




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

Eric W. Fox, Martin B. Short, Frederic P. Schoenberg, Kathryn D. Coronges & Andrea L. Bertozzi

To cite this article: Eric W. Fox, Martin B. Short, Frederic P. Schoenberg, Kathryn D. Coronges & Andrea L. Bertozzi (2016): Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes, Journal of the American Statistical Association, DOI: [10.1080/01621459.2015.1135802](https://doi.org/10.1080/01621459.2015.1135802)

To link to this article: <http://dx.doi.org/10.1080/01621459.2015.1135802>

 View supplementary material [↗](#)

 Accepted author version posted online: 02 Feb 2016.

 Submit your article to this journal [↗](#)

 Article views: 33

 View Crossmark data [↗](#)

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

Eric W. Fox<sup>1</sup>, Martin B. Short<sup>2</sup>, Frederic P. Schoenberg<sup>3</sup>, Kathryn D. Coronges<sup>4</sup>, Andrea L. Bertozzi<sup>5</sup>

<sup>1</sup> UCLA Department of Statistics, 8125 Math Sciences Bldg., Los Angeles, CA, 90095-1554.  
eric.fox@stat.ucla.edu.

<sup>2</sup> Department of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA  
30332-0160. mbshort@math.gatech.edu.

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

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

<sup>5</sup> UCLA Department of Mathematics, 520 Portola Plaza, Los Angeles, CA 90095-1555.  
bertozzi@math.ucla.edu.

## **Abstract.**

We propose various self-exciting point process models for the times when e-mails are sent between individuals in a social network. Using an EM-type approach, we fit these models to an e-mail network dataset from West Point Military Academy and the Enron e-mail dataset. We argue that the self-exciting models adequately capture major temporal clustering features in the data and perform better than traditional stationary Poisson models. We also investigate how accounting for diurnal and weekly trends in e-mail activity improves the overall fit to the observed network data. A motivation and application for fitting these self-exciting models is to use parameter estimates to characterize important e-mail communication behaviors such as the baseline sending rates, average reply rates, and average response times. A primary goal is to use these features, estimated from the self-exciting models, to infer the underlying leadership status of users in the West Point and Enron networks.

**Keywords:** conditional intensity, Hawkes process, IkeNet dataset, Enron e-mail dataset, social networks.

## 1 Introduction

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 paper 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 paper is organized as follows: In Section 2 we provide some descriptive statistics for the IkeNet dataset. In Section 3 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 4 we describe how to use our parameter estimates to characterize communication behaviors and predict leadership for the IkeNet social network. In Section 4 we also discuss model comparisons and diagnostics. In Section 5 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 1 we spell out the simulation algorithm we use to generate realizations of the IkeNet e-mail network from the fitted self-exciting models.

## 2 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 4.

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. 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](#) 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 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 4](#), a right skewed histogram which shows that on a typical day, fewer than thirty e-mails are sent within the network.

The e-mail network itself is shown in [Figure 5](#) 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 for sending the highest number of e-mails in the network. The network plot reveals 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) 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. [Figure 5](#) also illustrates the overall sparsity in e-mail communication on this closed network.

### 3 Self-Exciting Models for IkeNet E-mail Activity

Self-exciting point processes have their origins in seismology where models were developed to characterize the so-called branching structure of earthquakes, whereby 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  $t_k < t$ :

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

In this model mainshocks occur at a constant baseline rate  $\mu$  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. Besides seismology, self-exciting point processes have found application in many other areas such as 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](#)).

In this section we extend the Hawkes process 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 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 (officers) of the network to model the rate of sending e-mails. Other relational views as considered in [Perry and Wolfe \(2013\)](#) include, for instance, the modeling of dyadic interactions whereby the point processes are placed on the directed edges of the network to measure the rate of sending or receiving e-mails between pairs of officers.

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 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  $t$ :

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

The Hawkes process given by (1) 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  $r_k^i < t$ :

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

In the context of e-mails, the background rate  $\mu_i$  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,  $\mu_i$  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 (3), the background rate  $\mu_i$  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 1 and 2. Non-stationary forms for the background rate will be discussed subsequently in Section 3.1.

The exponential triggering function is perhaps not unreasonable. For instance, Figure 6 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 using estimated model (3). This plot indicates that the inter-event time distribution for the estimated model closely resembles that of the observed data. A description of the simulation procedure for model (3) is given in Appendix 1.

As an illustration of model (3), the top panel in Figure 7 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 (3) characterize general e-mail communication habits of each officer. For instance,  $\theta_i$  can be interpreted as the reply rate for officer  $i$ , since it is the expected number of reply e-mails<sup>1</sup> sent by officer  $i$  per e-mail received from another officer in the network, as

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

<sup>1</sup>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.

The integrated triggering function over a finite time period will be slightly less than  $\theta_i$ , but for the IkeNet data, where  $T = 8640$  hours and  $\omega^{-1} \ll T$  (see Table 1),  $\theta_i$  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  $\omega_i$ , with larger values of  $\omega_i$  indicating faster response times for officer  $i$ . Indeed,  $\omega_i^{-1}$  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 e-mail traffic:

$$\begin{aligned}\lambda_i(t) &= \nu_i \mu(t) + \sum_{r_k^i < t} g_i(t - r_k^i) \\ &= \nu_i \mu(t) + \theta_i \sum_{r_k^i < t} \omega_i e^{-\omega_i(t - r_k^i)},\end{aligned}\tag{4}$$

where  $\nu_i$  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 8); the details of this estimation procedure are given subsequently. Since  $\int_0^T \nu_i \mu(t) dt = \nu_i$ , the parameter  $\nu_i$  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  $m \in \{0, \dots, 59\}$  be the minute,  $h \in \{0, \dots, 23\}$  the hour, and  $d \in \{0, \dots, 6\}$  the day ( $Mon = 0, \dots, Sun = 6$ ) 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

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

$$w(d) = \sum_{k=1}^N P_k I(d_k = d), \quad (6)$$

and  $P_k$  is a probability weight that sums to one over  $k \in \{1, \dots, N\}$ , 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 (5) the standard kernel density estimate of the histogram of the number of e-mails sent by hour of day

(Figure 1). For this kernel smoothing we choose a gaussian kernel  $K(\cdot)$  with bandwidth  $\sigma$  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). Our initial estimate of the background rate density  $\hat{\mu}(t)$ , with equal probability weights, is plotted as the dashed curve in Figure 8. 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 3.3, 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 7 shows the estimated conditional intensity for officer 15 under model (4). The troughs in the estimated conditional intensity in Figure 7 correspond to times when few e-mails are sent and received.

### 3.2 Alternative Model

One shortcoming of models (3) and (4) is that the reply rate  $\theta_i$  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  $t$ :

$$\begin{aligned}\lambda_i(t) &= \nu_i \mu(t) + \sum_j \sum_{r_k^{ij} < t} g_{ij}(t - r_k^{ij}) \\ &= \nu_i \mu(t) + \sum_j \sum_{r_k^{ij} < t} \theta_{ij} \omega_i e^{-\omega_i(t - r_k^{ij})}.\end{aligned}\quad (7)$$

The triggering function,  $g_{ij}(t - r_k^{ij})$ , gives the contribution of the  $k^{\text{th}}$  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 a distinct  $\omega_{ij}$  and  $\nu_{ij}$  for each sender-recipient pair, 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 5) 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  $\theta_{ij}$ , 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 (4). (Instances when officers send e-mails to themselves have been removed, so the reply rate  $\theta_{ii}$  is not included in model (7).) A more in-depth comparison between models (4) and (7) is provided in Section 4.



