Modeling Messaging Activities in a Network: Enron Case Study

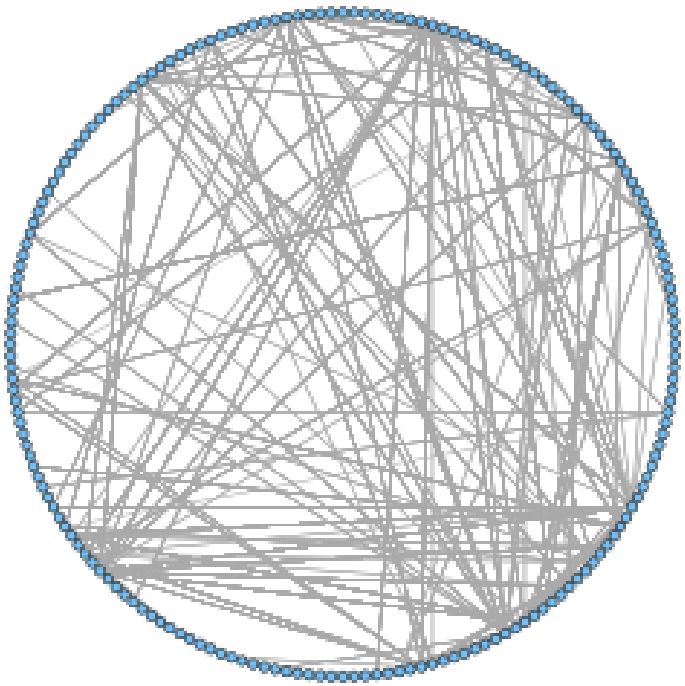
Svitlana Volkova, Karlo Perica, D. Michael Parrish

Dec, 2 2011

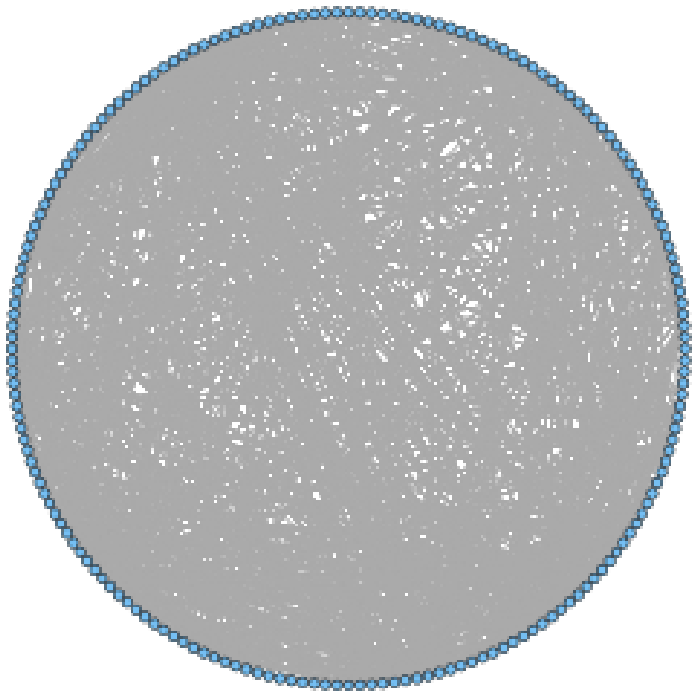
Enron Dataset

- 189 weeks from 1998 to 2002 (Nov, 13 1998 to Jun 21, 2002).
- 184 users, 125k messages.
- Data format: time t, sender i, receiver j.

1998 - 2000 year



2000 - 2001 year



Research Tasks and Solving Strategies

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 $\lambda_{ij} = const$);
  - inhomogeneous Poisson process ($\lambda_{ij}(t)$ is different for Δt).
- 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.

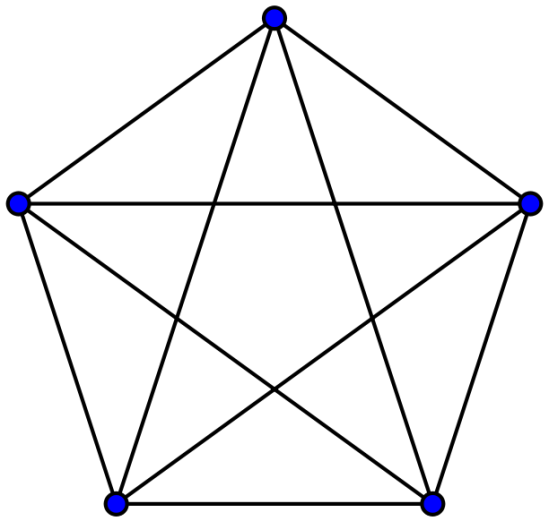
Definition (Modeling with IHPP)

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.

Simulation of Messaging Activities

Switching to a movie...

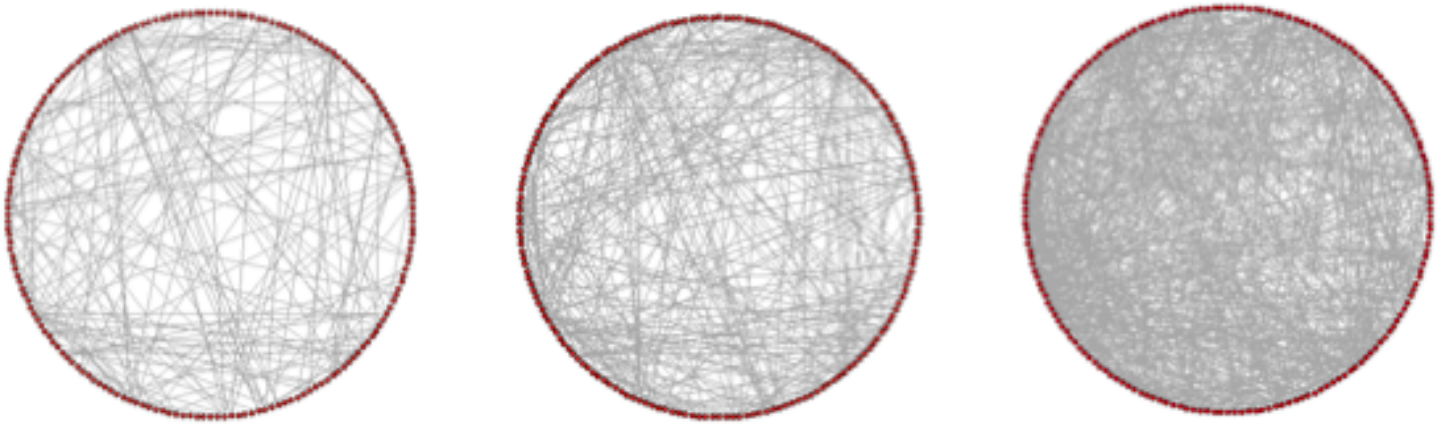
Evaluation: Summary Statistics for Graph Comparison



