Final Report

Modeling Messaging Activities in a Network: Enron Case Study

Submitted by

Svitlana Volkova, Project Manager, Final Report Coordinator
Karlo Perica, Experiment Designer, Midterm Report Coordinator
D. Michael Parrish, Statistic Analyst

to

Prof. N.H. Lee, Instructor
Francisco Sanchez Vega, Teaching Assistant

550.433 Monte Carlo Methods

Submitted 7 Dec 2011
1 Introduction

Due to recent advances in social networking and other on-line messaging activities, for example, blogging and tweeting, an understanding of a network structure, detecting abnormal activities and communities within a network are essential.

The main goal of this project is to study the structure of interactions in a social network by modeling messaging activities as a multivariate Poisson process. In this project we consider Enron corpus\(^1\), collected for a period of 189 weeks from 1998 to 2002, that contains highly accurate information about the time exchange messages between 184 corporate executives.

2 Previous Work

Priebe et. al. \cite{Priebe2003} apply scan statistics in effort to detect anomalies, conceived as statistically significant signals that stand out in a sea of noise. Scan statistics can be used to partition the community into smaller groups.

Pittel and Wozychnski \cite{Pittel2004} introduce the topic of network inference.

Malmgren et.al. \cite{Malmgren2005} apply a cascading non-homogeneous Poisson process, formulated as a double-chain Markov model, to analyze communication data from two different communities. They found common attributes among some individuals both within and between communities.

Perry and Wolfe \cite{Perry2006} model network interactions without reducing the data to binary has a relationship/does not have a relationship status. They also examine the question of whether shared attributes may be used to predict strength of relationship (based on rate of interaction) and whether communication from one individual to another can be used to predict the converse (i.e., communication in the other direction). Additionally, they model interactions between one sender and multiple receivers.

Rukhin and Priebe \cite{Rukhin2004} explore the relative statistical power of the size invariant (number of edges in the graph) and of the degree invariant in determining whether an observed graph is homogeneous or anomalous, and found that the maximum degree statistic is more powerful than the degree invariant statistic for graphs with smaller numbers of vertices (say up to 1024) while limit theory would suggest the opposite.

\(^{1}\)Enron Email Dataset -http://www.cs.cmu.edu/~enron/
3 Mathematical Background

3.1 Homogeneous Poisson process (HPP)

Assume events occur randomly over time and \( N(t) \) is number of events in a interval \([0, t]\).

**Definition:** A process \( N(t), t \geq 0 \) is said to be a Poisson process with rate \( \lambda, \lambda > 0 \), if:

1. \( N(0) = 0 \),
2. the number of events in disjoint intervals are independent (independent increments),
3. the distribution of the number of events in a certain interval depends only on the length of the interval and not on its position (stationary increments).

\[
\lim_{k \to 0} \frac{p(N(k) = 1)}{k} = \lambda. \tag{1}
\]

\[
\lim_{k \to 0} \frac{p(N(k) \geq 2)}{k} = 0. \tag{2}
\]

Some properties of homogeneous Poisson process are listed below:

1. The number of events in an interval of length \( t \) is a Poisson random variable with mean \( \lambda t \).
2. The inter-arrival times \( X_1, X_2, \ldots \) are independent \( Exp(\lambda) \) random variables.
3. The time of the \( n \)-th event is a \( \Gamma(\lambda, n) \) random variable.

3.2 Inhomogeneous Poisson process (IHPP)

The major weakness of homogeneous Poisson process is constant rate assumption meaning that the distribution of the number of events within an interval depends only on the length of the interval but not on its position.

**Definition:** A process \( N(t), t \geq 0 \) is a inhomogeneous Poisson process with rate function \( \lambda(t), t \geq 0 \), if:

\[\text{Notes - http://www.stat.sdu.dk/matstat/yuri-st505/w5slides1.pdf}\]
1. \( N(0) = 0, \)

2. the number of events in disjoint intervals are independent (independent increments).

\[
\lim_{k \to 0} \frac{p(N(t + k) - N(k) = 1)}{k} = \lambda(t). \tag{3}
\]

\[
\lim_{k \to 0} \frac{p(N(t + k) - N(k) \geq 2)}{k} = 0. \tag{4}
\]

The mean-value function \( m(t) \) is given by:

\[
m(t) = \int_0^t \lambda(k)dk, t \geq 0. \tag{5}
\]

Some properties of inhomogeneous Poisson process are listed below:

1. The number of events in the interval \((t, t + k], N(t + k) - N(t),\) is a Poisson random variable with rate \( m(t + k) - m(t) \).

