Email Mining Toolkit Supporting Law Enforcement Forensic Analyses

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

Shlomo Hershkop
Columbia University
500 West 120th St
New York, NY 10027
1-212-939-7078
shlomo@cs.columbia.edu

ABSTRACT

The Email Mining Toolkit (EMT) is a data mining tool that visualizes a very wide range of detailed analyses of email and email flows derived from an archive of email in a variety of formats. EMT may be leveraged in many applications. In this project we focus on providing support to detectives and analysts in law enforcement to develop powerful means of analyzing emails acquired under due process as evidence in various investigations.

Categories and Subject Descriptors

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

General Terms

Experimentation, Security, Algorithms

Keywords

Email Forensics, Data Mining, Analysis.

1. INTRODUCTION

Email is the ubiquitous internet application. Millions of people routinely use email in their daily lives. There has been much prior research on content analysis of email for various purposes, such as automated filing, spam detection and filtering, and even for direct marketing purposes. We conjecture that the analysis of email flows to and from a user's email account(s) reveals a tremendous amount of additional information about a person's interests, activities and behaviors that cannot be derived alone from content analyses of individual emails.

2. Email Mining Toolkit

In this project, we have developed the Email Mining Toolkit (EMT) [1, 2], a data mining tool that visualizes a very wide range of detailed analyses of email and email flows derived from an archive of email in a variety of formats. EMT may be leveraged in many applications. In this project we focus on providing support to detectives and analysts in law enforcement to develop powerful means of analyzing emails acquired under due process as evidence in various investigations.

Email forensic analysis is one area which would benefit from an automatic data mining tool such as EMT. Given a very large set of emails, investigators, typically don't have time to read each email and would greatly increase their productivity given a tool that automatically prioritizes the messages to support more timely and effective investigations. For example, in some cases the legal acquisition of email may present gigabytes of data to a detective. It is a daunting task to analyze and recover from this source information pertinent to an ongoing criminal investigation. EMT is designed to analyze sources on this scale in minutes and present to detectives and analysts email ranked and logically ordered for their direct inspection. EMT's power is based upon a number of advanced analytical techniques not previously reported for email sources.

3. Progress

Since the start of this grant, we have briefed and continue to work directly with the detectives of the New York Police Department. EMT software has continued to be updated to suit their needs. EMT provides a very large collection of analytics including behavior analyses of individual email accounts, group or social behaviors revealed in email flows, and various content analyses that allows a thread of communication to be automatically inferred and visually displayed.

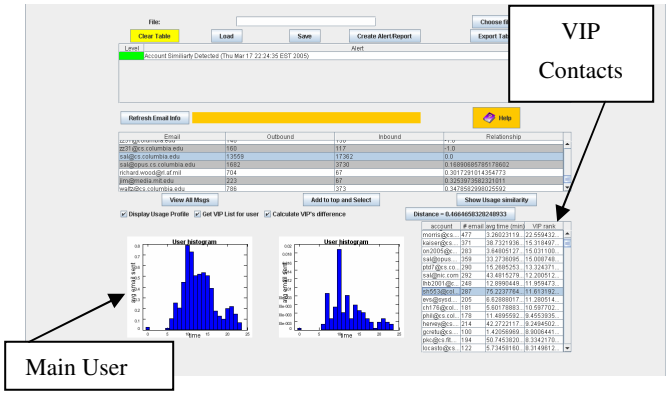

Figure 1 - Main Forensic Window

Of particular interest in our research work has been to devise sound and principled methods to correlate the various models computed by EMT in order to detect significant changes in behavior, or to infer information about individuals that would otherwise not be apparent from simple content analyses. For example, EMT profiles a user's typical email behavior and detects significant changes in that behavior by considering many emails sent and received over long periods of time. A user may send an average number of emails per hour over the course of a business week day. Changes in this profile may indicate a significant events worthy of deeper inpection. Furthermore, EMT measures the rate at which an individual email account responds to received

emails from each possible sender. The latter analysis displays the "average response rate" that indicates the relative importance of an email sender, and rank orders the senders accordingly. This information tends to infer who are the more important individuals in a social network, information that is of particular interest to law enforcement investigations.

