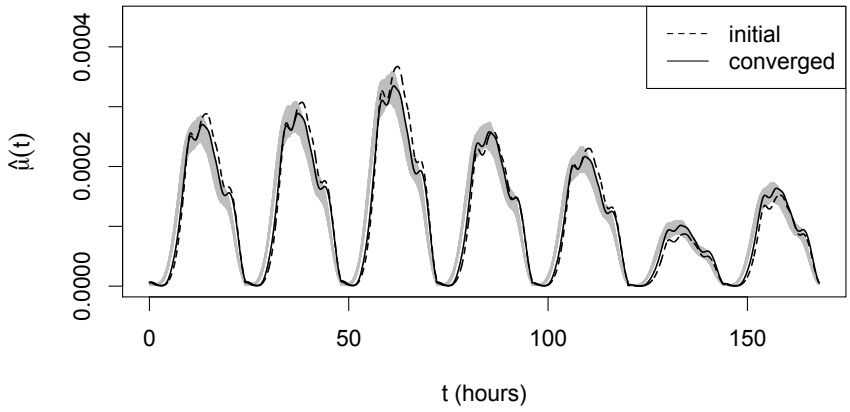

Figure 2.9. Estimated background rate density $\hat{\mu}(t)$ for the IkeNet e-mail network (solid black curve) using model (2.5) after convergence of the EM-type algorithm. The dashed curve is the initial estimate of the background rate density using equal probability weights. This figure only shows one period (i.e. one week, Mon.–Sun.) of $\hat{\mu}(t)$. A 95% simulation confidence envelope was formed by re-estimating the background rate for 100 simulated realizations of fitted model (2.5), and the pointwise 0.025 and 0.975 quantiles are plotted in gray.

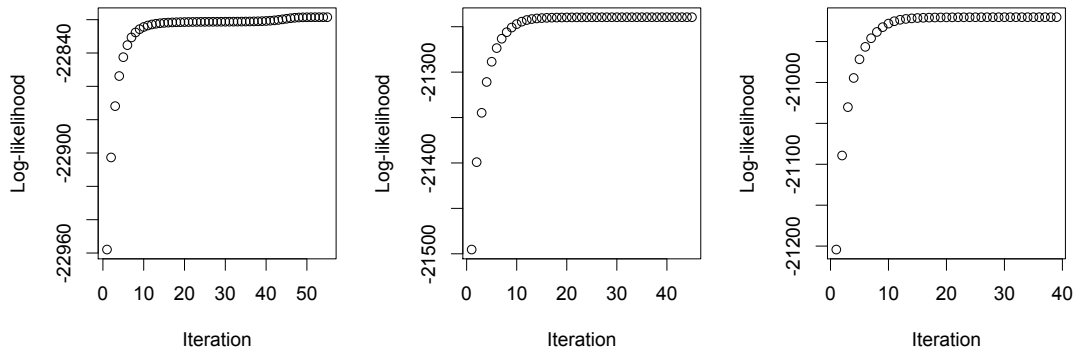

Figure 2.10. Scatter plots showing the convergence of the EM-type algorithm, in terms of log-likelihood, for estimating the self-exciting models (2.1, 2.2, and 2.5, respectively).

Table 2.1. Parameter estimates, standard errors, and maximum log-likelihood values for model (2.1). Standard errors are computed by the root-mean-square deviation from 100 simulations of the estimated model.

i	N_i^{send}	$\hat{\mu}_i$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l_i(\hat{\Omega}_i)$
1	94	0.009 (0.0010)	0.17 (0.04)	8.64 (2.54)	-464.2
2	260	0.014 (0.0013)	0.58 (0.05)	3.64 (0.39)	-732.8
3	301	0.021 (0.0017)	0.49 (0.05)	1.38 (0.19)	-1089.4
4	316	0.024 (0.0017)	0.43 (0.05)	2.93 (0.40)	-1126.4
5	179	0.012 (0.0013)	0.35 (0.04)	1.64 (0.25)	-702.9
6	207	0.014 (0.0013)	0.34 (0.04)	3.10 (0.40)	-752.5
7	276	0.016 (0.0015)	0.51 (0.04)	0.80 (0.10)	-989.0
8	355	0.025 (0.0014)	0.40 (0.04)	4.71 (0.49)	-1125.6
9	868	0.044 (0.0024)	0.54 (0.02)	6.68 (0.41)	-1620.0
10	155	0.012 (0.0012)	0.33 (0.05)	3.29 (0.54)	-635.4
11	687	0.034 (0.0020)	0.55 (0.03)	2.19 (0.15)	-1647.9
12	277	0.018 (0.0016)	0.43 (0.05)	1.35 (0.19)	-1018.5
13	876	0.038 (0.0024)	0.45 (0.02)	2.21 (0.14)	-2029.1
14	296	0.016 (0.0016)	0.57 (0.04)	2.87 (0.32)	-871.4
15	558	0.040 (0.0023)	0.53 (0.04)	1.75 (0.17)	-1717.8
16	181	0.014 (0.0012)	0.41 (0.06)	6.44 (1.09)	-683.6
17	295	0.019 (0.0015)	0.26 (0.02)	2.87 (0.38)	-1023.1
18	1181	0.059 (0.0028)	0.64 (0.03)	6.91 (0.32)	-1853.8
19	247	0.019 (0.0016)	0.53 (0.07)	0.83 (0.14)	-992.8
20	73	0.006 (0.0008)	0.26 (0.06)	3.17 (0.83)	-360.2
21	26	0.002 (0.0005)	0.21 (0.08)	0.73 (0.67)	-158.7
22	689	0.030 (0.0018)	0.73 (0.04)	3.52 (0.23)	-1223.4

Table 2.2. Parameter estimates, standard errors, and maximum log-likelihood values for model (2.2). Standard errors are computed by the root-mean-square deviation from 100 simulations of the estimated model.

i	N_i^{send}	$\hat{\nu}_i/N_i^{send}$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l_i(\hat{\Omega}_i)$
1	94	0.83 (0.05)	0.16 (0.05)	9.82 (6.47)	-430.1
2	260	0.47 (0.04)	0.56 (0.05)	4.06 (0.42)	-682.1
3	301	0.65 (0.04)	0.45 (0.05)	1.62 (0.23)	-1017.8
4	316	0.71 (0.03)	0.37 (0.04)	4.41 (0.64)	-1021.1
5	179	0.57 (0.05)	0.34 (0.04)	1.65 (0.28)	-690.7
6	207	0.59 (0.04)	0.32 (0.04)	3.50 (0.48)	-717.9
7	276	0.53 (0.05)	0.47 (0.05)	0.90 (0.11)	-932.6
8	355	0.63 (0.03)	0.38 (0.03)	5.52 (0.65)	-1060.1
9	868	0.50 (0.02)	0.49 (0.03)	10.18 (0.58)	-1464.4
10	155	0.70 (0.04)	0.31 (0.05)	4.63 (0.90)	-598.9
11	687	0.48 (0.03)	0.50 (0.03)	2.73 (0.24)	-1541.5
12	277	0.63 (0.04)	0.38 (0.04)	1.99 (0.29)	-973.4
13	876	0.44 (0.02)	0.40 (0.02)	2.76 (0.22)	-1908.7
14	296	0.50 (0.04)	0.54 (0.05)	3.31 (0.34)	-802.0
15	558	0.68 (0.03)	0.46 (0.04)	2.52 (0.25)	-1614.9
16	181	0.69 (0.04)	0.39 (0.05)	7.52 (1.27)	-640.2
17	295	0.61 (0.04)	0.23 (0.03)	4.17 (0.49)	-954.5
18	1181	0.48 (0.02)	0.59 (0.02)	9.80 (0.57)	-1629.8
19	247	0.71 (0.04)	0.46 (0.06)	1.25 (0.24)	-938.8
20	73	0.73 (0.05)	0.25 (0.06)	3.41 (1.14)	-341.6
21	26	0.72 (0.09)	0.20 (0.07)	0.75 (0.80)	-149.9
22	689	0.42 (0.02)	0.68 (0.04)	4.39 (0.29)	-1128.5

Table 2.3. Parameter estimates, standard errors, and maximum log-likelihood values for model (2.5) . The column labeled $\hat{\bar{\theta}}_i$ gives the estimated average reply rate for each officer $\hat{\bar{\theta}}_i = \sum_j \hat{\theta}_{ij} \cdot N_{ij}^{rec}/N_i^{rec}$. Standard errors are computed by the root-mean-square deviation from 100 simulations of the estimated model.

i	N_i^{send}	$\hat{\nu}_i/N_i^{send}$	$\hat{\bar{\theta}}_i$	$\hat{\omega}_i$	$l_i(\hat{\Omega}_i)$
1	94	0.82 (0.04)	0.16 (0.04)	9.62 (2.94)	-421.4
2	260	0.47 (0.04)	0.56 (0.05)	4.09 (0.41)	-668.1
3	301	0.65 (0.04)	0.44 (0.05)	1.74 (0.23)	-1003.9
4	316	0.71 (0.03)	0.37 (0.04)	4.53 (0.62)	-1013.9
5	179	0.56 (0.05)	0.35 (0.05)	1.50 (0.26)	-678.1
6	207	0.59 (0.04)	0.32 (0.04)	3.64 (0.55)	-703.3
7	276	0.53 (0.04)	0.47 (0.05)	0.91 (0.13)	-924.5
8	355	0.63 (0.03)	0.38 (0.04)	5.59 (0.54)	-1043.3
9	868	0.49 (0.02)	0.49 (0.03)	9.81 (0.61)	-1453.9
10	155	0.69 (0.05)	0.32 (0.05)	4.17 (0.73)	-586.8
11	687	0.48 (0.03)	0.50 (0.03)	2.76 (0.20)	-1522.4
12	277	0.64 (0.03)	0.37 (0.04)	2.21 (0.30)	-954.1
13	876	0.45 (0.03)	0.40 (0.02)	2.83 (0.22)	-1885.4
14	296	0.50 (0.03)	0.54 (0.04)	3.23 (0.35)	-793.1
15	558	0.69 (0.03)	0.43 (0.04)	2.90 (0.34)	-1594.5
16	181	0.68 (0.04)	0.39 (0.06)	7.40 (1.13)	-633.1
17	295	0.61 (0.03)	0.23 (0.02)	4.09 (0.53)	-935.2
18	1181	0.48 (0.02)	0.59 (0.02)	9.67 (0.47)	-1600.9
19	247	0.71 (0.04)	0.46 (0.07)	1.26 (0.22)	-931.6
20	73	0.72 (0.07)	0.26 (0.07)	3.17 (1.10)	-333.8
21	26	0.71 (0.11)	0.21 (0.09)	0.69 (0.53)	-143.0
22	689	0.42 (0.02)	0.68 (0.04)	4.60 (0.30)	-1095.8

2.3 IkeNet Analysis

2.3.1 Characterizing E-mail Communication Behavior

The parameter estimates in Table 2.2 provide insight into the communication habits of officers in the network. For instance, the estimated proportion of e-mails sent by officer i that are not replies (background events) is given by $\hat{\nu}_i/N_i^{send}$. In other words, $\hat{\nu}_i$ can be thought of as the estimated number of e-mail threads officer i initiated over the one-year observation period. For example, according to the fitted model (2.2), approximately 68% of e-mails sent by officer 15 are not replies and 48% of e-mails sent by officer 18 are not replies. Over the entire network, $\hat{\nu}_i/N_i^{send}$ ranges between 42% and 83%, and the estimated overall percentage of e-mails sent in the network that are not replies is $\sum_{i=1}^{22} \hat{\nu}_i/N \approx 55\%$, where N is the total number of observed messages for the network.

The estimated mean number of replies officer i sends in response to a typical e-mail received is given by $\hat{\theta}_i$ in Table 2.2. For example, officer 18 sends approximately 59 replies per 100 e-mails received, while officer 15 sends approximately 46 replies per 100 e-mails received. Note also that the estimated proportion of sent e-mails that are not replies ($\hat{\nu}_i/N_i^{send}$) is higher for officer 15 than 18. This suggests that officer 15 has a higher tendency to initiate e-mail conversations than officer 18, while officer 18 has a higher tendency to respond to e-mails than officer 15. Over the entire network, $\hat{\theta}_i$ ranges between 16% and 68%, and the estimated overall percentage of e-mails sent in the network that are replies is $\sum_{i=1}^{22} \hat{\theta}_i \cdot N_i^{rec}/N \approx 45\%$.

The speed at which officers send e-mails is governed by $\hat{\omega}_i^{-1}$, which can be interpreted as the estimated mean time it takes officer i to reply to an e-mail. By examining Table 2.2 we see that officers 18 and 9 are estimated to take about 6 minutes to reply to an e-mail. This is much faster than many of the other officers, such as officer 13, who takes an estimated 21 minutes, on average, to reply. The

matrix plot (Figure 2.5) shows that officers 9 and 18 communicate frequently with each other, which may account for their similar and speedy response times. The estimated mean response times for officers in the network ranges from about 6 to 80 minutes, and the estimated overall mean time it takes an officer to reply is $\sum_{i=1}^{22} N_i^{send} \cdot \hat{\omega}_i^{-1}/N \approx 0.307$ hours or 18.4 minutes.

2.3.2 Inferring Network Leadership

An important question is what properties of an e-mail network can best identify and rank the perceived leaders of that network. As mentioned in Section 2.1, each officer in the IkeNet dataset was asked in a survey to list up to five officers they considered to be strong team leaders, and up to five officers they considered to be strong military leaders. The distinction made in the survey was that a team leader is someone who is perceived as confident leading a business or research project, while a military leader is someone who is perceived as confident leading soldiers in combat. Figures 2.11 and 2.12 are scatter plots of the total number of e-mails sent versus the aggregate number of team and military leadership votes, respectively. The correlations in these scatter plots are weak to moderate, and an inspection reveals that sending a relatively large number of e-mails does not necessarily indicate that an officer is a top leader. For instance, officer 15 stands out for having the most votes for both team and military leadership, though this officer ranks below the 80th percentile in terms of the total number of e-mails sent (officers 18, 13, 9, 22, and 11 all sent more messages than officer 15). Moreover, officer 9 sent a large number of e-mails in the network, but ranks low in terms of team and military leadership votes. Clearly, total number of e-mails sent is a poor predictor of one's perceived leadership status within the network.

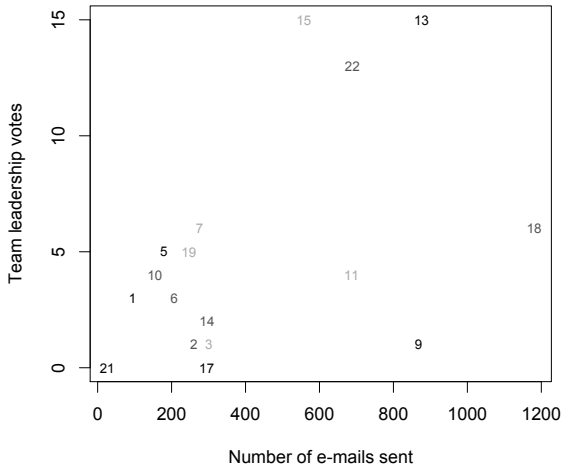

Figure 2.11. Plot of the total number of e-mails sent versus the aggregate number of votes each officer received for perceived team leadership ($r = 0.52$). Votes are based on a survey which asked each officer to list up to five other officers in the network that he or she considered to be a strong team leader.

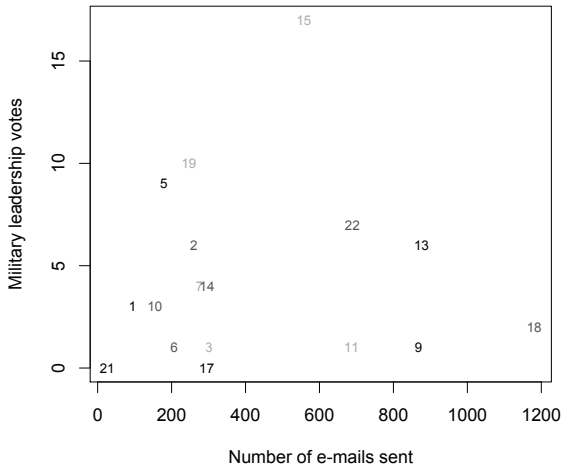

Figure 2.12. Plot of the total number of e-mails sent versus the aggregate number of votes each officer received for perceived military leadership ($r = 0.13$). Votes are based on a survey which asked each officer to list up to five other officers in the network that he or she considered to be a strong military leader.

Fortunately, the parameter estimates from the Hawkes process models quantify other features of e-mail communication which may be predictive of network leadership. Two particularly important features which we consider are the rate at which a user initiates e-mail threads (background rate), and the responsiveness of a user to e-mails received (reply rate). We capture these features in a potential predictor Y, which is defined for each officer i as the total number of other officers j for which officer i has an estimated mean reply rate ($\hat{\theta}_{ij}$) above threshold c_1, and sent an estimated number of non-reply e-mails ($\hat{\nu}_i N_{ij}^{send}/N_i^{send}$) above threshold c_2. That is

$$Y_i(c_1, c_2) = \sum_j I(\hat{\theta}_{ij} > c_1,\ \hat{\nu}_i N_{ij}^{send}/N_i^{send} > c_2), \tag{2.12}$$

where I denotes the indicator function, N_{ij}^{send} is the number of e-mails sent from officer i to j, and all fitted parameters are from model (2.5). Intuitively, officers that initiate many e-mail threads and are very response to e-mails received obtain a high value for predictor Y, and are therefore considered potential leaders.

