Analysis of Personal Email Networks using Spectral Decomposition

***Ungsik Kim*†**

†*Department of Electrical and Computer Engineering, University of Florida, Gainesville, 32608 U.S.A.*

Summary
We analyzed personal emails in forms of network data and proposed a new method for classifying spam and nonspam emails based on graph theoretic approaches. The proposed algorithm can distinguish between unsolicited commercial emails, so called spam and non-spam emails using only information in the email headers. We exploit the properties of social networks and spectral decomposition to implement our algorithm. In this paper, we mainly used the community structure in social network to classify non-spam and proposed a new method for edge partition of networks. We tested our method on one of author's mail box, and it classified 41% of all emails as spam or non-spam emails, with no error. And these results are obtained with only few subnetworks resulted from the proposed decomposition method. It requires no supervised training and soley based on properties of networks, not on the contents of emails.

Key words:
Spectral decomposition, Spam email, Laplacian matrix, eigenvector centrality, orthogonal projection

1. Introduction

We are facing the explosion of spam-unsolicited commercial email-everyday and having a spam wave-more like a tsunami. Recent study has shown that the volume of junk mail on the Internet at large began skyrocketing in 2006, after a lull in growth rates in late 2005. It also says that 63 billion spam messages were sent in October 2006, more the double the number of messages dispatched in October 2005. This crisis has demanded proposals for a broad range of potential solutions, such as the design of efficient anti-spam tools, calls for anti-spam laws [1].

We proposed an effective technique which can be easily implemented based on graph theoretic methods and spectral decomposition of networks. Main ideas of this algorithm are the unique characteristics in social networks and eigen-projection of matrix. In our social activities, almost all our contractual decisions depend heavily on information provided by our networks of friends. The reliability of the decisions we made, then, depends strongly on the trustworthiness of our social networks [1]. Usually, we seem to have developed the interaction strategies for generating of a trustworthy network. The common rules is that trust is built based on not only on how well you know a person, but also on how well that person is known to the other people in your existent network. This strategy results in community structure that is one of important issues in social network studies. It is also known as one of properties of small-world networks, And, this concept can be extended to the cyberspace as well, and can be used to find some features for spam fighting tool. The emails originating from person one of user's friend or friend's friends can be trustworthy or non-spam. After construction of personal emails network, then we can apply many network analysis techniques to provide an effective and automated algorithm. We propose a new spectral decomposition and eigen-projection for this purpose.

In next two sections, we discuss the construction of personal email network, analysis of network, and implementation. In section 4, we show that our algorithm can classify nearly half of all email messages with no error. This network-only-based algorithm leaves subnetworks of the messages unclassified. The remaining of unclassified are related to subnetworks which are too small size to allow the statistical and analytic determination. And the performance of this method can be enhanced with a simple book-keeping of recipient addresses in the sent box of user and combination with other anti-spam approaches.
Section 5 is devoted to concluding remarks of this study.

2. Email Network Data

In this work, we build email network data based on the information which available to one user of email system, specifically, the header of all the email messages in user's inbox. Every email header has a unique id, date information, the email address of one sender, the list of recipient addresses, referenced message, and in-reply-to

Manuscript received April 5, 2007
Manuscript revised April 25, 2007

message. These information are stored in the “message-id”, “date”, “From”, “To”, ”Reference”, and “In-reply-to” fields. We retrieve an email network by first creating nodes representing all the addresses in the “From” and “To” fields and consider the message-id in the “message-id”, “reference”, and “in-reply-to” fields as nodes. Edges are added between message-ids and addresses that appear in the same header. Then, all nodes representing owner’s email addresses are removed, because we are only interested in links among nodes that communicate via the user. Fig. 1 shows an example of this process.

