Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes

Eric W. Fox[1], Martin B. Short[2], Frederic P. Schoenberg[3], Kathryn D. Coronges[4], Andrea L. Bertozzi[5]

[1] UCLA Department of Statistics, 8125 Math Sciences Bldg., Los Angeles, CA, 90095-1554. eric.fox@stat.ucla.edu.

[2] Department of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160. mbshort@math.gatech.edu.

[3] **Corresponding author.** UCLA Department of Statistics, 8125 Math Sciences Bldg., Los Angeles, CA, 90095-1554. frederic@stat.ucla.edu.

[4] Department of Behavioral Sciences and Leadership, United States Military Academy at West Point, 606 Thayer Road, West Point, NY 10996. Kathryn.Coronges@usma.edu.

[5] UCLA Department of Mathematics, 520 Portola Plaza, Los Angeles, CA 90095-1555. bertozzi@math.ucla.edu.

Abstract.

Self-exciting point process models are used to model a social network dataset consisting of email communications between officers at West Point Academy during a one year period beginning in May 2010. The models appear to adequately capture major clustering features in the data, and features of the model may be used to predict perceived leadership status within the social network. The results suggest that such models may be used for simulation, understanding basic properties of, and perhaps even prediction of underlying leadership status of social communication networks.

Keywords: conditional intensity, epidemic-type aftershock sequence models, Hawkes process, IkeNet data, social networks.

Running title: Modeling email networks using self-exciting point processes.

1 Introduction

Self-exciting point processes describe random collections of events where the occurrence of one event increases the likelihood that another event occurs shortly thereafter. Models for self-exciting point processes have been used extensively in seismology to characterize the branching structure of earthquakes, where each mainshock potentially triggers its own aftershocks sequence (Ogata, 1988, 1998). The Hawkes process (Hawkes, 1971; Hawkes and Oakes, 1974) was one of the earliest models of the conditional intensity, $\lambda(t)$, for the expected rate at which earthquakes occur at time t, given all earthquakes that occurred previously at times $t_k < t$:

$$\lambda(t) = \mu + \sum_{t_k < t} g(t - t_k). \tag{1}$$

In this model mainshocks occur at a constant rate μ over time, and each earthquake at time t_k elevates the risk of future earthquakes (aftershocks) through the triggering function $g(t - t_k)$, which is often assumed power-law or exponential.

Self-exciting point processes have found application in many areas besides seismology. These include modeling the spread of invasive plant species (Balderama et al., 2011), insurgencies in Iraq (Lewis et al., 2011), and domestic crimes (Mohler et al., 2011) as well as retaliatory acts of violence between rival gangs in Los Angeles (Stomakhin et al., 2011; Hegemann et al., 2012). Within the field of communication, self-exciting models have been used to model face-to-face conversation sequences (Masuda et al., 2012) or for e-mail transactions between an employee and an employer (Halpin and De Boeck, 2013). However, the study of self-exciting point processes in the context of social networks is a relatively new topic.

In this paper we extend the Hawkes process to model e-mail activity on a social network. Like earthquakes, e-mail communications may be viewed as branching processes. The 'mainshocks' are the times when an individual initiates e-mail conversations, or starts new e-mail threads. The 'aftershocks' are the reply e-mails, which are sent in response to e-mails received from other individuals 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 account for circadian and weekly trends. A primary motivation in the present paper is to use estimated model parameters to identify and rank network leaders. In Section 2, we introduce an e-mail network data set from West Point Military Academy and present some descriptive statistics. In Section 3, we propose various self-exciting models for e-mail activity on a social network and fit these models to the network data. In Section

4, we describe how our estimated model parameters can be used to infer network behavior and leadership. In Section 4 we also discuss model comparison, model diagnostics, and a simulation analysis of e-mail network traffic.

2 Data and Descriptive Statistics

In our study we use e-mail data collected from twenty-two officers attending West Point Academy over a one year period beginning in May 2010. This dataset, coined the IkeNet dataset, presents a unique opportunity to analyze and model e-mail traffic on a closed and finite social network. The sender, receiver, timestamp, and id were recorded for each message sent between officers in the network. The officers were anonymized in the data for privacy, therefore we will refer to them by number (1–22) in lieu of name. 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 also asked in a survey to list the officers, within the network, whom they considered strong team and military leaders. This additional data enables us to look at connections between IkeNet e-mail communication and leadership. In many previous studies on e-mail activity, the data do not contain this sort of supplementary information about the participants (e.g. Barabási (2005); Malmgren et al. (2008)). This dataset offers a particularly unique opportunity to address questions such as how one might predict network leadership using only observations of network communication. In Section 4 we propose a novel way to address this issue using the parameters of our model for e-mail network traffic described in Section 3.

Descriptive statistics for the IkeNet dataset reveal daily, weekly and seasonal trends in e-mail traffic. Figure 1 is a histogram of the number of e-mails sent in the network each hour of the day, over the yearlong observation window. There was a clear diurnal rhythm in this plot: e-mails were most frequently sent mid-day and activity diminishes at night. Decreased activity during lunch and dinner is also visible, around noon and seven p.m. Figure 2 is a bar plot of the number of emails 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 3 is a time series plot of the number of e-mails sent in the network each day. The red smoother curve helps reveal monthly trends. For instance, there was a drop in network activity in January; this is probably due to the holidays and officers being out of town. In the time series plot two days have an unusually high amount of traffic: 02 May 2011 and 02 February 2011, with 166 and 162 e-mails sent on those days, respectively. These outliers are also present in Figure 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 by each officer and between pairs of officers. Officers 9, 18, 13, and 22 stand out for sending the largest number of e-mails in the network (Figure 5). There is also a high correlation ($r = 0.945$) between the number of e-mails sent and received by each officer (Figure 6). Figure 7(a) 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 plot reveals that officer pair (9,18) is the most prolific, as these officers sent a total of 1042 emails to each other.

3 Self-Exciting Models for E-mail Network Activity

Several studies on e-mail communication have shown that the times when e-mails are sent by an individual deviates from a stationary Poisson process (Barabási, 2005; Malmgren et al., 2008). A stationary Poisson process model assumes that the rate at which events are expected to occur is constant at all times. One property of this model is that the times between consecutive events (inter-event 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 e-mail traffic is highly clustered: short periods of intense activity are followed by long periods when relatively few e-mails are sent. To explain this observed activity, Barabási (2005) proposed a priority queue model, in which high priority e-mails are responded to before low priority e-mails. We take a different approach, and account for clustering by viewing e-mail traffic as a self-exciting point process, whereby 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.