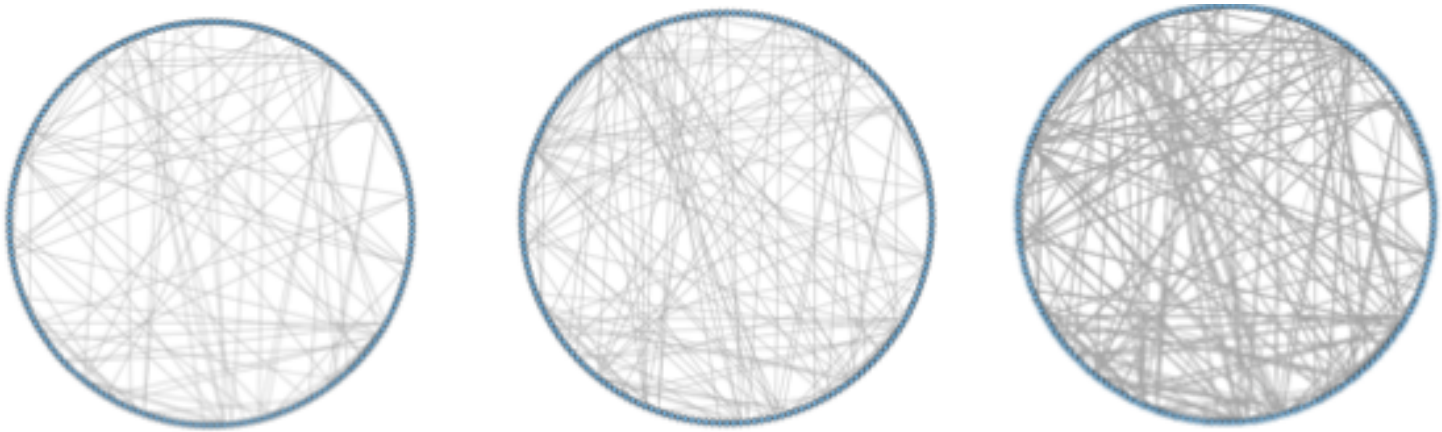
- graph size: $|E| = |E(G)| = 10$.
- graph maximum degree: $|\Delta| = |\Delta(G)| = 4$.
- Kolmogorov-Smirnov Test fits two degree distributions: p-value.

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:

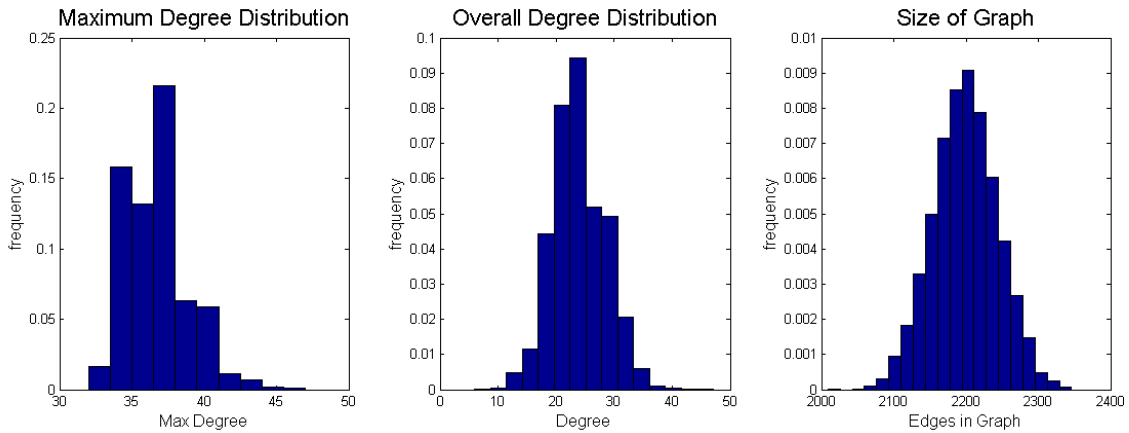
$$\hat{\lambda} = \frac{[N(t_{k+n}) - N(t_k)]}{(t_{k+n} - t_k)}$$

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

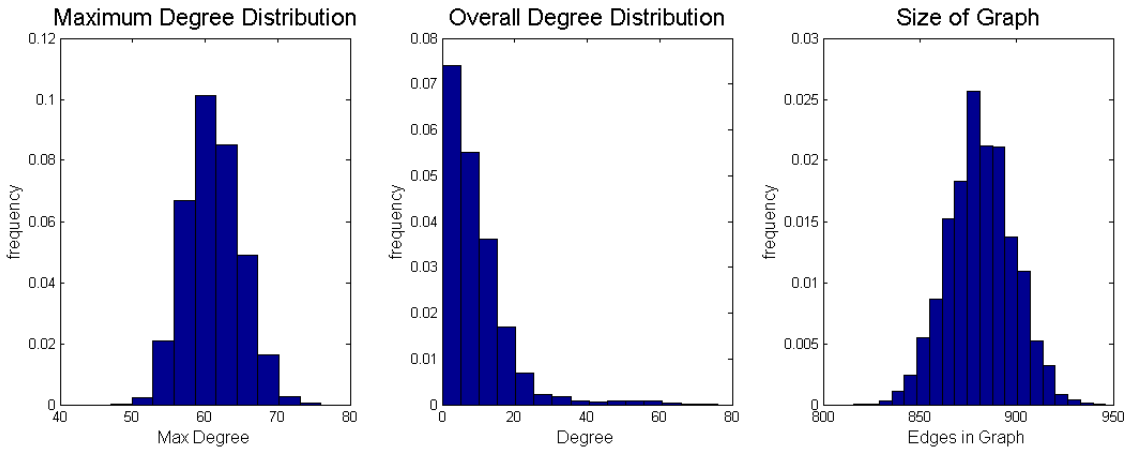
$$\hat{\lambda}_{ij} = \frac{[N_{ij}(t_{k+n}) - N_{ij}(t_k)]}{(t_{k+n} - t_k)}$$