### 3.3 Parameter Estimation

The parameters of models (3), (4), and (7) 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 (4). Models (3) and (7) can be estimated similarly. In particular, model (3) is just a special case of model (4) 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  $l^{\text{th}}$  e-mail was sent by officer  $i$ ,  $r_k^i$  be the time when the  $k^{\text{th}}$  e-mail was received by  $i$ , and  $N_i^{\text{send}}$  and  $N_i^{\text{rec}}$  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:

$$\psi_l^i = \begin{cases} 1 & \text{if } s_l^i \text{ is a non-reply message (background event)} \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

$$\chi_{kl}^i = \begin{cases} 1 & \text{if } s_l^i \text{ is a reply to message } r_k^i, \text{ where } s_l^i > r_k^i \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

The log-likelihood function (Ogata, 1978) for the conditional intensity defined in model (4) is given by

$$\begin{aligned} l_i(\Omega_i) &= \log L_i(\Omega_i) = \sum_{k=1}^{N_i^{\text{send}}} \log(\lambda_i(s_k^i)) - \int_0^T \lambda_i(t) dt \\ &= \sum_{k=1}^{N_i^{\text{send}}} \log(\lambda_i(s_k^i)) - \left( \nu_i + \theta_i \sum_{k=1}^{N_i^{\text{rec}}} [1 - e^{-\omega_i(T-r_k^i)}] \right), \end{aligned} \quad (10)$$

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 (10) 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:

$$l_i^c(\Omega_i) = \underbrace{\sum_{l=1}^{N_i^{send}} \psi_l^i \log(v_i \mu(s_l^i)) - \int_0^T v_i \mu(t) dt}_{l_i^\mu} + \underbrace{\sum_{k=1}^{N_i^{rec}} \left[ \sum_{\{l: s_l^i > r_k^i\}} \chi_{kl}^i \log(g_i(s_l^i - r_k^i)) - \int_{r_k^i}^T g_i(t - r_k^i) dt \right]}_{l_i^g}. \quad (11)$$

Since the true branching structure is unobserved, we estimate model parameters by maximizing the expected complete data log-likelihood, which is found by replacing  $\psi_l^i$  and  $\chi_{kl}^i$  in (11) with the estimated probabilities each event is either background or offspring:

$$B_l^i = \text{probability sent message } s_l^i \text{ is background} = \frac{\hat{v}_i \hat{\mu}(s_l^i)}{\hat{\lambda}_i(s_l^i)}, \quad (12)$$

$$O_{kl}^i = \text{probability receiving message } r_k^i \text{ triggers sending message } s_l^i = \begin{cases} \frac{\hat{g}_i(s_l^i - r_k^i)}{\hat{\lambda}_i(s_l^i)} & s_l^i > r_k^i \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

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 (5 and 6). This leads to the EM-type algorithm for estimating model (4):

Step 1. Initialize parameters estimates  $(\hat{v}_i^{(0)}, \hat{\theta}_i^{(0)}, \hat{\omega}_i^{(0)})$  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  $k \in \{1, \dots, N\}$  in (5) and (6). 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 (5) and (6), 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  $k \in \{1, \dots, N\}$  is non-reply (background) at iteration  $m + 1$ . The bandwidth  $\sigma$  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:

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

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

Step 5. Update  $m \leftarrow m + 1$  and repeat Steps 2–5 until convergence when

$$\left| \sum_i \left[ l_i(\hat{\Omega}_i^{(m+1)}) - l_i(\hat{\Omega}_i^{(m)}) \right] \right| < \epsilon \text{ for some small value } \epsilon \text{ (in practice we set } \epsilon = 10^{-3}\text{)}.$$

The estimators in Step 4 are found by setting the partial derivatives of the expected complete data log-likelihood (11), 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 (10). The convergence of this EM-type algorithm for the self-exciting models is apparent in Figure 9.

Parameter estimates, standard errors, and maximum log-likelihood values (10) for the Hawkes

process models (3, 4, and 7) are given in Tables 1, 2, and 3. Since estimated model (7) contains twenty-one reply rates  $\hat{\theta}_{ij}$  we instead present the average reply rate  $\hat{\theta}_i = \sum_j \hat{\theta}_{ij} \cdot N_{ij}^{rec} / N_i^{rec}$ , where  $N_{ij}^{rec}$  is the number of messages officer  $i$  received from  $j$ , for each officer in Table 3. Notice that the parameter estimates for models (4) and (7) presented in these tables are similar. This result is consistent with model (4) being contained within model (7) (it is the case with  $\theta_{ij} = \theta_i$  for each sender  $j$  and recipient  $i$  pair).

The standard errors in Tables 1, 2, and 3 are found by simulating each model 100 times (Appendix 1) 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 8). The gray error bound in this figure is formed by simulating fitted model (7) 100 times (Appendix 1) 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 1 and 2 reveals that model (4) outperforms model (3) 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 (7) (see Table 3) are greater than model (4) for each officer; however, due to the large number of parameters, model (7) does not outperform model (4) 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 4.4.

## 4 IkeNet Analysis

### 4.1 Characterizing E-mail Communication Behavior

The parameter estimates in Table 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  $\hat{v}_i/N_i^{send}$ . In other words,  $\hat{v}_i$  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 (4), 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,  $\hat{v}_i/N_i^{send}$  ranges between 42% and 83%, and the estimated overall percentage of e-mails sent in the network that are not replies is  $\sum_{i=1}^{22} \hat{v}_i/N \approx 55\%$ , 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. 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 ( $\hat{v}_i/N_i^{send}$ ) 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 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. Figure 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.

## 4.2 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, 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 10 and 11 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.

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 ( $\hat{v}_i N_{ij}^{send} / N_i^{send}$ ) above threshold  $c_2$ . That is

$$Y_i(c_1, c_2) = \sum_j \mathbf{1}\{\hat{\theta}_{ij} > c_1, \hat{v}_i N_{ij}^{send} / N_i^{send} > c_2\} \quad (14)$$

where  $\mathbf{1}$  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 (7). Intuitively, officers that initiate many e-mail threads and are very responsive 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 (14), denoted by  $Y^{(k)}$  for  $k = 1, \dots, 4$ . Let  $A = \{\hat{\theta}_{ij} | i \neq j\}$  be the set of estimated reply rates from officers  $i$  to  $j$ ,  $B = \{\hat{y}_i N_{ij}^{send} / N_i^{send} | i \neq j\}$  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 predictor  $Y^{(1)}$ , threshold  $c_1 = \bar{\theta} = 0.45$  and threshold  $c_2 = 4.79$  is the median of set  $B$ . For predictor  $Y^{(2)}$ , 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 4 and 5 lists several predictors of network leadership and the Pearson, 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 10 and 11). 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 4 and 5 show that predictor  $Y$ , for the four 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 four 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 12 represent the Pearson ( $r_p$ ), Spearman ( $r_s$ ), and Kendall ( $\tau$ ) correlations between the predictors ( $Y$ ,  $N^{send}$ , and  $N^{rec}$ ) and the leadership survey votes. Panel (a) shows that predictor  $Y$  has higher correlations with the team leadership votes than the naive predictors ( $N^{send}$  and  $N^{rec}$ ) for the four sets of thresholds considered.  $Y^{(1)}$  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. Panel (b) also shows that predictor  $Y$  has higher correlations with the military leadership votes 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  $N^{send}$ .  $Y^{(1)}$  and  $Y^{(2)}$  perform the best overall at predicting and ranking military leaders.

### 4.3 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 12 shows that for very reasonable threshold selections (i.e. means, medians, and third quartiles as discussed in Section 4.2), predictor  $Y$  performs much better at ranking and estimating leadership scores than the naive predictors  $N^{send}$  and  $N^{rec}$ . Table 4 also shows that  $Y$  is generally able to identify the top four teams leaders with slight variations in order. For all threshold values considered in Tables 4 and 5,  $Y$  does a better job than  $N^{send}$  or  $N^{rec}$  at identifying the top leaders.

In Figure 13 we further assess the sensitivity of  $Y(c_1, c_2)$  to the threshold values. Each panel shows 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 horizontal line in each panel is the respective correlation between predictor  $N^{send}$  and the leadership votes.

The correlations corresponding to predictor  $Y(c_1, c_2)$  typically fall above the horizontal line in each panel as the thresholds vary; this indicates that  $Y(c_1, c_2)$  is more strongly associated with the leadership votes than  $N^{send}$  for a wide variety of threshold combinations. 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 the 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.

#### 4.4 Model Comparison and Diagnostics

The maximized log-likelihoods for the network and corresponding AIC values are provided in [Table 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 Hawkes process models (3, 4, and 7) described in Section 3. The Hawkes process model (3) 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 (4) is higher than the model with the stationary background rate (3). 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  $\hat{\mu}(t)$  (4 and 7) with the completely parametric model (3). The Hawkes process model (7), which incorporates

pairwise interactions between officers, fits the data slightly more closely than model (4) 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 (7). Although, due to the overall sparsity in the IkeNet e-mail network (Figure 5), about 15% of the estimated parameters in (7) are equal to zero. Comparison of models (4) and (7) 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 (7) to predict and rank network leaders was shown in Section 4.2.

The simulation procedure described in Appendix 1 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 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 3).

Using the first 11 months ( $T = 7920$  hours) of e-mail data in the training set we estimate models (3), (4), and (7) with the EM-type algorithm described in Section 3.3. 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 (5) and (6) 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 1, 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 14, the 0.025 and 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 non-stationary background rates in models (4) and (7) 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 (3) with the stationary background term. Moreover, the similarity between the performances of models (4) and (7) 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

$$\tau_k^i = \Lambda(s_k^i) = \int_0^{s_k^i} \lambda_i(t) dt. \quad (15)$$

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 15 for the stationary Poisson process model and all Hawkes process models (3, 4, and 7) of e-mail network traffic considered in this paper. 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 (4) than the plot for model (3), and likewise when comparing models (7) and (4). This claim is supported by the decreasing values of the Kolmogorov-Smirnov test statistics in Table 6, which compare the transformation  $\{U_k\}$  for

each network model with the uniform distribution.

