Behavior Profiling of Email

Salvatore J. Stolfo, Shlomo Hershkop, Ke Wang, Olivier Nimeskern, and Chia-Wei Hu

Columbia University, New York, NY 10027, USA
{sal,shlomo,kewang,on2005,charlie}@cs.columbia.edu

Abstract. This paper describes the forensic and intelligence analysis capabilities of the Email Mining Toolkit (EMT) under development at the Columbia Intrusion Detection (IDS) Lab. EMT provides the means of loading, parsing and analyzing email logs, including content, in a wide range of formats. Many tools and techniques have been available from the fields of Information Retrieval (IR) and Natural Language Processing (NLP) for analyzing documents of various sorts, including emails. EMT, however, extends these kinds of analyses with an entirely new set of analyses that model "user behavior". EMT thus models the behavior of individual user email accounts, or groups of accounts, including the "social cliques" revealed by a user's email behavior.

1 Introduction

This paper describes the forensic and intelligence analysis capabilities of the Email Mining Toolkit (EMT) under development at the Columbia IDS Lab. EMT provides the means of loading, parsing and analyzing email logs, including content, in a wide range of formats. Many tools and techniques have been available from the fields of IR and NLP for analyzing documents of various sorts, including emails. EMT, however, extends these kinds of analyses with an entirely new set of analyses that model "user behavior". EMT thus models the behavior of individual user email accounts, or groups of accounts, including the "social cliques" revealed by a user's email behavior. EMT's design has been driven by the core security application to detect virus propagations, spambot activity and security policy violations. However, the technology also provides critical intelligence gathering and forensic analysis capabilities for agencies to analyze disparate Internet data sources for the detection of malicious users, attackers, and other targets of interest. This dual use is graphically displayed in Figure 1. For example, one target application for intelligence gathering supported by EMT is the identification of likely "proxy email accounts", email accounts that exhibit similar behavior and thus may be used by a single person. Although EMT has been designed specifically for email analysis, the principles of its operation are equally relevant to other Internet audit sources.

This data mining technology previously reported [4, 6, 7], and graphically displayed in Figure 2, has been proven to automatically compute or create both signature-based misuse detection and anomaly detection-based misuse discovery.

The application of this technology to diverse Internet objects and events (e.g., email and web transactions) allows for a broad range of behavior-based analyses including the detection of proxy email accounts and groups of user accounts that communicate with one another including covert group activities.

Data mining applies machine learning and statistical techniques to automatically discover and detect misuse patterns, as well as anomalous activities in general. When applied to network-based activities and user account observations for the detection of errant or misuse behavior, these methods are referred to as behavior-based misuse detection.

Behavior-based misuse detection can provide important new assistance for counter-terrorism intelligence. In addition to standard Internet misuse detection, these techniques will automatically detect certain patterns across user accounts that are indicative of covert, malicious or counter-intelligence activities. Moreover, behavior-based detection provides workbench functionalities to interactively assist an intelligence agent with targeted investigations and off-line forensics analyses.

Intelligence officers have a myriad of tasks and problems confronting them each day. The sheer volume of source materials requires a means of honing in on those sources of maximal value to their mission. A variety of techniques can be applied drawing upon the research and technology developed in the field of Information Retrieval. There is, however, an additional source of information available that can used to aid even the simplest task of rank ordering and sorting documents for inspection: behavior models associated with the documents can be used to identify and group sources in interesting new ways. This is demonstrated by the Email Mining Toolkit that applies a variety of data mining techniques for profiling and behavior modeling of email sources.

The deployment of behavior-based techniques for intelligence investigation and tracking tasks represents a significant qualitative step in the counter-intelligence "arms race". Because there is no way to predict what data mining will discover over any given data set, "counter-escalation" is particularly difficult.

Behavior-based misuse detection is more robust against standard knowledge-based techniques. Behavior-based detection has the capabilities to detect new patterns (i.e., patterns that have not been previously observed), provide early warning alerts to users and analysts, and automatically adapt to both normal and misuse behavior. By applying statistical techniques over actual system and user account behavior measurements, automatically-generated models and rules are tuned to the particular source material. This process, in turn, avoids the human bias that is intrinsic when misuse signatures, patterns and other knowledge-based models are designed by hand, as is the norm.