Summary Statistics For “Bulk" and Vertex Pair Dependent λ in HPP case

Summary statistics for “Bulk" $\hat{\lambda}$ message rate, $\theta = 1$:

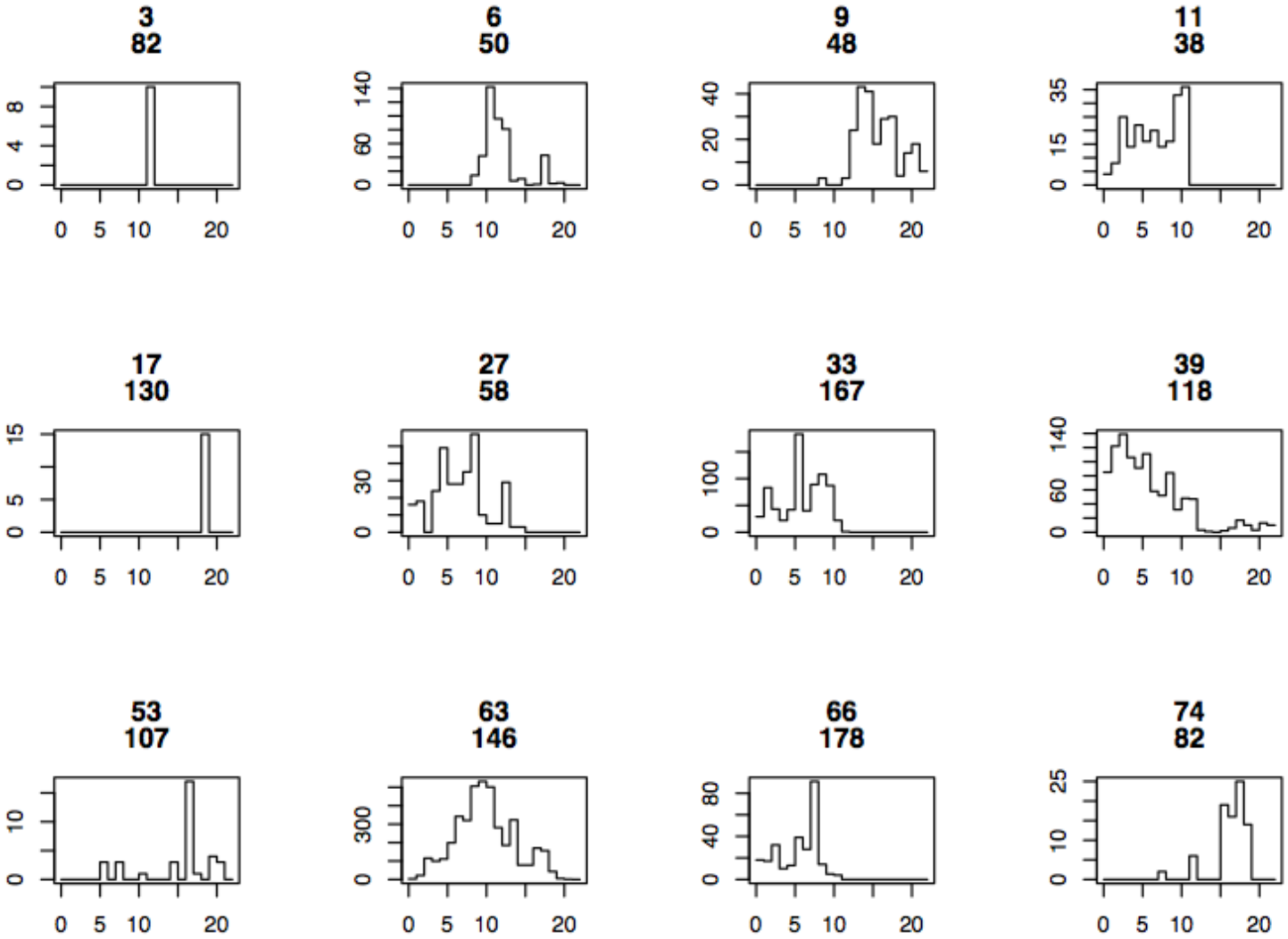


Summary statistics for vertex pair dependent $\hat{\lambda}_{ij}$, $\theta = 1$:



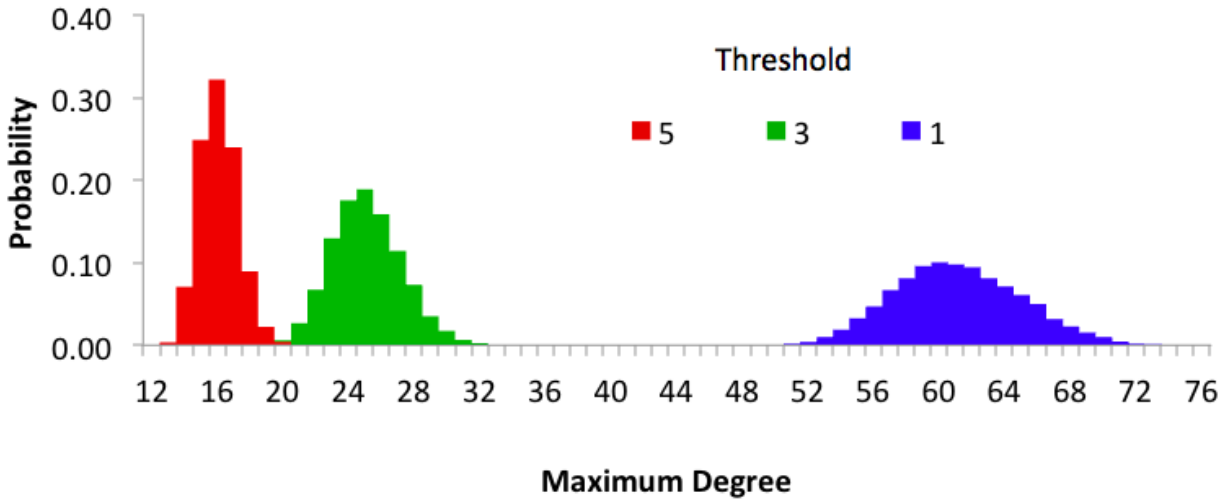
Estimating Intensities l_{ij} for IHPP Simulation

- Estimate the intensity functions l_{ij} for a pair of vertices i, j for 88 weeks (x: 4-week periods, y: number of messages).