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 email 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 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:

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{E[N(t, t + \Delta t)|H_t]}{\Delta t}. \tag{2}$$

An important conditional intensity model for a self-exciting point process is the Hawkes process (Hawkes, 1971; Hawkes and Oakes, 1974; Ogata, 1988) given by equation (1). This 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 $r^i_k < t$:

$$\begin{aligned}\lambda_i(t) &= \mu_i + \sum_{r^i_k<t} g_i(t - r^i_k) \\ &= \mu_i + \theta_i \sum_{r^i_k<t} \omega_i e^{-\omega_i(t-r^i_k)}. \end{aligned} \tag{3}$$

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^i_k elevates the overall rate of sending e-mails at time $t > r^i_k$, through the triggering function $g_i(t - r^i_k)$, which is assumed to be exponential.

In model (3), 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 nonstationarities suggested in Figures 1 and 2. Non-stationary forms for $\mu_i(t)$ will be discussed subsequently in Section 3.1.

The exponential triggering function is perhaps not unreasonable. Note that since the officers are using mobile devices (Blackberries) to send e-mails, they do not have to wait until they reach a computer, and can thus send quick replies. Moreover, Figure 8 shows that the survival function of the inter-event times for e-mails sent by each officer in the network falls reasonably close to the 95% confidence envelope for the self-exciting Hawkes process model. This indicates that the inter-event time distribution for the model closely resembles that of the observed data. The simulation procedure used to create the confidence envelop will be described in greater detail in Section 4.3.

As an illustration of model (3), the top panel in Figure 9 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 (red points) and received (blue points) are easily discerned and are characteristic of Hawkes point processes.

The parameters of model (3) 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 total expected number of replies, on average, 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 1), θ_i will be extremely close to the expected number of replies 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.

3.1 Non-stationary Background Rate

Model (3) 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 1 and 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 (3) of email traffic:

$$\lambda_i(t) = p_i N_i^{send} \hat{\mu}_0(t) + \theta_i \sum_{r_k^i < t} \omega_i e^{-\omega_i (t - r_k^i)}, \tag{4}$$

where the parameter p_i is the probability that a randomly selected email sent by officer i is not a reply, N_i^{send} is the number of e-mails sent by officer i to other officers, and $\hat{\mu}_0(t)$ is a non-parametrically estimated density curve obtained by kernel smoothing the emails sent by all officers, accounting for daily and weekly rhythms in IkeNet activity (Figure 11).

Technically, 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 $\hat{\mu}_0(t) = \hat{f}(h + m/60)w(d)$, where

$$\hat{f}(h + m/60) = \frac{1}{\sigma N} \sum_{k=1}^{N} K\left(\frac{h + m/60 - h_k}{\sigma}\right) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(h+m/60-h_k)^2}{2\sigma^2}}, \tag{5}$$

and

$$w(d) = \frac{7}{360N} \sum_{k=1}^{N} I(d_k = d). \tag{6}$$

In equation (5) we use kernel density estimation to get a smoothed version of the histogram of the number of e-mails sent by hour of day (Figure 10). We chose a normal smoothing kernel $K(\cdot)$ with bandwidth σ set to the default value suggested by Scott (1992); N is simply the total number of e-mails sent in the network. To account for weekly trends $\hat{f}(\cdot)$ is given a weight, $w(d)$, proportional to the total number of e-mails sent on day d (Figure 2). The constant of proportionality is chosen so that $\hat{\mu}_0(t)$ integrates to one over [0,T]. Note that $\hat{\mu}_0(t)$ is periodic, with period equal to one week (7 days / 168 hours); i.e. $\hat{\mu}_0(t+168) = \hat{\mu}_0(t)$. One period of $\hat{\mu}_0(t)$ is plotted in Figure 11. To illustrate the fitted model, the lower panel of Figure 9 shows the estimated conditional intensity for officer 15 under model (4). The troughs in the conditional intensity in Figure 9 correspond to times when few e-mails are sent and received.

3.2 Pairwise Model

One shortcoming of models (3) and (4) is that the reply rate θ_i for officer i does not depend on who sends an e-mail to i. According to these models, if two officers send officer i a message, officer i sends each of them the same expected number of replies. 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:

$$\begin{aligned}\lambda_i(t) &= p_i N_i^{send}\hat{\mu}_0(t) + \sum_{j\neq i}\sum_{r_k^{ij}<t} g_{ij}(t - r_k^{ij}) \\ &= p_i N_i^{send}\hat{\mu}_0(t) + \sum_{j\neq i}\sum_{r_k^{ij}<t} \theta_{ij}\omega_i e^{-\omega_i(t-r_k^{ij})}. \qquad (7)\end{aligned}$$

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 other than i. Note that one may also extend model (7) so that a distinct ω_{ij} is estimated for each email receiver i and each sender j, and thus essentially model each pairwise interaction individually, though with the current dataset this may not be advisable due to the the sparsity in the number of e-mail sent between certain pairs of individuals (Figure 7(a)) and the large number of additional parameters to estimate.

The parameters of model (7) 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 than model (4). A more in-depth comparison between models (4) and (7) is provided in Section 4.

3.3 Parameter Estimation

The parameters of models (3) and (4) can readily be estimated by maximizing the log-likelihood over the parameter space Ω_i. The log-likelihood equation (Ogata, 1978) is given by

$$\begin{aligned} l_i(\Omega_i) = logL_i(\Omega_i) &= \sum_{k=1}^{N_i^{send}} \log(\lambda_i(s_k^i)) - \int_0^T \lambda_i(t)dt \\ &= \sum_{k=1}^{N_i^{send}} \log(\lambda_i(s_k^i)) - \left(\int_0^T \mu_i(t)dt + \theta_i \sum_{k=1}^{N_i^{rec}} [1 - e^{-\omega_i(T - r_k^i)}] \right), \end{aligned} \tag{8}$$

where s_k^i is the time when the k^{th} e-mail was sent by officer i, r_k^i is the time when the k^{th} e-mail was received by i, and N_i^{rec} and N_i^{send} are the total number of e-mails received and sent by i, respectively. The integral of the background rate depends on whether it is stationary, $\mu_i(t) = \mu_i$, or non-stationary, $\mu_i(t) = p_i N_i^{send} \hat{\mu}_0(t)$. For the stationary background rate the integral is simply $\mu_i T$, while for the non-stationary background rate the integral is $p_i N_i^{send}$ since $\hat{\mu}_0(t)$ is a density. Standard errors for the maximum likelihood estimators, $\hat{\Omega}_i$, of equation (8) can be derived using asymptotic properties (Ogata, 1978).