Despite this, no general infrastructure has been developed for the systematic application of behavior-based (misuse) detection across a broad set of detection and intelligence analysis tasks such as fraudulent Internet activities, virus detection, intrusion detection and user account profiling. Today's Internet security systems are specialized to apply a small range of techniques, usually knowledge-based, to an individual misuse detection problem, such as intrusion, virus or

SPAM detection. Moreover, these systems are designed for one particular network environment, such as medium-sized network enclaves, and only tap into an individual cross-section of network activity such as email activity or TCP/IP activity. Behavior-based detection technology as proposed herein will likely provide a quantum leap in security and in intelligence analysis in both offline and online task environments.

EMT has been described in another publication, focusing on its use for security applications, including virus and spam detection, as well as security policy violations. In this paper, we focus on several of its features specific to intelligence applications, namely the means of clustering email by content based analyses, identification of "similar email accounts" based upon measuring similarity between account profiles represented by histograms, and clique analyses that are supported by EMT.

1.1 Applying Behavior-Based Detection to Email sources

Table 1 enumerates a range of behavior-based Internet applications. These applications cover a set of detection, security and marketing applications that exist within the government, commercial and private sectors. Each of these applications are within the capabilities of behavior-based techniques by applying data mining algorithms over appropriate audit data sources.

Our current research has applied behavior-based methods directly to the first six applications listed in Table 1: Fraud detection, malicious email detection, intrusion detection, user community discovery, behavior pattern discovery, and analyst workbench. Each of these are Internet security applications, applying to both outbound and inbound network- and email-based traffic.

Solving Internet security problems greatly assists surveillance intelligence activities. For example, the discovery of user account communities and the discovery and detection of certain community behavior patterns can be directed to uncover certain classes of covert, clandestine or espionage behavior performed with Internet resources. Furthermore, fraud detection in particular has direct benefit for an intelligence agency by profiling and identifying users and clusters of users that participate in such malicious Internet activities such as fraudulent activities.

Behavior-based detection has been proven against similar, analogous security applications. The finance, telecom and energy industries have protected their customers from fraudulent misuse of their services (e.g., fraudulent misuse of credit card accounts, telephone calling cards, stealing of utility service, etc.) by modeling their individual customer accounts and detecting deviations from this model for each of their customers. The behavior-based protection paradigm applied to the Internet thus has an historical precedent that is now ubiquitous and transparent as exemplified by the credit card in the reader's wallet or purse.

1.2 EMT as an Analyst Workbench for Interactive Intelligence Investigations

The "Malicious Email Tracking" (MET) [1] is an online system that uses email flow statistics to capture new virii, which are largely undetectable by the "signature" detection methods of today's state-of-the-art commercial virus detection systems. Specifically, all email attachments are tracked by tracing a private hash value, temporal statistics such as replication rate are recorded to trace the attachments' trajectory, e.g., across LANs, and these statistics directly inform the detection of self-replicating, malicious software attachments. MET has been developed and deployed as an extension to mail servers and is fully described elsewhere. MET is an example of an online "behavior-based" security system that defends and protects a system not solely by attempting to identify known attacks against a system, but rather by detecting deviations from a system's normal behavior. Many approaches to "anomaly detection" have been proposed, including research systems that aim to detect masqueraders by modeling user behaviors in command line sequences, or even keystrokes. However, in this case, MET is architected to protect user accounts by modeling user email flows to detect malicious email attachments, especially polymorphic viruses that are not detectable or traceable via signature-based detection methods.

The "Email Mining Toolkit" (EMT) on the other hand, is an offline system applied to email files gathered from server logs or client email programs. EMT computes information about email flows from and to email accounts, aggregate statistical information from groups of accounts, and analyzes content fields of emails. The EMT system provides temporal statistical feature computations and behavior-based modeling techniques, through an interactive user interface to enable targeted intelligence investigations and semi-manual forensic analysis of email files. Figure 1 illustrates the general architecture of a behavior-based system deploying dual functionality:

1. An online security detection application (in this case, MET for malicious email detection)
2. A general analyst workbench for intelligence investigations (EMT, for email source analysis)

As this figure illustrates, these functionalities share a great deal of overhead. With regard to the implementation, by deploying these dual functionalities, the audit module, computation of temporal statistics, user modeler and database of user models each serve for both functionalities. Moreover, with regard to the conceptual design, the particular set of temporal statistics and user model processes designed for one can improve the performance of the other. In particular, temporal features, as well as user account models and clusters, are representatively general "fundamental building blocks." EMT provides the following functionalities, interactively:

- Querying a database (warehouse) of email data and computed feature values, including:

- Ordering and sorting emails on the basis of content analysis (n-gram analysis, keyword spotting, and classifications of email supported by an integrated supervised learning feature using Nave Bayes classifier trained on user selected features)
- Historical features that profile user groups by statistically measuring behavior characteristics.
- User models that group users according to features such as typical emailing patterns (as represented by histograms over different selectable statistics), and email communities (including the "social cliques" revealed in email exchanges between email accounts.

– Applying statistical models to email data to alert on abnormal or unusual email events.

Table 1. Behavior-Based Internet Applications for Security and Beyond

Application:	Description and Variations:	Examples:	Audit Sources:
Fraud detection	Unauthorized outgoing email Unauthenticated email Unauthorized transactions	Console usurped Child attacks teacher Deceptive source Purchase/credit fraud	Email HTTP Transaction services.
Malicious email detection	Viruses Worms "SPAM"		Email
Intrusion detection	Network-based detection Host-based detection Application-based detection	Standard IDS Less standard IDS Future IDS	TCP/IP System logs App. logs
User community discovery	Closely connected user-base	Email 'circles'	Email
Behavior-pattern discovery	Account-based patterns Community-based patterns	Suspect activities Clandestine activities	All sources: Email, HTTP, Transaction services, TCP/IP, Telnet traffic, FTP traffic, cookiesEmail, FTP, Telnet
Analyst Workbench	Interactive forensic analysis	Targeted intelligence investigations	All sources
Account proxy detection	Accounts used by same user	Clandestine activities	All sources
Collaborative filtering	Website recommendations Purchase recommendations	Pageview prediction Music/movie choices	HTTP Transaction services
Policy violation detection	ISP or enclave security policies	User espionage Outgoing SPAM	All sources Email
Web-bot detection	Statistics/knowledge gathering Site maintenance Search-engine spider	Competitive analysis Finding broken links Google, Altavista	HTTP

EMT is also designed as a plug in to a data mining platform, originally designed and implemented at Columbia called the DW/AMG architecture (Data Warehouse/Adaptive Model Generation system). That work has been transferred to System Detection Inc (SysD http://www.sysd.com), a DARPA-spinout from Columbia who has commercialized the system as the Hawkeye Security Platform.

2 EMT Features

The full range of EMT features have been described elsewhere . For the present paper, we provide a brief overview of several of its key features of direct relevance to security analysis and intelligence applications, along with descriptive screenshots of EMT in operation.

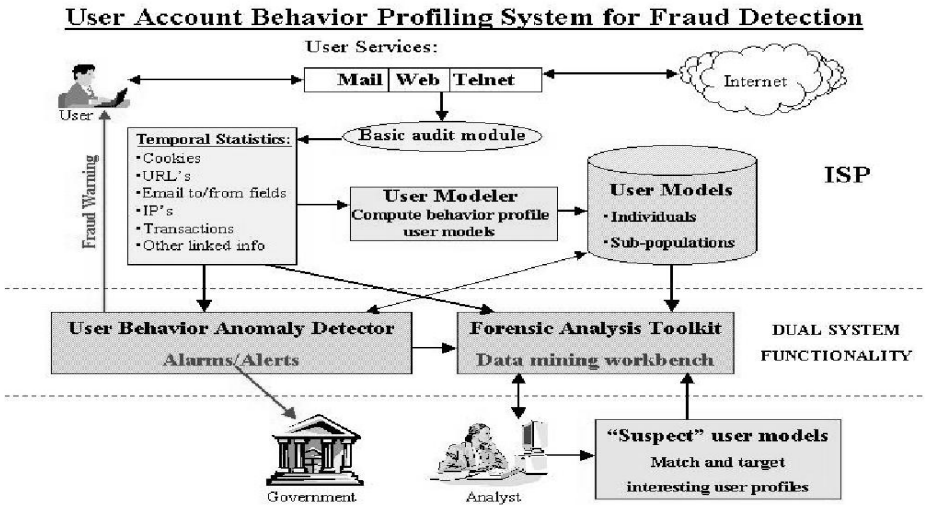

Fig. 1. User account profiling, dual use: online detection and offline analysis.