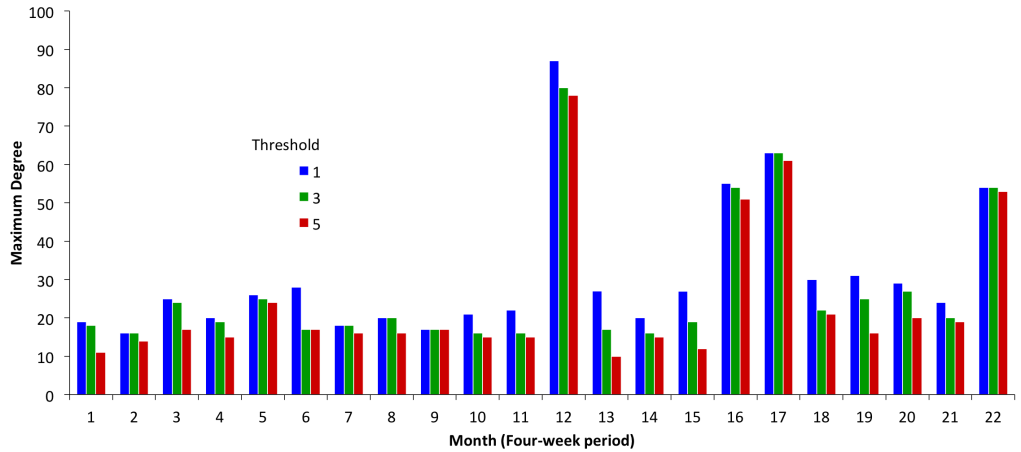


HPP Simulation: Maximum Degree Comparison

Maximum Degree Statistic for HPP simulated graphs:



Maximum Degree Statistic for Enron graphs:

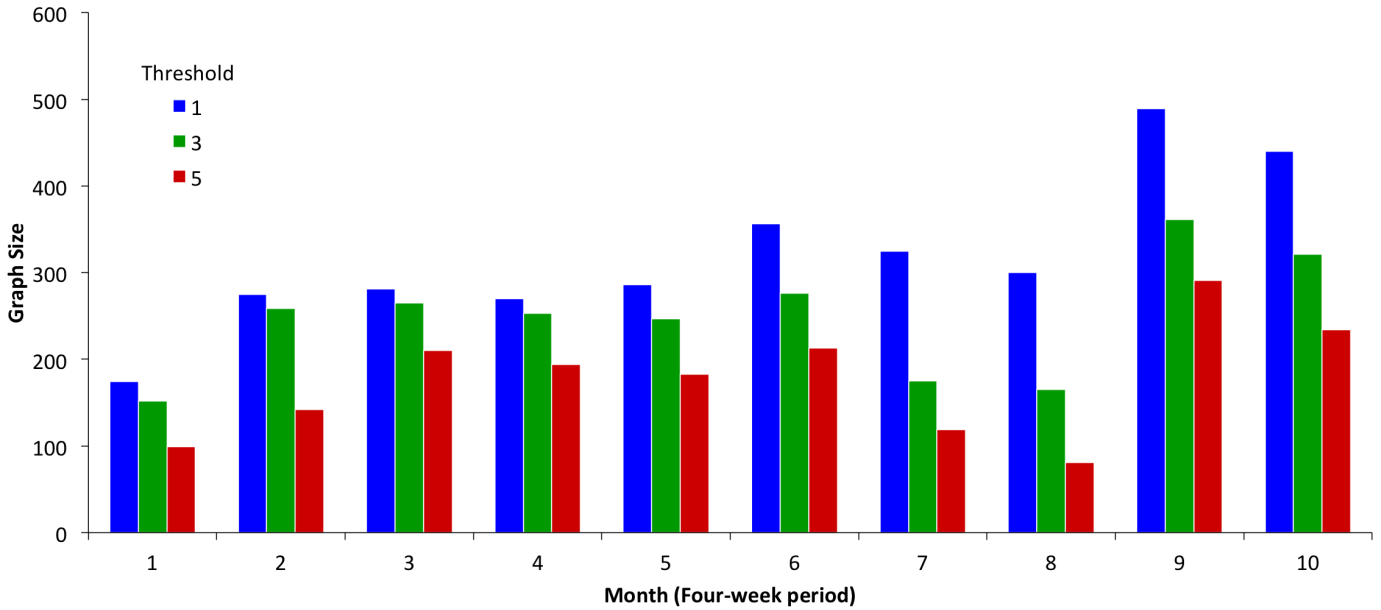


HPP Simulation: Graph Size Comparison

Percentiles for the Graph Size Statistic from Simulated Graphs:

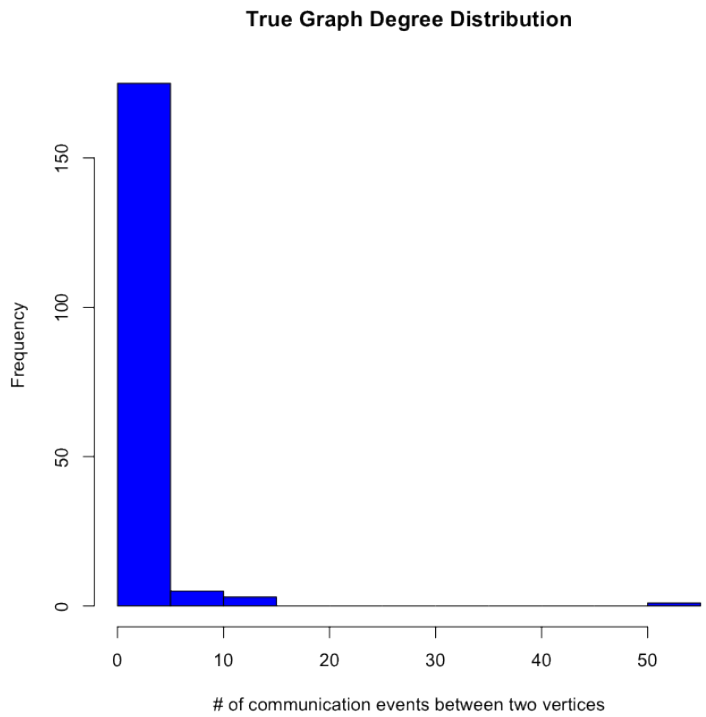
Percentile:	1	2.5	5	25	50	75	95	97.5	99
Threshold 1:	843	849	854	871	881	893	910	915	920
Threshold 3:	301	305	308	317	324	331	341	344	347
Threshold 5:	166	168	171	178	182	187	194	196	199

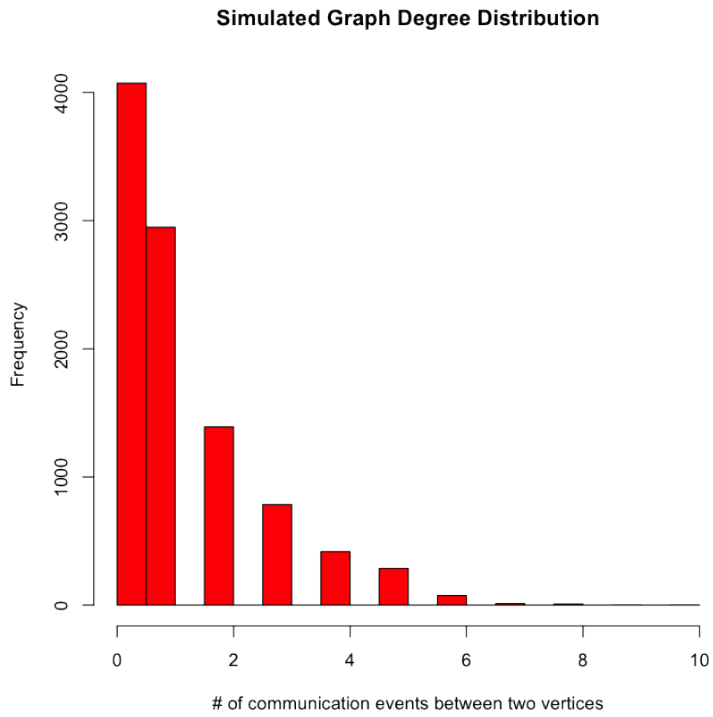
Graph Size Statistic for Enron graphs:



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 $p - value \leq 0.05$.

