A Temporal Based Forensic Analysis of Electronic Communication

Salvatore J. Stolfo
Columbia University
500 West 120th St
New York, NY 10027
1-212-939-7080
sal@cs.columbia.edu

Germán Creamer
Columbia University
500 West 120th St
New York, NY 10027
1-646-775-6068
gcreamer@cs.columbia.edu

Shlomo Hershkop
Columbia University
500 West 120th St
New York, NY 10027
1-646-775-6041
shlomo@cs.columbia.edu

ABSTRACT

Previous work [1] reported on our research in developing a data mining environment for analyzing email communication data. In this paper, we describe our extensions to EMT for applying forensic discovery over temporal email data. The goal is to produce a semi-automatic system to aid in evidence discovery and a host of other applications. We describe our research on profile stability, temporal search and clustering, and new social network dynamic algorithms.

Categories and Subject Descriptors

H.4 Information Systems, H.4.3 Communication Applications, Electronic mail and Information Browsers

General Terms Algorithms, Theory.

Keywords Social Networks, Email Mining, Histograms, Search, Email Visualizations.

1. INTRODUCTION

The analysis of *email flows* to and from a user's email account(s) reveals a tremendous amount of information about a person's interests, activities and behaviors that cannot be derived alone from content analyses of individual emails. In order to develop email forensic tools, we focus on several aspects of email flows: evidence discovery; profiling, interactions, and content communication.

The behavior analysis of individual users over time is one way to start an investigation dealing with email. By computing individual behavior over time, we can more accurately study changes in behavior for both individual users and groups of users, and find interesting points in time and discover influences of interest.

The second aspect of our work is to study how to apply forensics to social group interactions over time. Our approach is to visually allow the group features to be adjusted based on features which are important to the end user analyst.

Last, locating interesting emails through searching is a non-trivial task since the search terms are not always clear ahead of time. In a digital evidence framework, one would prefer a system of being able to search through emails using time as a basis but also to include content similarity across all messages. We have augmented EMT's current search capabilities to automatically expand and visualize search parameters based on time and word relationships.

2. EMT

The Email Mining Toolkit (EMT) has been in development at Columbia University since 2001 and featured in numerous publications. It is a Java-based data mining environment for large email collections focused on automatically extracting patterns of users, social group interactions, and attachment level analysis of email communication. EMT has been downloaded by dozens of organizations.

3. Profiling Account Use

User level analysis is based on profiling individual email accounts over time. Specific features are used to create a profile over a period of time ('baseline normal') and compare future behavior to this profile. This static approach has been useful in locating similar behaving accounts in a large email collection. The reality is that true behavior is dynamic over time, and a profile representation should be augmented by adjusting it over time to detect usage changes within the same account.

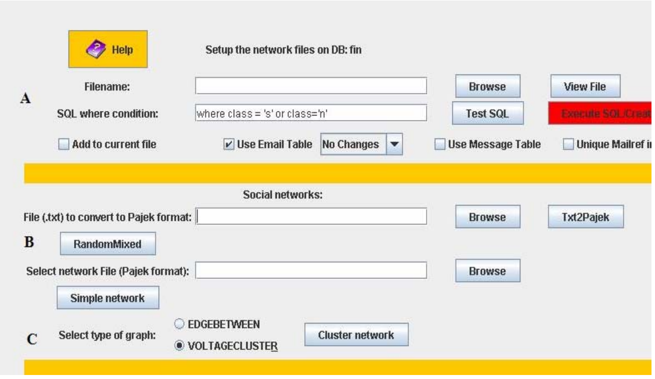
3.1 Rolling Histogram

A rolling histogram is a dynamic per user profile computed over time. We can measure profile stability over time, by sliding a set window over the behavior data, and computing a similarity measure for two adjacent time periods. For example, given a year of email data we can compute weekly or monthly profiles and compare the behavior of the account within these small increments.

We define account stability to be distance between periods under some average threshold. One can view the changes in histogram distance scores over time using the interface. Alerts are generated for those periods which signify unusual behavior allowing time periods to be further investigated (i.e. all emails can be examined over a specific period to see what triggered the alert).

4. Social Mining

EMT allows social networks to be extracted both on a per user basis and per enclave basis. We have extended the basic feature with the ability to generate clusters among related email accounts between senders and recipients (figure 1). It calculates the shortest path length (average distance) from a specific vertex to all vertices in the graph, and the cluster coefficient for each vertex. For every user, these indicators should be stable among a community of users. Therefore, outlier values may indicate suspicious patterns. This module is also able to discover communities of users based on voltage drops or weaker connections across networks (figure 2).