The log-likelihood equation (8) can also be used to simultaneously find the maximum likelihood estimates of the parameters, $\Omega_i = (p_i, w_i, \vec{\theta}_i)$, of model (7). The vector $\vec{\theta}_i$ contains twenty-one parameters θ_{ij} such that $j \neq i$. For model (7), the loglikelihood has the form

$$l_i(\Omega_i) = \sum_{k=1}^{N_i^{send}} \log(\lambda_i(s_k^i)) - \left(p_i N_i^{send} + \sum_{j \neq i} \sum_{k=1}^{N_{ij}^{rec}} \theta_{ij} [1 - e^{-\omega_i (T - r_k^{ij})}] \right), \qquad (9)$$

where N_{ij}^{rec} is the the number of messages officer i received from j. When $N_{ij}^{send} = 0$ or $N_{ij}^{rec} = 0$ we set $\theta_{ij} = 0$.

Parameter estimates and standard errors for the stationary (3) and non-stationary (4) Hawkes process models are given in Tables 1 and 2. The estimates were obtained using the general purpose optimization function, `optim()`, provided in the statistical software package R (R Core Team (2013)). We used the box-contrained optimization method described in Byrd et al. (1995) to maximize equation (8). The maximum log-likelihoods for each officer, $l_i(\hat{\Omega}_i)$, for models (3) and (4) are also provided in Tables 1 and 2. In terms of likelihood, model (4) outperforms model (3) since it has larger maximum log-likelihood values for every officer, and typically (as well as overall) by a statistically significant margin according to the Akaike Information Criterion (AIC) of Akaike (1974). The inclusion of the non-stationary background rate evidently provides a better fit to the network data.

4 Analysis

4.1 Inferring Network Behavior and Leadership

The parameter estimates in Table 2 provide insight into the communication habits of officers in the network. The estimator $\hat{p}_i$ is the estimated probability an e-mail sent by officer i is a background event, or non-reply e-mail. In other words, $N_i^{send}\hat{p}_i$ is the expected total number of e-mail threads initiated by officer i. Thus, according to the fitted model (4), we can infer for example that approximately 68% of e-mails sent by officer 15 are non-replies, and 48% of e-mails sent by officer 18 are non-replies. Over the entire network, $\hat{p}_i$ ranges between 42% and 83%, and the estimated overall percentage of non-reply e-mails for the entire network is $100 * \sum_{i=1}^{22} N_i^{send}\hat{p}_i/N \approx 55\%$.

The estimator $\hat{\theta}_i$ provides an estimate of the mean number of replies officer i sends in response to a typical e-mail. For instance, according to the model, officer 18 sends approximately 58 replies per 100 e-mails received, while officer 15 sends approximately 45 replies per 100 e-mails received. Note also that the estimate of $\hat{p}_i$ 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.

The speed at which officers send e-mails is governed by $\hat{\omega}_i^{-1}$, which according to the fitted model (4) is the estimated mean time it takes officer i to reply to an e-mail. By examining Table 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 7(a)) 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 less than six minutes to over seventy minutes.

An important question is what properties of an e-mail network can best identify the perceived leaders of that network. As indicated in Section 2, each officer in the

IkeNet data set was asked in a survey to list up to five people considered to be strong team leaders, and up to five people 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. The aggregate numbers of votes received by each officer for team and military leadership are shown in Figures 12 and 13, respectively. The differences between the aggregate counts in Figures 12 and 13 and the the total number of e-mails sent by each officer (Figure 5) are quite striking. 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). Clearly, total number of emails sent is a poor predictor of one's perceived leadership status within the network.

Fortunately, aspects of the fitted Hawkes process models (4) and (7) can perhaps be used to help identify and rank perceived leaders among the network. Table 3 lists five possible covariates for estimating leadership. The covariates N^{send} and N^{rec} are the number of e-mails respectively sent and received by an officer, and $\hat{\theta}$ is the reply rate (mean number of messages an officer sends per message received) from fitted model (4). We consider also the potential predictor Y_i which is defined, for each officer i, as the total number of other officers j for which officer i has an estimated mean email reply percentage ($\hat{\theta}_{ij}$) above the overall mean for the network, and a mean non-reply send rate $\hat{p}_i N_{ij}^{send}$ above the overall median for the network. That is

$$Y_i = \sum_{j \neq i} I(\hat{\theta}_{ij} > 0.45,\, \hat{p}_i N_{ij}^{send} > 5.09), \tag{10}$$

where I denotes the indicator function and all fitted parameters are from model (7). The threshold 0.45 is the estimated mean percentage of reply emails sent in the entire network ($\frac{1}{N}\sum_{i\neq j}\sum_j N_{ij}^{rec}\hat{\theta}_{ij}$), and the threshold 5.09 is simply the median of the set $\{\hat{p}_i N_{ij}^{send} | N_{ij}^{send} > 0\}$. Of course, many other thresholds are possible.

Tables 3 and 4 report the Spearman correlation, Pearson correlation, and root mean square error to measure how accurately each covariate estimates and ranks network leaders. The correlations are between the covariate of interest and the actual numbers of votes for team and military leadership shown in Figure 12 and Figure 13. The root mean square errors are found by fitting a simple linear regression of the leadership vote totals on the values of the covariates for each officer. The last column in both leadership tables gives the top three leaders predicted by each covariate.

Tables 3 and 4 show that Y, defined in equation (10), is much more highly correlated with team and military leadership votes than the number of messages received and sent by each officer. Y also has the lowest root mean square error when compared to all other covariates. Moreover, Y is the only covariate that correctly identifies the top three team leaders (officers 15, 13, and 22) clearly seen in Figure 12. Thus, not only is Y more closely associated with the true leadership vote totals, but it also better identifies the top leaders, which are not well estimated using N^{send} or N^{rec}. Comparison of Tables 3 and 4 reveals that the covariates considered are all more highly correlated with team leadership vote totals than with military leadership vote totals. Note also that the estimated reply rate $\hat{\theta}$ from model (4) correlates much more strongly with perceived military leadership when compared to the email totals.