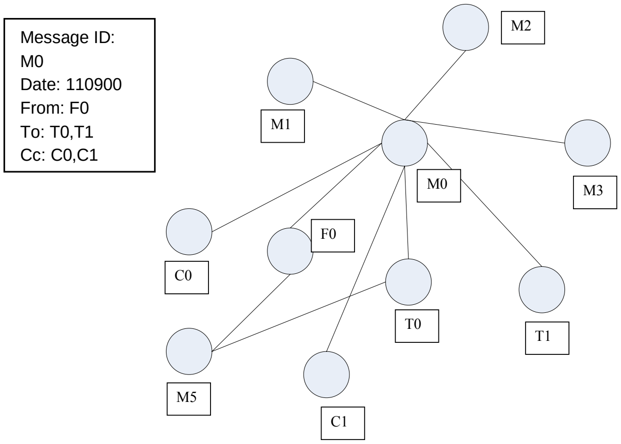
Message ID:
M0
Date: 110900
From: F0
To: T0,T1
Cc: C0,C1

Fig. 1: The subgraph resulting from an example message which has M0 ID

We consider and retrieve email network data as the simplest networks having undirected, unweighted single edges between pairs of vertices.

3. Analysis and Implementation

3.1 Properties of Network

Networks are the most common features linking diverse systems ranging from the technological, biological, economic, and social system. As one of example of technological systems, the internet is a complex network of computers and routers connected by various links. On social network, nodes are human beings and edges represent various social relationships[3]. As we mentioned, email networks can be considered as extension of social network. Because of the omnipresence of networks, many efforts have been given to uncover the organizing principles that govern the formation and evolution of various complex networks. As one of these efforts, the graph theory based approach have been attempted to analyze the complex networks using specific quantities such as degree distribution and clustering coefficients [4]. In this study, to find effective methods for spam filtering, we start analysis of network data by checking centrality measures, which are some of the most fundamental and frequently used measures of network structure. The centrality of a node in a network is a measure of the structural importance of the node. A person's centrality in a social network affects the opportunities and constraints that they face. The centralities used in this study are degree and eigenvector centralities[5]. The degree centrality is simply the number of nodes that a given node is connected to. If the network consists of who knows whom, degree centrality is the number of people that a given person knows. The eigenvector centrality is a measure if the importance of a node in a network. It assigns relative scores to all nodes in the network based on the connections to nodes having a high score contribute more to the score of the node in question[6]. The eigenvector centrality can expressed as:

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{N} A_{i,j}\, x_j \qquad (1)$$

Let $A_{i,j}$ be the adjacency matrix of the network Hence $A_{i,j} = 1$ if the i^{th} node is connected to the j^{th} node, and $A_{i,j} = 0$ otherwise and x_i is the i^{th} component of the eigenvector corresponding to the eigenvector λ gives the centrality score of the i^{th} node in the network.

Clustering is a common property of social networks that cliques from, representing circles of friends or acquaintances in which every member knows every other member[7]. As a qualitative measure of the closeness of a community, the clustering coefficient of a network is used[7]. When we focus on a selected node i in the network, having k_i edges which connect it to k_i other nodes. If the nearest neighbors of the original node were part of a clique, there would $k_i(k_i -1)/2$ edges between them. The ratio between the number E_i of edges that actually esist between these k_i nodes and the total number $k_i(k_i -1)/2$ gives the of the clustering coefficeint of node i. The clustering coefficient (also known as transitivity), C, of a graph can be expressed as:

$$C = \frac{2E_i}{k_i(k_i - 1)} \qquad (2)$$

$$C = \frac{3 \times (number\ of\ triangles\ in\ the\ graph)}{number\ of\ wedges} \qquad (3)$$

3.2 Spectral Decomposition of Network

The structure of networks can be described by the associated adjacency matrices. The adjacency matrices of undirected graphs are symmetric matrices with matrix elements, equal to number of edges between the given vertices. In this study, we can break down the entire email network to the summation of subnetworks.

$$\hat{P} = v\ v^T \tag{4}$$

P is orthogonal Projection on subspace[4]. And any matrix S can be represented as a combination of the weighted projection matrices.

$$S = \lambda_1 P_1 + \lambda_2 P_2 + ... + \lambda_m P_m \tag{5}$$

where λ_i is eigenvalue and *Pi* is projection on subspaces. In this work, we used Laplacian matrix for measure of eigenvector centrality and spectral decomposition of email network data instead of an adjacency matrix. The Laplacian matrix L is defined as:

$$L = Diag(A) - A \tag{6}$$

Where, Diag(A) is a diagonal matrix with the row-sums of A along the diagonal. A is an adjacency matrix.
The eigen-projection of using Laplacian matrix can be considered as a modified eigenvector centrality in Eq. (1). It has more advantages than conventional eigenvector centrality. It gives the importance of the node but also gives information of links between nodes.