EMT not only detects anomalous behavior, it also detects similar behavior patterns across accounts, a means of detecting "proxy accounts" used by a person to perhaps hide their identity.

EMT also analyzes the content and flows of email attachments. It automatically clusters attachments that are deemed similar to each other to identify possible breaches of security policy. For example, it may be policy that no "word document" should be emailed to an external email account. Such policy violations can assist detectives, and analysts to hone in on email behavior that may be the subject matter of ongoing investigation of criminal wrongdoing.

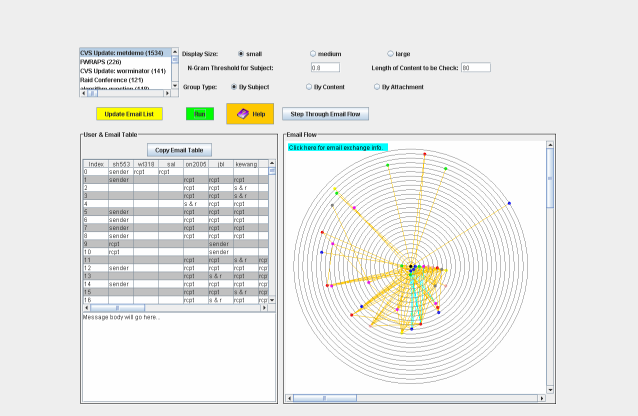

Figure 2 - Email Flow Window

Our ongoing research aims to develop a convenient and easy to use tool for law enforcement. We are also focused on basic research questions concerning the most effective and accurate means of correlating multiple models to optimize the presentation of analyses. The correlation functions under study include methods based upon Dempster-Shafer theory and classical control theory. The outcome of this work we believe will shed new light on automated data mining systems that optimize their own performance for a specific application.

In Figure 1, the forensic window allows an analysis to quickly find interesting email accounts based on both usage and communication behavior.

In Figure 2 a specific user's conversations can be displayed and replayed based on the email data collection.

4. Impact

This work has had implications in a few research areas. We have successfully shown how new behavior models can be used in the task of Spam detection [3]. In addition we found that combining the behavior models with content analyses increases coverage of both weak and strong spam detectors in [4]. In addition we have successfully used these ideas in the virus detection domain [5].

5. Future Work

Our ongoing research has produced a mature body of technology and we have explored several application areas where email is the core data under analysis. The NYPD has introduced us to other sources of structured information, crime reports, and seeks assistance to apply EMT analytics for the important purpose of identifying criminal organization patterns and behaviors. NYPD has recently completed the implementation of a major new IT infrastructure, a data warehouse bringing together a large collection of hither-to-fore unrelated and unintegrated police records. We shall work with NYPD detectives to reshape EMT to serve their analysis tasks. The analysis of these data sources has never before been attempted.

6. ACKNOWLEDGMENTS

We indebted to the professional staff of the New York Police Department. We also acknowledge the assistance of Kathy Wang, Olivier Nimeshkern, Wei-Jen Li and Charlie Hu for help in building the core EMT technology.

7. REFERENCES

[1] S. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern, and C.-W. Hu, "A Behavior-based Approach to Securing Email Systems," Mathematical Methods, Models and Architectures for Computer Networks Security, 2003.

[2] S. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern, and C.-W. Hu, "Behavior Profiling of Email," presented at 1st NSF/NIJ Symposium on Intelligence & Security Informatics(ISI 2003), Tucson, Arizona, 2003.

[3] S. Hershkop and S. J. Stolfo, "Identifying Spam without Peeking at the Contents," ACM Crossroads, 2004.

[4] S. Hershkop and S. J. Stolfo, "Combining Email Models for False Positive Reduction," March 2005 2005.

[5] S. Stolfo, W.-j. Li, S. Hershkop, K. Wang, C.-w. Hu, and O. Nimeskern, "Detecting Viral Propagations Using Email Behavior Profiles," TOIT, 2003.