4.2 Model Comparison and Diagnostics

The maximized log-likelihoods for the network and corresponding AIC values are provided in Table 5. The first row gives these values for a stationary Poisson model of e-mail 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

(3, 4, and 7) described in Section 2. The Hawkes process models fit the data significantly better than the stationary Poisson model in terms of AIC. Additionally, the AIC values for the models with non-stationary background rates (4, 7) are significantly lower than the AIC for the model with the stationary background rate (3). This indicates that taking diurnal and weekly trends into account provides a better fit to the data. The Hawkes process model that incorporates pairwise interactions between officers (7) fits the data more closely than model (4) as measured by the loglikelihood, but scores worse in terms of AIC. This is because the AIC penalizes for the large number of parameters in (7).

One simple diagnostic for model (7) is the inspection of whether the estimated numbers of reply and non-reply messages add up to the total number of messages sent by each officer. For the fitted model (7),

$$\frac{\hat{p}_i N_i^{send} + \sum_{j \neq i} \hat{\theta}_{ij} N_{ij}^{rec}}{N_i^{send}} \tag{11}$$

is indeed very close to unity for each officer i. In the expression above, $\hat{p}_i N_i^{send}$ is the expected number of non-reply messages and $\sum_{j \neq i} \hat{\theta}_{ij} N_{ij}^{rec}$ is the expected number of reply messages estimated under model (7). Similarly, we verified that $(\hat{p}_i N_i^{send} + \hat{\theta}_i N_i^{rec})/N_i^{send} \approx 1$ for each officer i under model (4).

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

$$\tau_k^i = \Lambda(s_k^i) = \int_0^{s_k^i} \lambda_i(t)dt. \tag{12}$$

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 interevent 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 14 for the Hawkes process model (4) and the stationary Poisson model of e-mail network traffic. A comparison of these plots reveals much less clustering around the perimeter for the Hawkes process model, 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 model.

4.3 Simulation

The fitted models (4) or (7) can be used to simulate IkeNet e-mail traffic. For instance, Figure 7(b) is a matrix plot from a simulation of the network using model (7), which consists of $N = 8490$ events (compare to the $N = 8397$ events in the observed network data). The model appears to reproduce the matrix of total e-mail communications sent between officers quite accurately.

By simulating the network repeatedly, one can form 95% confidence envelopes for the non-stationary background rate $\mu_0(t)$ (Figure 11). This error bound is formed by simulating model (7) 100 times and re-estimating the background rate for each simulation. The pointwise 0.025 and 0.975 quantiles of the simulated and re-estimated background rates are shown in cyan. Note that the background rate from the observed network data falls reasonably within the 95% confidence bands. A confidence envelope for the survival function of the inter-event times (Figure 8) can be formed similarly.

5 Discussion

Self-exciting point process models for e-mail network traffic clearly appear to outperform traditional stationary Poisson models, at least for the IkeNet data considered here. These Hawkes 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 significantly 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 $\hat{\omega}$, are easily interpretable and characterize important properties of the e-mail communication network, such as an individuals' tendencies to reply to e-mails and initiate new e-mail threads. These parameters also appear to be useful for both ranking and identifying team and military leaders, and thus may more accurately summarize the core structure of the social network, compared with simple email totals.

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 (7) quantified these additional features, which we attempted to combine into a simple measure (10) for estimating network leadership. This covariate was more highly correlated with the perceived leadership vote totals than simply the number of e-mails sent or received by each individual in the network.

These results may be seen as merely a first step towards attempting to characterize the leadership patterns from email occurrences in a social network. Obviously, many other measures of leadership status should be investigated, and the results tested on other social network datasets. Another future direction for this research is to consider different forms for the triggering function in models (3), (4), and (7). The authors of Halpin and De Boeck (2013), who fit a similar Hawkes process model for e-mails sent between an employee and employer, modeled the triggering function with a gamma kernel instead of an exponential kernel. A gamma triggering function may help account for the time it takes an individual to check their inbox and write a reply. Another option may be to consider a completely non-paramteric approach to estimating the background rate and triggering function as described in Marsan and Lengline (2008).

Lastly, it would be interesting to test our methods on other e-mail data sets. For example, Eckmann et al. (2004) utilized an e-mail data set extracted from the log files of a university mail server. Their network is substantially larger that the IkeNet, and consists of e-mails sent between thousands of users at a university over an 83 day period. Similar to the IkeNet, this data contains the sender, recipient, and timestamp for each e-mail. An additional variable, the size of each e-mail, is also provided, but Barabási (2005) showed that this does not correlate well with the reply times for e-mails. Also of interest is the Enron e-mail dataset (Klimt and Yang, 2004) since it contains the actual message content for each e-mail. Furthermore, within the Enron corporation there are established leaders, hence we have a means to test our covariate for identifying leaders and discover new covariates associated with network leadership. By continuing to explore and model such communication network data sets, one may hope ultimately to find robust ways to identify leaders of criminal or terrorist organization from e-mail communication patterns.

Acknowledgements

This research is supported by ARO grant W911NF1010472 and AFOSR MURI grant W911NF-11-1-0332.

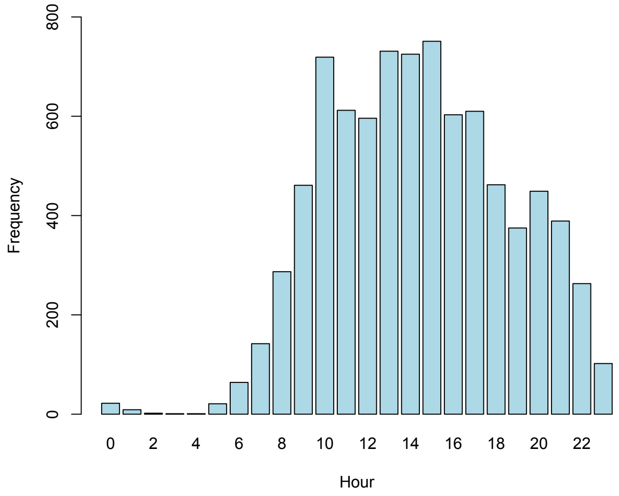

Figure 1. Histogram of the number of emails sent, each hour of the day, over the yearlong observation window.

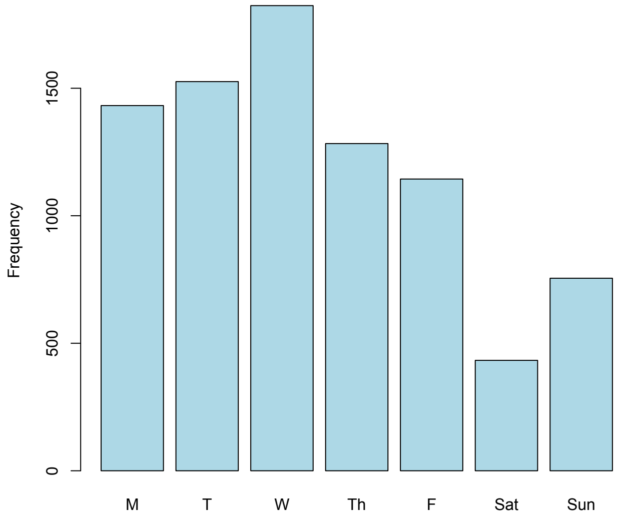

Figure 2. Number of emails sent by day of the week.