For classification of spam, we break the whole network into a set of subnetworks using Eq. (4), (5)., then classify each subnetwork according to metrics such as clustering coefficient and eigen-projections. In general, the subnetworks which have high clustering coefficient values can be classified as non-spam.

4. Applications and Results

We obtained empirical data from one of user's email box and emails have been chopped into 108 days period.

These emails contain 2500 messages and converted to a network which has 3755 nodes and 6930 edges. All nodes representing the user's own email addresses are removed, since we are interested only in the connections among nodes who communicate via the user. Fig. 2 shows a personal emails network for test.

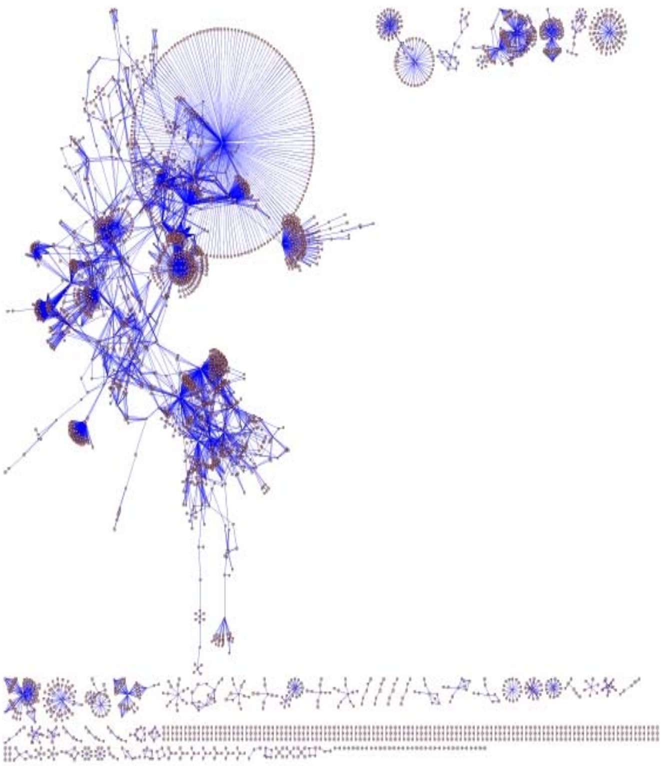

Fig. 2: A complete email network

We can obtain subnetworks from the the email network by the spectral decomposition in Eq. (4), (5). Fig. 3 shows the eigenvalue spectrum follows the power-law.

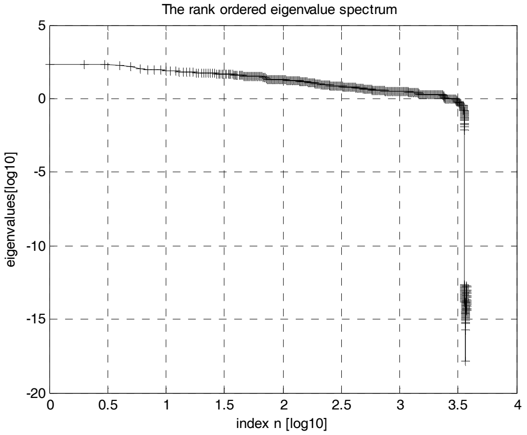

Fig. 3: The rank ordered eigenvalue spectrum of email network

Using eigenvalues and corresponding eigenvectors, we break the entire email network into the summation of subnetworks. Even the order of matrix is 3755, the most links are covered by small numbers of subnetworks. Fig. 4

shows that 90% of links in the network is covered by only 727 subnetworks.