For our analysis we consider four sets of thresholds for the predictor defined in (2.12), denoted by $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, and $Y^{(4)}$. Let $A = \{\hat{\theta}_{ij} | i \neq j\}$ be the the set of estimated reply rates from officers i to j, $B = \{\hat{\nu}_i N_{ij}^{send}/N_i^{send} | i \neq j\}$ be the set containing the estimates for the number of non-reply e-mails (background events) sent from officers i to j, and $\bar{\theta} = \frac{1}{N}\sum_i \sum_j N_{ij}^{rec}\hat{\theta}_{ij}$ be the estimated mean percentage of reply e-mails sent in the entire network. For covariate $Y^{(1)}$, threshold $c_1 = \bar{\theta} = 0.45$ and threshold $c_2 = 4.79$ is the median of set B. For covariate $Y^{(2)}$, threshold $c_1 = \bar{\theta} = 0.45$ and threshold $c_2 = 9.92$ is the mean of set B. The thresholds $(c_1, c_2) = (0.33, 4.79)$ selected for $Y^{(3)}$ are the respective medians of sets A and B. The thresholds $(c_1, c_2) = (0.52, 9.91)$ selected for $Y^{(4)}$ are the respective third quartiles of sets A and B. Of course, many other thresholds are possible, and the selected thresholds are just simple, easily computed candidates.

Tables 2.4 and 2.5 lists several predictors of network leadership and the Pear-

son, Spearman, and Kendall correlations between these predictors and the survey votes for team and military leadership. The Pearson correlation is between the predictor of interest and the total number of team or military leadership votes (Figures 2.11 and 2.12). The Spearman and Kendall correlations compare the predicted rankings with the rankings from the leadership survey votes. A value of 1 for Kendall's coefficient indicates that the rankings are perfectly concordant, 0 indicates that the rankings are independent, and -1 indicates the rankings are perfectly discordant (in reverse order). The last column in both tables gives the top four leaders identified by each predictor.

Tables 2.4 and 2.5 show that predictor Y, for the 4 selected sets of thresholds, is much more highly correlated with team and military leadership votes than the total number of messages sent (N^{send}) or received (N^{rec}) by each officer. Predictor Y also does a better job at identifying the top leaders than N^{send} and N^{rec}. For instance, $Y^{(1)}$, $Y^{(2)}$, and $Y^{(4)}$ all correctly identify the top 4 team leaders (13, 15, 22, and 18). Moreover, officer 15, the highest ranked officer in terms of team and military leadership votes, is identified by predictor Y as a top leader, while N^{send} and N^{rec} do not recover the importance of this officer.

The points in Figure 2.13 represent the Pearson (r_p), Spearman (r_s), and Kendall (τ) correlations between the predictors (Y, N^{send}, and N^{rec}) and the leadership survey votes. The plot labeled (a) gives the correlations with the team leadership votes, and the plot labeled (b) gives the correlations with the military leadership votes. The correlations corresponding to predictor Y are plotted in blue for the 4 sets of thresholds ($Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, and $Y^{(4)}$), while the correlations corresponding to the naive predictors (N^{send} and N^{rec}) are plotted in red. Plot (a) shows that predictor Y has higher correlations than the naive predictors (N^{send} and N^{rec}) for all sets of thresholds considered. In this plot, predictor $Y^{(1)}$ performs the best overall at predicting and ranking team leaders; $Y^{(3)}$ also does comparably well at ranking team leaders even though it has a lower Pearson correlation.

Plot (b) also shows that predictor Y has higher correlations than N^{send} and N^{rec}; this is true for all sets of thresholds considered, with $Y^{(4)}$ the only exception since it has approximately the same Spearman correlation as N^{send}. In this plot, $Y^{(1)}$ and $Y^{(2)}$ perform the best overall at predicting and ranking military leaders.

2.3.3 Sensitivity to Thresholds

The correlations between predictor $Y(c_1, c_2)$ and the leadership survey votes depend on the choice of thresholds c_1 and c_2. Figure 2.13 shows that for very reasonable threshold selections (i.e. means, medians, and third quartiles as discussed in Section 2.3.2), predictor Y performs much better at ranking and estimating leadership scores than the naive predictors N^{send} and N^{rec}. Table 2.4 also shows that Y is generally able to identify the top 4 teams leaders with slight variations in order. For all threshold values considered in Tables 2.4 and 2.5, Y does a better job than N^{send} or N^{rec} at identifying the top leaders.

In Figure 2.14 we further assess the sensitivity of $Y(c_1, c_2)$ to the threshold values. In each panel, the blue lines give the correlations (Pearson, Spearman, or Kendall, as indicated) between $Y(c_1, c_2)$ and the leadership votes as c_1 varies continuously between 0 and 0.52, and c_2 takes fixed values at the first quartile (1.8), median (4.8), and third quartile (9.9) for the number of background events (non-reply e-mails) sent between officers in the network. The upper three panels give the correlations between Y and the team leadership votes, and the lower three panels give the correlations between Y and the military leadership votes. The red horizontal line in each panel is the respective correlation between predictor N^{send} and the leadership votes.

In Figure 2.14 the blue lines typically fall above the red horizontal line in each panel; this indicates that, for a wide variety of thresholds, predictor $Y(c_1, c_2)$ is associated more strongly with the leadership votes than N^{send}. In the top three

panels, threshold $c_2 = 4.8$ (median) performs the best overall at ranking network officers, as indicated by the relatively high Spearman and Kendall correlations when this threshold value is chosen. In bottom three panels, there appears to be a peak when threshold c_1 is approximately 0.45, which is the estimated mean percentage of reply e-mails sent in the entire network ($\bar{\theta}$). Conclusively, in all panels it is apparent that for a wide variety of choices for thresholds we obtain quantitatively similar results.

Table 2.4. Predictors of team leadership.

Predictor	r_p	r_s	τ	Estimated top 4 leaders
N^{send}	0.52^{*}	$0.40^{\cdot}$	$0.29^{\cdot}$	$18, 13, 9, 22$
N^{rec}	0.49^{*}	$0.39^{\cdot}$	$0.29^{\cdot}$	$13, 18, 9, 11$
$Y^{(1)}$	0.68^{**}	0.66^{**}	0.52^{**}	$15, 18, 13, 22$
$Y^{(2)}$	0.64^{**}	0.50^{*}	0.40^{*}	$13, 15, 18, 22$
$Y^{(3)}$	0.53^{*}	0.60^{**}	0.47^{**}	$13, 18, 9, 15$
$Y^{(4)}$	0.66^{**}	0.45^{*}	0.36^{*}	$13, 18, 22, 15$

The significance values testing whether each correlation is different from zero are denoted by (·) at the 0.1 level, (*) at the 0.05 level, and (**) at the 0.01 level. In the event of ties in Y the tiebreaker is the number of e-mails sent in determining the top 4 leaders. The actual top 4 team leaders from the survey votes are officers 13, 15, 22, and 18.

Table 2.5. Predictors of military leadership.

Predictor	r_p	r_s	τ	Estimated top 4 leaders
N^{send}	0.13	0.29	0.21	$18, 13, 9, 22$
N^{rec}	0.02	0.20	0.15	$13, 18, 9, 11$
$Y^{(1)}$	0.48^{*}	0.44^{*}	0.34^{*}	$15, 18, 13, 22$
$Y^{(2)}$	0.45^{*}	0.45^{*}	0.37^{*}	$13, 15, 18, 22$
$Y^{(3)}$	$0.36^{\cdot}$	$0.41^{\cdot}$	0.32^{*}	$13, 18, 9, 15$
$Y^{(4)}$	0.32	0.27	0.24	$13, 18, 22, 15$

The significance values testing whether each correlation is different from zero are denoted by (·) at the 0.1 level, (*) at the 0.05 level, and (**) at the 0.01 level. In the event of ties in Y the tiebreaker is the number of e-mails sent in determining the top 4 leaders. The actual top 4 military leaders from the survey votes are officers 15, 19, 5, and 22.

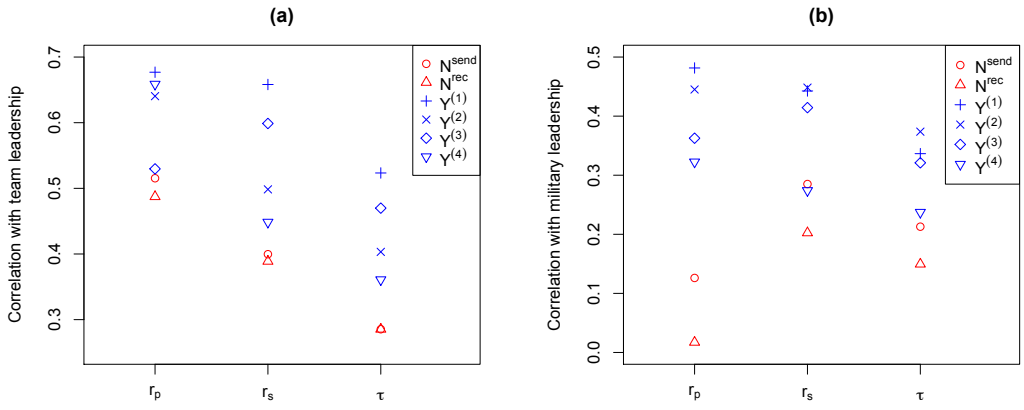

Figure 2.13. The points in each plot represent the Pearson (r_p), Spearman (r_s), and Kendall (τ) correlations between the predictor variables and the team (panel a) and military (panel b) leadership votes. The correlations corresponding to the naive predictors N^{send} (number of e-mails sent) and N^{rec} (the number of e-mail received) are plotted in red. The correlations corresponding to predictor $Y(c_1, c_2)$, defined in (2.12), are plotted in blue for various threshold selections c_1 and c_2. The specific thresholds chosen for $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, and $Y^{(4)}$ are discussed in Section 2.3.2.

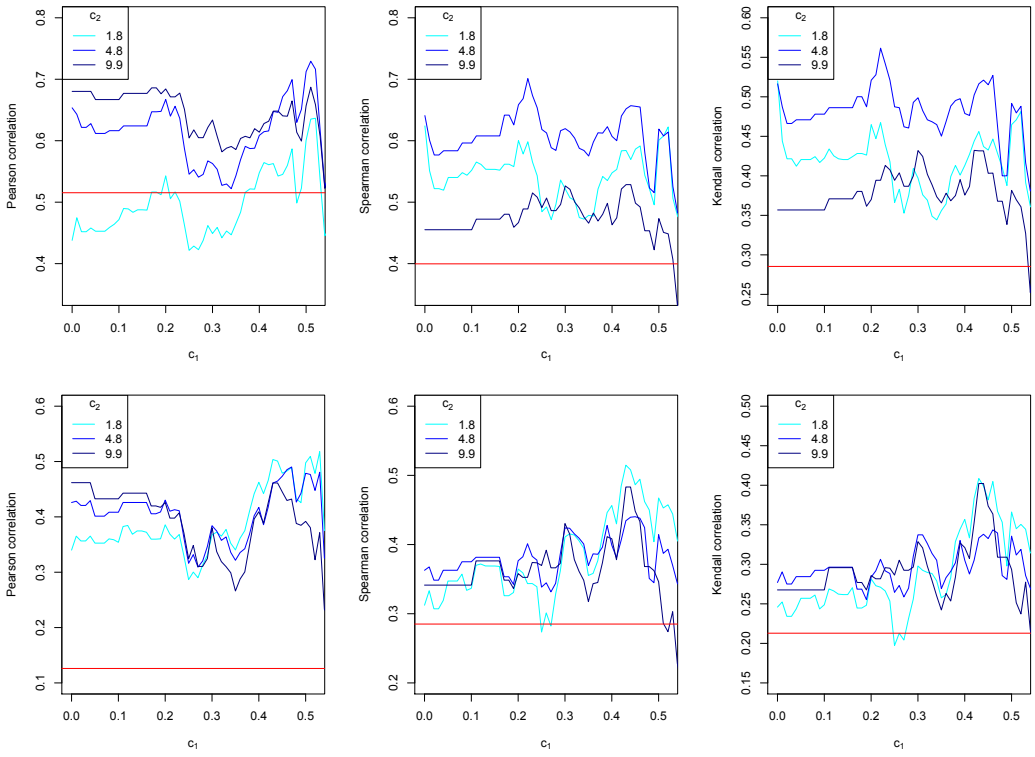

Figure 2.14. Sensitivity plots for the Spearman, Pearson, and Kendall correlations between predictor $Y(c_1, c_2)$ and the team leadership votes (upper three panels) and military leadership votes (lower three panels) for different values of thresholds c_1 and c_2. The lines in each plot give the correlations between $Y(c_1, c_2)$ and the leadership votes as c_1 varies continuously between 0 and 0.52, and c_2 takes fixed values at the first quartile (1.8), median (4.8), and third quartile (9.9) for the number of background events (non-reply e-mails) sent between officers in the network. The red horizontal line in each plot is the respective correlation between N^{send} (total number of e-mails sent by each officer) and the leadership survey votes. This plot shows that for a wide variety threshold values predictor $Y(c_1, c_2)$ is more strongly correlated with the leadership votes than the naive predictor N^{send}.

2.3.4 Model Comparison and Diagnostics

The maximized log-likelihoods for the network and corresponding AIC values are provided in Table 2.6. The first row gives these values for a stationary Poisson model of e-mail network traffic, where the rate at which each officer sends e-mails is constant and given by $\lambda_i(t) = \mu_i$. This model only has twenty-two parameters (the constant rate for each officer). The other three rows of this table are for the Hakwes process models (2.1, 2.2, and 2.5) described in Section 2.2. The Hawkes process model (2.1) fits the data significantly better than the stationary Poisson model according to the AIC. Additionally, the maximum log-likelihood value for the model with non-stationary background rate (2.2) is higher than the model with the stationary background rate (2.1). This indicates that taking diurnal and weekly trends into account provides an overall better fit to the network data. While the increase in maximum log-likelihood is noteworthy, it is not entirely justifiable to use the AIC to compare the models that include the nonparametrically estimated background density $\hat{\mu}(t)$ (2.2 and 2.5) with the completely parametric model (2.1). The Hawkes process model (2.5), which incorporates pairwise interactions between officers, fits the data slightly more closely than model (2.2) as measured by the maximum log-likelihood, but scores worse in terms of AIC. This is because the AIC penalizes for the large number of parameters in (2.5). Although, due to the overall sparsity in the IkeNet e-mail network (Figure 2.5), about 15% of the estimated parameters in (2.5) are equal to zero. Comparison of models (2.2) and (2.5) suggests that e-mail traffic is well modeled by few parameters, and adding in extra parameters to capture the differences in reply rates between officer pairs does not provide a significantly better fit to the data. However, the utility of model (2.5) to predict and rank network leaders was shown in Section 2.3.2.

The simulation procedure described in Appendix A can be used to evaluate how well the estimated Hawkes process models capture aspects of the observed data. For instance, one test of predictive performance is to split the data into a

Table 2.6. Number of parameters (ρ), AIC and maximum log-likelihood values for the Poisson and Hawkes process models of the IkeNet e-mail network. The value KS is the Kolmogorov-Smirnov test statistics comparing the transformed time to the uniform distribution.

	ρ	$l(\hat{\Omega})$	AIC	KS
Stationary Poisson	22	-32347.4	64738.9	0.39
Hawkes model (2.1)	66	-22818.5	45769.0	0.17
Hawkes model (2.2)	66	-21239.5	42611.0	0.15
Hawkes model (2.5)	506	-20920.2	42852.5	0.14

training and validation set and assess how well each model simulated many times from the parameters estimated from the training set is able to reproduce some characteristic of the validation set. For this diagnostic, the selected training set is the first 11 months ($T = 7920$ hours) of e-mail data, and the selected validation set is the last month (720 hours, between 13 April 2011 and 12 May 2011) of e-mail data. Here, we choose the portion of all e-mails sent attributed to each individual officer as our metric for the predictive performance of each model on the validation set. We have chosen to inspect each officer's portion of all e-mails sent rather than each officer's raw sent e-mail count since the overall rate of e-mail exchanges appears to be much higher during the final month of our dataset (the validation set) than is typical of the previous months, and our model cannot account for this change. This unusual spike in activity, occurring during the beginning of May, can be seen clearly in the time series plot (Figure 2.3).

Using the first 11 months ($T = 7920$ hours) of e-mail data in the training set we estimate models (2.1), (2.2), and (2.5) with the EM-type algorithm described in Section 2.2.3. To estimate the non-stationary background rate density, $\hat{\mu}(t)$, in Step 2 of the EM-type algorithm we use the weighted kernel density estimate in

(2.3) and (2.4) evaluated over the e-mail events occurring in the training set. For each self-exciting model, we use the parameters estimated from the training data to simulate the IkeNet e-mail network 100 times over a period of $T = 720$ hours (1 month). For the simulation procedure for the non-stationary background process (Appendix A, Algorithm A), the estimate $\hat{\mu}(t)$, from the training set, is evaluated over a 720 hour period that starts and ends on the same days as the validation set (only the start and end days matter since $\hat{\mu}(t)$ is periodic).