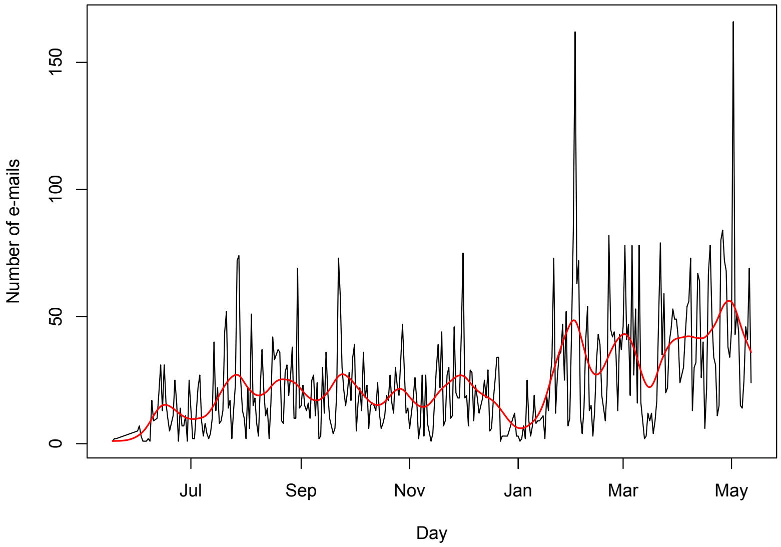

Figure 3. Number of e-mails sent by date.

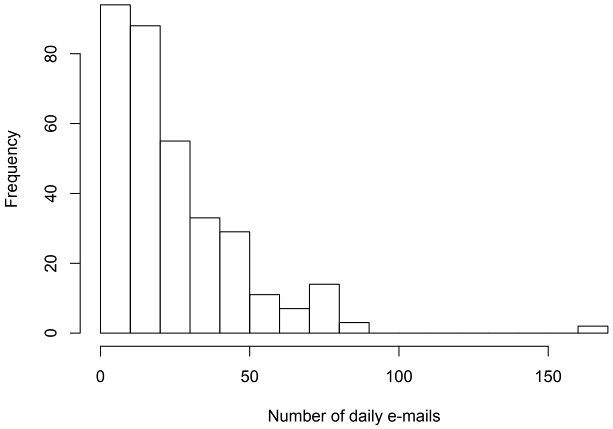

Figure 4. Histogram of the number of daily e-mails sent.

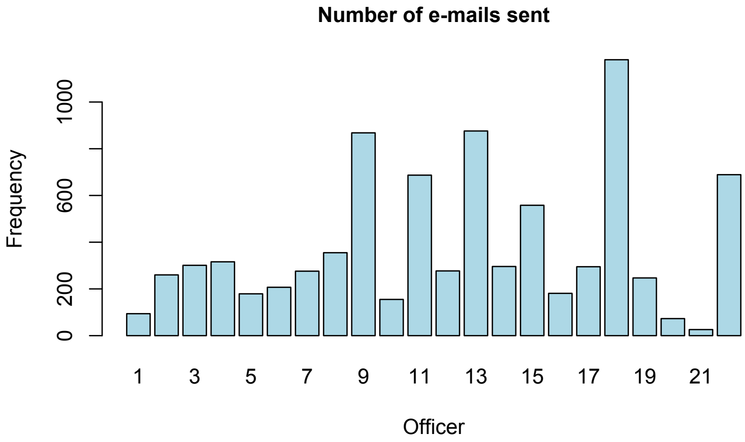

Figure 5. Number of e-mails sent by each officer over the yearlong observation window.

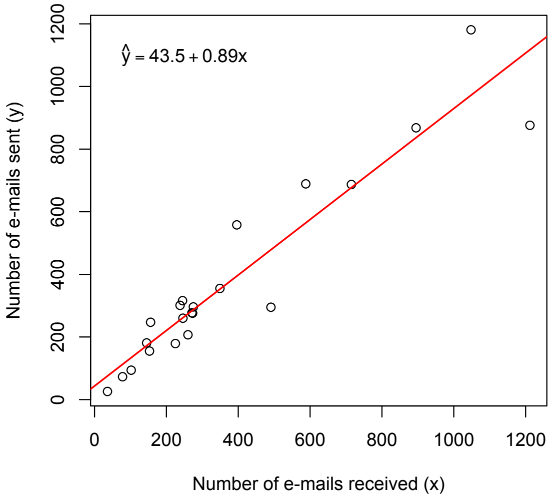

Figure 6. Scatter plot of the total number of emails received by each officer (x) versus the total number of e-mails sent (y). The scatter plot and regression line show a strong association between the number of e-mails sent and received ($r = 0.945$).

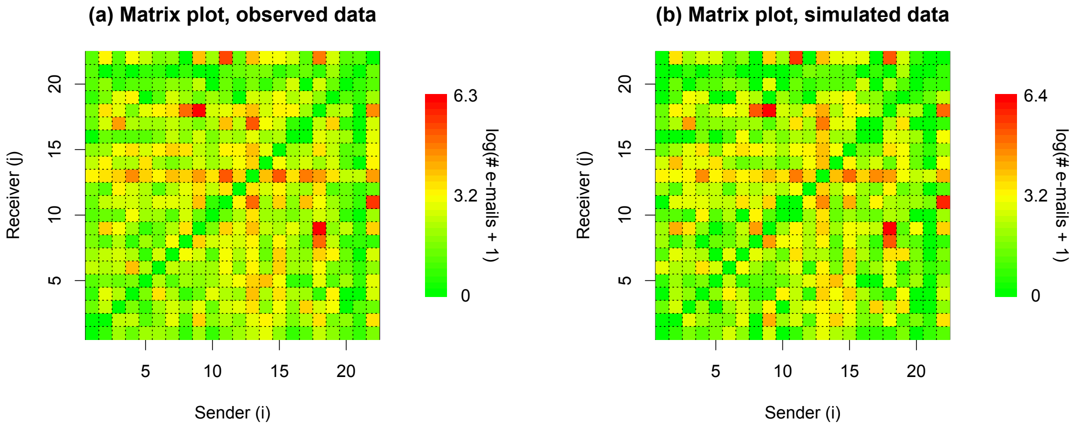

Figure 7. (a) Matrix plot of the logarithm of the number of e-mails sent from officer i (column) to j (row) from the observed data. The red and orange cells indicate pairs of officers that communicate frequently through e-mail. Likewise, the yellow and green cells indicates moderate to low communication between officer pairs. (b) Simulated matrix plot of the logarithm of the number of e-mail sent from officer i (column) to j (row). Data were synthesized from one simulation of the IkeNet e-mail network using the maximum likelihood estimates of the pairwise model (7).