## 5 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 16 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 pronounced peak in activity between the dates when Jeffrey 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 17 (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 (3, 4, and 7) to the Enron data using the EM-type algorithm described in Section 3.3. The maximum log-likelihood and AIC values for the network are provided in Table 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 (3) 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 (4). 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 (3). Due to the large number of parameters, the AIC for model (7) is much larger than model (4). However, like IkeNet, the Enron e-mail net-

work is sparse in the number of messages sent between pairs of individuals. In fact, approximately 94% of the estimated parameters for model (7) 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 4.4) indicate the Hawkes process models for the Enron network account for the clustering in the times when e-mails are sent significantly better than the stationary Poisson model.

Table 8 displays the mean percentage of reply and non-reply messages estimated from the self-exciting models (3, 4, and 7) of the Enron and IkeNet e-mail networks. These percentages are quite similar for both networks: model (3) 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-stationary background rate in models (4) and (7). Table 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 (4), 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.

## 5.1 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.

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 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 10 presents features of e-mail communication estimated from self-exciting models (4) and (7), 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 ( $\hat{y}/N^{send}$ ), the estimated mean reply rate ( $\hat{\theta}$ ), and the predictor  $Y$  (equation 14). 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 4.2).<sup>2</sup>

---

<sup>2</sup>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.

The features considered in [Table 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 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 consider simple binary classification rules. For instance, CEOs send far fewer e-mails, on average, than other Enron users (Table 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 18 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 18 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 Director / Managing Director (panel c).

The ROC curves corresponding to the binary classification of CEO status (panel a) indicate that the 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) indicate that predictors  $Y^{(1)}$  and  $Y^{(3)}$  perform better than the naive predictors. 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 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.

The ROC curves corresponding to the binary classification of 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 Director / Managing Director status.

While binary classification rules are a simple way to infer Enron leadership, it is somewhat unclear from the ROC plots which predictors perform the best, and whether there are any substantial differences in the performance of the various predictors. To better evaluate the proposed predictors of leadership, particularly for the Enron network, we consider a modeling approach in the next section.

## 5.2 Regression Models for Predicting Leadership

In this section we consider regression models for predicting IkeNet and Enron leadership status using predictors derived solely from the e-mail log data (sender, recipient, and timestamp) for each network. The response variables of interest are the IkeNet team and military survey leadership rankings, and the Enron leadership roles coded as binary variables indicating CEO, President / Vice President, and Director / Managing Director status. For example, the binary response variable for CEO is coded as 1 if the employee is a CEO, and 0 otherwise. Logistic regression is used to predict the Enron leadership roles, and standard least squares regression is used to predict the IkeNet leadership survey rankings.

A set of five user-specific predictor variables are used to build the leadership models:  $N^{send}$ ,  $N^{rec}$ ,  $Y$ , and two additional predictors named  $R$  and  $I$  which incorporate features from the fitted Hawkes process models but are simpler than  $Y$  and do not involve interactions. We define predictors  $R$  and  $I$  for each user  $i$  as

$$R_i(c_1) = \sum_j \mathbf{1}\{\hat{\theta}_{ij} > c_1\},$$

$$I_i(c_2) = \sum_j \mathbf{1}\{\hat{v}_i N_{ij}^{send} / N_i^{send} > c_2\}$$

for some choice of thresholds  $c_1$  and  $c_2$ . For predictors  $R(c_1)$ ,  $I(c_2)$ , and  $Y(c_1, c_2)$  we consider the same types of thresholds discussed in Section 4.2. Namely, the mean, median, and third quartile of the estimated reply rates and estimated number of non-reply e-mails (background events) between pairs of users in each network. A motivation for considering these additional predictors is that



perhaps in social networks with hierarchies as complex as Enron certain leadership roles are better quantified by either the responsiveness of the user to e-mails (as measured by  $R$ ) or the thread initiation rate (as measured by  $I$ ), and not a combined measure as quantified by  $Y$ . Also, these simpler predictors may be useful when considering multivariate models for leadership.

Figure 19 shows the AIC scores for simple and multiple regression models of IkeNet team and military leadership rankings, fit to all combinations of the five predictor variables  $N^{send}$ ,  $N^{rec}$ ,  $R$ ,  $I$ , and  $Y$ . For example, the AIC scores for the simple regression models of team leadership fit to  $N^{send}$  and  $N^{rec}$  are plotted in the first two rows of Figure 19. The three AIC scores for predictor  $R$  (third row) correspond to three different simple regression models for team leadership fit to  $R(c_1)$  using the three threshold  $c_1$  considerations. Similarly, the simple regression models fit to  $I$  and  $Y$  also have several AIC scores which correspond to different threshold selections. The sixth row of Figure 19 shows the AIC score for the bivariate regression model fit to  $N^{send}$  and  $N^{rec}$ . The other bivariate models involve predictors  $R$ ,  $I$ , and  $Y$  with different thresholds combinations. For example, there are six bivariate models fit to predictors  $R(c_1)$  and  $I(c_2)$  (row 13) using three threshold values  $c_1$  for  $R$  and two threshold values  $c_2$  for  $I$ . The last three rows of Figure 19 show the distribution of AIC scores for multiple regression models fit to all combinations of three, four, and five predictors.

The regression models for IkeNet leadership (Figure 19) which incorporate features from the fitted Hawkes process models (plotted as circles) generally perform better, in terms of AIC, than

the models with only the basic descriptive statistics  $N^{send}$  and  $N^{rec}$  (plotted as triangles). The univariate models with predictors  $R$ ,  $I$ , and  $Y$  perform relatively well and have the lowest AIC scores amongst all models for certain thresholds; this indicates that the descriptive statistics  $N^{send}$  and  $N^{rec}$  offer little additional information beyond these predictors. Moreover, many of the multivariate models also show substantial improvement over the basic descriptive statistic models in terms of AIC. For instance, the bivariate models for military leadership with predictors  $R$ ,  $I$ , and  $Y$  (rows 7-15) consistently perform better than the best fitting descriptive statistic model  $N^{send}$ . The same relationship also holds true for team leadership, with the only exception being the bivariate models with  $Y, N^{send}$  (row 11) and  $Y, N^{rec}$  (row 12), which perform nearly as well as  $N^{send}$  for two thresholds selections, and substantially better for the other two thresholds. The regression models with more than three predictors often do not perform as well as the univariate or bivariate models. The model with the highest AIC score has all five predictors for team leadership and four predictors ( $R$ ,  $Y$ ,  $N^{send}$ ,  $N^{rec}$ ) for military leadership. This is perhaps due to collinearity since the AIC penalizes for adding in redundant predictors.

The AIC scores for the logistic regression models of Enron leadership roles are plotted in [Figure 20](#). All predictors are log-transformed due to the overall sparsity of the Enron e-mail dataset. For CEO status, the bivariate logistic models with predictors  $I, N^{send}$  (row 9);  $I, N^{rec}$  (row 10); and  $I, Y$  (row 15) perform better than the best fitting descriptive statistic model  $N^{rec}$  for some threshold selections. Some multivariate models for CEO status with three or more predictors have the lowest AIC scores, however, the performance of these models appears sensitive to threshold selection. For

instance, the model with the highest AIC has all five predictors. Since there are only 4 CEOs in the Enron social network (out of 151 employees) it is difficult to train any classifier for CEO status due to the small and unbalanced sample.

The logistic regression models for President / Vice President status with predictors  $R$ ,  $I$ , and  $Y$  generally perform substantially better than the basic descriptive statistic models with  $N^{send}$  and  $N^{rec}$  in terms of AIC. Moreover, the univariate and bivariate models with predictor  $R$  (rows 3,7,8,13,14) consistently outperform the descriptive statistic models. Since the coefficients for  $R$  in these models are always positive and significant, this indicates that responsiveness is a strong predictor of Enron President / Vice President status. The model with the highest AIC has univariate predictor  $I(c_2)$  with  $c_2$  set to the mean thread initiation rate over all officer pairs. This indicates that thread initiation is not an important feature for the prediction of President / Vice President status.

The univariate logistic regression models for Director / Managing Director status generally perform the best in terms of AIC. The univariate model with predictor  $Y$  has a lower AIC score than the best fit descriptive statistics model  $N^{send}$  for most threshold selections, and appears to be the best classifier overall. The model with the the highest AIC score has all five predictors, and this is perhaps due to collinearity. Since Director / Managing Director status is further down the Enron hierarchy than President / Vice President status it is not surprising that there is less substantial improvement in modeling when considering predictors  $R$ ,  $I$ , and  $Y$ , as Directors / Managing Directors probably interact with employees more directly.