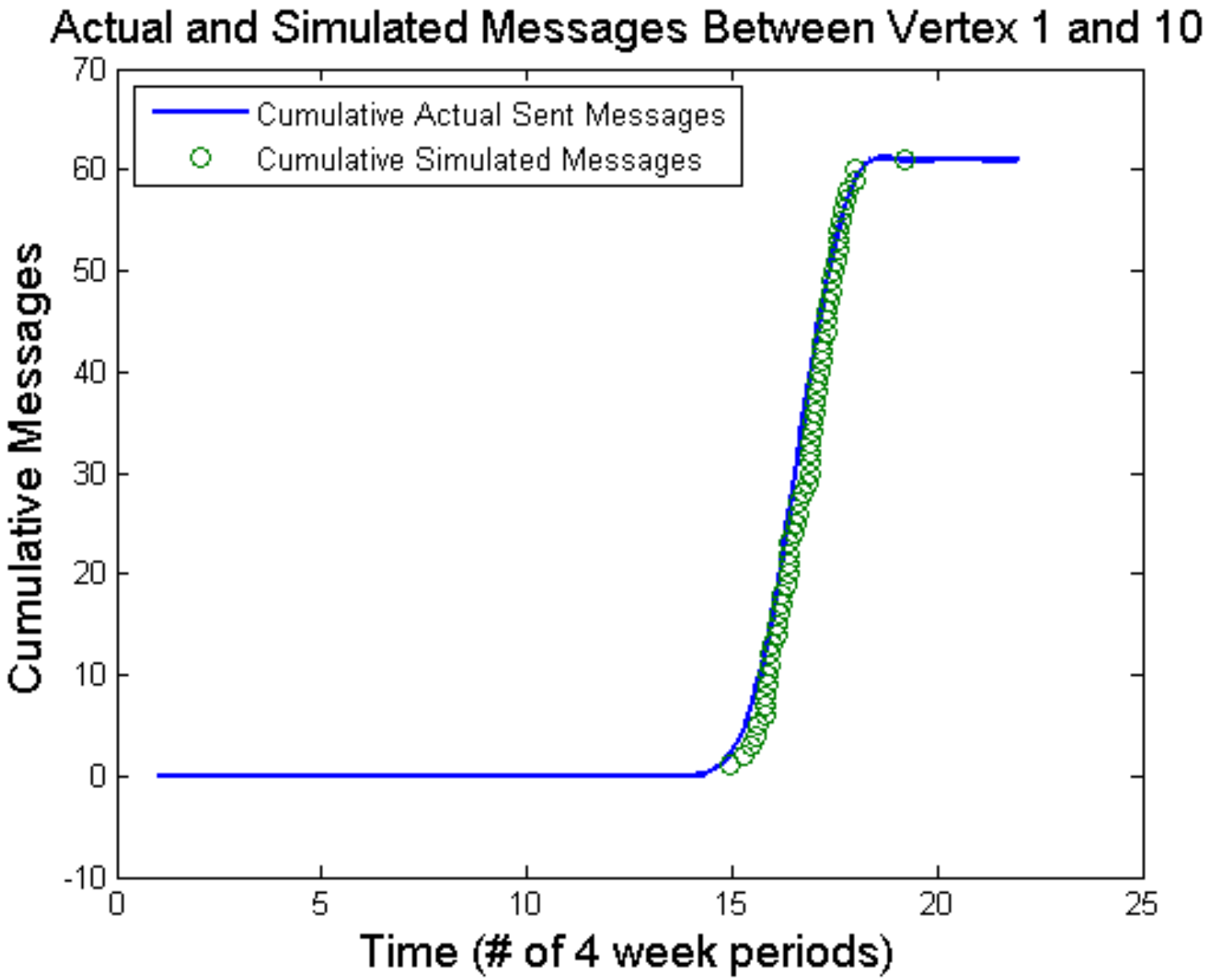




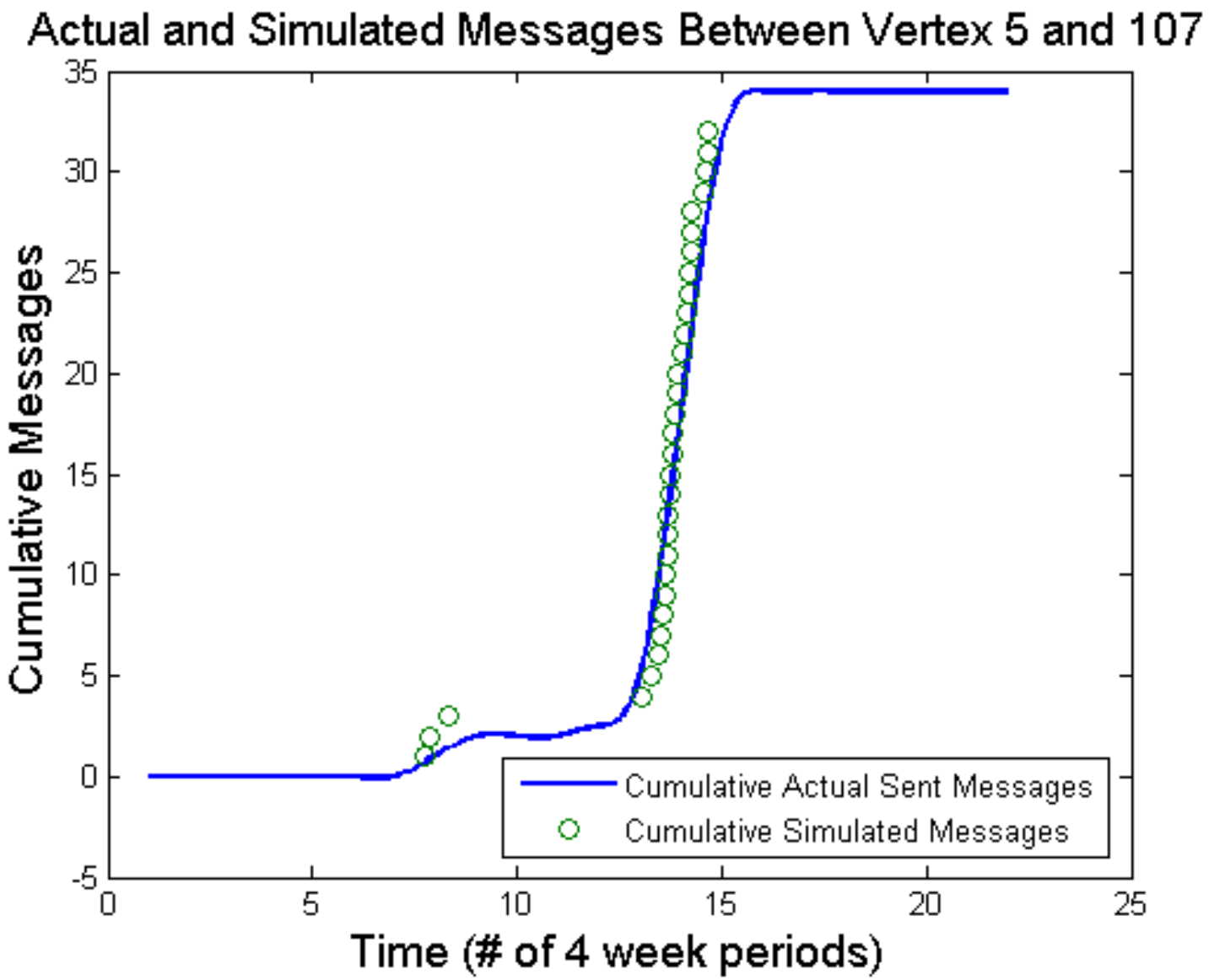
Limitations of HPP

- 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 1 and 10

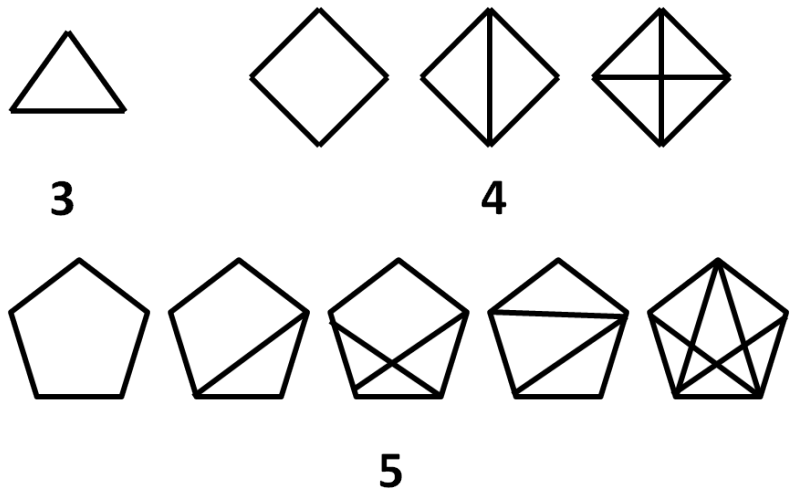


IHPP Simulation Example: Vertex 5 and 107