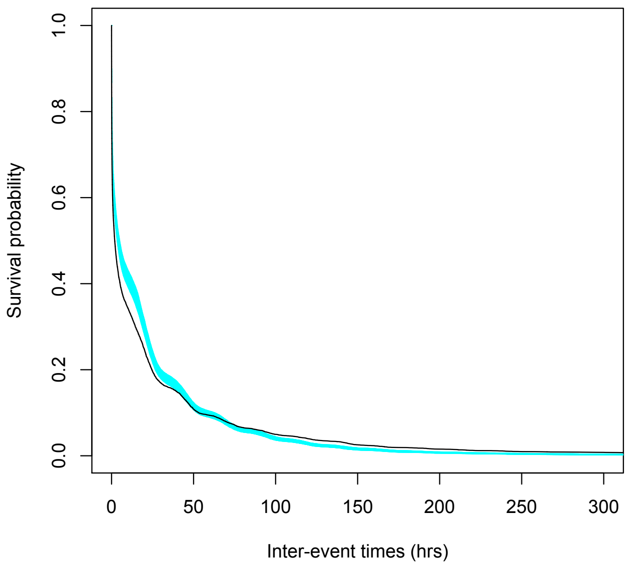

Figure 8. Survivor plot of the inter-event times for e-mails sent by each officer in the network (black line). A 95% confidence envelope was formed by simulating the network 100 times from the fitted model (7) and computing the survivor function for each realization. The pointwise 0.025 and 0.975 quantiles of the simulated survivor functions are plotted in cyan.

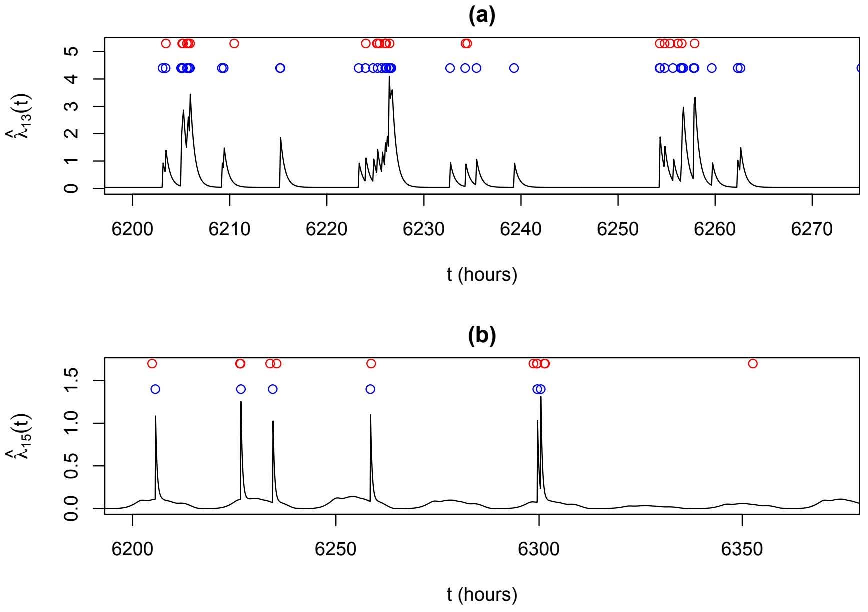

Figure 9. Top panel shows the estimated conditional intensity for officer 13 over a three day period using the Hawkes model with the stationary background rate (3). The bottom panel shows the estimated conditional intensity for officer 15 over the same three day period using the Hawkes model with the non-stationary background rate (4). The blue dots represent the times when messages are received, while the red dots represent the times when messages are sent.

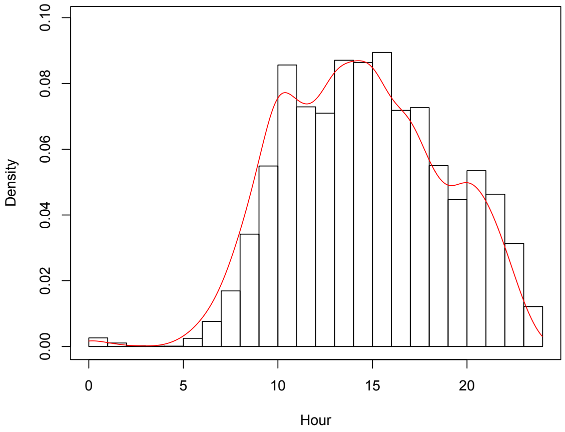

Figure 10. Kernel density estimation of the histogram of e-mails sent per hour of day, using the default fixed bandwidth suggested by Scott (1992).

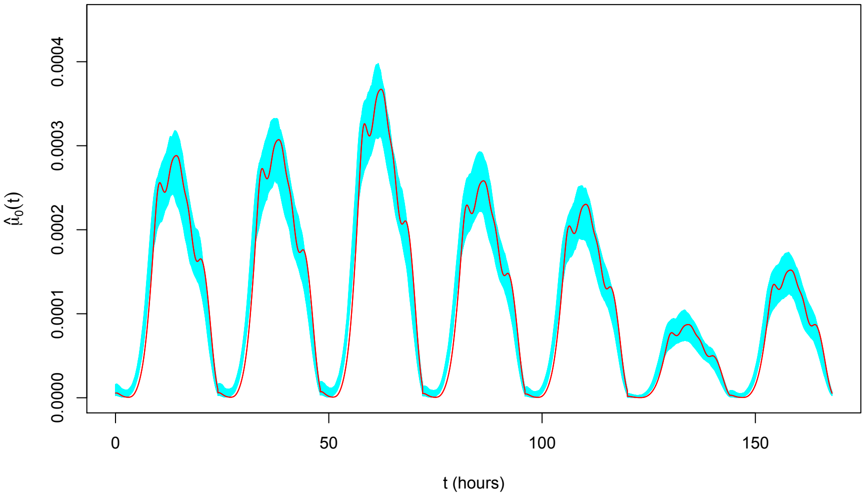

Figure 11. Estimated background rate density $\hat{\mu}_0(t)$ for the IkeNet data set (red). The background rate is periodic; the figure plots only one period (i.e. one week, Mon.-Sun.). A 95% confidence envelope is plotted in cyan. This is formed by simulating the network repeatedly 100 times.