## 6 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 (7) quantified these additional features, which we attempted to combine into a measure  $Y(c_1, c_2)$  (equation 14) 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 13).

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  $Y$ , indicating that the features of e-mail communication quantified by the fitted Hawkes process models help distinguish Presidents / Vice Presidents from other users in the Enron social network. 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.

Simple and multivariate regression models for IkeNet and Enron leadership were considered to evaluate and compare the performance of the fitted Hawkes process predictors ( $R$ ,  $I$ , and  $Y$ ) and the basic descriptive statistics ( $N^{send}$  and  $N^{rec}$ ). In terms of AIC scores, the regression models with the fitted Hawkes process predictors generally perform better than the regression models with only descriptive statistic predictors for IkeNet team and military leadership survey rankings and Enron President / Vice President status. For Enron CEO status, there is a slight improvement in AIC for

some multivariate models which incorporate the fitted Hawkes process predictors; although, it is difficult to say whether this improvement is meaningful since only 4 of the 151 Enron users are CEOs. The univariate logistic model with predictor  $Y$  is the best overall classifier for Enron Director / Managing Director status, in terms of AIC. Although, the fitted Hawkes process predictors more substantially improve the logistic models for Enron President / Vice President status than Director / Managing Director status. One possible explanation is that Directors / Managing Directors are further down the Enron hierarchy, and probably interact with the employees they supervise more directly.

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. Hence, 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.

Another important distinction between the IkeNet and Enron datasets is that leadership ground-truth for IkeNet is in the form of counts and rankings from the aggregated survey votes, while Enron leadership roles are binary. Therefore, the prediction problems and corresponding evaluations are slightly different. More examples on networks with these types of leadership and communication

data would be useful in the future to further elucidate how and when the proposed methods offers advantages for inferring leadership.

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 nonparametric self-exciting model, as described in [Marsan and Lengliné \(2008\)](#), would allow for more flexibility in estimating the background and triggering intensities. However, such models require more computational effort and are less easily interpretable than the exponential forms considered in this paper. 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.

## Acknowledgements

This research is supported by ARO MURI grant W911NF-11-1-0332, AFOSR-MURI grant FA9550-10-1-0569, NSF grants DMS-1045536 and DMS-0968309, and ONR grant N000141210838.

## Appendix 1: Simulation

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 (4) and (7) this can be done using the method of Poisson thinning (Lewis and Shedler, 1979) described in the following algorithm:

### Algorithm A

- Step 1. Let  $\mu^*$  be the maximum of  $\hat{\mu}(t)$  over  $[0, T]$ .
- Step 2. Draw  $N_b^*$  from  $Pois(\hat{\nu}_i \mu^* T)$  (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  $\{Z_l : l = 1, \dots, N_b^*\}$  from  $Unif(0,1)$  and set  $s_l^i = T \cdot Z_l$ .
- Step 4. For each event  $l = 1, \dots, N_b^*$  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  $G_i^{send}(0) = \{s_k^i : k = 1, \dots, N_i^{send}(0)\}$  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  $\{j : j \in \{1, \dots, 22\}, j \neq i\}$  with corresponding weights  $\{N_{ij}^{send} : j \in$



$\{1, \dots, 22\}, j \neq i\}$ , 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  $i = 1, \dots, 22$ . To simulate the background process (non-reply e-mail send times) for model (3) 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  $G_i^{rec}(v) = \{r_k^i : k = 1, \dots, N_i^{rec}(v)\}$  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 (3) and (4), 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  $\{Z_l : l = \eta + 1, \dots, \eta + n_k^{(v+1)}\}$  from  $Unif(0,1)$ .

Step 4. The reply times  $\{s_l^i : l = \eta + 1, \dots, \eta + n_k^{(v+1)}\}$  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  $G_i^{send}(v + 1) = \{s_l^i : l = 1, \dots, N_i^{send}(v + 1)\}$  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  $\{j : j \in \{1, \dots, 22\}, j \neq i\}$  with corresponding weights  $\{N_{ij}^{send} : j \in \{1, \dots, 22\}, j \neq i\}$ , where  $N_{ij}^{send}$  is the observed number messages sent from  $i$  to  $j$ .

Algorithm B is repeated for each officer  $i = 1, \dots, 22$  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 (7) 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 (7) the number of replies generated for each e-mail received by  $i$  depends on the sender  $j$ .

## 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.
- 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.

- 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, 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.
- 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.
- 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 théorème de knight. In *Séminaire de Probabilités 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.

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.

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.

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.

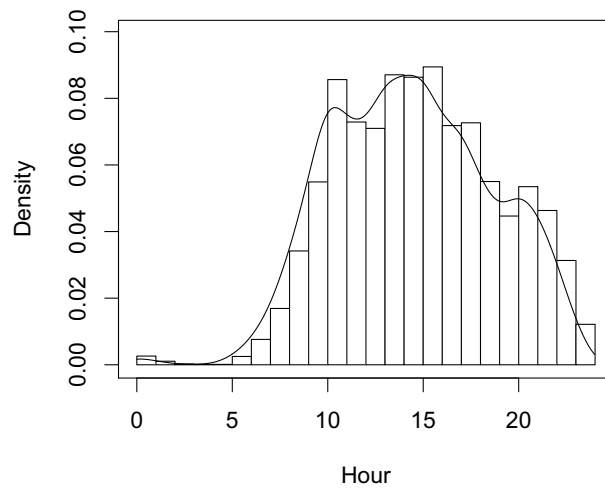


Figure 1. Histogram density of the number of e-mails sent each hour of the day over the one-year observation window. The smoother curve was formed using kernel density estimation with a fixed bandwidth (Scott, 1992).

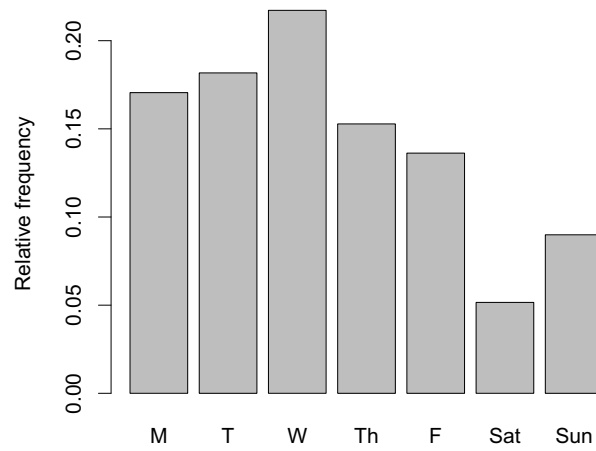


Figure 2. Proportion of e-mails sent each day of the week over the one-year observation window.



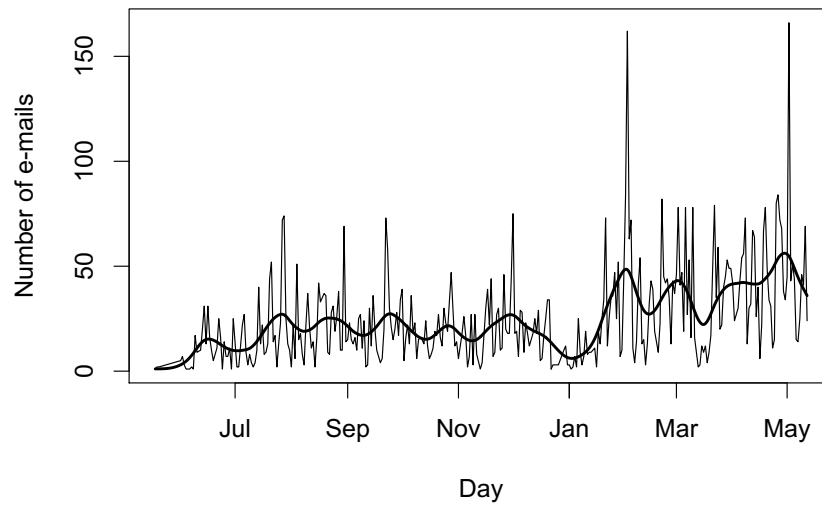


Figure 3. Time series plot of number of e-mails sent by date.

