Email Archive Analysis Through Graphical Visualization

Wei-Jen Li
Columbia University
weijen@cs.columbia.edu

Shlomo Hershkop
Columbia University
shlomo@cs.columbia.edu

Salvatore J. Stolfo
Columbia University
sal@cs.columbia.edu

ABSTRACT

The analysis of the vast storehouse of email content accumulated or produced by individual users has received relatively little attention other than for specific tasks such as spam and virus filtering. Current email analysis in standard client applications consists of keyword based matching techniques for filtering and expert driven manual exploration of email files.

We have implemented a tool, called the Email Mining Toolkit (EMT) for analyzing email archives which includes a graphical display to explore relationships between users and groups of email users. The chronological flow of an email message can be analyzed by EMT. Our design goal is to embed the technology into standard email clients, such as Outlook, revealing far more information about a user's own email history than is otherwise now possible. In this paper we detail the visualization techniques implemented in EMT. We show the utility of these tools and underlying models for detecting email misuse such as viral propagation, and spam spread as examples.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications applications – *electronic mails.*

General Terms

Design.

Keywords

Email, spam, virus.

1. INTRODUCTION

The Email Mining Toolkit (EMT) developed at Columbia University [3] is a data mining system that computes **behavior profiles** or **models** of user email accounts. These models may be used for a variety of forensic analyses and detection tasks. The toolkit is useful for report generation and summarization of email archives, as well as for detecting email security violations when incorporated within a real-time violation detection system, such as the MET system [1].

EMT provides the means of loading, parsing and analyzing email messages from a wide range of storage formats. It not only demonstrates the statistics of email account behavior, it also computes the volume and velocity of emails exchanged between parties, analyzes specific content and patterns, and explores social relationships between groups of users, and the relative rankings of importance of different individuals in an organization.

Moreover, EMT extends these kinds of analyses to model "user behavior" at a very fine granularity. It models the behavior of individual user email accounts or groups of accounts, and can be used to detect changes in behavior that may be of interest in forensic analyses. These features of EMT provide the means to detect fraudulent misuse and attacks such as viruses and Spam (unwanted) email.

EMT includes 15 different features and models. The statistical models that include stationary and non-stationary user profile are used to generate user behavior models. These models include

- *Message Table* where individual emails may be automatically classified by built in machine learning subsystems,
- *Usage Histogram* revealing a user's typical daily email behavior,
- *Similar Users* which identifies groups of emails users who behave in similar ways
- *Recipient Frequency* providing a detailed analysis of the typical communicants with a user and
- *Attachment Statistics* detailing attached files serving as a personal file system of a user, as well as the statistical analyses including the birth rate, lifespan, incident rate, prevalence, threat, spread, and death rate useful in identifying interesting attachments and viral attachments.

The analyses built in to EMT concerning groups of accounts and their communication is provided to detect violations of group behavior. These models include

- *Enclave Clique* groups of users who frequently pair wise exchange messages,
- *User Clique* the set of accounts a particular user typically emails as a group,
- *Email Flow* revealing how a single message produces a web of new communication throughout an organization and
- *Average Communication Time* that views a user's typical response rates to individuals, indicating the relative importance of communicants.

EMT's graphical user interface provides an easy to use interface to execute these functions and that visualizes results in tabular form with displays of plots and histograms that are easy to understand. In this paper, we introduce two important visualized

VizSEC/DMSEC'04, October 29, 2004, Washington, DC. USA.

graphical models –the visualization of *Cliques* and *Email Flows*. A quick inspection of these displays reveals a number of interesting pieces of information including how a viral propagation appears in a uniform (and quite appealing) flow pattern, and the relative importance of an individual user or an email message on the basis of how connected they are to other users or other emails generated within an organization.

2. CLIQUES

In order to study email flows between groups of users, we can compute a set of *cliques* based on the data present in an email archive. We seek to identify clusters or groups of related email accounts that participate with each other in common email communications. Conceptually, two broad types of cliques can be extracted from user email archives: *user cliques* and *enclave cliques*. In simple terms, *user cliques* can be inferred by looking at the email history of only a single user account, while *enclave cliques* are social groups that emerge as a result of analyzing traffic flows among a group of user accounts within an organization.