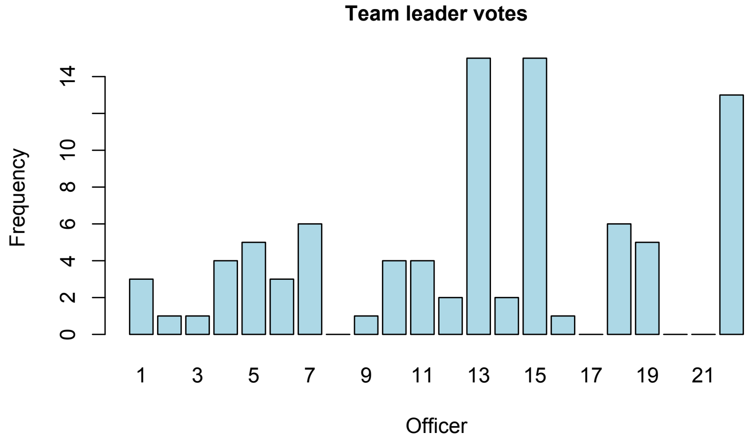

Figure 12. Barplot of the number of votes each officer received for perceived team leadership. Votes are based on a survey which asked each officer to list up to five officers that he or she considered to be a team leader.

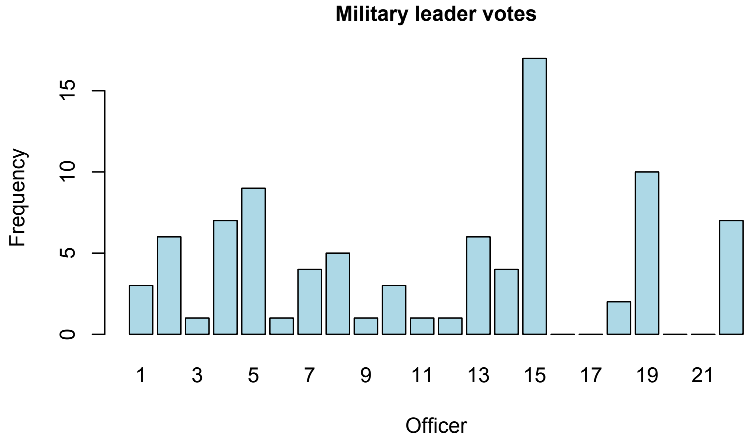

Figure 13. Barplot of the number of votes each officer received for perceived military leadership. Votes are based on a survey which asked each officer to list up to five officers that he or she considered to be a military leader.

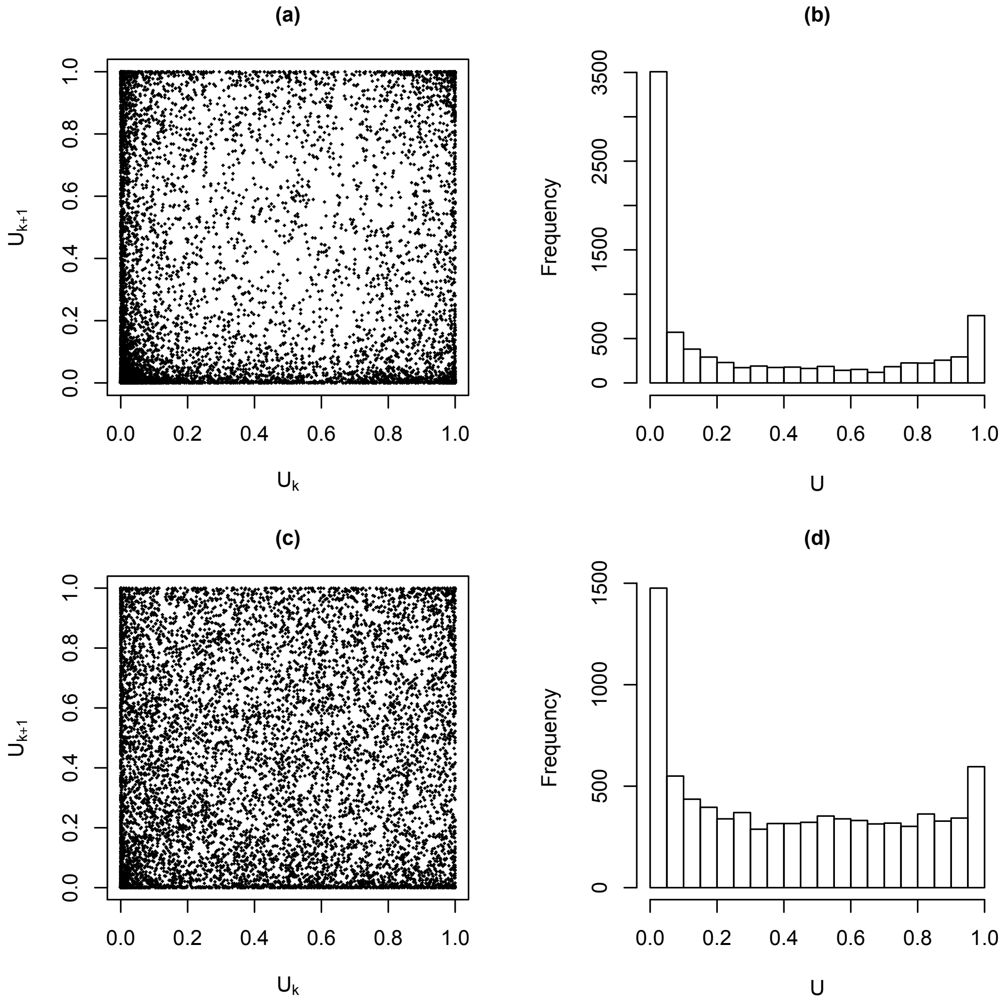

Figure 14. (a-b) Plot of U_{k+1} versus U_k for a stationary Poisson process model of e-mail activity on the network with corresponding histogram of transformed time values U. (c-d) Plot of U_{k+1} versus U_k for the Hawkes process model (4) of e-mail activity on the network with corresponding histogram of transformed time values U.