2. Suppose that events are occurring according to a Poisson process with rate \( \lambda \), and suppose that, independently of anything that occurred before, an event that happens at time \( t \) is counted with probability \( p(t) = \frac{\lambda(t)}{\lambda} \). Then the process of the counted events constitutes an inhomogeneous Poisson process with rate \( \lambda(t) = \lambda p(t) \), where \( \lambda = \max \lambda(t) \).

### 3.3 Random Graphs based on Multiple Poisson Processes

We obtain a random graph \( G = (V,E) \) over a vertex set \( V = [n] = \{1, \ldots, n\} \), where the random graph is obtained from \( \binom{n}{2} \) pairs of count process. For the model for the multiple count processes, we consider homogeneous and inhomogeneous Poisson processes.

1. The simplest model of random graph based on homogeneous Poisson process is Erdős - Rényi model \( ER(n,p) \) where each of \( \binom{n}{2} \) possible edges is included in a graph with constant probability \( p \).

2. An alternative model of random graph based on inhomogeneous Poisson process is a graph where an anomalous subset of \( k \) vertices \( K = \{i_1, i_2, \ldots, i_k\} \) exists. Edges are independently included in the graph, however, the edge \((i, j)\) connects two anomalous vertices \( i, j \in K \) with pair specific probability \( s_{ij} \geq p \); each of the remaining \( \binom{n}{2} - \binom{k}{2} \) edges are present with probability \( p \).
4 Proposed Mathematical Approach

4.1 Random Graph Simulation based on Homogeneous Poisson Process

For each vertex pair $i, j$, let $N_{ij}$ denote a Poisson process whose intensity function $\lambda = \lambda_{ij}, \forall t$, for some $\lambda > 0$. First, we construct a graph using homogeneous Poisson process for a real data such as Enron corpus. For that purposes, we have to estimate the parameters $\hat{\lambda}_{ij}$. Next, in order to analyze the structure of a random graph $G$, we construct a simple statistic $N_{ij}$ for each pair of vertices $i, j$ based on the random variables as shown below:

$$\{n_{ij}^1, n_{ij}^2, n_{ij}^3, \ldots, n_{ij}^k\}, \quad (6)$$

where $k_{ij}$ represents the number of messages between vertices $i$ and $j$.

$$n_{ij}^{k+1} = \int_{\tau_{k,ij}}^{\tau_{k+1,ij}} \lambda(s)ds = \lambda_{ij}(\tau_{k+1}^{ij} - \tau_{k}^{ij}). \quad (7)$$

We can take the mean of the statistic as shown below:

$$N_{ij} = \frac{1}{k} \sum_{k=0}^{k} n_{ij}^k. \quad (8)$$

For large number of events, e.g., $k$ is sufficiently large for statistic $N_{ij}$ from Equation (8) is distributed as $N_{ij} \sim N(1, 1)$.

As a result we can reduce messaging data to a weighted random graph by applying “pairwise threshold” approach. That is, for every pair of vertices, an edge between vertex $i$ and vertex $j$ is formed if the average number of messages $S_{ij}$ between these two vertices exceeds a certain threshold $\theta_0$. For instance, if $(S_{ij} \geq \theta_0) = p_0$ then simulated random graph $G$ is Erdős - Rényi graph $ER(n, p_0)$ where $n$ is the number of vertices and $p$ is the probability of an edge in a graph.

Once we have a graph $G$ based on $N_{ij}$, we consider some graph invariants and perform deeper analysis of a messaging pattern in a graph $G$. Similar to other Enron graph statistics\(^3\), our statistics include:

- graph maximum degree statistic $G_{\text{degree}} = |\Delta| = |\Delta(G)|$.
- graph size statistic $G_{\text{size}} = |E| = |E(G)|$.
- Kolmogorov-Smirnov statistic compares true and simulated distributions of the number of messaging activities $N_{ij}$ in period $\Delta t$.

\(^3\)Scan Statistic on Enron Graph - http://cis.jhu.edu/~textasiitildeparky/Enron/enron.html
4.2 Random Graph Simulation based on Inhomogeneous Poisson Process

A more complex but potentially more realistic model of social interactions accounts for inhomogeneous properties. For example, the probability of a message may be higher during work hours than at night. If the probability of an interaction is non-constant but still independent from previous events, it can be modeled as a inhomogeneous Poisson process with rate \( \lambda_{ij}(t) \). Thus, for each vertex pair \( i, j \), let \( N_{ij}(t) \) denote a inhomogeneous Poisson process whose intensity function \( \lambda(t) = \lambda_{ij}(t), \forall t \), for some \( \lambda > 0 \).