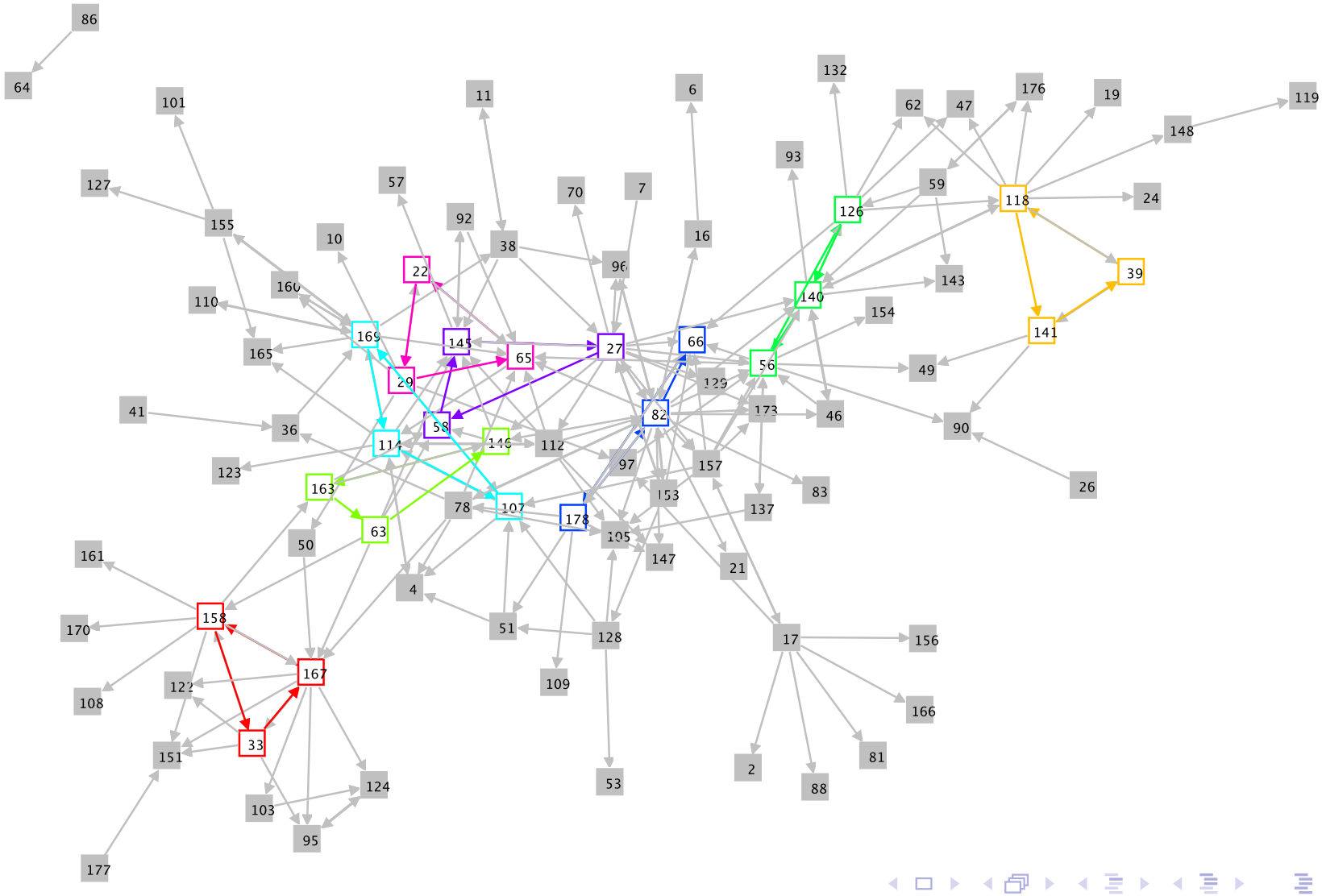


Anomaly Detection using Motifs Count

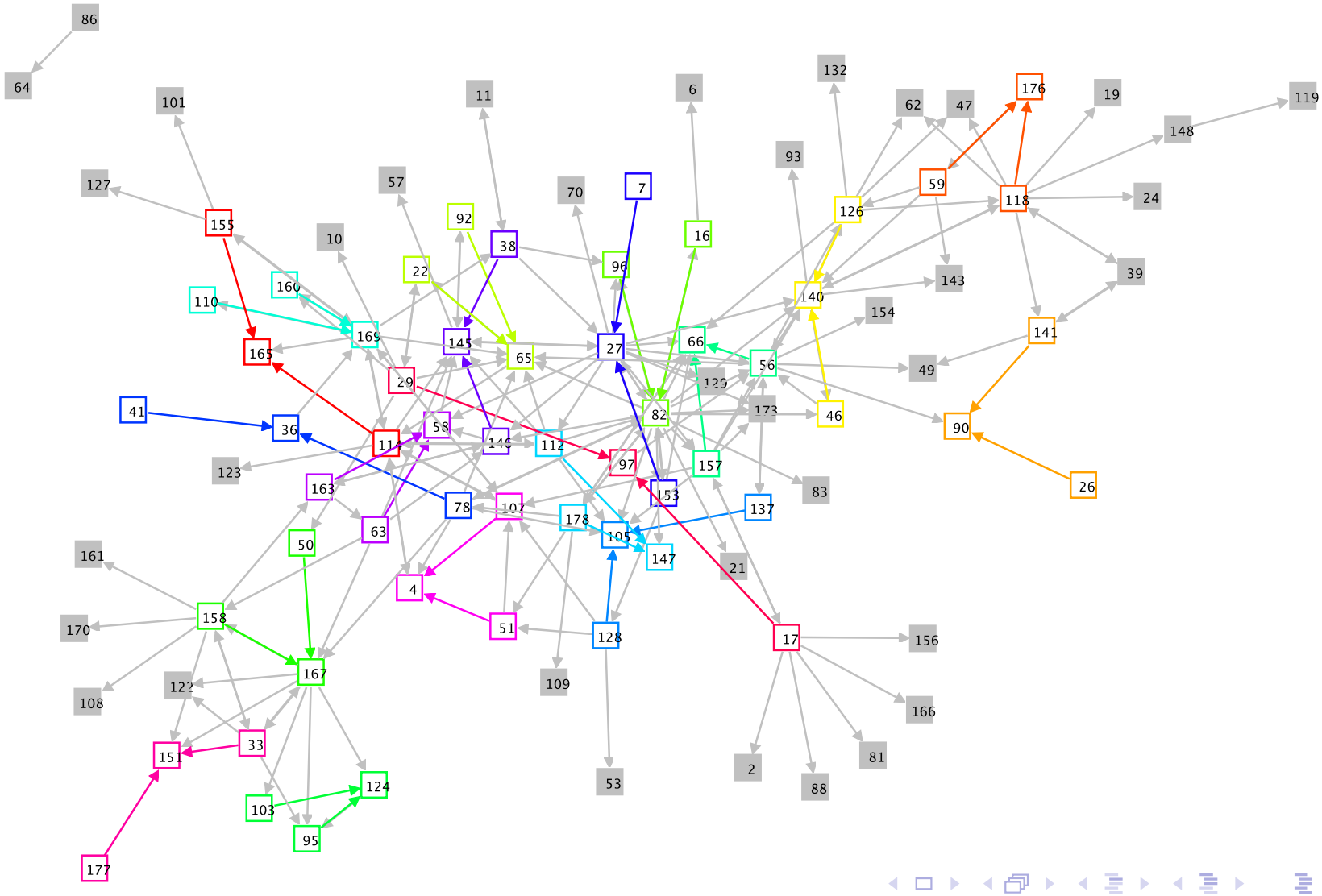
- Goal: detect small subgraphs occurring in a network more often than would be expected in a random graph.