In Figure 2.15, the 0.025 an 0.975 quartiles for the simulated proportions of e-mails sent by each officer in the network under each model are plotted as gray vertical lines. The observed proportion of e-mails sent by each officer in the validation set is also plotted in this figure as black horizontal lines. Most of these observed proportions are either contained within or fall near the simulated intervals for each officer. Only officers 10, 13, and 22 deviate significantly from the simulated outcomes. There also does not appear to be any major differences between the predictive performances of the considered models. However, this is not surprising since the the non-stationary background rates in models (2.2) and (2.5) only accounts for daily and weekly trends, and since we are simulating over a period of one month there should not be any major differences in the simulated number of messages for these models when compared to model (2.1) with the stationary background term. Moreover, the similarity between the performances of models (2.2) and (2.5) in this diagnostic is consistent with the log-likelihood analysis for these models.

Another goodness-of-fit diagnostic considered in Ogata (1988) is the transformed time $\{\tau_k^i\}$, which may be defined for each officer i as

$$\tau_k^i = \Lambda(s_k^i) = \int_0^{s_k^i} \lambda_i(t)dt. \tag{2.13}$$

If the model used in their construction is correct, then the transformed times should form a Poisson process with rate 1 (Meyer, 1971), and similarly the inter-

event times $\tau_k^i - \tau_{k-1}^i$ between the transformed times should follow an exponential distribution; hence $U_k^i = 1 - \exp\{-(\tau_k^i - \tau_{k-1}^i)\}$ should be uniformly distributed over $[0, 1)$. Thus, as suggested e.g. in Ogata (1988), if the main features of the data are well captured by the estimated model, a plot of U_{k+1}^i versus U_k^i should look like a uniform scatter of points. These plots are presented in Figure 2.16 for the stationary Poisson process model and all Hawkes process models (2.1, 2.2, and 2.5) of e-mail network traffic considered in this chapter. A comparison of these plots reveals much less clustering around the perimeter for the Hawkes process models, indicating that while the Poisson model clearly fails to account for the clustering in the data, this feature is noticeably less pronounced for the self-exciting models. Furthermore, there appears to be slightly less clustering in the plot for model (2.2) than the plot for model (2.1), and likewise when comparing models (2.5) and (2.2). This claim is supported by the decreasing values of the Kolmogrov-Smirnov test statistics in Table 2.6, which compare the transformation $\{U_k\}$ for each network model with the uniform distribution.

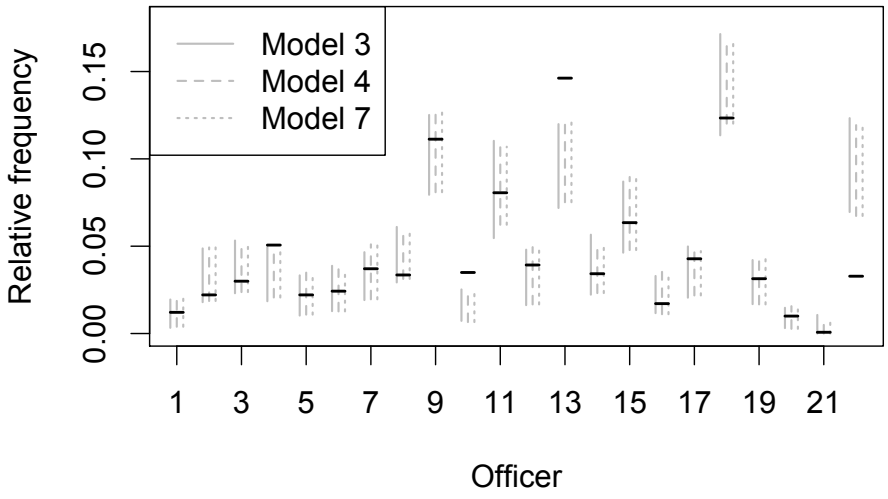

Figure 2.15. Comparison of the simulated and observed proportion of e-mails sent by each officer over a period of one month (720 hours). The gray vertical lines are the pointwise 0.025 and 0.975 quartiles for the proportions generated from 100 simulations of the IkeNet e-mail network using the models estimated from the training set (first 11 months of e-mail data). The black horizontal lines are the observed proportions from the validation set.

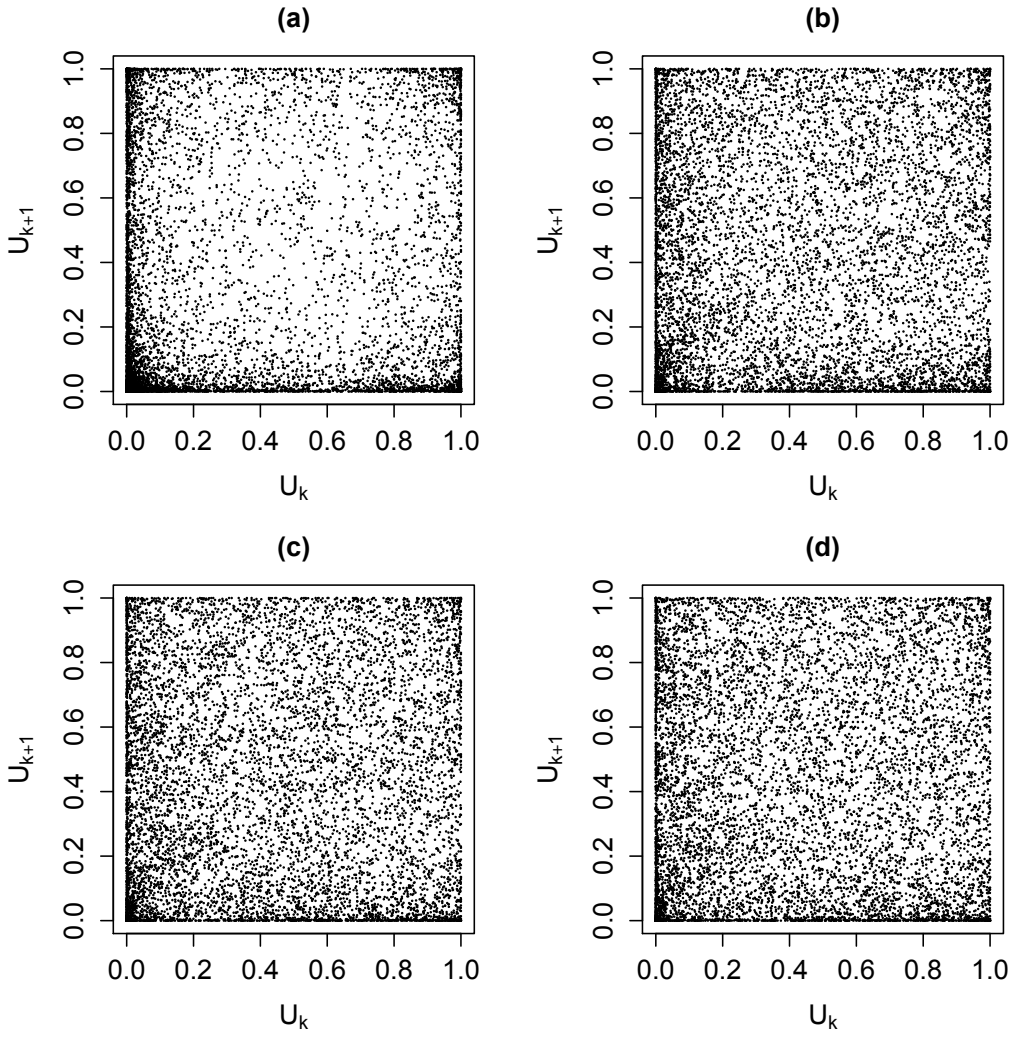

Figure 2.16. (a-d) Plot of U_{k+1} versus U_k for the stationary Poisson process model and Hawkes process models (2.1, 2.2, and 2.5) of e-mail activity on the network, respectively.

2.4 Comparative Analysis Using the Enron E-mail Dataset

E-mail datasets are difficult to find due to the many privacy concerns involved when making such data publicly available. The Enron e-mail corpus is one of the few large e-mail communication datasets readily available for public research. The corpus was originally released in 2002 by the Federal Energy Regulatory Commission (FERC) during the scandal. William Cohen (CMU) distributed a version of the original corpus containing about 517,430 e-mails from 151 users on 3500 folders (Cohen, 2009). Shetty and Adibi (USC) cleaned Cohen's versions of the dataset and organized the corpus in a MySQL database containing 252,759 messages collected from 151 users (Shetty and Adibi, 2004).

We consider the sender, recipient, and timestamp of each message in a closed version of the Enron e-mail network of Shetty and Adibi (2004) containing messages sent between the 151 users. Once duplicates and messages individuals sent to themselves are removed, the corpus is reduced to 14,959 sent messages and 24,705 received messages. Approximately 27.7% of e-mails sent in the closed network have multiple recipients. Each sent message is coded as a single sent message, regardless of the number of recipients, and in this way the number of receiving and sending messages are allowed to vary for each user. When defining N_i^{send} and $\sum_j N_{ij}^{send}$ for the Enron dataset, a multicast e-mail sent by i to 10 recipients, for example, would contribute 1 to N_i^{send} and 10 to $\sum_j N_{ij}^{send}$.

Figure 2.17 is a time series plot of the number of e-mails sent each month in the closed Enron e-mail network over the three year period between May 1999 and June 2002. There is a pronouced peak in activity between the dates when Jeffery Skilling abruptly resigned as CEO (August 2001) and Enron filed for bankruptcy (December 2001). E-mail usage steadily declined to a zero level during the months after January 2002. The scatter plot in Figure 2.18 (right panel) shows that there is a strong association ($r \approx 0.72$) between the natural logarithms of the number of

messages sent and received by each user in the closed Enron network. This result is similar to the IkeNet dataset (left panel), which shows a very high correlation ($r \approx 0.95$) between the raw number of incoming and outgoing messages. We apply the logarithmic transform to the Enron data since it is more skewed than IkeNet.

We fit the Hawkes process models (2.1, 2.2, and 2.5) to the Enron data using the EM-type algorithm described in Section 2.2.3. The maximum log-likelihood and AIC values for the network are provided in Table 2.7. The results presented in this table are quite similar to IkeNet, indicating that perhaps our models generalize well to other larger e-mail networks. The self-exciting model (2.1) fits the Enron network data significantly better than the stationary Poisson model according to the AIC. Additionally, there is a substantial increase in the maximum log-likelihood values for the network with the inclusion of the non-stationary background rate in model (2.2). Hence, it appears that the modeling of diurnal and weekly periodicities in e-mail network activity provides a better fit to the Enron data than the stationary background rate in (2.1). Due to the large number of parameters, the AIC for model (2.5) is much larger than model (2.2). However, like IkeNet, the Enron e-mail network is sparse in the number of messages sent between pairs of individuals. In fact, approximately 94% of the estimated parameters for model (2.5) of the Enron dataset are equal to zero. Enron e-mail traffic is well captured by a few parameters for each node in the network, and incorporating parameters to model pairwise connections between users does not significantly improve the overall fit to the data. The values of the Kolmogorov-Smirnov test statistic (Section 2.3.4) indicate the Hawkes process models for the Enron network are also accounting for the clustering in the times when e-mails are sent significantly better than the stationary Poisson model.

Table 2.8 displays the mean percentage of reply and non-reply messages estimated from the self-exciting models (2.1, 2.2, and 2.5) of the Enron and IkeNet e-mail networks. These percentages are quite similar for both networks: model

(2.1) estimates that approximately half of the e-mails sent in each network are non-replies, and this percentage increases with the inclusion of the non-statationary background rate in models (2.2) and (2.5). Table 2.8 also reveals that the estimated reply times are much higher for the Enron dataset than the IkeNet dataset. For instance, according to estimated model (2.2), the middle 50% of estimated reply times ($\hat{\omega}_i$) are between 13.2 and 28.8 minutes for the IkeNet e-mail network, and between 1.63 and 60.52 hours for the Enron e-mail network. One explanation is that IkeNet officers are using mobile devices to send e-mails, and are thus able to reply to messages quickly, within an hour, while individuals in Enron are using personal desktops, and therefore take much longer to reply.

2.4.1 Describing and Inferring Enron Leadership Roles

The prediction of the leadership and hierarchy underlying the Enron corporation from the e-mail corpus data is an important problem, and there are various techniques in the literature proposed for this task. Shetty and Adibi (2005) use a graph entropy model to find prominent and influential individuals in the Enron e-mail dataset. Nodes (e-mail users) that cause the greatest change in graph entropy for the network once removed are ranked highest and regarded as most important. Creamer et al. (2009) use a SNA (Social Network Analysis) approach to extracting social hierarchy information from the Enron dataset. These authors rank and group e-mail users according to a social score, which is defined as a weighted sum of user specific statistics such as number of messages, number of cliques, degree and betweenness centrality. McCallum et al. (2007) proposed the Author-Recipient-Topic model which learns topic distributions conditioned on the senders and receivers of e-mail messages; the topic distributions estimated from the Enron e-mail corpus are used to predict the roles of individuals in the network.

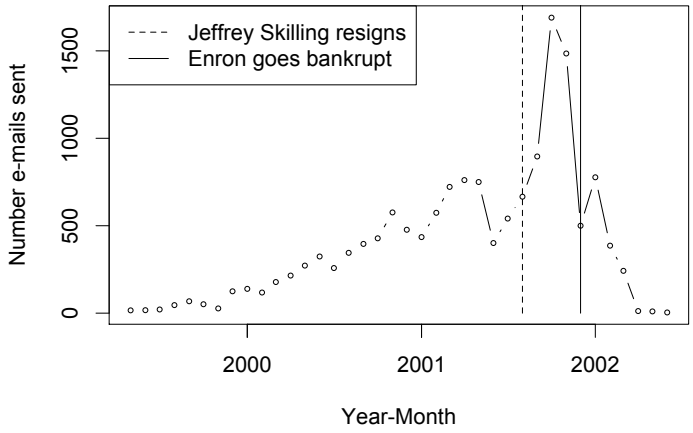

Figure 2.17. Time series plot of number of e-mails sent each month between May 1999 and June 2002 in the Enron dataset.

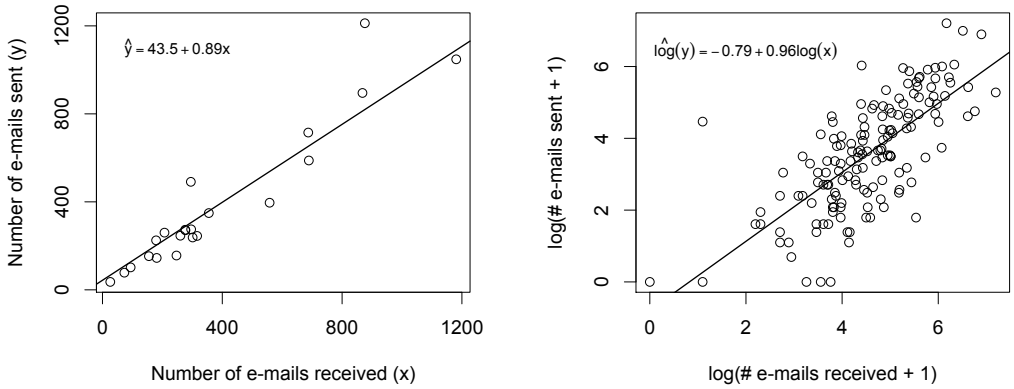

Figure 2.18. Left Panel: Scatter plot of the total number of e-mails received (x) versus the total number of e-mails sent (y) by each officer in the IkeNet dataset. The scatter plot and regression line show a strong association between the raw number of e-mails sent and received ($r = 0.95$). Right Panel: Scatter plot of the natural logarithm of total number of e-mails received versus the natural logarithm of the total number of e-mails sent by each user in the Enron dataset. The scatter plot and regression line show a strong association between the natural logarithm of number of e-mails sent and received ($r = 0.72$).

Table 2.7. Number of parameters (ρ), AIC and maximum log-likelihood values for the Poisson and Hawkes process models of the Enron e-mail network. The value KS is the Kolmogorov-Smirnov test statistics comparing the transformed time to the uniform distribution.

	ρ	$l(\hat{\Omega})$	AIC	KS
Stationary Poisson	151	-85605.0	171512.0	0.42
Hawkes model (2.1)	453	-75031.4	150968.8	0.28
Hawkes model (2.2)	453	-70721.7	142349.4	0.27
Hawkes model (2.5)	22952	-68925.9	183755.9	0.25

Table 2.8. Mean percent non-reply messages ($\sum_i \hat{\nu}_i/N$), mean percent reply messages ($\sum_i \hat{\theta}_i \cdot N_i^{rec}/N$), average reply time ($\sum_i N_i^{send}\hat{\omega}_i^{-1}/N$) in hours, and first and third quartiles for reply times estimated from the Hawkes process models of the Enron and IkeNet e-mail networks.

Dataset	Model	% Non-reply	% Reply	Mean reply time
IkeNet	Hawkes model (2.1)	50.2%	49.8%	0.4 (0.28, 0.6)
	Hawkes model (2.2)	54.4%	45.6%	0.31 (0.22, 0.48)
	Hawkes model (2.5)	54.6%	45.4%	0.31 (0.22, 0.43)
Enron	Hawkes model (2.1)	50%	50%	68.47 (1.69, 111.28)
	Hawkes model (2.2)	59.5%	40.5%	48.5 (1.63, 60.52)
	Hawkes model (2.5)	54.6%	45.4%	61.19 (1.53, 49.16)

For the actual positions of the users in the Enron e-mail network we draw from the classification of Shetty and Adibi (2004) of workers into nine categories: CEO, President, Vice President, Managing Director, Director, Manager, Lawyer, Trader, and Employee. The position Employee refers to individuals that serve non-managerial roles such as associates, analysts, and administrative assistants. In order to fill in the position data missing in Shetty and Adibi's classification we cross-referenced Creamer et al. (2009) and the actual legal documents released during the Enron scandal (Congress, 2003). Using all three sources we determined the positions of 150 of the 151 users in the Enron e-mail network.