Similarly to the homogeneous Poisson model detailed in the previous section, we propose to study a random graph simulation based on inhomogeneous Poisson process.

4.3 Message Rate Estimation from Enron Data

In order to construct a graph based Poisson processes for a real data we have to estimate the message rate \( \hat{\lambda}_{ij} \). We estimate two different message rates: bulk message \( \hat{\lambda} \) rate and vertex pair dependent rate \( \hat{\lambda}_{ij} \).

For bulk message rate \( E[N(\Delta t)] \) equals to the actual number of messages divided by the corresponding interval:

\[
\hat{\lambda} = \frac{[N(t_{k+n}) - N(t_k)]}{(t_{k+n} - t_k)}.
\]

For vertex pair dependent message rates we calculate the number of messages for each sender/receiver pair during a period of 88 weeks. This was accomplished by scanning the data and totaling the number of messages between each pair. The number of message for each pair was divided by the selected interval of 88 weeks to arrive at mean message rates for each pair, \( \lambda_{ij} \) where \( i \) and \( j \) are indices of senders and receivers and \( i < j \) (i.e., ignoring the sender/receiver status of each individual):

\[
\hat{\lambda}_{ij} = \frac{[N_{ij}(t_{k+n}) - N_{ij}(t_k)]}{(t_{k+n} - t_k)}.
\]

We present messaging rates (counts) in Appendix A for sixteen arbitrarily selected pairs of individuals \( (i, j) \). The message rates seem to change in about the middle of the period studied.
5 Enron Corpus

Enron data includes messaging communication between 184 users for a period on 189 weeks (from Friday, 13 Nov 1998 to Friday, 21 Jun 2002). For our experiments, we only consider the subset of Enron data - 88 week period from 961891200 sec to 1015004254 sec (from June 2000 to Feb 2002).

At the preprocessing stage we clean the data and remove messages where the sender and the receiver is the same. After cleaning the data, we have 22 symmetric square binary matrices that represent messaging communication between 184 users. These adjacency matrices are binary because we are working with undirected unweighted graphs and draw an edge between two vertices \(i\) and \(j\) when there is a message between two of them. In other words, we do not take into account who is the sender and who is the receiver.

Moreover, we use different thresholds and draw an edge between two vertices \(i\) and \(j\) if they meet the threshold \(\theta_1 = 1, \theta_2 = 3, \theta_3 = 5\). The examples of true undirected unweighted Enron graphs constructed from symmetric square binary adjacency matrices are shown in Figure 1. We use true Enron graphs for validation of our simulation model in Section 7.

![Enron Graphs](image)

(a) \(t = 24, \theta = 1\)  (b) \(t = 24, \theta = 3\)  (c) \(t = 24, \theta = 5\)

(d) \(t = 48, \theta = 1\)  (e) \(t = 48, \theta = 3\)  (f) \(t = 48, \theta = 5\)

Figure 1: True Enron graphs constructed for 20 - 24 and 44 - 48 week period (\(\Delta t = 4\) weeks) starting from mid 1999 year, \(t_{start} = 961891200\) sec.
6 Random Graph Simulation

6.1 Random Graph Simulation based on Homogeneous Poisson Process

As described previously and schematized in Figure 2, random graphs were created from a simulated homogeneous Poisson process. Briefly, the Poisson simulation generated a sequence of vertex communication events with arrival times drawn from a Poisson distribution (top left). The number of events between a vertex pair was then counted for each 4 week interval over the two year period (top right). To create a random graph, a threshold of one was chosen. Thus, in a given 4 week interval, an edge was drawn between a vertex pair if any communication took place between them (bottom right), allowing generation of the random graph (bottom left).

**Figure 2:** Workflow for generating random graphs: simulated messages $\rightarrow$ message frequency per interval $\rightarrow$ adjacency matrix $\rightarrow$ random graph.

The algorithm used to simulate messaging activities distributed via homogeneous Poisson process with rate $\lambda$ up to time $T$ is described below.
In brief, the times between events are drawn from an exponential distribution, such that the number of arrival events in any given interval is Poisson distributed. After a time is drawn for an event to occur, the vertex pair which transmitted the message is chosen uniformly. Since the graph is undirected, but self-vertices are not allowed, the order in which the vertex pair is drawn is not important, as long as the two vertices are unique. Thus, the algorithm generates a series of messaging events with associated times at which they occurred and the vertex pair that sent the message.