2.1 Attachment models

MET was initially conceived to statistically model the behavior of email attachments in real time flowing through an enclave's email server, and support the coordinated sharing of information among a wide area of email servers to identify malicious attachments and halt their propagation before saturation. In order to properly share such information, each attachment must be uniquely identified, which is accomplished through the computation of an MD5 hash of the entire attachment.

EMT runs an analysis on each attachment in the database to calculate a number of metrics. These include, birth rate, lifespan, incident rate, prevalence, threat, spread, and death rate. They are explained fully in [1], and are displayed graphically in Figure 3.

Rules specified by a security analyst using the alert logic section of EMT are evaluated over the attachment metrics to issue alerts to the analyst. This analysis may be done to archived email logs by EMT offline, or at runtime in MET while sniffing real-time email flows. The initial version of MET provides the means of specifying alerts in rule form as a collection of Boolean expressions applied to thresholds compared to each of the calculated statistics. As an example, a basic rule might check for each attachment seen if its birth rate is greater than some specified threshold AND sent from at least users. The flow statistics of each

[1] A paper entitled "A Behavior-based Approach to Securing Email Systems" has been prepared for submission to a technical conference and is under review. That paper describes the use of EMT for virus and spam detection. There is a minor overlap with that paper in presentation material of some of EMT's features described herein.

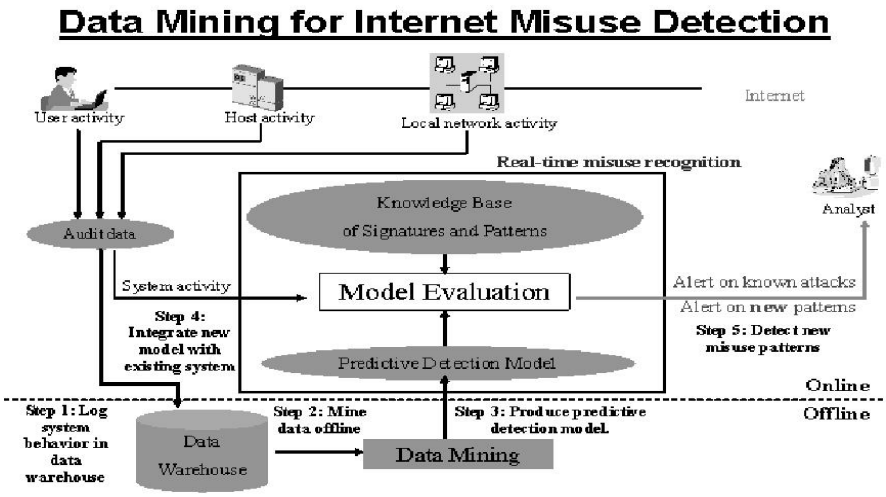

Fig. 2. Overview of data mining based detection system.

email attachment are computed by EMT, as well as the list of specific emails the attachment appears in, to identify recipients of those attachments. The primary detection task MET was designed for includes virus propagation and mitigation. Intelligence applications of this particular feature would include infosec security policy violations, and general evidence gathering in forensic analyses.

2.2 Email Content and Classification

Figure 4 illustrates EMT's main messages tab that provides an analyst with the means to inspect, cluster and sort email messages under analysis. Emails can be selected for review and analysis on the basis of time, sender or recipient account. This data may be labeled directly by an analyst for further data mining analysis supported by other feature tabs in EMT. Interestingly, EMT also provides the means of classifying attachments by way of the fully embedded EMF system, a supervised machine learning feature. In the earliest work on MEF (Malicious Email Filter [7]), the Naïve Bayes classifier was computed over user selected training sets of attachments. The features extracted include "n-grams" and their frequencies, extracted and computed directly from the attachment irrespective of its mime type. Hence, in addition to using flow statistics and attachment classifications to classify an email message, EMT uses the email body as a content-based feature. The two features supported are n-gram [8] modeling and a calculation of the frequency of a set of words [9] from the body of the email.

An n-gram represents the sequence of any n adjacent characters or tokens that appear in a document. An n-character wide window is passed over the entire

email body, one character at a time, and a count is computed on the number of occurrences of each n-gram. This results in a hash table that uses the n-gram as a key and the number of occurrences as the value for each email; this we refer to as the document vector.