Table 2.9 presents mean counts and standard deviations for the number of messages sent and received by individuals within each of the nine occupational categories for Enron's corporate hierarchy. Inspection of this table reveals that the Enron CEOs have the lowest average number of messages sent and received when compared to all other job categories. Lawyers and Vice Presidents stand out for sending and receiving the highest mean number of e-mails. However, the standard deviations indicate that there is much variability between individuals within each group. Hence, the discrimination of user roles within the Enron corporate hierarchy based purely on the counts for the number of messages sent and received would be difficult; this motivates looking at additional features of e-mail users' communication behaviors supplied by the parameter estimates from the Hawkes process models.

Table 2.10 presents features of e-mail communication estimated from self-exciting models (2.2) and (2.5), averaged over the users belonging to each of the nine occupational categories of Enron's corporate hierarchy. The features considered in this table are the estimated mean proportion of sent e-mails that are not replies ($\hat{\nu}/N^{send}$), the estimated mean reply rate ($\hat{\theta}$), and the predictor Y (equation 2.12). Three sets of thresholds are considered for $Y(c_1, c_2)$, denoted by $Y^{(1)}$, $Y^{(2)}$, and $Y^{(3)}$, which are defined similarly as the threshold selections for the

IkeNet dataset (Section 2.3.2).[2]

The features considered in Table 2.10 characterize general communication behaviors for each occupational position. For example, an estimated 84% of e-mails sent by the four Enron CEOs are not replies to e-mails they received from individuals in the network. Moreover, the CEOs have an estimated mean reply rate of 0.1 and thus only send an average of 10 reply messages per 100 messages received. When compared to all other occupational categories, CEOs send the the highest estimated percentage of e-mails that are not replies and have the lowest estimated reply rate. Hence, an interesting feature of CEOs revealed by the self-exciting models is that, on average, they are not responsive to e-mails received and tend to initiate e-mail conversations or threads. This is in contrast to the 14 Enron Managers, who have the highest estimated mean reply rate (0.34) and sent the lowest estimated mean proportion of e-mails that are not replies (0.26). Individuals with the job title Employee fall in-between CEOs and Managers in terms of these features. In general, it appears that as we travel down the Enron hierarchy, the average reply rate increases and the average proportion of sent e-mails that are not replies decreases. The major exception to this are the Traders which are more similar to CEOs than Employees in terms of these features.

Predictor $Y(c_1, c_2)$, which performed well for identifying IkeNet leaders, has large average values for Presidents and Vice Presidents in the Enron network. The standard deviations for values of Y are also large, although this is not surprising since there can be wide disparities in use of e-mails within groups (as seen in Table 2.9 as well). Lawyers also seem to be a class of their own, having large values for Y relative to other occupational categories.

One way to infer the leadership status of users in the Enron network is to

[2]Due to the overall sparsity of the Enron e-mail network the median and third quartiles for the set of estimated reply rates and set containing the number of background events sent between officers are zero. Thus $Y^{(3)} = Y^{(4)}$ since both have trivial thresholds $c_1 = c_2 = 0$, and we only consider $Y^{(3)}$ in the subsequent analysis of Enron.

consider simple binary classification rules. For instance, CEOs send far fewer e-mails, on average, than other Enron users (Table 2.9). Hence, to infer CEO status we can consider a cutoff value for N^{send} and classify all users that sent a total number of e-mails below the cutoff as CEOs, and non-CEOs otherwise. For any particular cutoff value we can compute the true positive rate (the percentage of CEOs correctly classified as CEOs) and the false positive rate (the percentage of non-CEOs that are incorrectly classified as CEOs). Similar binary classification rules can be constructed using the other predictors (N^{rec}, $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$) as well. Figure 2.19 panel (a) shows the Receiver Operating Characteristic (ROC) curves constructed by plotting the true positive versus false positive rates for all possible cutoff values for each predictor variable for classifying users as CEOs or non-CEOs. The other panels in Figure 2.19 show the ROC curves generated from similar binary classification rules for predicting whether or not each user is a Vice President / President (panel b) and Manager / Director / Managing Director (panel c).

The ROC curves corresponding to the binary classification of CEO status (panel a) indicate that the number of e-mails sent and received by each officer are the main distinguishing features for CEOs. These naive predictors (N^{send} and N^{rec}) perform generally as well as Y. Thus the additional features of e-mail communication estimated from the Hawkes process models do not contribute much to inferring CEO status, beyond what is already provided for by simple messages count totals. The large amount of variability between the true positive rates corresponding to each predictor is due to the small sample size of 4 CEOs in the Enron network.

The ROC curves corresponding to the binary classification of President / Vice President status (panel b) all perform rather similarly, but $Y^{(3)}$ appears to perform the best overall. For example, for a fixed false positive rate of 0.05, the true positive rates for each predictor are 0.07 for N^{send}, 0.1 for N^{rec}, 0.19 for $Y^{(1)}$, 0.09

for $Y^{(2)}$, and 0.21 for $Y^{(3)}$. Hence, there is noticeable improvement in predictive performance when using $Y(c_1, c_2)$ to distinguish Presidents / Vice Presidents status from the rest of the Enron users. However, this improvement only holds for the thresholds selected for $Y^{(1)}$ and $Y^{(3)}$, while $Y^{(2)}$ performs only as well as the naive predictors. Thus the ability for Y to distinguish Enron Vice President / President status is moderately sensitive to threshold selection.

The ROC curves corresponding to the binary classification of Manager / Director / Managing Director status (panel c) are all very close to the line $y = x$ (true positive rate equal to false positive rate) for false positive rates less than 0.3. Therefore, the binary classifiers constructed from each predictor variable are not doing any better than random chance at these values. For larger false positive rates (greater than 0.3) $Y^{(1)}$ and $Y^{(2)}$ appear to perform better than the other predictor variables (N^{send}, N^{rec}, $Y^{(3)}$) at discriminating Manager / Director / Managing Director status.

2.5 Discussion

Self-exciting point process models for e-mail networks clearly outperform traditional stationary Poisson models for both the IkeNet and Enron datasets considered here. These Hawkes process models, which appear to properly account for the clustering in the times when e-mails are sent and the overall branching structure of e-mail communication, are improved by accounting for diurnal and weekly rhythms in e-mail traffic in the background rate component. The estimated parameters of these models, such as $\hat{\theta}$ and $\hat{\nu}$, are easily interpretable and characterize important properties of e-mail communication, such as an individual's tendency to reply to e-mails and initiate new e-mail threads.

A network leader may possess more qualities than simply sending and receiving many messages. One attribute of a leader may be his or her responsiveness

to messages received from others in the network. Furthermore, a leader may initiate many e-mail conversations, and not rely on others to start projects and make decisions. The parameters of the Hawkes process model (2.5) quantified these additional features, which we attempted to combine into a measure $Y(c_1, c_2)$ (equation 2.12) for inferring network leadership. The results of our analysis of the IkeNet social network reveal that predictor Y is much more strongly correlated with the leadership survey votes and rankings than the naive predictors N^{send} (total number of e-mails sent) and N^{rec} (total number of e-mails received) for several reasonable threshold considerations. Moreover, an analysis of the sensitivity of $Y(c_1, c_2)$ to thresholds c_1 and c_2 demonstrates that we get quantitatively similar results for a wide variety of threshold selections as well (Figure 2.14).

For the Enron dataset we observed that CEOs, the highest ranked individuals within the network, send and receive far fewer e-mails, on average, than users in other occupational categories within the Enron hierarchy. Moreover, the estimated Hawkes process parameters also reveal that CEOs have a much higher tendency to initiate e-mail conversations (high background rate) than send replies (low reply rate). One possible explanation is that CEOs may be older than most other users in Enron and rely more on forms of communication besides e-mail (e.g. telephone, verbal, mail), or that many of the messages they received were low priority due to their high status within the organization. Enron Presidents and Vice Presidents are much more active within the e-mail network than CEOs since they send and receive a high volume of messages. Moreover, these users generally have relatively high values for predictor Y. This indicates, perhaps, that a discriminating feature for Presidents / Vice Presidents is the initiation of many e-mail threads and responsiveness to messages received from other users within the corporation. Note that Enron is merely one company, and a troubled one at that, so we hesitate to generalize our results to communication within other corporations, and further study is needed to verify if our findings apply to other

companies as well.

A main difference between the IkeNet and Enron networks is that the IkeNet social network is relatively flat (all officers in the network have the same military rank and are enrolled in the same academic program at West Point), while Enron has a complex leadership hierarchy that spans across multiple departments and positions. There is also much variability in e-mail usage and behavior between individuals with roughly the same role and position in the Enron social network. Therefore, it is a more straightforward process to identify and rank leaders within the IkeNet social network than to infer Enron leadership roles using various features of e-mail communication estimated from sender, recipient, and timestamp fields of e-mail logs.

In order to infer leadership on the Enron social network we constructed simple binary classification rules using the same predictor variables (N^{send}, N^{rec}, Y) considered in the IkeNet analysis. All predictors performed similarly in this analysis since the corresponding ROC curves are close together. Our analysis also suggests that no single predictor variable stands out as being able to best predict all the different leadership roles. For instance, the number of e-mails sent and received appears to be the main distinguishing feature for Enron CEOs, while predictor Y appears to perform slightly better than the naive predictors (N^{send} and N^{rec}) at distinguishing Vice President / President status. However, the applicability of $Y(c_1, c_2)$ to inferring leadership roles in Enron is less robust to threshold selection than in the IkeNet social network.

Table 2.9. Mean number of messages sent and received by users at different positions in Enron's corporate hierarchy.

Position	n	N^{send}	N^{rec}	Total
CEO	4	27.5 (39.1)	45.2 (36.4)	72.8 (26.3)
President	4	112 (124.7)	254.5 (195.5)	366.5 (303.8)
Vice President	25	162.1 (206.9)	267 (298.6)	429.1 (456.8)
Managing Director	5	59.6 (40.9)	105.6 (30.7)	165.2 (58.6)
Director	19	112.1 (312.4)	145.2 (130.9)	257.2 (421.3)
Manager	14	62 (58.2)	136.2 (184.7)	198.2 (208.6)
Lawyer	9	315.8 (325)	413.2 (302.4)	729 (520.3)
Trader	36	58.6 (97)	103.7 (94.3)	162.3 (170.8)
Employee	34	61.6 (66.2)	123 (137.3)	184.6 (191.7)

Note: The values for n are the number of individuals belonging to each occupational category. The values in the other columns are the means of the specified variables evaluated over the users belonging to each position, with corresponding standard deviations given in parenthesis.

Table 2.10. Features from the estimated Hawkes process models for describing e-mail communication behaviors at different positions in Enron's corporate hierarchy.

Position	n	$\hat{\nu}/N^{send}$	$\hat{\theta}$	$Y^{(1)}$	$Y^{(2)}$	$Y^{(3)}$
CEO	4	0.84 (0.36)	0.1 (0.05)	0.8 (1)	0.2 (0.5)	3 (4.8)
President	4	0.6 (0.16)	0.18 (0.13)	5.8 (7.5)	5.2 (6.6)	13.5 (15.2)
V. President	25	0.56 (0.3)	0.27 (0.27)	4.4 (3.3)	2.8 (2.3)	9.7 (6)
M. Director	5	0.65 (0.28)	0.2 (0.14)	2.6 (2.7)	1.6 (2.5)	6.4 (4)
Director	19	0.55 (0.2)	0.34 (0.4)	2.3 (3.8)	1.8 (3.9)	4.5 (4.9)
Manager	14	0.26 (0.34)	0.34 (0.53)	2 (1.8)	1 (0.9)	5.1 (4.1)
Lawyer	9	0.68 (0.12)	0.24 (0.18)	5.2 (3.4)	5 (3.4)	10.1 (4)
Trader	36	0.78 (0.15)	0.13 (0.12)	1.6 (1.9)	1.3 (1.8)	3.2 (3.5)
Employee	34	0.52 (0.28)	0.24 (0.22)	2.2 (2.5)	1.7 (2.2)	4.5 (4.2)

Note: The values in the columns are the estimated means of the specified variables evaluated over the individuals belonging to each position, and the standard deviations of the estimates for each variable are given in parenthesis. The table values for $\hat{\nu}/N^{send}$ and $\hat{\theta}$ are calculated as a weighted average and weighted standard deviation, with weights proportional to the number of e-mails sent and received by each individual, respectively. Mean values and standard deviations for $Y^{(1)}$, $Y^{(2)}$. and $Y^{(3)}$ are not weighted. The thresholds for $Y^{(1)}$, $Y^{(2)}$, and $Y^{(3)}$ are defined similarly for the Enron and IkeNet datasets (Section 2.3.2).

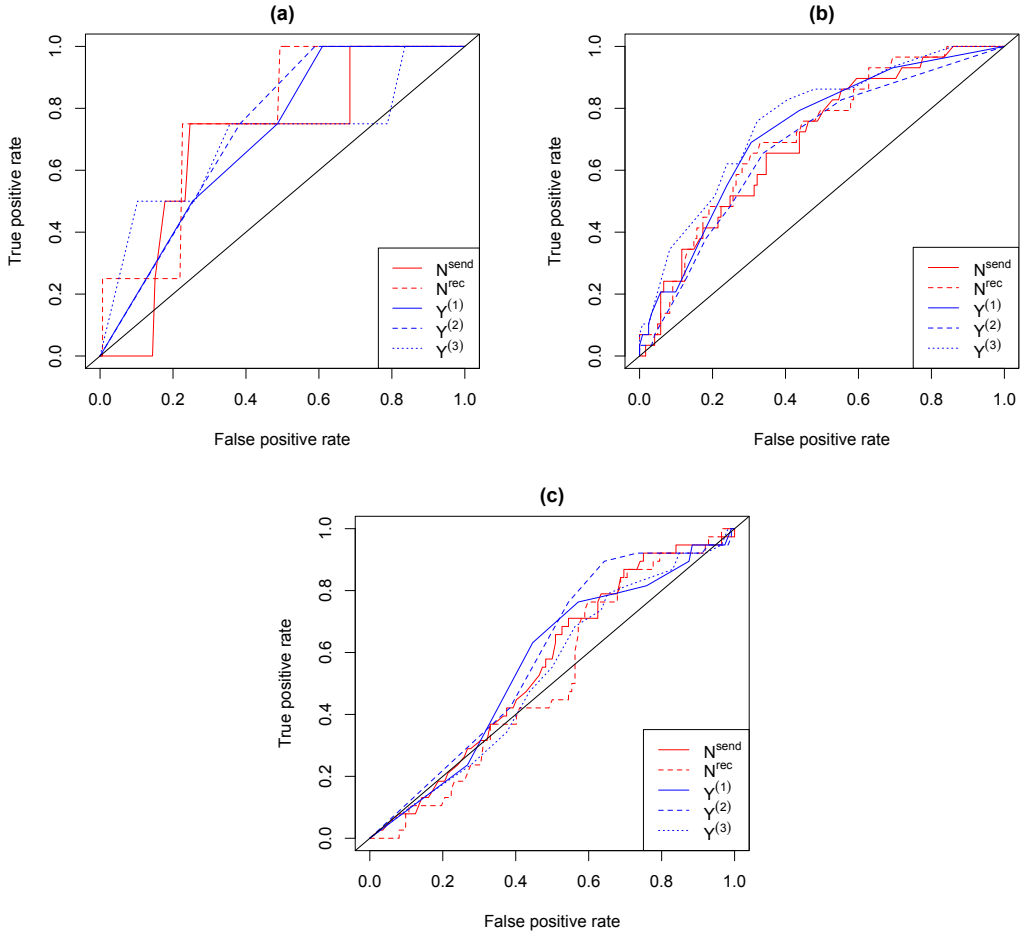

Figure 2.19. ROC curves corresponding to the binary classification of different Enron leadership roles. For each predictor of leadership (N^{send}, N^{rec}, Y) a cut-off value is chosen to classify each user as either a leader or non-leader. The ROC curves are constructed by considering all possible cut-off values for each predictor variable and plotting the corresponding true positive and false positive rates. The ROC curves in panels (a), (b), and (c) are for the classification rules for predicting whether or not each user is a CEO, President / Vice President, and Manager / Director / Managing Director, respectively.

CHAPTER 3

Nonparametric Methods for Estimating Point Process Models of Seismicity

Point process models of earthquake seismicity usually rely heavily on parametric assumptions about the triggering function for the spatial-temporal rate of aftershock activity following an earthquake. Some important examples are the parametric forms of the Epidemic Type Aftershock Sequences (ETAS) model of Ogata (1998). Marsan and Lengliné (2008) proposed a more flexible nonparametric approach for estimating point process models of seismicity which makes no a-priori assumptions about the shape of the triggering function, and provides a data-driven estimate instead. Their method, named Model Independent Stochastic Declustering (MISD), is an iterative algorithm that alternates between first estimating the probability each earthquake in the catalogue is either a mainshock or aftershock and second, updating a stationary background rate for mainshock activity and a probability weighted histogram estimate for the triggering function.