Algorithm (Discrete Event Simulation):

1. Generate a uniform random variable $U \sim Uniform(0, 1)$.
2. Generate exponential random variable by inversion:
   
   $$ t = t + \left(\frac{-1}{\lambda} \cdot \ln(U)\right). $$

   If $t > T$, then stop.
3. Uniformly select a vertex pair associated with that time.
4. Go to step (1).

Output:

Vertex pairs $V_i, V_j$ with associated times $t_{ij}$.

6.2 HPP Simulation Using Bulk Message Rate

As described in Section 4.3, the lambda parameter for Poisson simulation was calculated from MLE estimation of the Enron corpus message rate. The initial approach involved the calculation of a single bulk rate drawn from all vertex pairs over the entire time period under investigation. That is, a single lambda parameter was calculated for the entire corpus, representing the rate of messaging events between any vertex pair. This rate was determined to be messages/second between all vertices, for a rate of 0.14 messages in 4 weeks between any pair of the 184 vertices. A representative random graph simulated with this parameter is shown below in Figure 3.

The graph is very dense, with a large number of nodes sending at least one message in a given 4 week interval. This can also be seen from the distribution of summary statistics over 10,000 random graph simulations. These statistics are described in previous sections, and include the maximum degree of a graph, the total number of edges (graph size), and the overall degree distribution, as shown in Figure 4.

As can been seen in the middle panel, each simulated vertex sent approximately 20-30 messages during the 4 week interval, with about 4000-5000
total communications in the entire network. This is significantly higher than the rates observed in the real Enron Corpus.

A brief examination of the real messaging activities suggests the reason. Messaging rates are unevenly distributed: certain vertex pairs communicate frequently, 100's of times in a given interval, while other vertex pairs never communicate at all. This is likely to be representative of corporate email networks in general, where individual departments or teams may communicate much more frequently than the average employee. The simulation strategy employed does not account for this inhomogeneity. When a single bulk rate is calculated, frequently messaging pairs have their messaging behavior distributed across all vertices. Thus, many vertex pairs send at least one message, which is the threshold for drawing an edge in the graph.

Figure 3: Sample undirected unweighted random graph with bulk lambda that represent simulated communication activities for 4 week period.

Figure 4: Summary statistics (maximum degree, overall degree and graph size) for random graph with bulk lambda.
6.3 HPP Simulation using Vertex Dependent Message Rate

In order to ameliorate this problem while still using a computationally efficient homogenous Poisson model, messaging rates were estimated between each vertex pair (see Section 4.3). Since a majority of the vertices in the Enron Corpus do not communicate (true graphs are sparse), the messaging rate between most vertex pairs in the simulated graphs will also be zero. On the other hand, certain vertex pairs that communicate frequently will have very large lambdas representing frequent messaging behavior. The messaging rate between vertices varies from 0 messages/4 weeks to approximately 169 messages/4 weeks. The simulation procedure remains the same, but is now done for each vertex pair with a rate calculated for that pair.

Random graphs simulated using vertex dependent rates, shown in Figure 5, are more sparse than those in the previous section, a result of the many vertices that did not send any messages.

![Figure 5: Sample random graph for vertex dependent lambda that represent communication activities for 4 week period.](image)

This is easier to see with summary statistic distributions shown in Figure 6. As shown in the middle panel, the degree distribution is profoundly left shifted, with most vertices having few edges, accounting for the sparser graph. However, the expected maximum degree in a given random graph is higher, reflecting the influence of vertex pairs with higher rates.

The appropriateness of this model for describing the Enron Corpus and detecting anomalies will be discussed in later sections. However, it is already clear that further improvements can be made to the model to more accurately reflect messaging patterns. The more accurate the underlying model, the better the detection algorithm.
6.4 Setting Thresholds for Edge Creation

A parameter that can be modified is the threshold for drawing a random graph edge. Since initial simulations resulted in extremely dense graphs, higher thresholds were selected. In this section, 3 or 5 messages in the 4 week period were required from the simulation in order to draw an edge between vertices. The resulting graphs are shown in Figure 7, and as expected they are indeed sparser than those with threshold 1.

![Figure 7: Random graphs that represent messaging activities within 4 week period simulated using different thresholds 5, 3, 1 shown from left to right.](image)

The summary statistics shown in Figures 8 and 9 show that as expected, with higher thresholds the number of edges, maximum edges in a random graph, and degree distribution are all left shifted to smaller values.

However, thresholding does not address the underlying issue driving unexpectedly dense graphs in the previous simulations. The density results from brief periods of increased messaging between small numbers of ver-
vertices. Thresholding focuses instead on vertices that rarely send messages, and removes them from the simulation. However, an inappropriately high messaging rate derived from brief periods of activity still drives the homogeneous model to assign messaging behavior to more vertices than in the real corpus.