The graph clique finding algorithm is described in [2]. The algorithm initially counts the number of emails exchanged between any two given users, regardless of the direction of the traffic flow. This absolute number is compared to a threshold. If the count is above the threshold, then the link between the two underlying accounts is established and a graph begins to form. This is computed over all emails and all email accounts appearing in all messages, whether sender or receiver of an email. At this point, we have a list of cliques of size 2, called dyads. The dyad members in each clique are sorted lexicographically, and the cliques themselves are sorted in increasing order also. We employ the hierarchical algorithm at this point and build lists of cliques of size n, with n increasing by 1 with each iteration. Throughout the execution of this algorithm, the list of cliques at level n is sorted, both among the sets of cliques, and within a set. The hierarchical algorithm is a an iterative procedure with a two-step process – the first step is to generate a candidate set and the second step is to remove candidates that do not meet the clique definition. When the algorithm terminates a completely connected graph is formed and ready for display.

2.1 Visualization of Cliques

In order to be able to run the analysis quickly using low memory overhead on typical desktop machines, we resorted to using two 2-D graphs and simple icons to solve both problems.

Figure 1, which is called the *Clique Panel,* displays the cliques and their connectivity. Links between nodes indicate that there is a common member of both cliques. Each node is itself a clique email accounts. The detailed information, such as the members of the clique, the sharing users between cliques and the most common words appearing in the subject lines, are shown in another pop-up window. The nodes are small polygons, and the number of edges of these icons represent the number of members of the clique. For example, a triangle represents a 3-clique, a rectangle represents a 4-clique, and a 2-clique is circle.

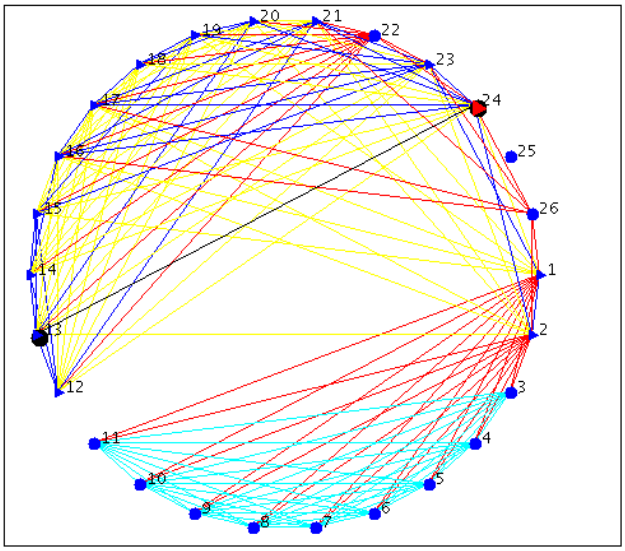

Figure 1. Visualization of Clique, the Clique Panel.

The edges in this graph represent the relationship of clique members. If there are two cliques with some common member, EMT displays a link between the two cliques. We use different colors to represent the percentage of users that the two cliques share, indicating how similar the cliques are to each other. The color mapping is shown in Table 1. For example, cliques A and clique B share 30% of users, and hence they are linked with a red edge. When the user clicks on an edge, the common users appearing in both cliques is displayed.

Table 2. The degree of sharing users

Color	Percentage
Orange	<10%
Yellow	> 10%, <= 20%
Red	> 20%, <= 30%
Cyan	> 30%, <= 40%
Blue	> 40%, <= 50%
Magenta	> 50%, <= 60%
Green	> 60%, <= 70%
Gray	> 70%, <= 80%
Pink	> 80%, <=90%
Black	> 90%

Figure 2, which is called the *User Panel*, demonstrates the relationship between users and their clique membership. The blue nodes aligned as the left most column of the display are the (indexed) cliques displayed in figure 1. The black nodes are the users. Each distinct black node corresponds to one distinct email address. They are placed in different columns depending upon the number of cliques they belong to. For example, a user who belongs to only one clique is in the first column immediately

adjacent to the blue colored clique nodes. A user who is part of two cliques is in the second column, and so forth. The users aligned at the right-most column are those who are members of the most cliques, and may be regarded as very significant well connected individuals in an organization.

In this panel, an edge connects a user to all of the cliques they belong to. The color of the edges in the graphical display makes the graph easier to be read but provides no additional information. The alignment reveals the "density" of the user.