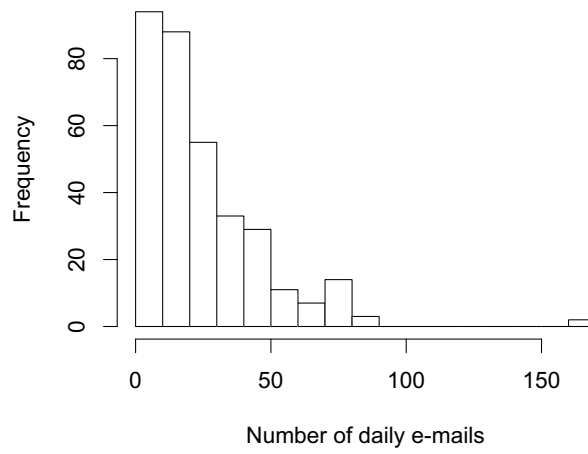


Figure 4. Histogram of the number of daily e-mails.

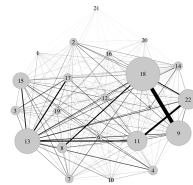


Figure 5. Plot of the IkeNet e-mail network with node sizes proportional to the number of e-mails sent by each officer, and edge widths proportional to the number of e-mails sent between officers.

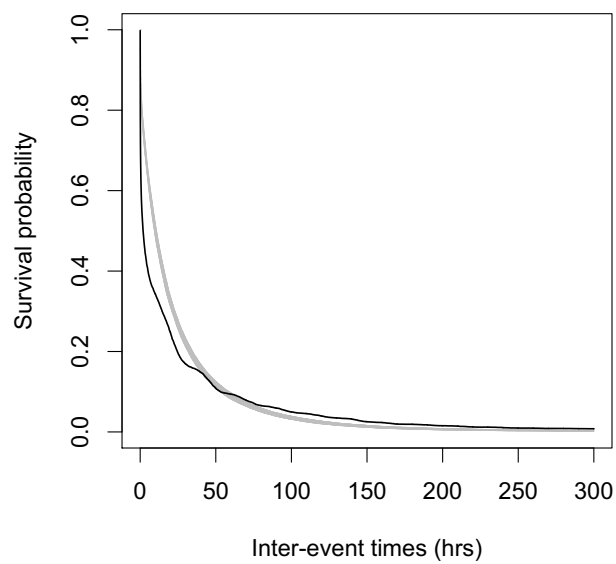


Figure 6. Survivor plot of the inter-event times for e-mails sent by each officer in the network (black line). A 95% confidence envelope was formed by simulating the network 100 times from the fitted model (3) and computing the survivor function for each realization. The pointwise 0.025 and 0.975 quantiles of the simulated survivor functions are plotted in gray.

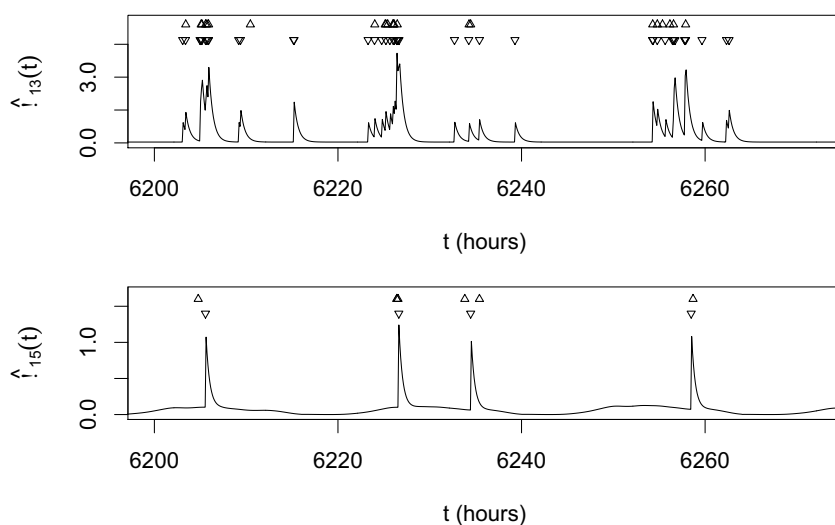


Figure 7. Top panel shows the estimated conditional intensity for officer 13 over a three-day period using the Hawkes model with the stationary background rate (3). The bottom panel shows the estimated conditional intensity for officer 15 over the same three-day period using the Hawkes model with the non-stationary background rate (4). The downward triangles represent the times when messages are received, while the upward triangles represent the times when messages are sent.

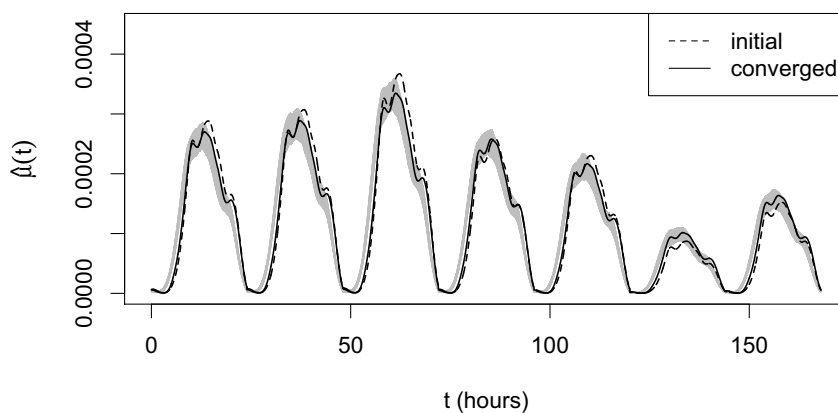


Figure 8. Estimated background rate density  $\hat{\mu}(t)$  for the IkeNet e-mail network (solid black curve) using model (7) after convergence of the EM-type algorithm. The dashed curve is the initial estimate of the background rate density using equal probability weights. This figure only shows one 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 (7), and the pointwise 0.025 and 0.975 quantiles are plotted in gray.

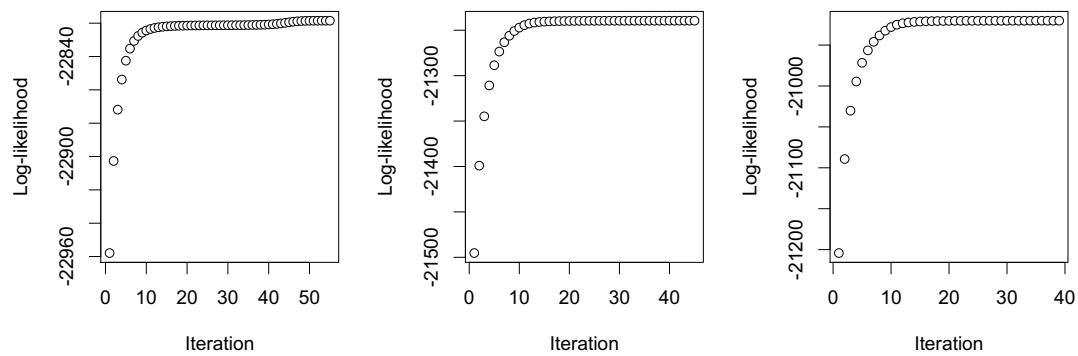


Figure 9. Scatter plots showing the convergence of the EM-type algorithm, in terms of log-likelihood, for estimating the self-exciting models (3, 4, and 7, respectively).

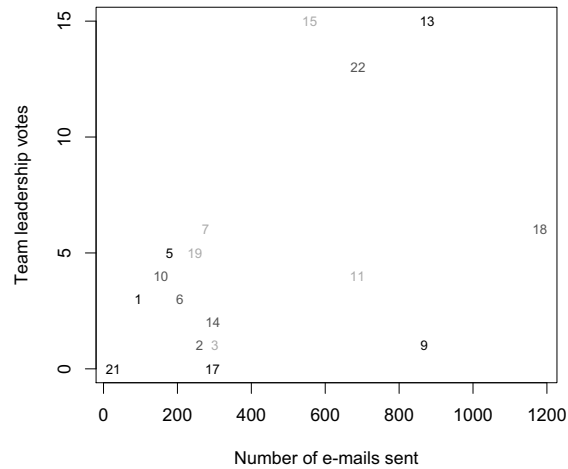


Figure 10. Plot of the total number of e-mails sent versus the aggregate number of votes each officer received for perceived team leadership ( $r = 0.52$ ). Votes are based on a survey which asked each officer to list up to five other officers in the network that he or she considered to be a strong team leader.