Figure 8: Summary statistics (maximum degree, overall degree and graph size) for random graph simulated using threshold 3.

Figure 9: Summary statistics (maximum degree, overall degree and graph size) for random graph simulated using threshold 5.
6.5 Random Graph Simulation based on Inhomogeneous Poisson Process

As described previously, a HPP does not capture a number of the characteristics of messaging activity. Communication activity is distinctly inhomogeneous, with messaging rates that increase during the workweek and during the day, or when vertex pairs are assigned to the same project. Thus, the graphs shown in the previous sections are still likely to be too dense. For example, a brief period of messaging activity right before a project deadline will be distributed evenly across the entire two year period by a homogeneous model, resulting in more edges drawn in a given interval than expected.

Capturing this type of behavior requires a inhomogeneous Poisson behavior. A process for simulating inhomogeneous Poisson processes is presented in this section. The exact approach for determining message rates between two vertices over time $\Delta t$ is discussed in 4.3. The example message rates for random vertex pairs $ij$ as shown in Appendix A.

A thinning algorithm for simulating messaging activities via inhomogeneous Poisson Process with rate $\lambda(t)$ that is bounded by $\lambda^*$ is presented below. The algorithm generates a non-stationary Poisson process up to desired time $T$ to get the messaging events during the interval.

**Algorithm:**

1. Generate a uniform random variable $U \sim Uniform(0, 1)$.
2. Generate an exponential random variable by inversion:
   \[ t = t + \left( \frac{-1}{\lambda^*} \cdot \ln(U) \right) \text{.} \] If $t > T$, then stop.
3. Generate a random variable $U \sim Uniform(0, 1)$.
4. If $U \leq \frac{\lambda(t)}{\lambda^*}$, uniformly select a vertex pair for messaging.
5. Go to step (1).

**Output:**

Vertex pairs $V_i, V_j$ with associated times $t_{ij}$.

In Figure 10 we present the cumulative number of messages between random pair of vertices simulated based on inhomogeneous Poisson process compared to true number of messages estimated from the data.
Figure 10: Actual and cumulative messages between vertices $v_i$ and $v_j$. 

(a) $v_i = 24, v_j = 1$

(b) $v_i = 24, v_j = 3$
7 Analysis of Graph Statistics for Anomaly Detection

Three summary statistics including graph size, maximum graph degree, and Kolmogorov-Smirnov test of graph degrees are used to compare true graphs from the Enron Corpus to graphs derived from the Poisson messaging simulation. True graphs that deviate significantly from modeled behavior may be an indicator for aberrant or unexpected activity in the network.

7.1 HPP: Maximum Degree Statistic

Maximum degree, the maximum of the number of edges attached to each vertex in the graph, is shown in Figure 11 for thresholds of 1, 3, and 5 messages. The histograms show that the maximum degree is very sensitive to the threshold near a threshold of 1. As one increases the number of messages required in order to draw an edge, the expectation and mode of the maximum degree decreases.

![Figure 11: Maximum degree statistic for simulated graphs of threshold 1, 3, 5](image)

For a threshold of 1, the 90th percentile of maximum degree is at 65 to 66 messages. For a threshold of 3, it is 27 to 28 messages, and for a threshold of 5, it is 17 to 18 messages. Similarly, the 10th percentiles for thresholds of 1,3, and 5 are, respectively, 55 to 56, approx. 22, and 14 to 15.

Maximum degree for the true Enron Corpus, on the other hand, is shown in Figure 12 (for thresholds of 1, 3, and 5 messages) as a time series. There are spikes in maximum degree during periods 12, 16, 17, and 22.
For thresholds 3 and 5, all of the spikes are in the extreme upper tail of the distribution expected under a homogeneous Poisson process. For threshold 1, the spike at period 12 is also in the extreme upper tail, while the spikes at periods 16 and 22 are in the lower tail, and the spike at period 17 is at about the 73rd percentile.

Thus, a threshold of 1 appears to be in a range where the maximum degree is too sensitive to threshold to provide meaningful results: the time series of maximum degree seems to follow the same trend regardless of threshold, yet detection of anomalous behavior based upon the maximum degree statistic would seem to have results for threshold 1 that are inconsistent with other thresholds.

### 7.2 HPP: Graph Size Statistic

The graph size statistic is the total count of edges in a given graph. This statistic would be able to detect sudden increases or decreases in activity across the whole network, rather than aberrant activity from any particular vertex. The percentiles for the distribution of graph sizes from simulated random graphs first shown in Figures 6, 8, and 9 are summarized in Table 1, for each of the three different thresholds described in Section 6.4.

