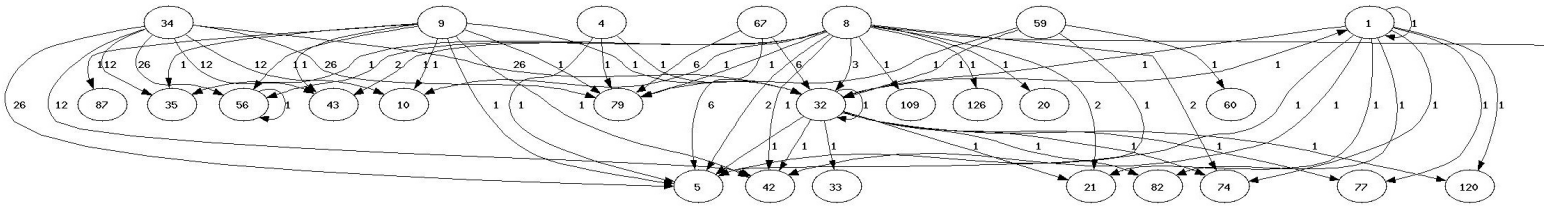
Communities of Interest in Email

Lisa Johansen, Michael Rowell, Kevin Butler, Patrick McDaniel

Email Connectivity



This connectivity graph represents a small subset of email communication within Penn State's CSE Department. This graph is part of a study of the characteristics of email communication namely, communities of interest. Because email has become an integral and sometimes overwhelming part of users' personal and professional lives, the application of the knowledge gained in this study can lead to valuable email processing tools.

Communities of Interest

Communities of interest are sets of entities which share a common interest. They have been examined within both the telephone network and data networks. The research in these fields has lead to extensive applications including "guilt by association" and anomaly detection. Our research is the first to examine these communities within an email network.

Email COI Applications

A number of email-specific applications use elements of social network dynamics to improve their functionality. By learning more about communities of interest within email, we may also be able to use them as a tool within these applications. Examples include spam filtering, email classification/prioritizing, and virus identification.

Factors of Email COI

This research examines email log files which are the most concise and easily accessible source of information about email traffic. Unlike email studies which use complex and hard-to-get email data, using only email logs simplifies the development and deployment of communities of interest. The communication characteristics provided by this limited data are discussed here.

Number of Emails

If emails are sent or received between two contacts, there is an obvious indication of some type of connection between the contacts. The more interesting problem is discovering how many emails signifies a meaningful relationship. This would be useful in many possible applications.

Frequency of Emails

Email communication frequency may indicate strength of relationship between two contacts. The decline of email frequency may also indicate a change in COI. Thus, COIs can be added to when more frequent communication is seen and removed from when frequencies decrease.

Collaboration

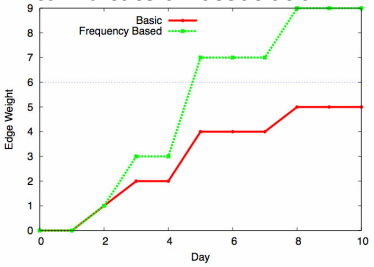
Collaboration involves sharing COIs among associated contacts. Thus, even though two users are not directly linked, they may be indirectly linked through members of their individual COIs. This concept, commonly seen in graph problems, may increase the usefulness of COIs within applications.

Experiments & Results

Our initial research consisted of collecting and organizing Penn State CSE email log file data, constructing a connectivity graph, analyzing the link weights (sent and received emails), studying the frequency of emails, and constructing basic COI models. These models examined different numbers of emails, frequencies, and collaboration. We tested the accuracy of our models by comparing their classification results with actual human classification data.

Our initial studies yielded the following results:

- Sent emails are a stronger indication of an association than received emails
- COI membership is not stable over time
- COI associations are transitive
- Frequent emails are highly likely to indicate an association



- Users exit COIs frequently