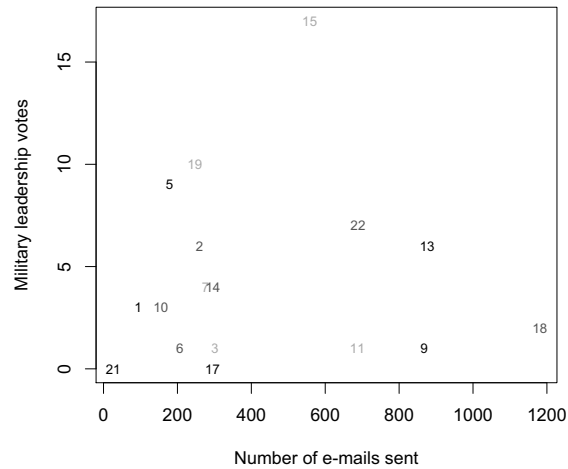


Figure 11. Plot of the total number of e-mails sent versus the aggregate number of votes each officer received for perceived military leadership ( $r = 0.13$ ). Votes are based on a survey which asked each officer to list up to five other officers in the network that he or she considered to be a strong military leader.

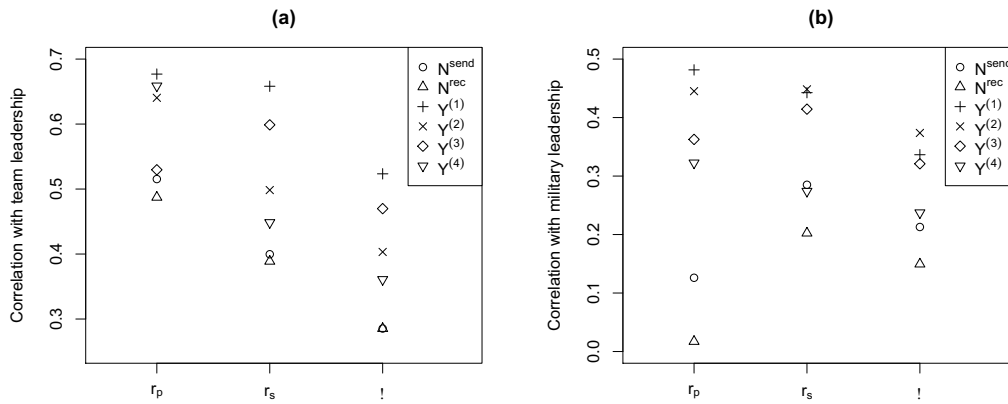


Figure 12. Pearson ( $r_p$ ), Spearman ( $r_s$ ), and Kendall ( $\tau$ ) correlations between the predictor variables and the team (panel a) and military (panel b) leadership votes.  $Y^{(k)}$  for  $k = 1, \dots, 4$  denotes predictor  $Y(c_1, c_2)$  for the four sets of thresholds  $c_1$  and  $c_2$  discussed in Section 4.2. Both panels show that predictor  $Y$  is more strongly correlated with the leadership votes than  $N^{send}$  and  $N^{rec}$ .

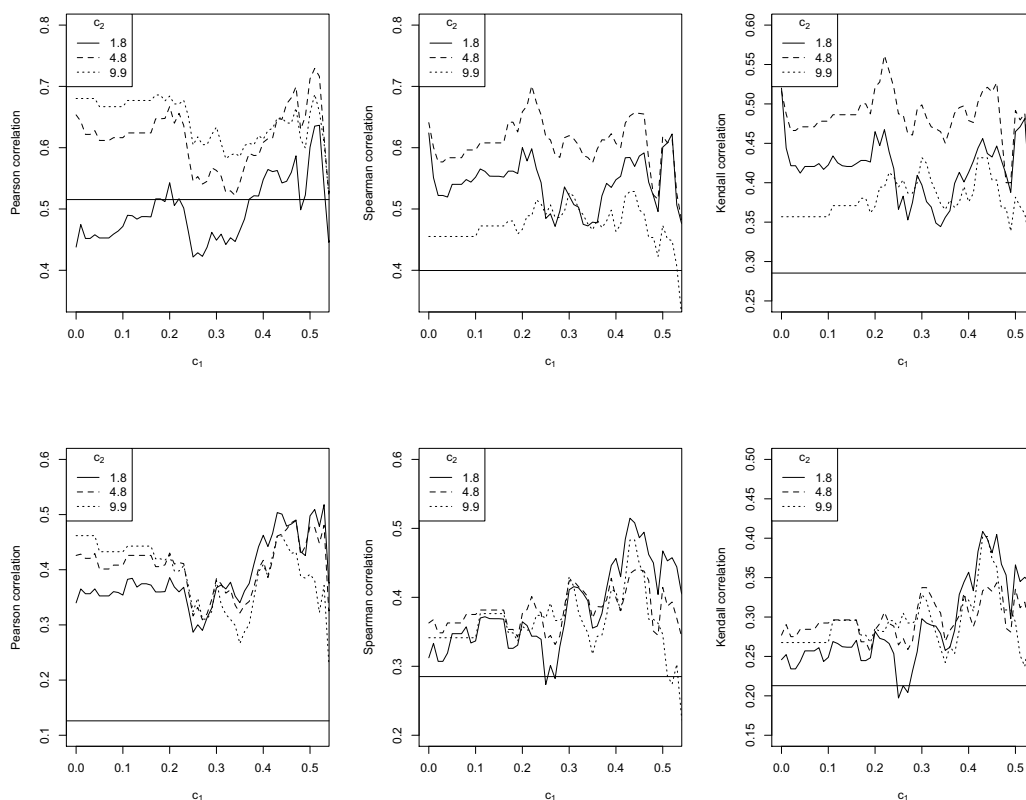


Figure 13. Sensitivity plots for the Spearman, Pearson, and Kendall correlations between predictor  $Y(c_1, c_2)$  and the team leadership votes (upper three panels) and military leadership votes (lower three panels) for different values of thresholds  $c_1$  and  $c_2$ . The lines in each plot give 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 horizontal line in each panel 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  $N^{send}$ .

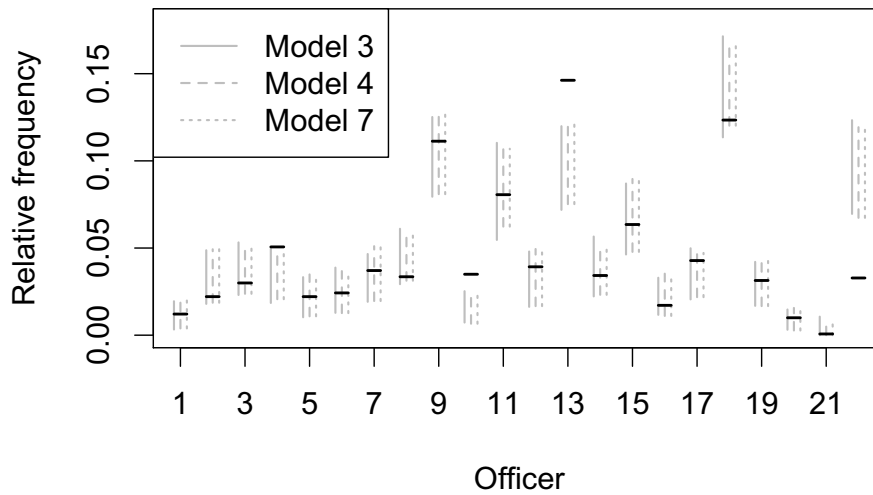


Figure 14. Comparison of the simulated and observed proportion of e-mails sent by each officer over a period of one month (720 hours). The gray vertical lines are the pointwise 0.025 and 0.975 quartiles for the proportions generated from 100 simulations of the IkeNet e-mail network using the models estimated from the training set (first 11 months of e-mail data). The black horizontal lines are the observed proportions from the validation set.