Nonparametric methods for estimating point process models have shown a wide rage of applications, especially in situations where the form of the intensity function is unknown and difficult to determine. Using wavelets Brillinger (1998) described a technique for estimating the conditional intensity and second order intensity with applications to neurophysiology and seismology. Adelfio and Chiodi (2013, 2015) considered a semi-parametric estimation procedure that simultaneously estimates a nonparametric background rate and parametric triggering function for a space-time Hawkes process model of seismicity. Marsan and

Lengliné (2008) applied the fully nonparametric MISD method to a Southern California earthquake catalogue to estimate the spatial-temporal rates aftershock activity following an earthquake of given magnitude. They also demonstrated the application of their routine for stochastically declustering earthquake catalogues to isolate mainshocks and remove aftershock clusters. Nichols and Schoenberg (2014) used MISD as a diagnostic tool to evaluate the dependency between the magnitude of an earthquake and the magnitudes of its aftershocks. By repeatedly applying the MISD algorithm to stochastically assign earthquakes as either mainshocks or aftershocks they created confidence intervals for the average magnitude of aftershocks following an earthquake of given magnitude. In an application to criminology, Mohler et al. (2011) developed a Monte-Carlo based nonparametric method similar to MISD to estimate a space-time point process model for the occurrence rate of burglaries in a Los Angeles district. They demonstrated that this approach leads to improved hotspot maps for flagging times and locations where burglaries are likely to occur. An interesting result of this study is that crimes spur other crimes nearby in space and time, much as earthquakes trigger local aftershocks sequences.

The focus of this work is the improvement and assessment of the nonparametric method of Marsan and Lengliné (2008) for estimating space-time Hawkes point process models of earthquake occurrences. Along these lines, our primary goals are,

1. The proposal of novel ways to incorporate a non-stationary background rate into the MISD algorithm.

2. Adding error bars to the histogram estimates of the triggering function which quantify the sampling variability and bias in the estimation of the underlying seismic process.

The original MISD algorithm assumes that the background rate for main-

shocks is a stationary Poisson process in time and space. While an estimate of the mean mainshock rate over an observation region is useful, the expansion of MISD to incorporate a non-stationary background component is an important next step and improvement by allowing for localized estimates mainshock activity related to variations in the underlying tectonic field and the locations major faults. Moreover, an estimate of a spatially varying background rate can be used to identify regions with a persistent and heightened incidence of large seismic events, independent of aftershock clustering features which diminish over time.

Nichols and Schoenberg (2014) proposed a way to adjust MISD to incorporate non-stationarities in the background process by initially kernel smoothing over all events in the catalogue and weighing each event by its corresponding kernel estimate. However, a main shortcoming of this approach is that non-stationaries in the background rate are only defined on the observed data and not at each pixel of the observation window. Moreover, the authors of this work were primarily interested in applying the method to evaluate the dependence between the magnitudes of earthquakes and their aftershocks, and the explicit assessment or validation of the proposed estimation technique was not addressed.

In this work, we propose two novel ways to incorporate a spatially varying background rate into the MISD method. First, we discuss a histogram estimator approach, which is a natural extension of the stationary rate estimator of Marsan and Lengliné (2008). Second, we apply the variable kernel estimator, used by Zhuang et al. (2002) for semi-parametric estimation, into the context of MISD. We validate and assess new methodology by simulating earthquake catalogues from a space-time model (ETAS) and evaluating the ability of each method to recover the true form of the non-stationary background rate and triggering function governing the simulations.

Simulation is also a powerful tool for understanding the statistical properties of the histogram estimators of the triggering function. By repeatedly simulating

and re-estimating an earthquake model error bars can be computed which capture the sampling distributions of the estimates. An easily calculated analytic approximation for the error bars found through simulation is discussed at the end of the chapter.

This chapter is organized as follows: In Section 3.1, we provide an overview of space-time point process models of seismicity. In Section 3.2, we describe our modified version of the MISD algorithm, and propose a couple new ways to incorporate a non-stationary background rate. In Section 3.3, we validate and assess our methods with simulation studies, and discuss boundary issues. In Section 3.4, we apply our method to an earthquake catalogue from Japan. In the Discussion Section we summarize and speculate about our results and suggest future directions and applications for this research.

3.1 Space-time Point Process Models in Seismology

Consider a marked space-time point process $N(t, x, y)$ representing the times, locations, and magnitudes, $\{(t_i, x_i, y_i, m_i) : i = 1 \cdots, N\}$, of earthquake occurrences. In seismology, one typically models the corresponding conditional intensity (1.1) as a Hawkes-type self-exciting point process taking the following form:

$$\begin{aligned} \lambda(t, x, y, m|H_t) &= J(m)\lambda(t, x, y|H_t) \\ \lambda(t, x, y|H_t) &= \mu(x, y) + \sum_{\{i:t_i<t\}} \nu(t - t_i, x - x_i, y - y_i; m_i). \end{aligned} \tag{3.1}$$

For example, models of this type, referred to as Epidemic Type Aftershock Sequences (ETAS) models, were introduced by Ogata (1988) for the description of earthquake catalogs. Such models categorize earthquake occurrences into two types: mainshocks and aftershocks. The rate of mainshocks occurring over a spatial region is modeled by the background intensity $\mu(x, y)$, which is assumed a non-stationary Poisson process in space and stationary in time. The rate of af-

tershock activity following an earthquake occurring at (t_i, x_i, y_i) with magnitude m_i is modeled by the triggering function ν, which is often assumed Gaussian or power-law in parametric models. The summation term gives the contribution of all previously occurring events in the catalog to the overall rate of seismicity at time t and location (x, y). The distribution of earthquake magnitudes $J(m)$ is typically assumed independent of all other model components, and follows an exponential distribution according to the well known magnitude frequency law of Gutenberg and Richter (1944). Note that model (3.1) specifies a space-time branching process since any earthquake occurrence (including an aftershock) is capable of triggering its own aftershock sequence.

Ogata (1998) considered many parameterizations of the response function of (3.1) which take the standard following form:

$$\nu(t - t_i, x - x_i, y - y_i; m_i) = \kappa(m_i)g(t - t_i)f(x - x_i, y - y_i; m_i). \tag{3.2}$$

Here $\kappa(m_i)$ is the magnitude productivity function, which gives the expected number of aftershocks following an earthquake of magnitude m_i. The temporal component g is a probability density function governing the rate of aftershocks following an earthquake at time t_i. The spatial component f is a probability density function for the spatial distribution of aftershocks occurring around an earthquake with epicenter (x_i, y_i). The dependence of the spatial response function on the magnitude m_i is built into some models.

One example of a parametrization of the triggering function for ETAS is given by:

$$\kappa(m) = Ae^{\alpha(m-m_c)}, \tag{3.3}$$

$$g(t) = (p-1)c^{(p-1)}(t+c)^{-p}, \tag{3.4}$$

$$f(x, y) = \frac{(q-1)d^{q-1}}{\pi}(x^2 + y^2 + d)^{-q}, \tag{3.5}$$

where m_c is the magnitude cut-off for the catalogue, $t > 0$, and (A, α, p, c, q, d) are parameters to be estimated. Here g corresponds to the modified Omori formula

(see Utsu et al. (1995) for details), and f is isotropic (rotation invariant) with a long range power-law decay rate.

The parameters of model (3.1) can be estimated by maximizing the log-likelihood function (1.4). The non-stationary background component, $\mu(x, y)$, is typically estimated with nonparametric techniques such as bi-cubic B-splines (Ogata, 1998) or kernel smoothing (Zhuang et al., 2002; Musmeci and Vere-Jones, 1992). The techniques for estimating $\mu(x, y)$ are often implemented in conjunction with a declustering algorithm used to isolate mainshocks.

Marsan and Lengliné (2008) proposed the MISD algorithm to nonparametrically estimate the triggering function ν and stationary background rate $\mu(x, y) = \mu$ for the space-time branching process model (3.1). Marsan and Lengliné (2010) showed that their method is an EM-type algorithm under the assumption that the background rate is stationary and the triggering function is piecewise constant. For the E-step, the branching structure of the process is estimated by computing the probabilities, for each pair (i, j) of earthquakes, of earthquake i having directly triggered earthquake j, as well as the probability of being a mainshock for each observed earthquake. For the M-step, the estimated branching structure is used to update an estimate of the stationary background rate and triggering function with probability weighted histogram estimators. The two-step procedure is repeated until the algorithm converges. A similar method is discussed in Mohler et al. (2011) using a Monte-Carlo based approach that alternates between sampling a realization of the estimated branching structure and updating estimates of the background rate and triggering function using kernel density estimation on the sampled data.

3.2 Nonparametric Methods

This section discusses the nonparametric method of Marsan and Lengliné (2008) to estimate the space-time Hawkes process model (3.1) using histogram estimators. We make the following modifications to the original algorithm:

1. We incorporate a non-stationary background rate;
2. We assume the separability of the triggering function into components for magnitude, time, and distance;
3. We perform histogram density estimation on the temporal and spatial triggering components $g(t)$ and $f(r)$, where $r = \sqrt{x^2 + y^2}$.

The above modifications make the method consistent with estimating the standard form of the triggering function in (3.2). As in Marsan and Lengliné (2008), we assume the spatial triggering component is isotropic, that is $f(x, y) = f(x^2+y^2)$; this means the rate of aftershock activity following an earthquake only depends on the distance r from the earthquake's epicenter and not direction (circular aftershock regions). Also, to be consistent with model (3.1), the background component $\mu(x, y)$ is assumed non-stationary in space and stationary in time.

3.2.1 Histogram Estimators

Let P be a $N \times N$ lower triangular probability matrix with entries,

$$p_{ij} = \begin{cases} \text{probability earthquake } i \text{ is an aftershock of } j, & i > j \\ \text{probability earthquake } i \text{ is a mainshock}, & i = j \\ 0, & i < j \end{cases} \tag{3.6}$$

$$P = \begin{bmatrix} p_{11} & 0 & 0 & \cdots & 0 \\ p_{21} & p_{22} & 0 & \cdots & 0 \\ p_{31} & p_{32} & p_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ p_{N1} & p_{N2} & p_{N3} & \cdots & p_{NN} \end{bmatrix} \qquad P^{(0)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1/2 & 1/2 & 0 & \cdots & 0 \\ 1/3 & 1/3 & 1/3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1/N & 1/N & 1/N & \cdots & 1/N \end{bmatrix}$$

The only constraint for matrix P is $\sum_{j=1}^{N} p_{ij} = 1$. The rows must sum to 1 since each earthquake in the branching process is either a mainshock or an aftershock of a previously occurring earthquake. $P^{(0)}$ is one possible initialization. For this matrix, $\sum_{i=1}^{N} p_{ii}$ can be interpreted as the estimated number of mainshocks, while $\sum_{i=1}^{N} \sum_{j=1}^{i-1} p_{ij}$ (sum of the non-diagonal elements) is the estimated number of aftershocks.

Below is the MISD algorithm of Marsan and Lengliné (2008) with the modifications specified in the beginning of this section. For the spatial component, we specify a histogram density estimator of $h(r) = 2\pi r f(r)$ since $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy = \int_{0}^{\infty} 2\pi r f(r)dr = 1$; here $h(r)$ represents the underlying probability density function for the distance r between an earthquake and its aftershock.

Algorithm 1.

1. Initialize $P^{(0)}$, set iteration index $v = 0$.

2. Estimate non-stationary background rate $\mu(x, y)$:

$$\mu_{k,l}^{(v)} = \frac{1}{T\Delta x \Delta y} \sum_{D_{k,l}} p_{ii}^{(v)}, \quad k = 0, \cdots, n_x^{bins} - 1;\ l = 0, \cdots, n_y^{bins} - 1.$$

3. Estimate triggering components $\kappa(m)$, $g(t)$, and $h(r)$:

$$\kappa_k^{(v)} = \frac{\sum_{A_k} p_{ij}^{(v)}}{N_k^{mag}}, \quad k = 0, \cdots, n_m^{bins} - 1;$$
$$g_k^{(v)} = \frac{\sum_{B_k} p_{ij}^{(v)}}{\Delta t_k \sum_{i=1}^{N} \sum_{j=1}^{i-1} p_{ij}^{(v)}}, \quad k = 0, \cdots, n_t^{bins} - 1;$$
$$h_k^{(v)} = \frac{\sum_{C_k} p_{ij}^{(v)}}{\Delta r_k \sum_{i=1}^{N} \sum_{j=1}^{i-1} p_{ij}^{(v)}}, \quad k = 0, \cdots, n_r^{bins} - 1.$$

4. Update probabilities $P^{(v+1)}$, letting r_{ij} be the epicentral distance between earthquakes i and j and $f^{(v)}(r_{ij}) = h^{(v)}(r_{ij})/(2\pi r_{ij})$:

$$p_{ij}^{(v+1)} = \frac{\kappa^{(v)}(m_j) g^{(v)}(t_i - t_j) f^{(v)}(r_{ij})}{\mu^{(v)}(x_i, y_i) + \sum_{j=1}^{i-1} \kappa^{(v)}(m_j) g^{(v)}(t_i - t_j) f^{(v)}(r_{ij})} \text{ for } i > j,$$
$$p_{ii}^{(v+1)} = \frac{\mu^{(v)}(x_i, y_i)}{\mu^{(v)}(x_i, y_i) + \sum_{j=1}^{i-1} \kappa^{(v)}(m_j) g^{(v)}(t_i - t_j) f^{(v)}(r_{ij})}.$$

5. If $\max_{i,j} |p_{ij}^{(v+1)} - p_{ij}^{(v)}| < \epsilon$, where $i \leq j$, then the algorithm has converged (in practice we take $\epsilon = 10^{-3}$). Otherwise, set $v \leftarrow v + 1$ and repeat steps 2–5 until convergence.

For step 2 of Algorithm 1, the notation is defined as follows:

- n_x^{bins} and n_y^{bins} are the number of bins along the x and y axis for the 2-dimensional histogram estimator of $\mu(x, y)$ ($n_x^{bins} \cdot n_y^{bins}$ bins total).
- $D_{k,l} = \{i : k\Delta x < x_i \leq (k+1)\Delta x, l\Delta y < y_i \leq (l+1)\Delta y\}$ where Δx and Δy are the fixed bin widths along the x and y axes.

For step 3 of Algorithm 1, the notation is defined as follows:

- n_m^{bins}, n_t^{bins}, and n_r^{bins} are the number of bins for the the histogram estimators of the magnitude κ, temporal g, and spatial h components of the triggering function.

- $A_k = \{(i,j) : \delta m_k < m_j \leq \delta m_{k+1}, i > j\}$ is the set of indices of all pairs of earthquakes whose mainshock magnitudes fall within the k^{th} bin $(\delta m_k, \delta m_{k+1}]$ of the histogram estimator for $\kappa(m)$, where $\Delta m_k = \delta m_{k+1} - \delta m_k$ is the bin width.

- $N_k^{mag} = \sum_{j=1}^{N} I(\delta m_k < m_j \leq \delta m_{k+1})$ is the number of earthquakes whose magnitudes fall within the interval $(\delta m_k, \delta m_{k+1}]$.

- $B_k = \{(i,j) | \delta t_k < t_i - t_j \leq \delta t_{k+1}, i > j\}$, is the set of indices of all pairs of earthquakes whose time differences fall within the k^{th} bin $(\delta t_k, \delta t_{k+1}]$ of the histogram estimator for $g(t)$, where $\Delta t_k = \delta t_{k+1} - \delta t_k$ is the bin width.

- $C_k = \{(i,j) | \delta r_k < r_{ij} \leq \delta r_{k+1}, i > j\}$ is the set of indices of all pairs of earthquakes whose epicentral distances r_{ij} fall within the k^{th} bin $(\delta r_k, \delta r_{k+1}]$ of the histogram estimator for $h(r)$, where $\Delta r_k = \delta r_{k+1} - \delta r_k$ is the bin width.

In step 2 of Algorithm 1 the non-stationary background rate is estimated with a histogram estimator which is a generalization of the stationary estimator in the original MISD algorithm. In our modified method, the spatial observation window S is partitioned into equally sized cells of width Δx and height Δy. The estimated rate within each cell is given by the sum of the background probabilities, p_{ii}, corresponding to earthquakes occurring within that cell, and then dividing the sum by $\Delta x \cdot \Delta y \cdot T$ to give the rate of mainshocks per unit area per unit time. Note, the histogram estimator in step 2 reduces to the stationary case in Marsan and Lengliné (2008) when $n_x^{bins} = n_y^{bins} = 1$ and $\Delta x \cdot \Delta y = S$ (i.e. only one cell equal to the spatial observation window is specified). Also note that the estimator of g is itself a density since $\sum_{k=0}^{n_t^{bins}-1} \Delta t_k \hat{g}_k = 1$, and similarly for the histogram estimator of h.

The assumption of separability allows for robust computation of model components by substantially reducing the number of bins needed to estimate the model

(only a one-dimensional support is needed for the histogram estimator of each triggering component). Furthermore, since we perform histogram density estimation on g and f the output of Algorithm 1 has meaningful interpretation as in Ogata (1998). For instance, the histogram estimate of the magnitude productivity $\hat{\kappa}_k(m)$ has the natural interpretation as the estimated mean number of aftershocks directly triggered by an earthquake with magnitude m falling in the k^{th} magnitude bin $(\delta m_k, \delta m_{k+1}]$.

3.2.2 Variable Kernel Estimation

A shortcoming of the histogram method for estimating the background rate in Algorithm 1 is the implicit assumption of constancy within each bin. If a large mainshock occurs, then the contribution of that event to the background seismicity is limited to the bin in which the event is contained. If a bin does not contain any earthquake events, then the estimated rate of mainshocks in that bin is zero. Hence, the method does not allow for the estimate to vary smoothly over the spatial observation region and is highly dependent on the choice of the partition. This motivates considering a kernel smoothing approach, where the background rate estimate only depends on the choice of the smoothing parameter (bandwidth) and varies continuously over the pixels in the spatial window.