Given a set of training emails, the arithmetic average of the document vectors can be computed as the centroid for the set. Given an instance of an email, we compute the cosine distance [8] against the centroid created during training. If the cosine distance is equal to 1, then the two documents are deemed identical. The smaller the value of the cosine distance, the more different the two documents are. These content-based methods are integrated into the machine learning models for classifying sets of emails for further inspection and analysis. An analyst therefore has the means of honing in on a set of potentially relevant emails by first classifying and clustering sets of emails using the EMT GUI.

Using a set of normal email and spam we collected, we did some initial experiments over our own email sets to test the efficacy of the approach. We used half of the labeled emails, both normal and spams, as training data, and used the other half as the test set. The accuracy of the classification using n-grams and word tokens varies from 70% to 94% when using different parts as training and testing sets.

In the spam classification experiment, we noticed some spam emails did not vary much from normal emails. For example a spam that would be a single link to a non-threatening website. To improve accuracy we also used weighted key-words and removal of stop-words. For example, the spam email set noticeably contain the words: free, money, big, lose weight, etc in a much higher frequency than regular emails. Users can empirically assign stop-words and keywords and give higher weight to their frequency count. We continue to evaluate these content based approaches further; experiments and analysis are ongoing.

2.3 Account Statistics and Alerts

This mechanism has been extended to provide alerts based upon deviation from other baseline user and group models. EMT computes and displays three tables of statistical information for any selected email account. The first is a set of stationary email account models, i.e. statistical data represented as a histogram of the average number of messages sent over all days of the week, divided into three periods: day, evening, and night. EMT also gathers information on the average size of messages for these time periods, and the average number of recipients and attachments for these periods. These statistics can generate alerts when values are above a set threshold as specified by the rule-based alert logic section of EMT.

Stationary User Profiles - Histograms over discrete time intervals Histograms are used to model the stationary behavior of a user's email account. Figure 8 displays an example for one particular user account. Histograms are

compared to find similar behavior or abnormal behavior between different accounts, and within the same account (between a long-term profile histogram, and a recent, short-term histogram).

A histogram depicts the distribution of items in a given sample. EMT employs a histogram of 24 bins, for the 24 hours in a day. Email statistics are allocated to different bins according to their outbound time. The value of each bin can represent the daily average number of emails sent out in that hour, or daily average total size of attachments sent out in that hour, or other features defined over an of email account computed for some specified period of time.

Two histogram comparison functions are implemented in the current version of EMT, each providing a user selectable distance function. The first comparison function is used to identify groups of email accounts that have similar usage behavior. The other function is used to compare behavior of an account's recent behavior to the long term profile of that account. The histogram comparison functions also may be run "unanchored", meaning, the histograms are shifted to find the best alignment with minimum distance; thus accounting for time zone changes.

Similar Users - Histogram distance Similar behaving user accounts may be identified by computing the pair-wise distances of their histograms (eg., a set of accounts may be inferred as similar to given known or suspect account that serves as a model). The histogram distance functions were modified for this detection task. First, we balance and weigh the information in the histogram representing hourly behavior with the information provided by the histogram representing behavior over different aggregate periods of a day. This is done since measures of hourly behavior may be too low a level of resolution to find proper groupings of similar accounts. For example, an account that sends most of its email between 9am and 10am should be considered similar to another that sends emails between 10am and 11am, but perhaps not to an account that emails at 5pm. Given two histograms representing a heavy 9am user, and another for a heavy 10am user, a straightforward application of any of the histogram distance functions will produce erroneous results.

Thus, we divide a day into four periods: morning (7am-1pm), afternoon (1pm-7pm), night (7pm-1am), and late night (1am-7am). The final distance computed is the average of the distance of the 24-hour histogram and that of the 4-bin histogram, which is obtained by regrouping the bins in the 24-hour histogram.

Second, because some of the distance functions require normalizing the histograms before computing the distance function, we also take into account the volume of emails. Even with the exact distribution after normalization, a bin representing 20 emails per day should be considered quite different from an account exhibiting the emission of 200 emails per day. Figure 6 graphically displays the EMT analysis showing the target user account and a list of the most similar accounts found by EMT's histogram analysis.