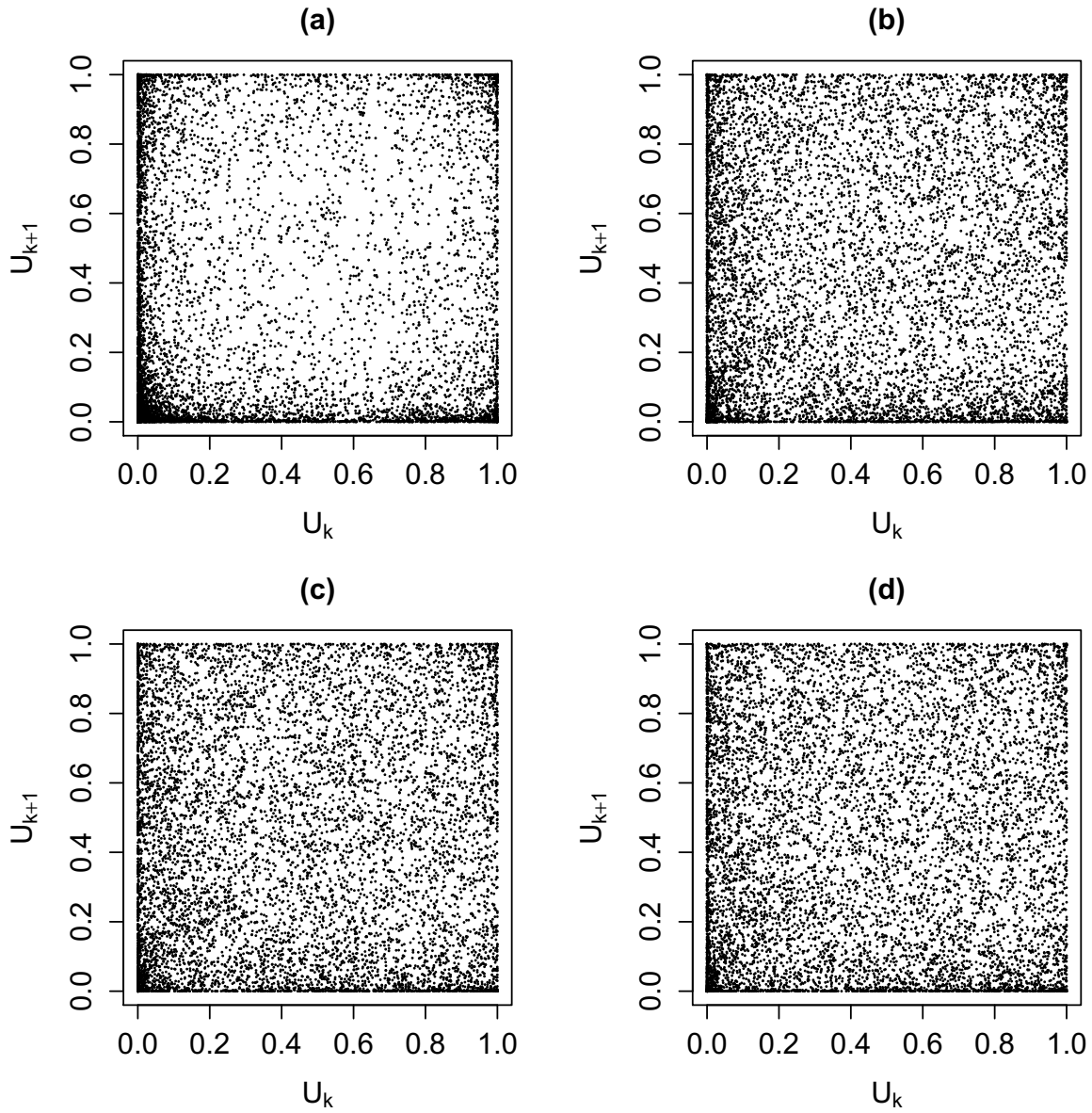


Figure 15. (a-d) Plot of  $U_{k+1}$  versus  $U_k$  for the stationary Poisson process model and Hawkes process models (3, 4, and 7) of e-mail activity on the network, respectively.

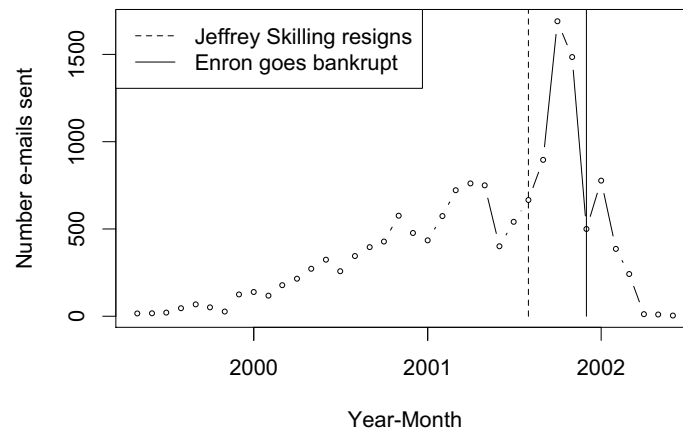


Figure 16. Time series plot of number of e-mails sent each month between May 1999 and June 2002 in the Enron dataset.

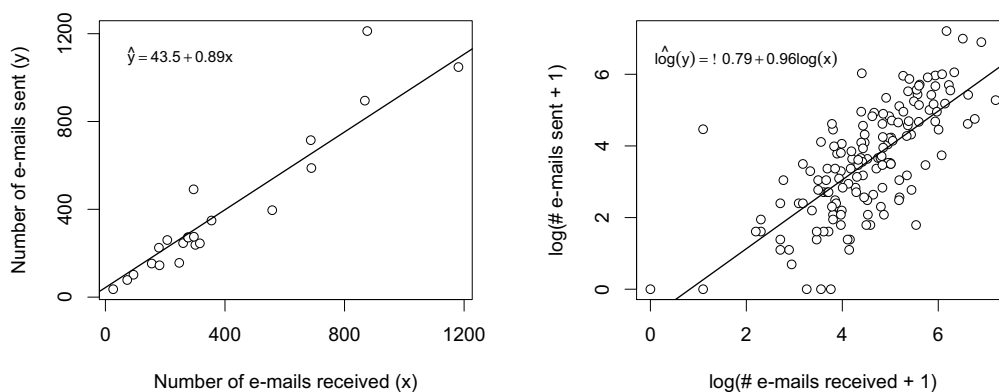


Figure 17. Left Panel: Scatter plot of the total number of e-mails received ( $x$ ) versus the total number of e-mails sent ( $y$ ) by each officer in the IkeNet dataset. The scatter plot and regression line show a strong association between the raw number of e-mails sent and received ( $r = 0.95$ ). 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 ( $r = 0.72$ ).

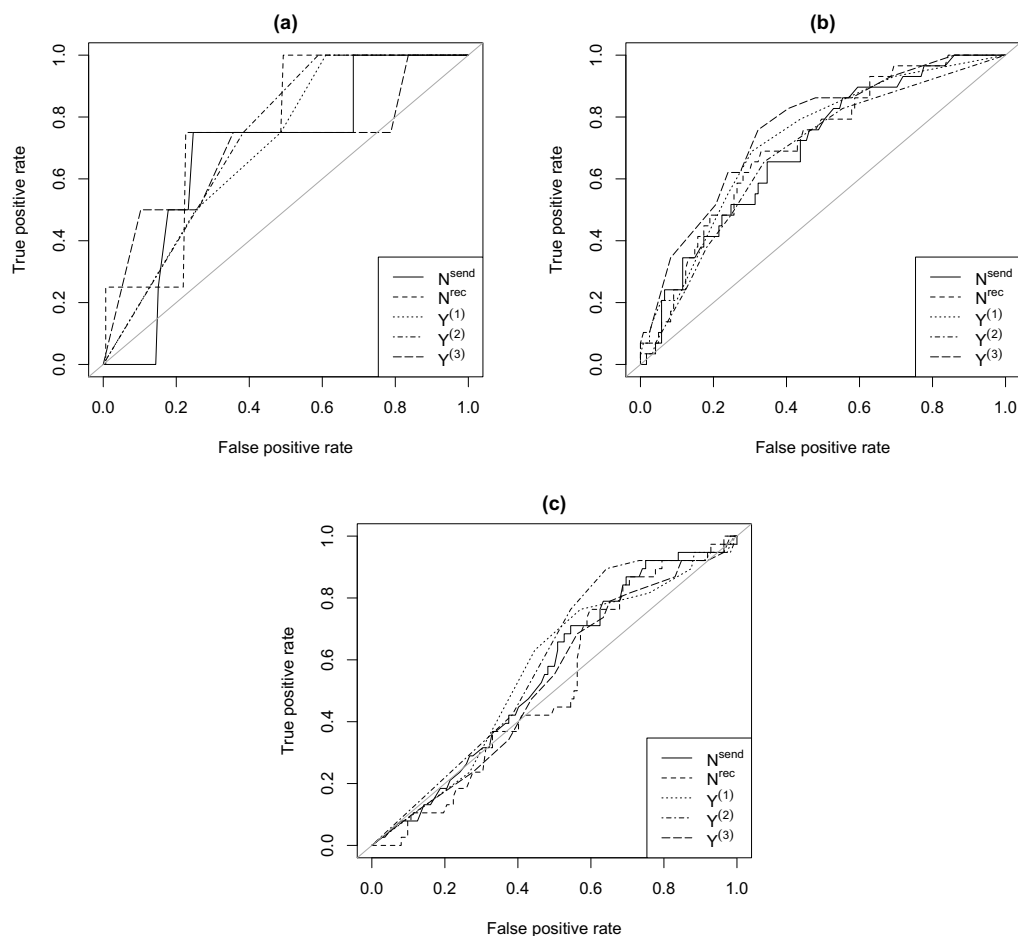


Figure 18. ROC curves corresponding to the binary classification of different Enron leadership roles. For each predictor of leadership ( $N^{send}$ ,  $N^{rec}$ ,  $Y$ ) a cut-off value is chosen to classify each user as either a leader or non-leader. The ROC curves are constructed by considering all 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 Director / Managing Director, respectively.