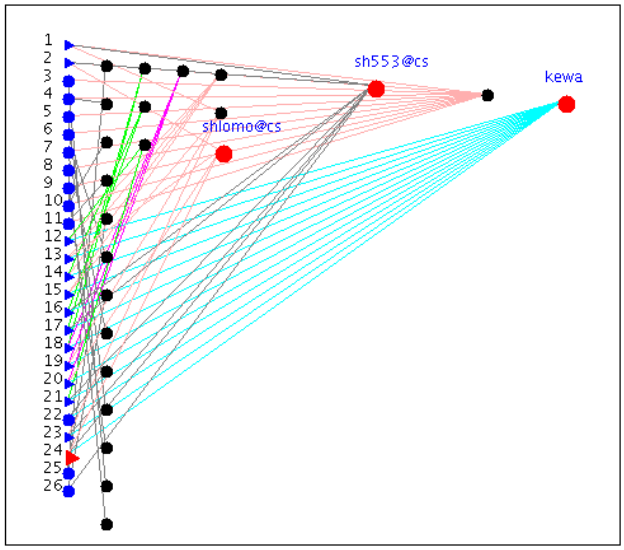

Figure 2. Visualization of Clique, the User Panel.

To combine these *Clique* and *User Panel*, we can visualize a 3-D graph. Think about three axes, x, y and z. The *Clique Panel* is the x-z plane and the *User Panel* is the y-z plane. Instead of plotting a 3-D graph, we use the combination of the two 2-D graphs to reveal the same information but at far less cost.

Because all pairs of cliques may share users (e.g. have edges between them), all of the pairs of nodes may have an edge that connects them. We need an algorithm that arranges the edges without any overlap. Think about a regular polygon. If we draw an edge between each pair of nodes, there are no overlapped edges. Therefore, the algorithm we implemented firsts draw a big circle and then places all of the *n* clique nodes on this circle equally spaced. These nodes form a regular *n*-polygon. An example is displayed in figure 1. All of the nodes are placed around the circle and the edges are drawn between nodes within this circle. By using this representation, the graph is clear and easy to read without any overlapping edges.

In the *User Panel*, we also reduce the overlapping edges. However, it's harder to eliminate all possible overlaps. We use two simple ways to make the graph clear, even so there are still a few overlapping edges. First, each column of user nodes (e.g. the black nodes) has a distinct color of edges to connect to the clique node (e.g. the blue nodes). Second, the location of the nodes in a column is lower than the nodes in the previous column. For example, see the black nodes in Figure 2. The location of each node of the second column is lower then the corresponding node of the first column.

These two panels visualize the general structure of clique analysis. They demonstrate almost all of the information cliques reveal. These graphs can be used not only for email cliques, but also for Instant Messages (IM) and any kind of messaging, even IP connections between host computers in a network. A version of EMT is under development for these purposes revealing how client computers interact with each other and with servers.

2.2 Anomaly Detection Using Clique Violations

The clique information can be used to identify unusual email behavior that violates typical group behavior. Clique violations may also indicate email security policy violations internal to a secured enclave. For example, members of the legal department of a company might be expected to exchange many Word attachments containing patent applications. It would be highly unusual, and probably unwise, if members of the marketing department, and HR services would likewise receive these attachments.

The technique is that we can infer the composition of related groups by analyzing normal email flows to compute the naturally occurring cliques. These cliques are defined to be the normal behavior of an organization. Then we use the learned cliques to alert when new email communications between groups of users violate that clique behavior. Clique violation is a useful anomaly detection tool. We have performed several experiments demonstrating the power of this technique reported in [3].

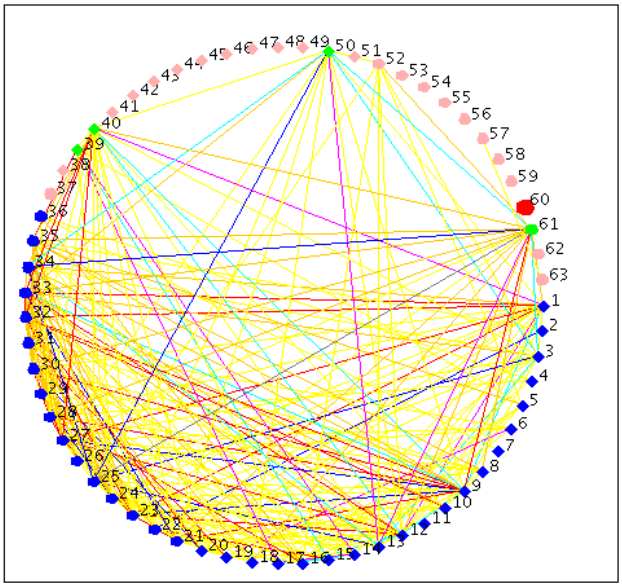

Figure 3. A Clique Violation Example.

Figure 3 demonstrates a clique violation test. The cliques in this experiment are grouped by subject. The blue nodes (#1 to #36) are

normal cliques that are computed for a set of emails identified as the training period. The pink and green nodes (#37 to #63) are cliques that are identified as testing data. (EMT allows the user to choose any arbitrary set of data to profile and test.) The green nodes (#38, #39, #49 and #61) are labeled as normal cliques and the pink ones are alerts.