Abnormal User Account Behavior EMT may apply these distance functions to one target email account. (See Figures 6.) A long term profile period is first selected by an analyst as the "normal" behavior period. The histogram computed for this period is then compared to another histogram computed for a more recent period of email behavior. If the histograms are very different (i.e., they have a high distance), an alert is generated indicating possible account misuse. We use the weighted Mahalanobis distance function for these profiles.

The long term profile period is used as the training set, for example, a single month. We assume the bins in the histogram are random variables that are statistically independent. When the distance between the histogram of the selected recent period and that of the longer term profile is larger than a threshold, an alert will be generated to warn the analyst that the behavior "might be abnormal" or is deemed "abnormal". The alert is also put into the alert log of EMT.

The histograms described here are stationary models; they represent statistics at discrete time frames. Other non-stationary account profiles are provided by EMT, as described next.

Non-Stationary User Profiles - Histograms over blocks of emails Another type of modeling considers the changing conditions over time of an email account. Most email accounts follow certain trends, which can be modeled by some underlying distribution. As an example of what this means, many people will typically email a few addresses very frequently, while emailing many others infrequently. Day to day interaction with a limited number of peers usually results in some predefined groups of emails being sent. Other contacts with whom the email account owner interacts with on less than a day to day basis have a more infrequent email exchange behavior.

The recipient frequency is used as a feature to study this concept of underlying distributions. Four behavior analysis graphs for any selected e-mail account are created by EMT for this model. These graphs display the address list size and average outgoing e-mail account spread over time, as well as the number of outgoing e-mails to each destination account.

Every user of an email system develops a unique pattern of email emission to a specific list of recipients, each having their own frequency. Modeling every user's idiosyncrasies enables the EMT system to detect malicious or anomalous activity in the account. This is similar to what happens in credit card fraud detection, where current behavior violates some past behavior patterns. Figures 5 provides a screenshot of the non-stationary model features in EMT, that are fully described elsewhere.

In a nutshell, The Profile tab in Figure 5 provides a snapshot of the account's activity in terms of recipient frequency. It contains three charts and one table. The various profile statistics selected by the analyst specify an empirical distribution that may then be compared by the analyst with a set of built-in metrics including Chi-square, and Hellinger distance [10]. Rapid changes in email

emissions among accounts can then be discerned which may have particular intelligence value.

2.4 Group Communication Models: Cliques

In order to study the email flows between groups of users, EMT provides a feature that computes the set of cliques in an email archive.

We seek to identify clusters or groups of related email accounts that frequently communicate with each other, and then use this information to identify unusual email behavior that violates typical group behavior, or identify similar behaviors among different user accounts on the basis of group communication activities.

Clique violations may also indicate internal email security policy violations. 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 if members of the marketing department, and HR services would likewise receive these attachments. EMT can infer the composition of related groups by analyzing normal email flows and computing cliques (see Figure 7), and use the learned cliques to alert when emails violate clique behavior. An analyst may simply wish to compute these cliques and rank order all associated emails of the clique members for direct inspection.

EMT provides the clique finding algorithm using the branch and bound algorithm described in [2]. We treat an email account as a node, and establish an edge between two nodes if the number of emails exchanged between them is greater than a user defined threshold, which is taken as a parameter (Figure 7 is displayed with a setting of 100). The cliques found are the fully connected subgraphs. For every clique, EMT computes the most frequently occurring words appearing in the subject of the emails in question which often reveals the clique's typical subject matter under discussion.

Chi Square + cliques The Chi Square + cliques (CS + cliques) feature in EMT is the same as the Profile window described above in 2.3.4, with the addition of the calculation of clique frequencies.

In summary, the clique algorithm is based on graph theory. It finds the largest cliques (group of users), which are fully connected with a minimum number of emails per connection at least equal to the threshold (set at 50 by default). In this window, each clique is treated as if it were a single recipient, so that each clique has a frequency associated with it. Only the cliques to which the selected user belongs will be displayed. Some users don't belong to any clique, and for those, this window is identical to the normal Chi Square window.

If the selected user belongs to one or more cliques, each clique appears under the name $clique_i$ i:=1,2... and is displayed in a cell with a green color in order to be distinguishable from individual email account recipients. (One can double click on each clique's green cell, and a window pops-up with the list of the members of the clique.)