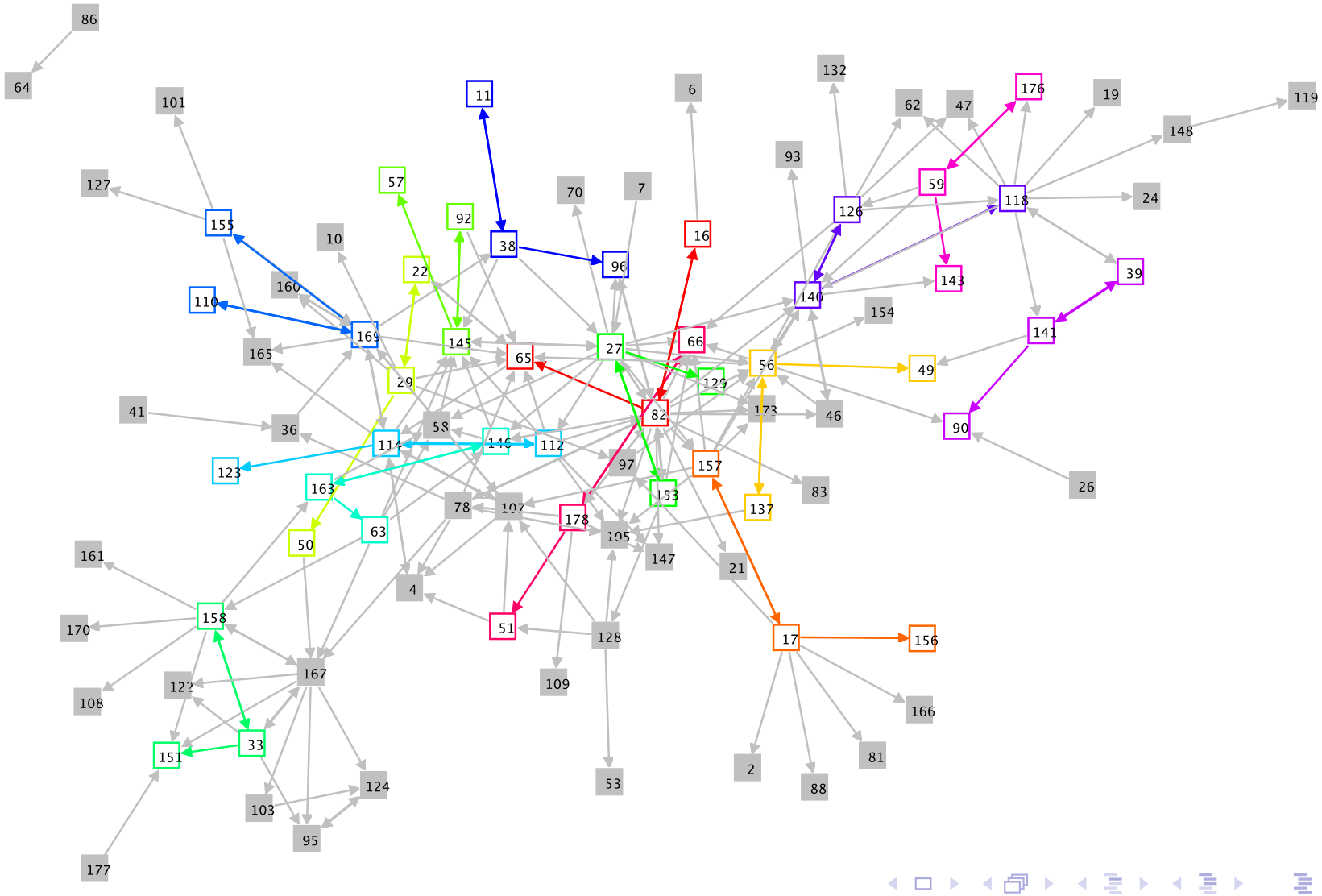
Motifs Count Statistic: GOX tuples in 4 week period



Motifs Count Statistic: GCR tuples in 4 week period



Motifs Count Statistic: F8X tuples in 4 week period



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.

References

R. Dean Malmgren, Jake M. Hofman, Luis A.N. Amaral, and Duncan J. Watts.

Characterizing individual communication patterns.
In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 607–616, New York, NY, USA, 2009. ACM.

Patrick O. Perry and Patrick J. Wolfe.

Point process modeling for directed interaction networks.
CoRR, abs/1011.1703, 2010.

B. Pittel and W. A. Woyczynski.

A graph-valued markov process as rings-allowed polymerization model: subcritical behavior.
SIAM J. Appl. Math., 50:1200–1220, June 1990.

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.
Journal of Statistical Planning and Inference, 141(2):1041 – 1046, 2011.

Cohen W.W.

Enron email dataset, 2009.