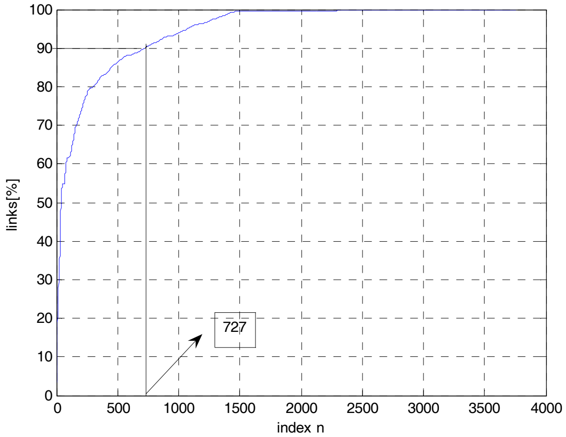

Fig. 4: Cumulative number of links in subnetworks

Modified eigenvector centrality based on Laplacian matrix is used to construct the subnetwork and classify spam.
Fig. 5 shows one of eigen-projections that can give the importance of node and link information. In this case, two nodes have high positive values act as an important actor in the networks.

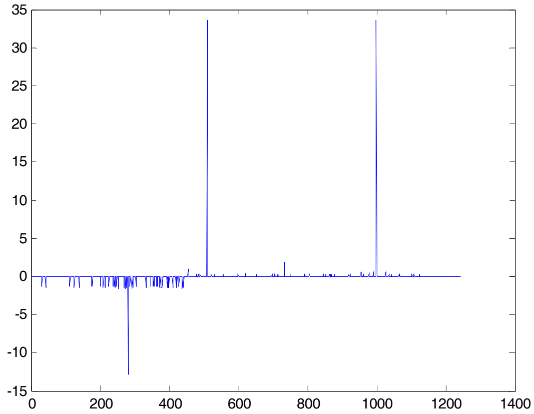

Fig. 5. One of Eigen-projections

We can distinguish 43% of non-spam and 36% of spam without error. It is interesting to note that 43% of non-spam can be obtained only using 141 subnetworks and 36% of spam can be classified only using 10 subnetworks. It also proves the effectiveness of our decomposition method. If we consider the addresses of recipient in the sent box, the performance of this method can be easily enhanced.

Table 1: Result of the algorithm

Date	Classified/total
Total	2844/6930
non-spam	2285/5346
spam	559/1586

5. Conclusion

We have proposed an algorithm based on the properties of social networks and spectral decomposition to distinguish spam and non-spam emails. Since, the only information necessary for this method is available in the user's email headers, the algorithm can be easily implemented and combined with other filtering process. The best content-based filters achieve approximately 99.9% accuracy, but require users to provide a training set of spam and non-spam message. This algorithm can automatically generate an accurate training set for learning of more sophisticated content-based filters. In this paper, we also proposed a new edge partitioning method and a measure of centrality using the eigenvector of well-known Laplacian matrix. The overall performance of this method can be enhanced with a simple book-keeping- considering the addresses of recipients in the sent box to classify.

Acknowledgment

The author would like to thank Dr. Boykin for his valuable advice.

References

[1] P. Oscar Boykin, and Vwani P. Roychowdhury, "Leveraging Social Networks to Fight Spam", IEEE Computer, Vol. 38, No. 4, page 61-68, Apr 2005.
[2] Michelle Girvan, and M.E. J. Newman, "The structure and function of complex networks", SIAM Reviews 45, 167-256, 2003.
[3] Michelle Girvan, and M.E. J. Newman, "Community structure in social and biological networks", 2001. arXiv:cond-mat/0112110 v1
[4] Edwin Olson, Matthew Walter, Seth Teller, and John Leonard,"Single-Cluster Spectral Graph Partitioning for Robotics Application"
[5] Phillip Bonacich,"Power and Centrality: A Family of Measures", The American Journal of Sociology, Vol 92, No. 5, pp 1170-1182, 1987
[6] S. Wasserman and K. Faust, *Social Network Analysis.* Cambridge University Press, Cambridge, 1994
[7] R Albert and A.L. Barabasi, "Statistical mechanics of complex networks", Rev Mod. Phys. 74, 47-97, 2002.