The detection algorithm is quite simple and is based upon a threshold, ***T***, governing the percentage of clique members shared between the test and training cliques. If a comparison between the test clique and a trained clique is shown to share more than ***T***% of their respective members, the test clique is regarded as normal, otherwise and alert is generated and the corresponding nodes are displayed in the appropriate color identifying likely clique violations.

Since our models use statistical approaches requiring sufficient training data to estimate empirical distributions, the system cannot easily detect spam messages or propagating viruses that appear only once. Sufficient statistics are needed before such email events become easily noticed. In any event, when such events do occur, they are easy to spot when visualized in the manner displayed in Figure 3.

3. EMAIL FLOW

The email flow analysis in EMT visualizes how an individual chosen email message permeates a communication event through an organization and shows how people relate to each other via these flows. We relate message flows containing the same content (via EMT N-Gram analysis described in [4]) as flow networks starting from a target email message chosen by the analyst using EMT. This kind of analysis may be useful in understanding whether or not some email user or their messages have any impact on an organization, or are just summarily ignored and have no effect. Mangers and industrial engineers may find such a tool useful in organizing teams of employees, and identifying those who are held in high regard and thus impact the organization and its communication flow.

The flow starts at an individual email message, and follows the email flow either by subject, attachment or by similar content between messages appearing later in time emanating from recipients of the original message. We first gather all of the equivalent or similar emails and display these in the window details including the relationship of the senders and recipients, their content, and the flow over time.

We can visualize the flow of a message and easily find cliques between the email accounts via visually exploring and manipulating this graph. Thus, the flow pattern indicated by nodes (email accounts) and edges (message exchanges between accounts) define interesting “content based cliques” within an organization.

3.1 Features of Email Flow and Examples

In Figure 4, the panel displays the relationship between users. Each node is a distinct sender or a recipient. An edge between two nodes represents a connection or multiple connections, and the different colors represent one-way or two-way connection(s).

The most important feature in this panel is that it shows a series of concentric rings. The inner-most ring and node, corresponds to the original target email. Time is depicted by stepping outward ring by ring. The nodes in the next closest ring are accounts that have exchanged email in the next time step. The spread and number of edges and nodes provides a view of how an individual message effected communication over a broad set of email accounts and the time frame that this spread occurred.

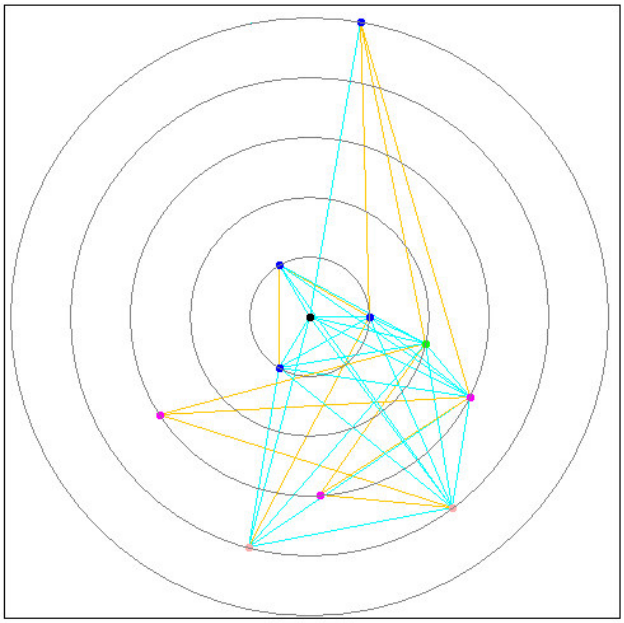

Figure 4. A Normal Discussion Via Emails.