Graph sizes for the real Enron Graphs, on the other hand, are given in Table 2. This table gives the total edge counts in graphs over a number of illustrative 4 week intervals in the analysis period for each of the three thresholds.

The simulated graphs have significantly more edges than the true graphs derived from the Enron Corpus, no matter the threshold chosen. As de-
Table 1: Percentiles for the Graph Size Statistic from Simulated Random Graphs based on HPP

<table>
<thead>
<tr>
<th>Percentile:</th>
<th>1</th>
<th>2.5</th>
<th>5</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>95</th>
<th>97.5</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold 1 Edges:</td>
<td>843</td>
<td>849</td>
<td>854</td>
<td>871</td>
<td>881</td>
<td>893</td>
<td>910</td>
<td>915</td>
<td>920</td>
</tr>
<tr>
<td>Threshold 3 Edges:</td>
<td>301</td>
<td>305</td>
<td>308</td>
<td>317</td>
<td>324</td>
<td>331</td>
<td>341</td>
<td>344</td>
<td>347</td>
</tr>
<tr>
<td>Threshold 5 Edges:</td>
<td>166</td>
<td>168</td>
<td>171</td>
<td>178</td>
<td>182</td>
<td>187</td>
<td>194</td>
<td>196</td>
<td>199</td>
</tr>
</tbody>
</table>

Table 2: Edge Counts for Real Enron Graphs Over Selected 4 Week Intervals

<table>
<thead>
<tr>
<th>Time (weeks):</th>
<th>4</th>
<th>12</th>
<th>20</th>
<th>28</th>
<th>36</th>
<th>44</th>
<th>52</th>
<th>60</th>
<th>68</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold 1:</td>
<td>174</td>
<td>275</td>
<td>281</td>
<td>270</td>
<td>286</td>
<td>356</td>
<td>325</td>
<td>300</td>
<td>489</td>
<td>440</td>
</tr>
<tr>
<td>Threshold 3:</td>
<td>152</td>
<td>259</td>
<td>265</td>
<td>253</td>
<td>247</td>
<td>276</td>
<td>175</td>
<td>165</td>
<td>361</td>
<td>321</td>
</tr>
<tr>
<td>Threshold 5:</td>
<td>99</td>
<td>142</td>
<td>210</td>
<td>194</td>
<td>183</td>
<td>213</td>
<td>119</td>
<td>81</td>
<td>291</td>
<td>234</td>
</tr>
</tbody>
</table>

scribed in Section 6, inhomogeneities in messaging rates can contribute to overestimates in the messaging rates. Thresholding out infrequent communicators does not address this underlying misspecification of the model. This summary statistic would label every 4 week interval as anomalous, and is thus not a useful statistic for detection until the model is refined.

7.3 HPP: Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test is a nonparametric test for comparing two probability distributions by determining the maximum separation between two distributions as shown below:

\[ D = \max_{x_{\min} \leq x \leq x_{\max}} \left| F_{\text{true}}(x) - F_{\text{sim}}(x) \right|. \]

We compare true Enron graph degree distribution \( F_{\text{true}}(x) \) with simulated graph degree distributions \( F_{\text{sim}}(x) \) for different thresholds. The null hypothesis says \( H_0 \): the results of the simulation come from the true Enron graph distribution \( F_i(x) \). The null hypothesis \( H_0 \) is rejected if at level \( \alpha = 5\% \) if \( p-value \leq 0.05 \).

Based on the results from Table 3 we conclude that messages \( N_{ij} \) between vertex \( i \) and \( j \) in simulated and true graphs are not drawn from the same distribution for all threshold values \( \theta_i \). All p-values are less that 0.05 except one p-value for the threshold \( \theta = 5 \), 60-64 week period. We show the corresponding histograms of true graph statistic and simulated statistic distribution for the case in Figures 13 and 14.
As can be seen, KS statistic would label every 4 week period as anomalous, and thus it is not a useful statistic for anomaly detection until the model is enhanced.
7.4 IHPP: Maximum Degree and Graph Size Statistics

Graph size and maximum degree statistics derived from the simulations (mean of 100 simulated values) follow the general trend of the true statistics. However, there are marked differences most particularly at the spikes in maximum degree, and in the difference between true and simulated graph size for threshold 1.

Qualitatively speaking, the maximum degree statistic would seem to be capturing aberrant behavior during months 12, 16, 17, and 22, since the differences between simulated and true data for those periods are greatest. However, the graph size statistic seems unable to capture the same aberrant behavior, as the simulated data follow the trend of the true data fairly well.
We show graph size and maximum degree statistic results for true Enron graphs vs. simulated graphs based on inhomogeneous Poisson process in Figures 15 and 16 respectively.