As an alternative to the histogram approach (Algorithm 1, step 2) for estimating the non-stationary background rate, we adopt the variable bandwidth kernel estimator used by Zhuang et al. (2002):

$$\mu(x, y) = \gamma\tau(x, y), \tag{3.7}$$

$$\tau(x, y) = \frac{1}{T}\sum_{i=1}^{N} p_{ii} k_{d_i}(x - x_i, y - y_i). \tag{3.8}$$

Here the index i runs through all the events in the catalogue, γ is a scaling factor,

and k is the Gaussian kernel function,

$$k_{d_i}(x,y) = \frac{1}{2\pi d_i^2} \exp\left(-\frac{x^2+y^2}{2d_i^2}\right).$$

The kernel is weighted by p_{ii}, the probability that event i is a mainshock, and has a varying bandwidth d_i specified for each event in the catalogue. The bandwidth d_i is computed by finding the radius of the smallest disk centered at (x_i, y_i) that contains at least n_p other events, and is greater than some small value ϵ representing the location error. Zhuang et al. (2002) suggest taking n_p between 15–100 and $\epsilon = 0.02$ degrees. A variable bandwidith estimate is preferred since a large fixed bandwidth over-smooths areas with clustered events, and a small fixed bandwidth under-smooths areas with sparsely located events.

In Zhuang et al. (2002) the estimate (3.7) is part of a semi-parametric model for ETAS, with parameters estimated via maximum likelihood. Since our approach is completely nonparametric, the scaling factor γ for the estimate of the background rate needs to be carefully defined. This leads to the following algorithm for estimating the space-time Hawkes process model (3.1) with a variable kernel estimator for the background seismicity:

Algorithm 2.

1. Initialize $P^{(0)}$ and compute d_i for each event $i = 1, \cdots, N$.

2. Estimate non-stationary background rate $\mu(x,y)$:

$$\mu^{(v)}(x,y) = \frac{\sum_{i=1}^{N} p_{ii}^{(v)}}{Z^{(v)}} \tau^{(v)}(x,y).$$

3. Follow Steps 3–5 in Algorithm 1.

The normalizing factor $Z^{(v)}$ at iteration v is chosen so that

$$\frac{1}{Z^{(v)}} \int_0^T \int\int_S \tau^{(v)}(x,y)dxdydt = 1,$$

and consequently,

$$\int_0^T \int\int_S \mu^{(v)}(x,y)dxdydt = \sum_{i=1}^{N} p_{ii}^{(v)},$$

where $\sum_i p_{ii}^{(v)}$ is the estimated number of mainshocks occurring in the space-time observation window. In practice, $Z^{(v)}$ can be found by first computing $\tau^{(v)}(x,y)$ as defined in (3.8) at each pixel, and then evaluating the integral of $\tau^{(v)}(x,y)$ over $S \times [0,T]$ with a Reimann sum over those pixels.

3.3 Simulation Results

3.3.1 Histogram Estimator Method

In this section we assess the performance of the nonparametric method described in Algorithm 1 to recover an earthquake model from synthetic catalogues. For this study, earthquake occurrences are simulated from the ETAS model with parametric triggering function given by (3.3, 3.4, 3.5). The parameter values are the maximum likelihood estimates $(A, \alpha, p, c, d, q) = (0.322,\ 1.407,\ 1.121,\ 0.0353,\ 0.0159,\ 1.531)$ from Table 2, row 8 of Ogata (1998) (parameters estimated from earthquake data over a $36 \sim 42°$N latitude and $141 \sim 145°$E longitude region off the east coast of Tohoku District, Japan with time span 1926–1995). Earthquake magnitudes are generated independently of other model components according to an exponential density $J(m) = \beta e^{-\beta(m-m_c)}$ with $\beta = ln(10)$ (equivalent to a Gutenberg-Richter b-value equal to 1). The observation window for the simulation is $S \times T = [0,4] \times [0,6] \times [0,25000]$, and the magnitude cut-off is $m_c = 0$. The non-stationary background rate is specified by partitioning the spatial observation window S into 4 equally sized cells with the varying rates shown in Figure 3.2(a). An example of a simulated realization is shown in Figure 3.1. For a description of the simulation procedure for ETAS please see Algorithm C from Zhuang et al. (2004).

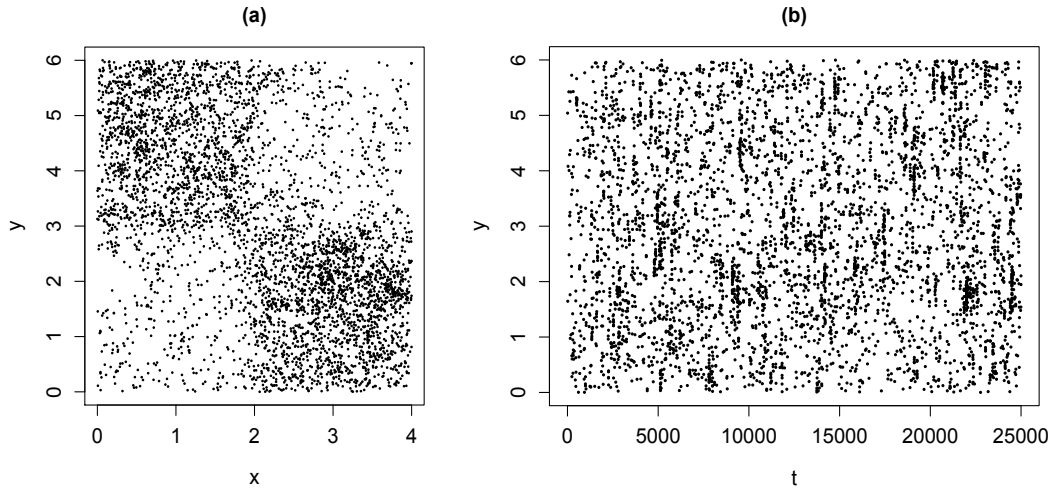

Figure 3.1. Simulated realization of ETAS model (3.3–3.5) with background rate varying in each quadrant; (a) epicentral locations, and (b) space-time plot of simulated earthquakes.

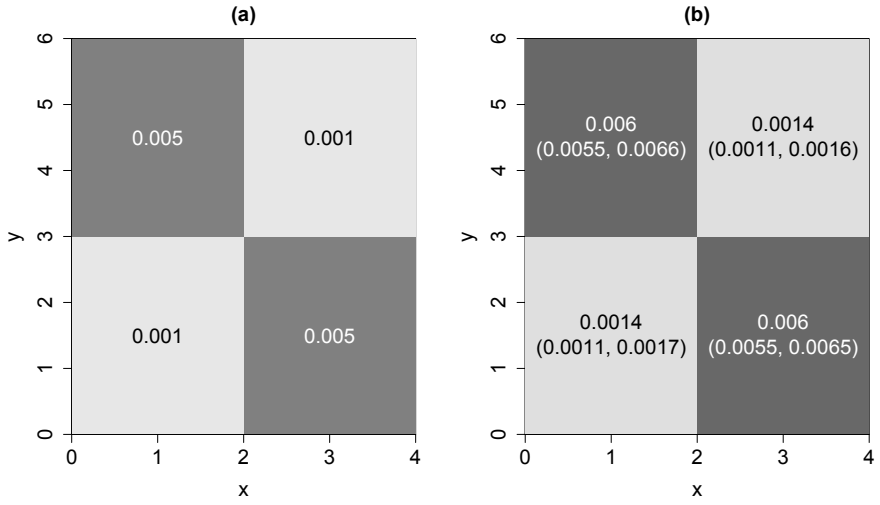

Figure 3.2. (a) True background rate for simulation study in Section 3.3.1. (b) Results for estimating the background rate with Algorithm 1 from 200 simulations of ETAS. The means of the estimates printed in each cell correspond to the grey scale levels; the intervals are the 0.025 and 0.975 quantiles for the estimates in each cell.

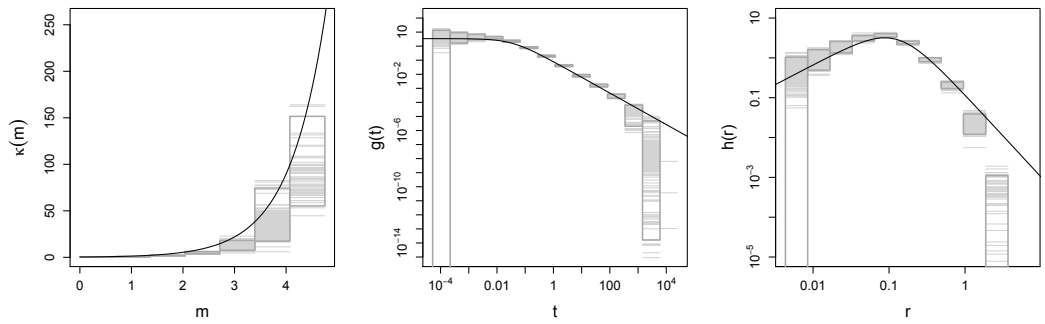

Figure 3.3. Magnitude, temporal, and distance components for triggering function from the simulation study in Section 3.3.1. The black solid lines are the true model components from which the data is generated. The light grey horizontal lines in each bin are the histogram estimates from the 200 simulations of ETAS; the solid grey boxes are the 95% coverage intervals (error bars) for the estimates in each bin (i.e. pointwise 0.025 and 0.975 quantiles).

We simulate and re-estimate the ETAS model 200 times to assess the variability in the estimates over multiple realizations of earthquake catalogues from the specified model. The results for the estimation of the non-stationary background rate are shown in Figure 3.2(b). Each cell in Figure 3.2(b) shows the 0.025 and 0.975 quantiles of the estimates, and the mean of the estimates which correspond to the cell's grey scale level. Figure 3.2(b) reveals that the nonparametric method (Algorithm 1) is able to recover the sharp differences between the rates in each cell with reasonably small errors. While the means of the estimates are close to the true rates (Figure 3.2(a)), a bias is apparent, as the 95% coverage intervals fall consistently above the true values. In the next section we show that this over-estimation is due to boundary effects induced by excluding aftershocks that occur outside the space-time observation window.

The results for the estimation of the magnitude, temporal, and distance components of the triggering function from the 200 realizations of ETAS are shown in Figure 3.3. The histogram density estimates of $g(t)$ and $h(r)$ are plotted on log-log

scales with equally spaced logarithmic bins since the true model components are power-law. The bins for the histogram estimator of the magnitude productivity $\kappa(m)$ are also equally spaced.

The method is able to recover the form of each component of the triggering function since the true value governing the simulations is contained within the 95% coverage error bars for most bins. The error bars for the estimates of $g(t)$ and $h(r)$ reveal that the estimation is most accurate in the middle range. The high variability in the estimates for bins corresponding to small time differences t and distances r is not surprising since the partition is logarithmically scaled, and therefore these bin widths are very small. The error bars at the right-tail ends of the distributions of $g(t)$ and $h(r)$ do not cross and underestimate the true densities. In the next section we show that this estimation bias is due to boundary issues. The error bars for the estimates of the magnitude productivity function $\kappa(m)$ increase with magnitude, although this is expected since earthquake magnitudes are exponentially distributed and therefore only a few large magnitude events occur in each simulation.

3.3.2 Boundary Issues

When simulating earthquake catalogues from the ETAS model the mainshocks are restricted to occur within the space-time observation window $S \times [0, T]$. However, the times and locations of aftershocks, simulated from the triggering function components g and f, may occur outside of this boundary. In the last section, we neglected boundary effects, and only used simulated data occurring within the space-time observation window to estimate the model using Algorithm 1.

To evaluate the boundary effects on the estimation we include simulated aftershocks which occur within a distance ϵ_r of the spatial boundary and a time ϵ_t of the temporal boundary, i.e. all aftershocks occurring within $[-\epsilon_r, 6 + \epsilon_r] \times$

$[-\epsilon_r, 4+\epsilon_r] \times [0, 25000+\epsilon_t]$. We then run Algorithm 1 on the expanded simulation data, and slightly modify step 4 so that $\mu(x_i, y_i) = 0$ if event (t_i, x_i, y_i, m_i) falls outside of $S \times [0, T]$.

To measure the change in performance of Algorithm 1 on estimating the non-stationary background rate as we increase ϵ_r and ϵ_t we use the root-mean-square deviation (RMSD):

$$\sqrt{\frac{1}{n_x^{bins} n_y^{bins}} \sum_{i,j} (\hat{\mu}_{ij} - \mu_{ij})^2}. \tag{3.9}$$

Here $\hat{\mu}_{ij}$ and μ_{ij} are the estimate and true value for the background rate in the (i, j) cell respectively. We simulate the ETAS model 10 times using the same parameters and background rate as in Section 3.3.1, with mainshocks again restricted to $S \times [0, T] = [0, 4] \times [0, 6] \times [0, 25000]$, but aftershocks allowed to occur outside that region. For each simulation, the RMSD is computed for increasing values of ϵ_r and ϵ_t. Figure 3.4 shows the mean RMSD from the 10 realizations at selected values of ϵ_r and ϵ_t; the vertical lines represent a standard deviation in RMSD above and below the mean. The incorporation of aftershocks falling outside the space-time observation window significantly improves the performance of the estimation of the background rate. The RMSD appears to level off when $\epsilon_r = 10^{0.5} = 3.16$ and $\epsilon_t = 10000$.

Figure 3.5 shows the results from simulating and re-estimating ETAS with Algorithm 1 200 times with a boundary correction of $\epsilon_r = 1000$ and $\epsilon_t = 10^6$. Again, we simulate events with the same parameters and space-time window as Section 3.3.1. The only difference is that in the estimation we use aftershocks occurring within a distance $\epsilon_r = 1000$ and time $\epsilon_t = 10^6$ of the boundary of the observation window. Since the temporal and distance components of the triggering function used to generate the data are power-law it is possible for aftershocks to occur at very far distances and times from the observation window.

The background rate estimate in Figure 3.5 is a substantial improvement over

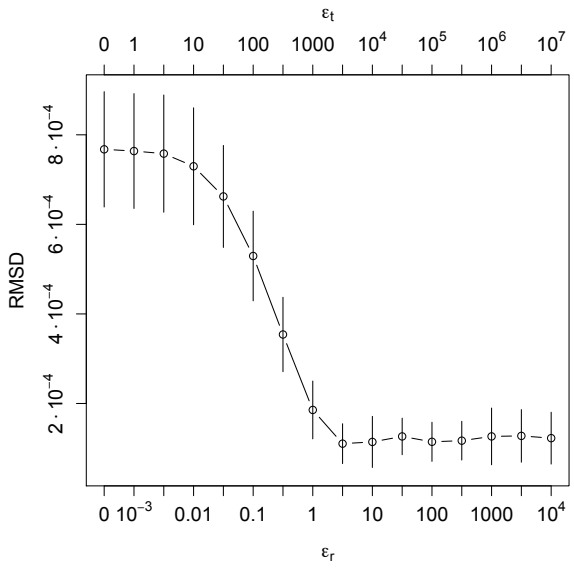

Figure 3.4. RMSD of the background rate, equation (3.9), for increasing values of ϵ_r and ϵ_t. RMSDs are averaged from 10 realizations of ETAS; the vertical bars cover one standard deviation above and below the mean.

the estimate in Figure 3.2, which neglected boundary effects. The true rates are contained in the 95% coverage intervals for each cell in Figure 3.5. Moreover, the consistent over-prediction of the rates evident in Figure 3.2 is no longer present, and the results suggest the bias in the cell means is negligible once the boundary effects are accounted for. The error bars for the triggering function components in Figure 3.5 also show substantial improvement when compared to Figure 3.3. The histogram estimates for $g(t)$ and $h(r)$ contain the true density values (solid line) for large time differences t and distances r. Moreover, accounting for boundary effects expands the reach of the estimation (histogram estimates at bins beyond $r = 10$ and $t = 25000$) and reduces the error at the tail ends. Lastly, the error bars for the estimates of the magnitude productivity $\kappa(m)$ in Figure 3.5 appear more centered around the true value than in Figure 3.3.

