Modeling Messaging Activities in a Network: Enron Case Study

Svitlana Volkova, Karlo Perica, D. Michael Parrish

Dec, 2 2011
Enron Dataset

- 184 users, 125k messages.
- Data format: time $t$, sender $i$, receiver $j$. 
Task:
- Understand the structure of interactions and the communication patterns in a corporate network.
- Detect abnormal activities and identify communities within a corporate network.

Approach:
- Construct random graphs that model messaging activities in a network via:
  - homogeneous Poisson process (messaging rate $\lambda_{ij} = \text{const}$);
  - inhomogeneous Poisson process ($\lambda_{ij}(t)$ is different for $\Delta t$).
- Compare simulated and true Enron graphs using different statistics.
Simulation: Homogeneous vs. Inhomogeneous PP

**Definition (Modeling with HPP)**

Each edge \((i, j)\) from the set of all \(\binom{n}{2}\) edges is included in the graph \(G\) with constant probability \(p\).

**Definition (Modeling with IHPP)**

An edge \((i, j)\) from the subset of \(\binom{k}{2}\) edges is drawn between two vertices \(i, j \in K\) with probability \(s \geq p\) and each of the remaining edges \(\binom{n}{2} - \binom{k}{2}\) are drawn with probability \(p\).
Switching to a movie...
Evaluation: Summary Statistics for Graph Comparison

- graph size: $|E| = |E(G)| = 10$.
- graph maximum degree: $|\Delta| = |\Delta(G)| = 4$.
- Kolmogorov-Smirnov Test fits two degree distributions: p-value.

Svitlana Volkova, Karlo Perica, D. Michael Parrish
Modeling Messaging Activities in a Network: Enron Case Study
Thresholding The Number Of Messages $N_{ij}$

Simulated graphs with thresholds 5, 3, 1 (from left to right):

True graphs with thresholds 5, 3, 1 (from left to right):
Estimating Message Rates $\hat{\lambda}$ for HPP Simulation

- Estimate a “bulk” message rate $\hat{\lambda}$ using MLE estimator:

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

- Estimate a constant message rate $\hat{\lambda}_{ij}$ for a pair of vertices $i, j$ using MLE estimator:

$$\hat{\lambda}_{ij} = \frac{[N_{ij}(t_{k+n}) - N_{ij}(t_k)]}{(t_{k+n} - t_k)}$$
Summary statistics for “Bulk” $\hat{\lambda}$ message rate, $\theta = 1$:

Summary statistics for vertex pair dependent $\hat{\lambda}_{ij}$, $\theta = 1$:
Estimating Intensities $I_{ij}$ for IHPP Simulation

- Estimate the intensity functions $I_{ij}$ for a pair of vertices $i, j$ for 88 weeks ($x$: 4-week periods, $y$: number of messages).
Maximum Degree Statistic for HPP simulated graphs:

Maximum Degree Statistic for Enron graphs:
Percentiles for the Graph Size Statistic from Simulated Graphs:

<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:</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:</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:</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>

Graph Size Statistic for Enron graphs:
HPP Simulation: Kolmogorov-Smirnov Test

- Compare true Enron graph degree distribution $F_t(x)$ with simulated graph degree distributions $F_s(x)$.
- $H_0$: the results of the simulation come from the true Enron graph distribution $F_t(x)$.
- $H_0$ is rejected if at level $\alpha = 5\%$ if $p - value \leq 0.05$.
Limitations of HPP

- HPP model consistently predicts an excessively high rate of messaging due to inhomogeneous bursts of messaging activity from a few highly active vertices.
- Each summary statistics 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.
IHPP Simulation Example: Vertex 5 and 107

Actual and Simulated Messages Between Vertex 5 and 107

Cumulative Messages

0 5 10 15 20 25 30 35
Cumulative Actual Sent Messages
Cumulative Simulated Messages

Time (# of 4 week periods)
0 5 10 15 20 25

Svitlana Volkova, Karlo Perica, D. Michael Parrish
Modeling Messaging Activities in a Network: Enron Case Study
Goal: detect small subgraphs occurring in a network more often than would be expected in a random graph.
Motifs Count Statistic: GCR tuples in 4 week period

Svitlana Volkova, Karlo Perica, D. Michael Parrish
Modeling Messaging Activities in a Network: Enron Case Study
Motifs Count Statistic: F8X tuples in 4 week period

Svitlana Volkova, Karlo Perica, D. Michael Parrish

Modeling Messaging Activities in a Network: Enron Case Study
Future Work

- Model messaging behavior using self-exciting Poisson process.
- Take into account additional parameters during the simulation, e.g., message content, communicant gender, age.
- Explore more other statistics, e.g., motif count analysis.
- Apply similar approach of modeling messaging behavior to other datasets, e.g., Twitter.
R. Dean Malmgren, Jake M. Hofman, Luis A.N. Amaral, and Duncan J. Watts.
Characterizing individual communication patterns.

Patrick O. Perry and Patrick J. Wolfe.
Point process modeling for directed interaction networks.

B. Pittel and W. A. Woyczynski.
A graph-valued markov process as rings-allowed polymerization model: subcritical behavior.

Carey E. Priebe, John M. Conroy, David J. Marchette, and Youngser Park.
Scan statistics on enron graphs.
2005.

Andrey Rukhin and Carey E. Priebe.
A comparative power analysis of the maximum degree and size invariants for random graph inference.

Cohen W.W.
Enron email dataset, 2009.