Figure 15: True and simulated graph size statistic based on inhomogeneous Poisson process.

Future work might include consideration of months 12, 16, 17, and 22 as separate from the other months of the study period, and computing difference-detecting statistics for the two subsets.

7.5 IHPP: Kolmogorov-Smirnov Test

Recall that Kolmogorov-Smirnov test compares true Enron graph degree distribution and simulated degree distribution for every period $\Delta t$. We did not find this test useful when we applied it to the results obtained from inhomogeneous Poisson process simulation. We found all $p$-values to be less than 0.05 that means that we reject the null hypothesis which says that both true and simulated graphs come from the same distribution. Similarly to the homogeneous Poisson process case we conclude that KS test is not an applicable statistic for anomaly detection in messaging activities.
Figure 16: True and simulated graph size statistic based on inhomogeneous Poisson process.

More precisely, we did not find KS test useful due to several reasons.

- First, it looks like the simulation is not creating as many vertices of very high degree. One possible explanation is that messaging is not independent between vertices. For example, if vertex $i$ sends an email to vertex $j$, it is also more likely to send an email to vertex $k$. One could model this as some sort of exciting process but we did not do this in our project. Therefore, our simulation does not create quite as many high degree vertices.

- Next, the p-values are very small because we have so many simulations.

- Last, there is a major aspect of messaging behavior the simulation is still missing.
Finally, a motif count statistic is proposed for graph analysis. Motifs are broadly defined as small structural units or subgraphs occurring in a network, particularly those that are found more often than would be expected in a random graph of a particular type. The occurrence of such substructures often gives clues to network function. For example, in a communication network like the Enron Corpus, one might expect dense subgraphs when members are assigned to a particular project and frequently communicate with each other, or when a small group of co-conspirators is plotting to take advantage of California’s deregulated energy markets.

We use MAVisto\textsuperscript{4} package to run preliminary experiments on motif count analysis. This software generates random graphs in a similar way we did using homogeneous Poisson process simulation. Then, MAVisto counts the most frequent motifs and reports the results in terms of p-values.

Even we worked with unweighted undirected graphs before, we decided to use both directed and undirected settings for motif count analysis. Here are three major steps we perform using MAVisto package for motif analysis:

- take true unweighted, undirected Enron graphs for different time periods from week 4 to week 88;
- generate 1,000 Erdős - Rényi random graphs;
- analyze the motifs that we found to be more frequent in true graphs compared to the random graphs.

\textsuperscript{4}http://mavisto.ipk-gatersleben.de/
We show the results of our preliminary experiments on motif count analysis for the first 4 weeks in our period which corresponds to June 2000 in Figures 18, 19 and 20 for the motifs that consist of 3 vertices. For instance, 3 vertex motifs that were found to be frequent in June 2000 include:

- \( v_1 \rightarrow v_2 \leftarrow v_3 \);
- \( v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_1 \);
- \( v_1 \leftrightarrow v_2 \rightarrow v_3 \).

In previous section, we discovered anomalous messaging activities for periods 12, 16, 17 and 22 based on maximum degree statistic. We want to investigate these periods deeper, and for that purpose we run the experiments on motifs count within these suspicious periods. The results of 3 vertex motif counts are shown in corresponding figures in Appendix B:

- the graph with \( v_1 \rightarrow v_2 \leftarrow v_3 \) motifs for period 12 (May 2001) is shown in Figure 22; the graph with \( v_1 \rightarrow v_2 \rightarrow v_3 \) motifs for period 12 (May 2001) is shown in Figure 23; the graph with \( v_1 \leftarrow v_2 \rightarrow v_3 \) motifs for period 12 (May 2001) is shown in Figure 24 (v82 messaging behavior seems to be very abnormal during this period);
- the graph with \( v_1 \rightarrow v_2 \leftarrow v_3 \) motifs for period 16 (Sept 2001) is shown in Figure 25; the graph with \( v_1 \leftarrow v_2 \rightarrow v_3 \) motifs for period 16 (Sept 2001) is shown in Figure 26;
- the graph with \( v_1 \rightarrow v_2 \leftarrow v_3 \) motifs for period 17 (Oct 2001) is shown in Figure 27; the graph with \( v_1 \rightarrow v_2 \rightarrow v_3 \) motifs for period 17 (Oct 2001) is shown in Figure 28; the graph with \( v_1 \leftarrow v_2 \rightarrow v_3 \) motifs for period 17 (Oct 2001) is shown in Figure 29;
- the graphs with \( v_1 \rightarrow v_2 \leftarrow v_3 \) motifs for period 22 (Feb 2002) are shown in Figures 30 and 31; the graphs with \( v_1 \rightarrow v_2 \rightarrow v_3 \) motifs for period 22 (Feb 2002) are shown in Figures 32 and 33; the graphs with \( v_1 \leftarrow v_2 \rightarrow v_3 \) motifs for period 22 (Feb 2002) are shown in Figures 34 and 35. This period seems to be especially anomalous because we found a set of vertices that are separate from all other vertices in the graph (we have not notice anything like that for any other periods).
Figure 18: 3 vertex motif count $v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_1$ for June 2000.
Figure 19: 3 vertex motif count $v_1 \rightarrow v_2 \leftarrow v_3$ for June 2000.
Figure 20: 3 vertex motif count $v_1 \leftrightarrow v_2 \rightarrow v_3$ for June 2000.
8 Conclusions and Future Work

