Spectral Analysis of Enron Email Data

Akhil Ketkar Arjun Sanghvi
akhilketkar@g.harvard.edu asanghvi@g.harvard.edu

December 2, 2014

1 Introduction

The Enron email dataset is interesting because it contains real email data from employees at a major organization that was involved in a massive fraud. The dataset contains a large amount of information that can be used to answer a number of interesting questions in areas such as Social Network Analysis, Organizational Behavior, such as: who are the key actors in the information network, are there communities in within the network, how do these features of the network evolve over time, does information flow over a network look different in a "crisis" etc. In addition to network or graph theoretic techniques, the dataset can be analyzed from an NLP perspective.

2 Brief Background on Enron

3 Data and Resulting Graphs

What dataset was used
Preprocessing done
Kind of graphs produced
Basic metrics on the graph such as degree distributions, diameter, components etc.

4 Centrality Measures

4.1 Eigenvalue Based Centrality Measures

Describe how the methods work and why they are a good measure of centrality to begin with

1. Eigenvalue Centrality

3. PageRank

4.2 Other measures of Centrality

Why might they be useful

1. In and Out degree
2. Betweenness centrality

5 Communities

Brief background on homophily and measures of it on a graph. Idea of modularity

5.1 Spectral Partitioning

Laplacian of the graph. Why is it important.
How can we use the second lowest eigenvector of the laplacian to split the partition.
Application to the Enron dataset

5.2 Modularity Maximization Using Spectral Techniques

Modularity matrix
Maximization Problem and solution using Lagrangian
Communities found in the Enron Dataset

5.3 Singular Value Decomposiiton

Low Rank Approximation of Martix
Community Detection using SVD

6 Evolution Across Time

How various features have evolved over time

7 Conclusion