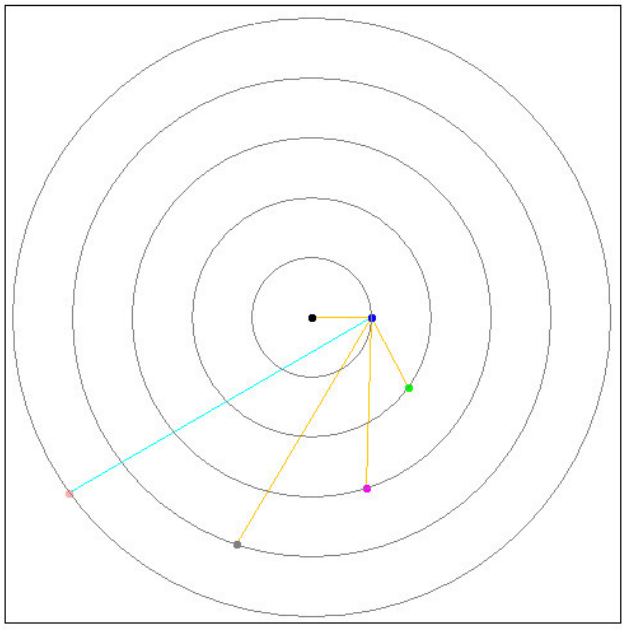

Figure 5. A Series of Emails Generated by Spambot.

Figure 4 displays a normal email discussion. The emails are grouped by subject. We can see that people exchange messages with one another. Although these 11 nodes don't form a full clique, we can conclude that they are a strongly related people in a group. Figure 5 displays the flow of a spam email. Some email account blasted out quite a few copies of the same email to many users in rapid succession. Notice the regularity of the plot indicating a uniform flow pattern logically produced by an automatic procedure, probably a spambot. The name of the sender is the same in each message and hence each node is linked to the center node, even though the domain names were found to have been changed in each message. There are five nodes that send emails to one recipient. The plot looks so highly regular it is obviously different than normal human communication patterns. Figure 6 displays the flow of the propagation of a virus (SoBig.F) and is grouped by attachment similarity. Hence, EMT provides a means of equating email messages on the basis of similar content, similar subject line and similar attachment content. The virus creates a random source address of the email, and propagates itself relatively quickly. The plot would appear very similar to the spam flow plot had we had enough examples in the email archive of the viral propagation.

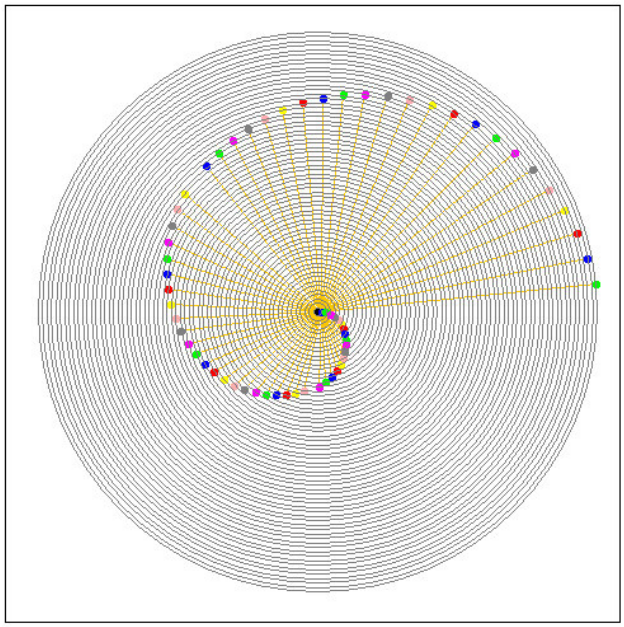

Figure 6. An Email Flow of Virus Propagation.

4. CONCLUSION

EMT is a data mining system that computes behavior profiles of user email accounts. The visualization techniques of clique and email flows provide an overview of email exchanges and user relationships within an organization. The techniques not only can demonstrate email information, but also these can be used to detect email misuse and policy violations. These models provide the behavior of user account and the behavior of group of users, such as the employees in a company or students in an institution.

EMT is a work in progress. In the future, we will develop other analyses germane to different applications and develop methods to efficiently combine and visualize the rich variety of models. Moreover, we will also try to implement these models for other similar internet applications which can be modeled in the same way email is modeled.

5. REFERENCES

[1] Manasi Bhattacharyya, Shlomo Hershkop, Eleazar Eskin, and Salvatore J. Stolfo. *MET: An Experimental System for Malicious Email Tracking*. In Proceedings of the 2002 New Security Paradigms Workshop (NSPW-2002). Virginia Beach, VA: September 23rd - 26th, 2002.

[2] C. Bron, and J. Kerbosch. 1973. Finding all cliques of an undirected graph. Comm. ACM 16(9), pp. 575-577.

[3] Salvatore J. Stolfo, Wei-Jen Li, Shlomo Hershkop, Ke Wang, Chia-Wei Hu, Olivier Nimeskern, Detecting Viral Propagations Using Email Behavior Profiles, ACM Transactions on Internet Technology (TOIT), May 2004.

[4] Damashek, Gauging similarity with n-grams: language independent categorization of text. Science.