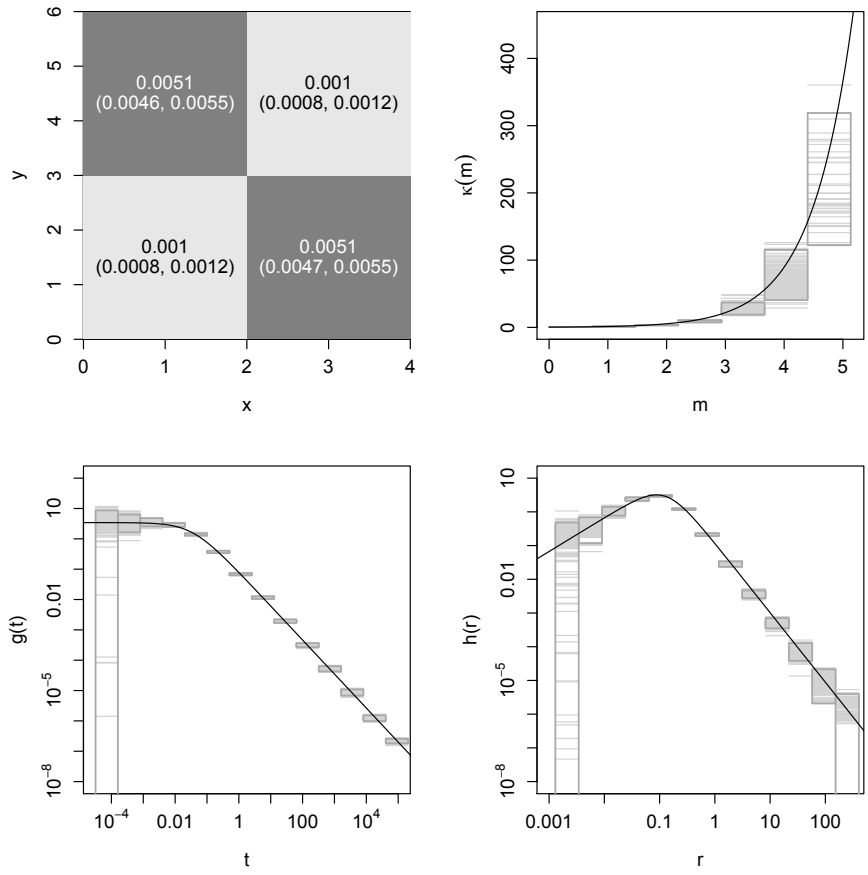

Figure 3.5. Estimates of the background rate and triggering function components from 200 ETAS simulations, with boundary correction for aftershock activity $\epsilon_r = 1000$ and $\epsilon_t = 10^6$.

3.3.3 Variable Kernel Estimation Method

In this section we use simulation to assess the ability of Algorithm 2 to recover the components of the space-time Hawkes process model (3.1) with a smoothly varying background rate. Here we simulate from a parametric ETAS model with the same triggering function and parameter values as Section 3.3.1. However, instead of the background rate in Figure 3.2(a) with stationary rates in each cell on a 2×2 grid, we simulate from the smoother background rate shown in Figure 3.7(a). This non-stationary background rate was generated by performing fixed bandwidth kernel density estimation over the locations of 883 earthquakes of magnitude 5.0 or greater, longitude $141 \sim 145°$E, latitude $36 \sim 42°$N, and time between 16 Jan 2007 to 28 Dec 2014.[1] To simulate from the kernel smoothed background rate in Figure 3.7(a) we use the thinning procedure of Lewis and Shedler (1979) and set the expected number of background events equal to 2000.

Figure 3.7(b) shows the probability weighted variable kernel estimate (Algorithm 2, step 2) of the non-stationary background rate from a single simulated realization of the ETAS model. The epicentral location and space-time plots of the simulated earthquake data used for this estimate are shown in Figure 3.6. The kernel estimate of the background rate depends on the smoothing parameter n_p (Section 3.2.2). Here we choose $n_p = 50$, since this value gives the lowest RMSD (3.9) for $n_p \in \{10, 15, \cdots, 95, 100\}$. The kernel estimates are evaluated on a 100×100 pixel grid (making $n_x^{bins} = n_y^{bins} = 100$ when evaluating (3.9)).

As discussed in Section 3.2.2, the nonparametric estimation of ETAS is sensitive to boundary effects. As a boundary correction for the estimation with Algorithm 2, we allow for aftershocks occurring within $\epsilon_r = 3$ degrees and $\epsilon_t = 3000$ days of the space-time boundary $S \times [0, T] = [0, 4] \times [0, 6] \times [0, 25000]$. Note that the selected values, $\epsilon_r = 3$ and $\epsilon_t = 3000$, correspond to where the RMSD in Fig-

[1] Data gathered from `http://www.quake.geo.berkeley.edu/anss/catalog-search.html` with spatial observation window the same as Ogata (1998).

ure 3.4 begins to level off. The panels in Figure 3.6 show the boundary (dashed rectangles) and simulated aftershocks occurring in the specified region outside the boundary.

This estimate in Figure 3.7(b) resembles the overall form of the true background intensity Figure 3.7(a) and recovers many of the mainshock hotspots. However, near location (2.06, 2.33), a hotspot appears to have been erroneously estimated, i.e. a false positive has been identified. This is due to the large magnitude event ($m > 4$) that occurred in the simulation at this location, as denoted by the asterisk in Figure 3.6(a). The mean of 200 estimates of the background rate from 200 simulated realizations of ETAS is shown in Figure 3.7(c), and appears to closely resemble the true background rate. Hence, while there may be discrepancies for estimates from a single realization due to sampling variation, the variable kernel estimator appears to be unbiased since the mean of the estimates from repeated simulation is close to the true background intensity. Moreover, the pointwise 0.025 and 0.975 quantiles for 200 estimates of the number of background events, given by $(1943.7, 2219.4)$, contains the true value of 2000 background events specified for the simulation.

The histogram estimates and corresponding 95% coverage error bars in Figure 3.8 appear to successfully describe the true triggering components. This demonstrates the ability of Algorithm 2 to recover the non-stationary background rate with a variable kernel estimator and triggering function with histogram estimators. There are slight discrepancies between the histogram estimates of the triggering function and the true values due to boundary effects. Most noticeably, the 95% coverage error bars for the estimates of $g(t)$ and $h(r)$ do not contain the true density values at the right-tail ends of the distributions. Boundary correction values larger than $\epsilon_r = 3$ and $\epsilon_t = 3000$ may result in more accurate estimates, as in the asymptotic case shown in Figure 3.5. However, the selected values seem sufficient for estimating the background intensity.

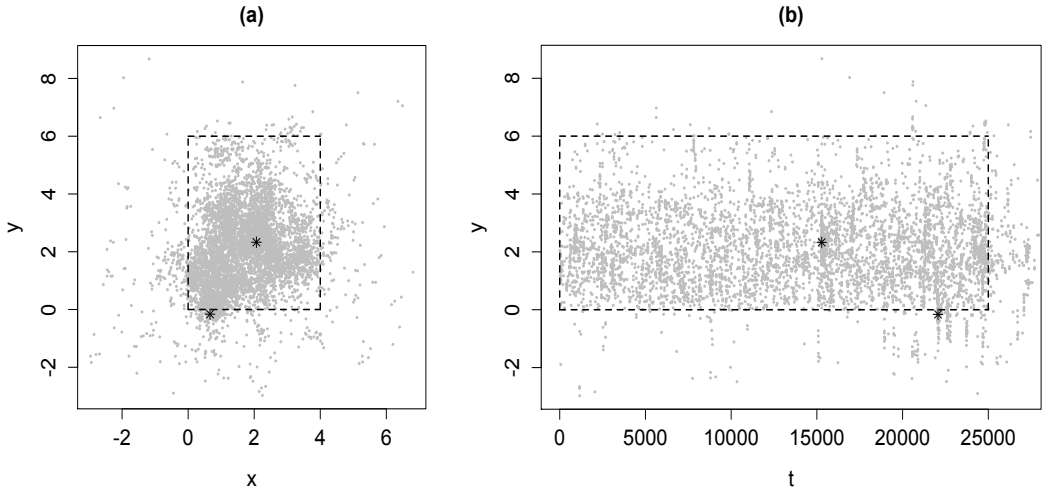

Figure 3.6. Simulated realization of ETAS model (3.3–3.5) with smooth non-stationary background rate; (a) epicentral locations, and (b) space-time plot of simulated earthquakes. The dotted rectangles in each plot are the spatial and temporal boundaries for the observation window $S \times [0, T] = [0, 4] \times [0, 6] \times [0, 25000]$. Aftershocks occurring within a distance $\epsilon_r = 3$ and time $\epsilon_t = 3000$ of the boundary are plotted outside the rectangle. The asterisks denote events with magnitudes $m > 4$.

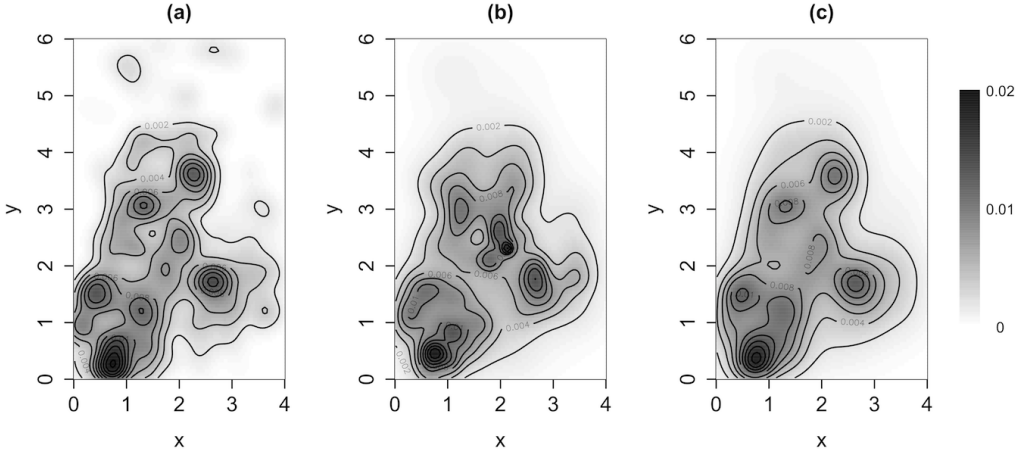

Figure 3.7. (a) True background rate for simulation study in Section 3.3.2. (b) Estimate of background rate from one simulated realization of ETAS and, (c) mean estimate from 200 realizations.

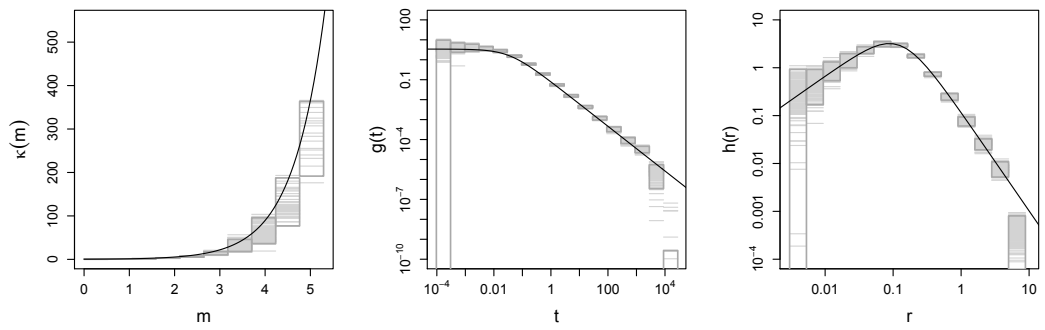

Figure 3.8. Magnitude, temporal, and distance components for triggering function from the simulation study in Section 3.3.3. The black solid lines are the true model components from which the data is generated. The light grey horizontal lines in each bin are the histogram estimates from the 200 simulations; the solid grey boxes are the 95% coverage error bars for the estimates.

3.4 Application to Japan Dataset

We apply the MISD algorithm (Algorithm 2) to earthquake data from the ANSS catalogue `http://www.quake.geo.berkeley.edu/anss/catalog-search.html`. The dataset contains 6075 earthquakes of magnitude 4.0 or greater occurring over a 10 year period between 5 Jan 2005 – 31 Dec 2014. The spatial widow is a 141 $\sim$ 145°E longitude and 36 $\sim$ 42°N latitude region off the east coast of the Tohoku District in northern Japan. This is the same spatial region analyzed in Ogata (1998), although the time window in this study is different. An epicentral and space-time plot of the data is show in Figure 3.9, with the asterisk corresponding to the 2011 magnitude 9.0 Tohoku earthquake.

The variable kernel estimate of the background rate (Algorithm 2, step 2) is shown in Figure 3.10. Here we chose the smoothing parameter $n_p = 50$, corresponding to the best choice for the simulation study in Section 3.3.3. Figure 3.10 is an important plot for assessing seismic risk since it shows the estimate of the underlying spatial Poisson processes $\mu(x, y)$ for maishock activity which persists

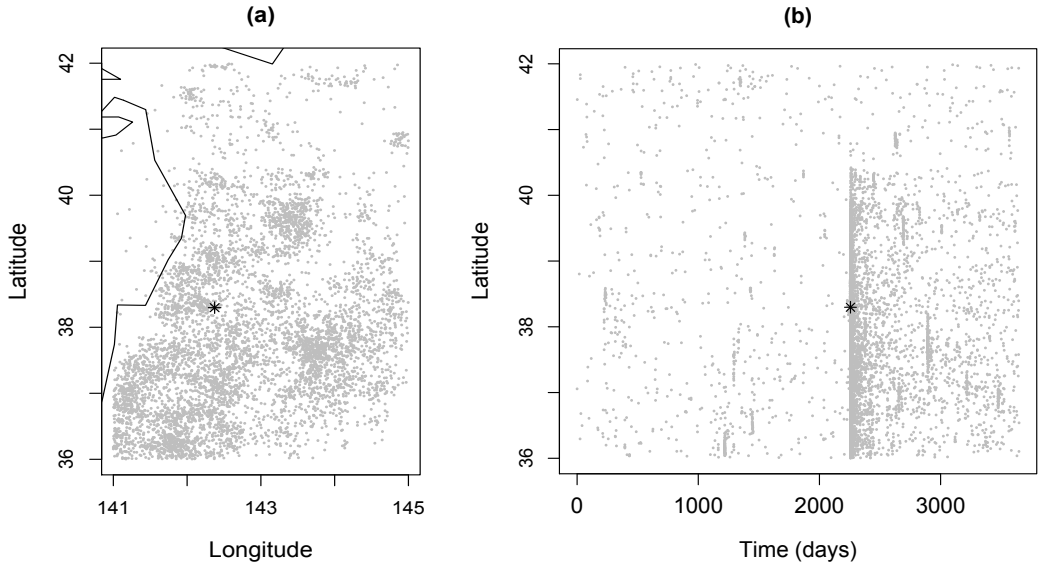

Figure 3.9. Epicentral locations (a) and space-time plot (b) of earthquakes, magnitude 4.0 or greater, occurring off the east coast of the Tohoku District, Japan. The asterisk corresponds to the 2011 Tohoku earthquake of magnitude 9.0.

over time in the region. In total, the algorithm estimated there to have been 809 mainshocks, or 13.3% of the total seismicity; this suggests that most of the events in the dataset are aftershocks, temporally and spatially linked to previously occurring earthquakes.

The histogram estimates of the components of the triggering function are shown in Figure 3.11. The grey error bars approximate ± 2 standard errors, capturing the sampling variation for the histogram estimates in each bin (see Appendix B for the derivation of the analytic standard errors). The estimates of $g(t)$ and $h(r)$ both exhibit power-law type distributions, and the error bars appear similar to the ones obtained in the simulation study (Section 3.3.3). Note, the estimates at the right-tail ends of these distributions ($t > 1000$ days and $r > 1$ degree) are perhaps unreliable and underestimate the truth due to boundary effects, as demonstrated in the simulation study (Section 3.3.2).

The estimate of the magnitude productivity function $\kappa(m)$ appears to follow an exponential form. The error in the estimation of the productivity increases with magnitude, as also demonstrated in the simulation study (Section 3.3.3). In the dataset there are only 3 events of magnitude 7.4 or greater, and hence large sampling variation for the estimates of the mean productivity for large magnitude events. The estimate in the last bin was estimated with only one event, namely the magnitude 9.0 Tohoku earthquake. It appears that the magnitude productivity for this event is underestimated; perhaps this is due to boundary effects since many of the aftershocks may have occurred outside the observation window.

Superimposed on Figure 3.11 are the parametric estimates of the ETAS model (3.3, 3.4, 3.5) for this same region from Table 2, row 11 of Ogata (1998). Amazingly, the parametric and nonparametric estimates agree closely. This suggests that seismicity in this region is well captured by an ETAS model with power-law $g(t)$ and $f(r)$, and exponential $\kappa(m)$. Since our dataset was gathered over a different time window than Ogata (1998), the results also suggest that properties of aftershock sequences in this region are rather invariant over time.

Note that in Figure 3.11 the nonparametric estimate of the triggering density $g(t)$ is slightly higher than what Ogata previously estimated for small time intervals t . This could perhaps be attributable to increased accuracy of seismometers in this region detecting aftershocks occurring shortly after large earthquakes more accurately than previously.

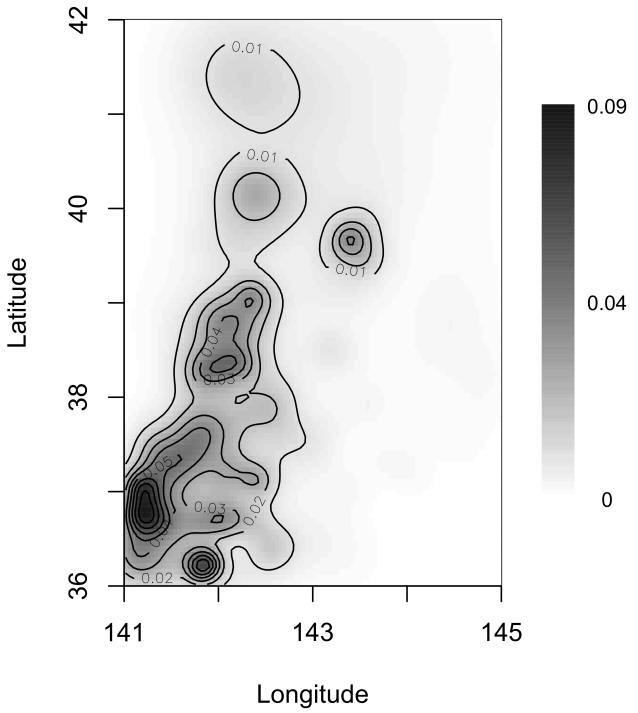

Figure 3.10. Estimate of background rate (Algorithm 2, step 2) for Japan earthquake dataset (Section 3.4). Rate values are in $events/day/degree^2$.