Table 1. Parameter estimates, standard errors and log-likelihood values for model (3)

i	N_i^{send}	$\hat{\mu}_i$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l(\hat{\Omega}_i)$
1	94	0.009 (0.0010)	0.17 (0.04)	8.65 (2.52)	-464.2
2	260	0.014 (0.0013)	0.58 (0.05)	3.64 (0.51)	-732.8
3	301	0.021 (0.0016)	0.49 (0.05)	1.38 (0.19)	-1089.4
4	316	0.024 (0.0017)	0.43 (0.05)	2.93 (0.51)	-1126.4
5	179	0.012 (0.0012)	0.35 (0.04)	1.64 (0.29)	-702.9
6	207	0.014 (0.0013)	0.34 (0.04)	3.10 (0.63)	-752.5
7	276	0.016 (0.0015)	0.51 (0.05)	0.80 (0.11)	-989.0
8	355	0.025 (0.0018)	0.40 (0.04)	4.71 (0.95)	-1125.6
9	868	0.045 (0.0026)	0.54 (0.03)	6.68 (0.73)	-1620.0
10	155	0.012 (0.0013)	0.33 (0.06)	3.38 (1.74)	-635.4
11	687	0.034 (0.0023)	0.55 (0.03)	2.19 (0.24)	-1647.9
12	277	0.018 (0.0018)	0.43 (0.05)	1.34 (0.42)	-1018.5
13	876	0.038 (0.0025)	0.45 (0.02)	2.21 (0.20)	-2029.1
14	296	0.016 (0.0014)	0.57 (0.05)	2.87 (0.34)	-871.4
15	558	0.040 (0.0024)	0.53 (0.04)	1.75 (0.27)	-1717.8
16	181	0.014 (0.0013)	0.41 (0.05)	6.49 (1.36)	-683.6
17	295	0.019 (0.0017)	0.26 (0.03)	2.87 (0.66)	-1023.1
18	1181	0.059 (0.0029)	0.64 (0.03)	6.91 (0.59)	-1853.8
19	247	0.019 (0.0017)	0.53 (0.08)	0.83 (0.25)	-992.8
20	73	0.006 (0.0008)	0.26 (0.06)	3.18 (0.92)	-360.2
21	26	0.002 (0.0004)	0.32 (0.12)	0.67 (0.27)	-159.7
22	689	0.030 (0.0020)	0.73 (0.04)	3.52 (0.35)	-1223.4

Table 2. Parameter estimates, standard errors, and log-likelihood values for model (4)

i	N_i^{send}	$\hat{p}_i$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l(\hat{\Omega}_i)$
1	94	0.83 (0.09)	0.16 (0.04)	10.01 (3.25)	-430.2
2	260	0.47 (0.04)	0.56 (0.05)	4.12 (0.59)	-681.6
3	301	0.65 (0.05)	0.44 (0.05)	1.66 (0.26)	-1018.7
4	316	0.71 (0.05)	0.37 (0.05)	4.70 (1.20)	-1017.6
5	179	0.58 (0.06)	0.34 (0.04)	1.67 (0.32)	-692.2
6	207	0.59 (0.06)	0.32 (0.04)	3.53 (0.64)	-716.7
7	276	0.54 (0.05)	0.47 (0.05)	0.91 (0.15)	-930.9
8	355	0.63 (0.04)	0.38 (0.04)	5.52 (0.96)	-1059.0
9	868	0.50 (0.03)	0.49 (0.03)	10.35 (1.05)	-1456.6
10	155	0.70 (0.07)	0.31 (0.05)	4.70 (1.07)	-599.3
11	687	0.48 (0.03)	0.50 (0.03)	2.77 (0.26)	-1534.0
12	277	0.57 (0.06)	0.44 (0.06)	1.04 (0.40)	-971.8
13	876	0.45 (0.03)	0.40 (0.02)	2.85 (0.30)	-1904.5
14	296	0.50 (0.04)	0.54 (0.05)	3.34 (0.40)	-800.8
15	558	0.68 (0.04)	0.45 (0.04)	2.58 (0.40)	-1616.1
16	181	0.69 (0.06)	0.39 (0.05)	7.69 (1.54)	-638.5
17	295	0.61 (0.05)	0.23 (0.02)	4.28 (0.69)	-953.0
18	1181	0.48 (0.02)	0.58 (0.03)	10.00 (0.83)	-1626.2
19	247	0.71 (0.06)	0.46 (0.06)	1.28 (0.26)	-936.5
20	73	0.73 (0.10)	0.25 (0.06)	3.45 (1.01)	-340.6
21	26	0.73 (0.17)	0.20 (0.08)	0.77 (0.36)	-148.9
22	689	0.42 (0.03)	0.68 (0.04)	4.51 (0.45)	-1127.7

Table 3. Covariates for estimating team leadership

Covariate	r_s	r_p	RMSE	Estimated top 3 leaders
N^{send}	0.40	0.52	4.03	$18, 13, 9$
N^{rec}	0.39	0.49	4.11	$13, 8, 9$
$N^{send} + N^{rec}$	0.40	0.51	4.05	$18, 13, 9$
$\hat{\theta}$	0.41	0.39	4.34	$22, 18, 2$
Y	0.62	0.70	3.37	$15, 13, 22$

Table 4. Covariates for estimating military leadership

Covariate	r_s	r_p	RMSE	Estimated top 3 leaders
N^{send}	0.29	0.13	4.28	$18, 13, 9$
N^{rec}	0.20	0.02	4.31	$13, 8, 9$
$N^{send} + N^{rec}$	0.24	0.07	4.30	$18, 13, 9$
$\hat{\theta}$	0.36	0.28	4.14	$22, 18, 2$
Y	0.42	0.50	3.73	$15, 13, 22$

Table 5. Number of parameters (ρ), AIC and maximum log-likelihood values for the Poisson and Hawkes process models of the e-mail network

	ρ	$l(\hat{\Omega})$	AIC
Stationary Poisson	22	-32347.4	64738.9
Hawkes model (3)	66	-22819.4	45770.9
Hawkes model (4)	66	-21201.8	42535.5
Hawkes model (7)	506	-20887.1	42786.2

References

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

Balderama, E., Schoenberg, F., Murray, E., and Rundel, P. (2011). Applications of branching models in the study of invasive species. *Journal of the American Statistical Association (to appear)*.

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

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

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.

Eckmann, J.-P., Moses, E., and Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337.

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*.

Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer.

Lewis, E., Mohler, G., Brantingham, P. J., and Bertozzi, A. L. (2011). Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264.

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

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

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.

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.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

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

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