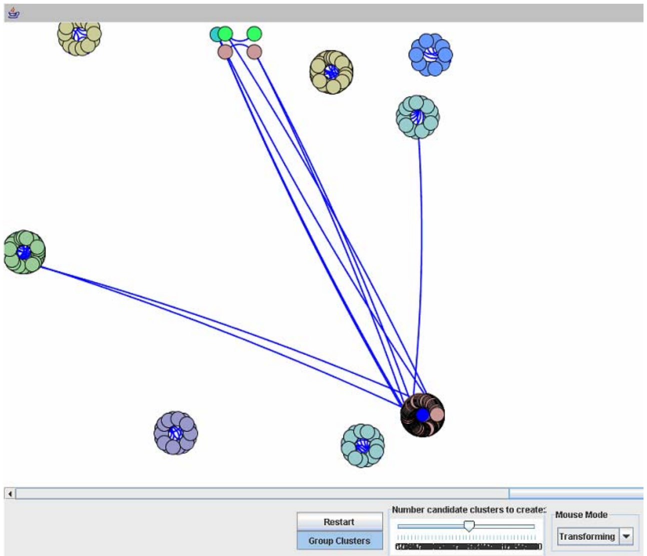
SQL where condition: where class = 's' or class='n'

□ Add to current file ☑ Use Email Table No Changes □ Use Message Table □ Unique Mailref i

Select type of graph: ○ EDGEBETWEEN ● VOLTAGECLUSTER

Figure 1 - Email social network window. (A) Network file can be generated with sender and recipient email accounts. (B) Text files can be converted to Pajek format. (C) Several types of social networks and clusters can be generated with email or external files.

The social network module has the capacity to process different social structures from external sources or email accounts. It has the flexibility to receive and generate text files with the structure of a network and convert it onto the Pajek format. The Pajek format is a well-known format in the social network literature that is used to generate new networks or clusters.



Mouse Mode: Transforming

Figure 2 Clusters of users created according to the quality of connection between networks.

5. Search Modules

Unusual communication patterns are a focus of this research, but equally as important is anomalous communication content which would be of interest to a forensic investigator.

We have been exploring extending the basic search functionality with a two step discovery process. We pre-compute a search index over all emails. When a search term is executed, we highlight a time line calendar with all relevant results and display a visualization of the contextual relationship of the number of emails per calendar entry of the results. We then use a thesaurus to grab similar words and unionize the results of those searches to find similar email communication.

6. Progress Report

In our last report [1] we noted our efforts in establishing a trial run of EMT with the New York Police Department. Unfortunately, our contact retired before being able to start the pilot study. Currently EMT is being evaluated by a number of commercial and research entities and has been offered to NARA in their new digital archive initiative managed by Lockheed Martin.

7. Future Directions

The flexibility of the social network module is useful for other security related tasks that may or may not be able to be detected through emails.

Our most recent emphasis is on extracting a user's social network given the user's email inbox [2] or on visualizing the social networks generated by email interactions [3], we use average distances and cluster coefficients to identify unusual relationships among members of different groups.

Some of the areas that this module can be used for include the creation of a "Chinese wall" among financial analysts and investment bankers. The social network module might be able to detect unauthorized activity between these groups. In addition it could also be applied to detecting links or conflict of interests among directors, and financial analysts that may bias financial reports. These indicators could also be used as part of a financial information system to predict stock prices or corporate performance.

We would like to thank both the National Science Foundation and DARPA for funding this work. In particular NSF grant - Email Mining Toolkit Supporting Law Enforcement Forensic Analysis from the Digital Government research program, No. 0429323.

8. REFERENCES

[1] Hershkop, S. *Behavior-based Email Analysis with Application to Spam Detection.* Ph.D. Thesis, Columbia University, New York, NY, 2005.

[2] Culotta, A., Bekkerman, R. and McCallum, A. *Extracting Social Networks and Contact Information from Email and the Web*. First Conference on Email and Anti-Spam (CEAS), Mountain View CA, 2004.

[3] Boyd D. and Potter, J. *Social Network Fragments: an interactive tool for exploring digital social connections.* SIGGRAPH, San Diego CA, 2003.

[4] Stolfo, S. *Email Mining Toolkit Supporting Law Enforcement Forensic Analyses NSF Final Report.* DG.o 2005 Atlanta, GA. May 2005.