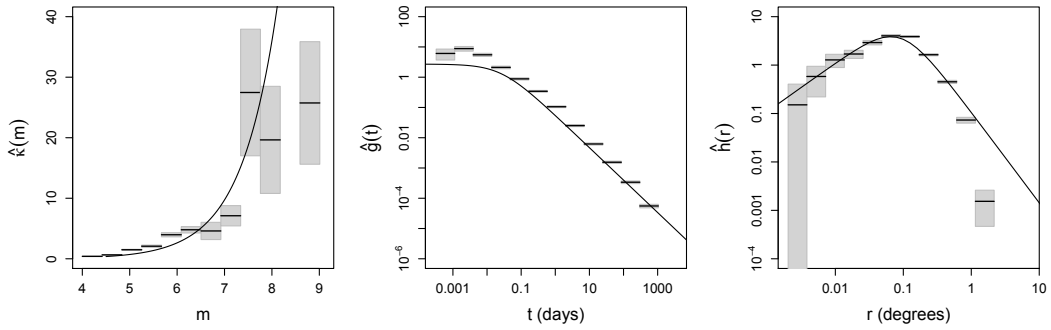

Figure 3.11. Magnitude, temporal, and distance components for triggering function estimated from the Japan earthquake dataset (Section 3.4). The black solid horizontal lines are the estimates in each bin. The grey boxes are the error bars covering ±2 standard errors. The solid black curves are the parametric estimates from Ogata (1998) in the same region.

3.5 Discussion

The results of this article demonstrate that the MISD algorithm performs remarkably well at nonparametrically estimating space-time Hawkes process models (3.1) for earthquake occurrences. By repeatedly simulating and re-estimating a known parametric earthquake model (ETAS), we verified and evaluated novel ways to incorporate a non-stationary background rate into the method. Moreover, the error bars added to the histogram estimates of the triggering function captured the true values and showed reasonable sampling variation in the estimates over most bins. While the simulation results showed bias in the estimation of the background rate and right-tail ends of the triggering function components, this problem became noticeably less severe once boundary effects were taken into account.

A striking result in the application to earthquake data from the Tohoku region in Japan is that the nonparametric estimate matched closely with a previously estimated parametric form of the ETAS model. This further justifies the ETAS model as an adequate model of seismicity for the selected region in Japan.

The parametric forms for point process models in seismology are the result of many decades of refinement. However, for any given seismic region, a multitude of different parameterizations of ETAS may be considered. The nonparametric methods discussed in this chapter can serve as a diagnostic to assess which parametrization is a good fit to the data. In other applications of self-exciting point processes, such as crime or finance, there is a less established literature on parametric models. In such applications, nonparametric estimation can be a powerful exploratory tool in determining a suitable parameterization of the triggering function. The error bars on the histograms estimates can be used for statistical inference, and to identify places where the nonparametric estimate is more or less reliable as either a diagnostic or exploratory tool.

CHAPTER 4

Future Directions

4.1 Point Process Models and Inference for E-mail Networks

One future direction for this research is to consider different types of point process models to better account for the observed clustering in e-mail traffic. For instance, a completely nonparameteric approach, as in Marsan and Lengliné (2008), would allow for a more flexible and data-driven estimation of the Hawkes process models. Moreover, the applications of such methods may suggest different types of parameterizations for the triggering function than the exponential forms considered in Chapter 2, and a background rate estimate which incorporates more effects than the observed hourly and daily periodicities. Also of interest are other types of parametric point process models, besides the Hawkes process, such as the Cox multiplicative intensity model considered in Perry and Wolfe (2013), which can be used to model dyadic and triadic effects, and homophily in e-mail network activity. Another possibility for future work is using the subject lines of e-mails to verify how well the latent branching structure of discussion chains are detected with the EM-type algorithm. Lastly, beyond looking at the temporal statistics and a point process analysis of e-mail communication networks, one may also consider using techniques from social network analysis and machine learning to help build predictors of network leadership using the content of e-mails or texts. Ultimately, through continuing with such research, we hope to improve methods

for inferring the leadership and hierarchy of criminal or terrorist organizations from communication patterns.

4.2 Nonparametric Methods for Point Processes

The simulation study in Chapter 3 demonstrated the ability of the considered nonparametric methods to recover the components of a space-time Hawkes process model. One future direction is to further analyze the statistical properties of the Marsan-Lengliné histogram estimators of the triggering function, and the non-stationary background rate estimators from Section 3.2. Such properties as asymptotic unbiasedness and consistency can be analyzed through simulation or more rigorous analytic means. Also of interest are considering ways to optimally select bin widths for equally spaced or adaptive partitions.

A common assumption for ETAS is that the spatial triggering component is isotropic, i.e. $f(x, y) = f(r)$. However, previous studies have shown that earthquake aftershock sequences tend to cluster around faults and have more elliptical shapes (Utsu, 1970; Wong and Schoenberg, 2009). One possible application of MISD is to estimate a spatial triggering component with form $f(r, \theta)$, where θ is the angle to the causative earthquake's fault plane. A nonparametric fit which incorporates directionality can be used as a diagnostic or exploratory tool for anisotropic extensions of ETAS.

Lastly, a nonparametrically fit model can be applied towards forecasting earthquake occurrences. The Collaboratory for the Study of Earthquake Predictability (CSEP) provides a framework for evaluating forecasting performance and comparing different models of seismicity.

APPENDIX A

E-mail Network Simulation Algorithm

In this appendix we describe a procedure for simulating IkeNet e-mail network activity using the estimated Hawkes process models. We start by simulating the background events, or non-reply e-mails, sent by each officer i over $[0, T]$. For models (2.2) and (2.5) this can be done using the method of Poisson thinning (Lewis and Shedler, 1979) described in the following algorithm:

Algorithm A

Step 1. Let μ^* be the maximum of $\hat{\mu}(t)$ over $[0, T]$.

Step 2. Draw N_b^* from $Pois(\hat{v}_i \mu^* T)$ (this is an upper bound on the number of background or non-reply e-mails for network member i).

Step 3. Draw an i.i.d. sample $\{Z_l : l = 1, \cdots, N_b^*\}$ from Unif(0,1) and set $s_l^i = T \cdot Z_l$.

Step 4. For each event $l = 1, \cdots, N_b^*$ at time s_l^i, retain that event within our simulated background set with probability $p_l = \hat{\mu}(s_l^i)/\mu^*$, otherwise remove it from our background set.

Step 5. Let $N_i^{send}(0)$ denote the number of events selected in step 4 and $G_i^{send}(0) = \{s_k^i : k = 1, \cdots, N_i^{send}(0)\}$ be the set of event times selected in step 4, which we will refer to as generation 0.

Step 6. Choose receivers for the events in $G_i^{send}(0)$ by drawing a sample of size $N_i^{send}(0)$ with replacement from the set $\{j : j \in \{1, \cdots, 22\}, j \neq i\}$ with

corresponding weights $\{N_{ij}^{send} : j \in \{1, \cdots, 22\}, j \neq i\}$, where N_{ij}^{send} is the observed number messages sent from i to j.

In order to generate all the non-reply e-mails sent in the entire network Algorithm A is repeated for each officer $i = 1, \cdots, 22$. To simulate the background process (non-reply e-mail send times) for model (2.1) we simply simulate a stationary Poisson process with rate $\hat{\mu}_i$ for each officer, and the receivers of e-mails are selected the same way as in Algorithm A.

After laying down the background events (non-reply e-mails) we simulate the reply e-mails. Let $G_i^{rec}(v) = \{r_k^i : k = 1, \cdots, N_i^{rec}(v)\}$ be the set of times when i received e-mails during generation v and $N_i^{rec}(v)$ be the number of simulated messages i received during generation v. Each message $r_k^i \in G_i^{rec}(v)$ received by officer i at generation v triggers reply messages on $(r_k^i, T]$ according to the non-stationary Poisson process $\hat{g}_i(t - r_k^i) = \hat{\theta}_i \hat{\omega}_i e^{-\hat{\omega}_i (t - r_k^i)}$. To generate these reply times for each officer i, using models (2.1) and (2.2), we apply the following algorithm (Lewis and Shedler, 1979):

Algorithm B

Step 1. Set $k = 1$ and $\eta = 0$.

Step 2. Draw $n_k^{(v+1)}$ from $Pois(\hat{\theta}_i)$, this is the number of reply messages i sends in response to receiving message $r_k^i \in G_i^{rec}(v)$ in the previous generation v.

Step 3. If $n_k^{(v+1)} = 0$ there are no replies and go to step (5), otherwise draw an i.i.d sample $\{Z_l : l = \eta + 1, \cdots, \eta + n_k^{(v+1)}\}$ from Unif(0,1).

Step 4. The reply times $\{s_l^i : l = \eta + 1, \cdots, \eta + n_k^{(v+1)}\}$ for message $r_k^i \in G_i^{rec}(v)$ are given by:

$$Z_l = \frac{1}{\hat{\theta}_i} \int_{r_k^i}^{s_l^i} \hat{g}_i(t - r_k^i) dt \implies s_l^i = \frac{ln(1 - Z_l)}{-\hat{\omega}_i} + r_k^i.$$

Step 5. Update $\eta \leftarrow \eta + n_k^{(v+1)}$ and $k \leftarrow k+1$.

Step 6. Repeat steps (2) – (5) until $k = N_i^{rec}(v) + 1$.

Step 7. Let $N_i^{send}(v+1) = \sum_{k=1}^{N_i^{rec}(v)} n_k^{(v+1)}$ denote the number of simulated e-mails sent by officer i in generation $v+1$ and $G_i^{send}(v+1) = \{s_l^i : l = 1, \cdots, N_i^{send}(v+1)\}$ be the corresponding set of times when officer i replies to messages sent during the previous generation v.

Step 8. Choose receivers for the events in $G_i^{send}(v+1)$ by drawing a sample of size $N_i^{send}(v+1)$ with replacement from the set $\{j : j \in \{1, \cdots, 22\},\ j \neq i\}$ with corresponding weights $\{N_{ij}^{send} : j \in \{1, \cdots, 22\},\ j \neq i\}$, where N_{ij}^{send} is the observed number messages sent from i to j.

Algorithm B is repeated for each officer $i = 1, \cdots, 22$ to generate all reply e-mails at generation v. Algorithm B is applied to each generation $v \geq 1$ until we reach a generation v^* such that $N_i^{send}(v^*) = 0$ for all officers i. The procedure for simulating reply e-mails for model (2.5) is similar Algorithm B, essentially we are substituting r_k^{ij} and $\hat{\theta}_{ij}$ in for r_k^i and $\hat{\theta}_i$. In other words, under estimated model (2.5) the number of replies generated for each e-mail received by i depends on the sender j.

APPENDIX B

Analytic Error Bars

Here we provide a derivation of an analytic approximation for computing standard errors for the histogram estimators of the triggering function components (error bars in Figure 3.11, Section 3.4). We proceed by first deriving an approximation of the standard error for the histogram estimator of $g(t)$, and then note that the standard errors for the histogram estimators of $\kappa(m)$ and $h(r)$ can be approximated similarly. Please use Section 2.1 as a reference for much of the notation in this appendix, and note that p_{ij} refers to the triggering probability (3.6) after Algorithm 2 has converged.

Let $t \in (\delta t_k, \delta t_{k+1}]$ and $\hat{g}(t) = g_k$ be the histogram density estimator of $g(t)$. Now suppose S_k is a random variable representing the number of triggered events in bin k, i.e. the number of aftershocks occurring between $(\delta t_k, \delta t_{k+1}]$ days after the earthquakes that directly trigger them. Then S_k follows a binomial distribution with number of trials n_t equal to the true number of triggered events (aftershocks) for the process, and success probability θ_k^g equal to the true probability an aftershock occurs between $(\delta t_k, \delta t_{k+1}]$ days after the earthquake that directly triggers it. Since we do not know the true values for the binomial parameters we estimate them with $\hat{n}_t = \sum_{i=1}^{N} \sum_{j=1}^{i-1} p_{ij}$ and $\hat{\theta}_k^g = \sum_{B_k} p_{ij}/\hat{n}_t$. Hence, an approximation of the variance of the histogram density estimator $g_k = S_k/(\Delta t_k n_t)$ is given by:

$$\widehat{Var}(g_k) = \frac{(\hat{\theta}_k^g)(1 - \hat{\theta}_k^g)}{\hat{n}_t \Delta t_k^2}$$

Similarly, we can approximate the variances for the other histogram estimators:

$$\widehat{Var}(\kappa_k) = \frac{\hat{n}_t(\hat{\theta}_k^{\kappa})(1-\hat{\theta}_k^{\kappa})}{(N_k^{mag})^2}$$
$$\widehat{Var}(h_k) = \frac{(\hat{\theta}_k^{h})(1-\hat{\theta}_k^{h})}{\hat{n}_t \Delta r_k^2}$$

where $\hat{\theta}_k^{\kappa} = \sum_{A_k} p_{ij}/\hat{n}_t$ and $\hat{\theta}_k^{h} = \sum_{C_k} p_{ij}/\hat{n}_t$.

Bibliography

Adelfio, G. and Chiodi, M. (2013). Mixed estimation technique in semi-parametric space-time point processes for earthquake description. In *Proceedings of the 28th International Workshop on Statistical Modelling 8-13 July, 2013, Palermo*, volume 1, pages 65–70.

Adelfio, G. and Chiodi, M. (2015). Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment*, 29(2):443–450.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

Brillinger, D. R. (1998). Some wavelet analyses of point process data. In *ASILOMAR CONFERENCE ON SIGNALS SYSTEMS AND COMPUTERS*, volume 2, pages 1087–1091. COMPUTER SOCIETY PRESS.

Cohen, W. W. (2009). Enron email dataset. `http://www.cs.cmu.edu/~enron/`.

Congress (2003). Report of investigation of enron corporation and related entities regarding federal tax and compensation issues, and policy recommendations, appendix d vii materials relating to pre-bankruptcy bonuses. `http://www.gpo.gov/fdsys/pkg/GPO-CPRT-JCS-3-03/content-detail.html`.

Creamer, G., Rowe, R., Hershkop, S., and Stolfo, S. J. (2009). Segmentation and automated social hierarchy detection through email network analysis. In *Advances in Web Mining and Web Usage Analysis*, pages 40–58. Springer.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume 1: Elementary Theory and Methods*. Springer, New York, second edition.

Gutenberg, B. and Richter, C. F. (1944). Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188.

Halpin, P. F. and De Boeck, P. (2013). Modelling dyadic interaction with hawkes processes. *Psychometrika*, pages 1–22.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11:493–503.

Hegemann, R., Lewis, E., and Bertozzi, A. (2012). An estimate & score algorithm for simultaneous parameter estimation and reconstruction of missing data on social networks. *Security Informatics*.

Lewis, E. and Mohler, G. (2010). A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 00(00):1–16.

Lewis, P. A. and Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413.

Malmgren, R. D., Stouffer, D. B., Motter, A. E., and Amaral, L. A. (2008). A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158.

Marsan, D. and Lengliné, O. (2008). Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079.

Marsan, D. and Lengliné, O. (2010). A new estimation of the decay of aftershock density with distance to the mainshock. *J. Geophys. Res.*, 115.

Masuda, N., Takaguchi, T., Sato, N., and Yano, K. (2012). Self-exciting point process modeling of conversation event sequences. *arXiv preprint arXiv:1205.5109*.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Intell. Res.(JAIR)*, 30:249–272.

Meyer, P. (1971). Démonstration simplifiée d'un théoréme de knight. In *Séminaire de Probabiliés V Université de Strasbourg*, volume 191 of *Lecture Notes in Mathematics*, pages 191–195. Springer Berlin Heidelberg.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Musmeci, F. and Vere-Jones, D. (1992). A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44(1):1–11.

Nichols, K. and Schoenberg, F. P. (2014). Assessing the dependency between the magnitudes of earthquakes and the magnitudes of their aftershocks. *Environmetrics*, 25(3):143–151.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849.

Rathbun, S. L. and Cressie, N. (1994). Asymptotic properties of estimators for the parameters of spatial inhomogeneous poisson point processes. *Advances in Applied Probability*, pages 122–154.

Schoenberg, F. P. (2013). Facilitated estimation of etas. *Bulletin of the Seismological Society of America*, 103(1):601–605.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Shetty, J. and Adibi, J. (2004). The enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report, University of Southern California*, 4.

Shetty, J. and Adibi, J. (2005). Discovering important nodes through graph entropy the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81. ACM.

Stomakhin, A., Short, M., and Bertozzi, A. (2011). Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27.

Tyler, J. R., Wilkinson, D. M., and Huberman, B. A. (2005). E-mail as spectroscopy: Automated discovery of community structure within organizations. *The Information Society*, 21(2):143–153.

Utsu, T. (1970). Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(3):129–195.

Utsu, T., Ogata, Y., and Matsu'ura, R. S. (1995). The centenary of the omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1):1–33.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Wong, K. and Schoenberg, F. P. (2009). On mainshock focal mechanisms and the spatial distribution of aftershocks. *Bulletin of the Seismological Society of America*, 99(6):3402–3412.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurences. *Journal of the American Statistical Association*, 97(458):369–380.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 109(B5).

Zipkin, J. R., Schoenberg, F. P., Coronges, K., and Bertozzi, A. L. (2015). Point-process models of social network interactions: parameter estimation and missing data recovery. *in revision Eur. J Appl. Math.*