This report develops a method for analyzing the dynamic messaging behavior of social communication networks. Briefly, the Enron corpus of email messages is converted to a graph by drawing an edge between vertices that sent a certain threshold of messages in a given interval. The resulting graphs are then compared to those derived from a homogenous and inhomogeneous Poisson processes of communication behavior, with message rates estimated from the Enron corpus.

For homogeneous Poisson process model, each of the three graph summary statistics examined (graph size, maximum graph degree, and Kolmogorov-Smirnov test of graph degrees) does not yield useful measures of aberrant or normal messaging behavior. Calculating vertex-depending messaging rates and setting higher graph thresholds does not sufficiently improve the model.

The inhomogeneous Poisson process model was used to account for brief burns of messaging between small numbers of vertices. For IHPP model, graph size and maximum graph degree statistics more closely follow the true data. Moreover, maximum graph degree allows capturing the aberrant messaging behavior during periods 12, 16, 17 and 22. We did not find KS statistic to be useful for anomaly detection in a network.

Finally, both HPP and IHPP models consistently predict an excessively high rate of messaging due to inhomogeneous and self-excitng bursts of messaging activity from a few highly active vertices. Therefore, there is some future work can be done in order to develop a better framework for modeling communication activities and detecting abnormal behavior in a network such as:

- Model messaging behavior using self-exciting Poisson process.
- Take into account additional parameters during the simulation, e.g., message content, message topic, communicant gender, age etc.
- Explore other statistics, e.g., motif count analysis.
- Apply similar approaches of modeling messaging activities to different datasets, e.g., Twitter, reviews, commentary data.
A Appendix

Figure 21: The message rates for 22 periods of 4-week for randomly selected vertex pairs $ij$. 
Figure 22: 3 vertex motif count $v_1 \rightarrow v_2 \leftarrow v_3$ for May 2001.
References


Figure 23: 3 vertex motif count $v_1 \rightarrow v_2 \rightarrow v_3$ for May 2001.
Figure 24: 3 vertex motif count $v_1 \leftarrow v_2 \rightarrow v_3$ for May 2001.
Figure 25: 3 vertex motif count $v_1 \rightarrow v_2 \leftarrow v_3$ for September 2001.
Figure 26: 3 vertex motif count $v_1 \leftarrow v_2 \rightarrow v_3$ for September 2001.
Figure 27: 3 vertex motif count $v_1 \rightarrow v_2 \leftarrow v_3$ for Oct 2001.
Figure 28: 3 vertex motif count $v_1 \rightarrow v_2 \rightarrow v_3$ for Oct 2001.
Figure 29: 3 vertex motif count $v_1 \leftarrow v_2 \rightarrow v_3$ for Oct 2001.
Figure 30: 3 vertex motif count (subgraph 1) $v_1 \rightarrow v_2 \leftarrow v_3$ for Feb 2002.
Figure 31: 3 vertex motif count (subgraph 2) $v_1 \rightarrow v_2 \leftarrow v_3$ for Feb 2002.
Figure 32: 3 vertex motif count (subgraph 1) $v_1 \rightarrow v_2 \rightarrow v_3$ for Feb 2002.
Figure 33: 3 vertex motif count (subgraph 2) $v_1 \rightarrow v_2 \rightarrow v_3$ for Feb 2002.
Figure 34: 3 vertex motif count (subgraph 1) $v_1 \leftarrow v_2 \rightarrow v_3$ for Feb 2002.
Figure 35: 3 vertex motif count (subgraph 2) $v_1 \leftarrow v_2 \rightarrow v_3$ for Feb 2002.