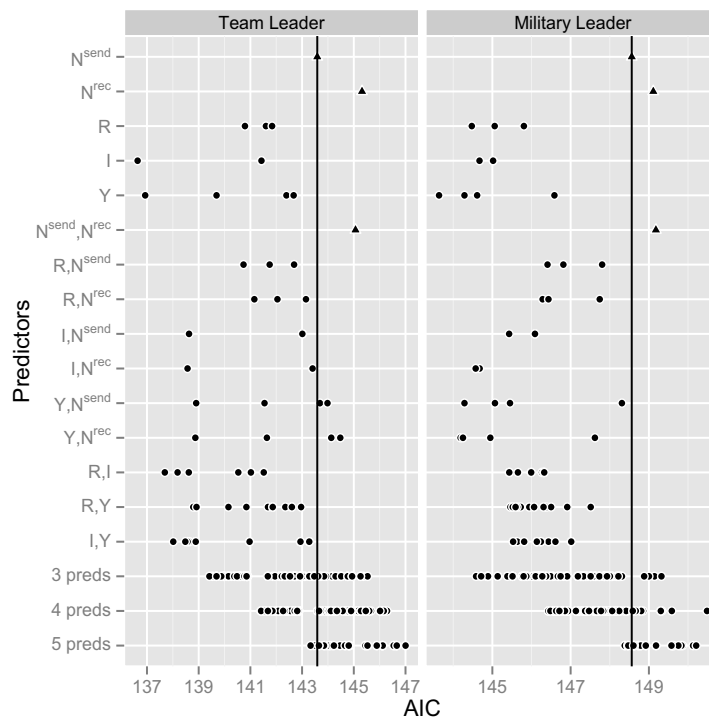


Figure 19. AIC scores for regression models for predicting IkeNet team and military survey leadership rankings. Considered are all combinations of predictors  $N^{send}$ ,  $N^{rec}$ ,  $R$ ,  $I$ , and  $Y$ . The different thresholds discussed in Section 4.2 are used to build the models with predictors  $R$ ,  $I$ , and  $Y$ . For example, the four points in the row for predictor  $Y$  are the AIC scores for four simple regression models using four different sets of thresholds. The last three rows show the distribution of AIC scores for multiple regression models built with all combinations of three, four, and five predictors. The triangles correspond to regression models constructed with only basic descriptive statistic predictors ( $N^{send}$  and  $N^{rec}$ ), and the vertical line indicates the best model fit to the descriptive statistic predictors. The circles correspond to regression models with at least one fitted Hawkes process predictor ( $R$ ,  $I$ , and  $Y$ ). Note, only circles are used in the last three rows since all models with three or more predictors are built with at least one fitted Hawkes process predictor.

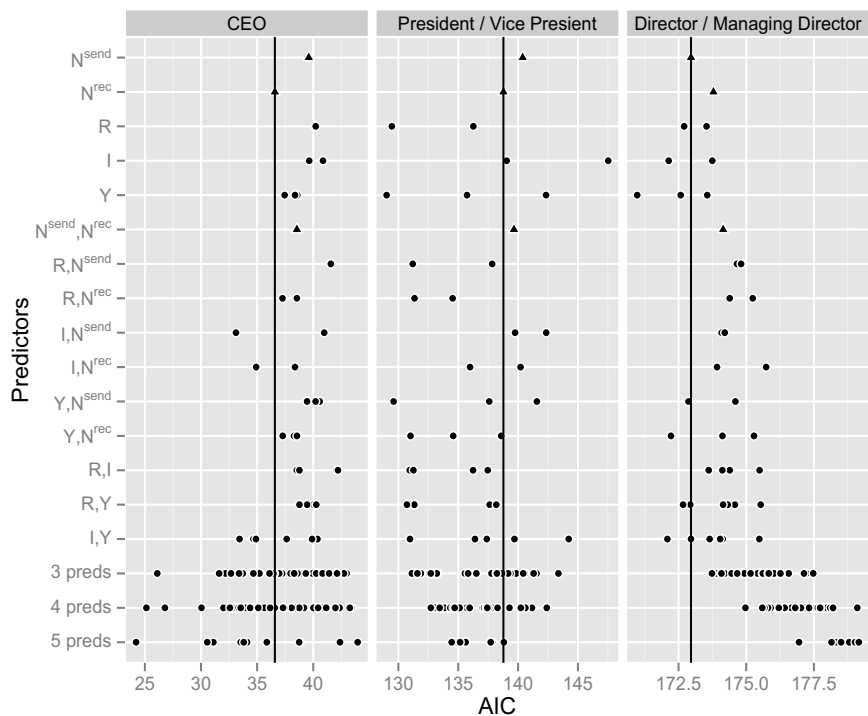


Figure 20. AIC scores for logistic regression models for predicting Enron leadership roles. Considered are all combinations of log-transformed predictors  $N^{send}$ ,  $N^{rec}$ ,  $R$ ,  $I$ , and  $Y$ . The different thresholds discussed in Section 4.2 are used to build the models with predictors  $R$ ,  $I$ , and  $Y$ . The triangles correspond to regression models constructed with only basic descriptive statistic predictors ( $N^{send}$  and  $N^{rec}$ ), while the circles correspond to regression models with at least one fitted Hawkes process predictor ( $R$ ,  $I$ , and  $Y$ ).

Table 1. Parameter estimates, standard errors, and maximum log-likelihood values for model (3). 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. Parameter estimates, standard errors, and maximum log-likelihood values for model (4). Standard errors are computed by the root-mean-square deviation from 100 simulations of the estimated model.

$i$	$N_i^{send}$	$\hat{v}_i/N_i^{send}$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l_i(\hat{\Omega}_i)$
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 3. Parameter estimates, standard errors, and maximum log-likelihood values for model (7). The column labeled  $\hat{\theta}_i$  gives the estimated average reply rate for each officer  $\hat{\theta}_i = \sum_j \hat{\theta}_{ij} \cdot N_{ij}^{rec} / N_i^{rec}$ . Standard errors are computed by the root-mean-square deviation from 100 simulations of the estimated model.

$i$	$N_i^{send}$	$\hat{v}_i / N_i^{send}$	$\hat{\theta}_i$	$\hat{\omega}_i$	$l_i(\hat{\Omega}_i)$
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

Table 4. Predictors of team leadership.

Predictor	$r_p$	$r_s$	$\tau$	Estimated top 4 leaders
$N^{send}$	0.52*	0.40·	0.29·	18, 13, 9, 22
$N^{rec}$	0.49*	0.39·	0.29·	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 5. Predictors of military leadership.

Predictor	$r_p$	$r_s$	$\tau$	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	0.41	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.

Table 6. Number of parameters ( $\rho$ ), 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.

	$\rho$	$l(\hat{\Omega})$	AIC	KS
Stationary Poisson	22	-32347.4	64738.9	0.39
Hawkes model (3)	66	-22818.5	45769.0	0.17
Hawkes model (4)	66	-21239.5	42611.0	0.15
Hawkes model (7)	506	-20920.2	42852.5	0.14



Table 7. Number of parameters ( $\rho$ ), 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.

	$\rho$	$l(\hat{\Omega})$	AIC	KS
Stationary Poisson	151	-85605.0	171512.0	0.42
Hawkes model (3)	453	-75031.4	150968.8	0.28
Hawkes model (4)	453	-70721.7	142349.4	0.27
Hawkes model (7)	22952	-68925.9	183755.9	0.25

Table 8. Mean percent non-reply messages ( $\sum_i \hat{v}_i / N$ ), mean percent reply messages ( $\sum_i \hat{\theta}_i \cdot N_i^{rec} / N$ ), average reply time ( $\sum_i N_i^{send} \hat{\omega}_i^{-1} / N$ ), 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 (hrs)
IkeNet	Hawkes model (3)	50.2%	49.8%	0.4 (0.28, 0.6)
	Hawkes model (4)	54.4%	45.6%	0.31 (0.22, 0.48)
	Hawkes model (7)	54.6%	45.4%	0.31 (0.22, 0.43)
Enron	Hawkes model (3)	50%	50%	68.47 (1.69, 111.28)
	Hawkes model (4)	59.5%	40.5%	48.5 (1.63, 60.52)
	Hawkes model (7)	54.6%	45.4%	61.19 (1.53, 49.16)

Table 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 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)
Vice President	25	0.56 (0.3)	0.27 (0.27)	4.4 (3.3)	2.8 (2.3)	9.7 (6)
Managing 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 4.2).