Cliques tend to have high ranks in the frequency table, as the number of emails corresponding to cliques is the aggregate total for a few recipients. These metrics are a first step to model user's behavior in terms of group email emission frequency. A larger database will enable us to refine them, and to better understand the time-continuous stochastic process taking place. The Chi square test may be modified or completed with finer measures.

The Chi Square tests if the frequencies of emission are constant for a given user. In the preliminary results that we ran on our collected database, the Chi Square test has tended to reject quite often the hypothesis that the frequencies were the same between training and testing periods, indicating that the frequencies are not stable. They change quite dynamically under short time frames, as new recipients and cliques become more or less popular over time. Any new model should take into account this dynamic evolution.

Enclave cliques v.s. User cliques Conceptually, two types of cliques can be formulated and both are supported by EMT. The one described in the previous section can be called enclave cliques because these cliques are inferred by looking at email exchange patterns of an enclave of accounts. In this regard, no account is treated special and we are interested in email flow pattern on the enclave-level. Any flow violation or a new flow pattern pertains to the entire enclave. On the other hand, it is possible to look at email traffic patterns from a different viewpoint altogether. Consider we are focusing on a specific account and we have access to its outbound traffic log. As an email can have multiple recipients, these recipients can be viewed as a clique associated with this account. Since a clique could be subsumed by another clique, we defined a user clique as one that is not a subset of any other cliques. In other words, user cliques of an account are its recipient lists that are not subsets of other recipient lists.

User clique computation provides an intelligence analyst with the means of quickly identifying groups directly associated with a target email account, and may be used to group emails for inspection based upon various clique analyses. This is an active area of our ongoing research. Preliminary experiments have been performed using these graph theoretic features for spam and virus detection. In both cases, the clique models provide interesting new evidence to improve the accuracy of detection beyond what is achievable with pure content-based features of emails.

3 Conclusion

It is important to note that testing EMT and MET in a laboratory environment is not particularly informative of its performance on specific tasks and source material. The behavior models are naturally specific to a site or particular account(s) and thus performance will vary depending upon the quality of data available for modeling, and the parameter settings and thresholds employed. EMT is designed to be as flexible as possible so an analyst can effectively explore the space of models and parameters appropriate for their mission. An

analyst simply has to take it for a test spin. (EMT has been deployed and is being tested and evaluated by external organizations.)

One of the core principles behind EMT's design may be stated succinctly: there is no single monolithic model appropriate for any detection or forensic analysis task. Hence, EMT provides a pallet of models and profiling techniques (specialized to email log files) that may be combined in interesting ways by an analyst to meet their own mission objectives. It is also important to recognize that no single modeling technique in EMT's repertoire can be guaranteed to have no false negatives, or few false positives. Rather, EMT is designed to assist an analyst or security staff member architect a set of models whose outcomes provide evidence for some particular detection task. The combination of this evidence is specified in the alert logic section as simple Boolean combinations of model outputs; and the overall detection rates will clearly be adjusted and vary depending upon the user supplied specifications of threshold logic.

The Email Mining Toolkit is a work in progress. This paper has described the core concepts underlying EMT, and its related Malicious Email Tracking system, and the Malicious Email Filtering system. We have presented the features of the system currently implemented and available to a analyst for various security and intelligence applications. The GUI allows the user to easily automate many complex analyses. We believe the various behavior-based profiles computed by EMT will significantly improve analyst productivity. We are continuing our research to broaden the range of features and models one may compute over email logs. For example, the notion of clique may be over-constrained, and may be relaxed in favor of other kinds of models of communication groups. Further, we are actively exploring stochastic models of long-term user profiles, with the aim to compute these models efficiently when training such profiles. Histograms computed in fixed time periods is very efficient, but likely insufficient to model a user's true dynamic behavior.

References

1. M. Bhattacharyya, S. Hershkop, E. Eskin, and S. 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, 2002.
2. C. Bron, J. Kerbosch Finding all cliques of an undirected graph Comm. ACM 16(9) (1973) 575–577.
3. E. Eskin, A. Arnold, M. Prerau, L. Portnoy and S. J. Stolfo. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data Data Mining for Security Applications. Kluwer 2002.
4. George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Pages 338-345, 1995
5. Wenke Lee, Sal Stolfo, and Kui Mok. Mining Audit Data to Build Intrusion Detection Models In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98), New York, NY, August 1998

6. Wenke Lee, Sal Stolfo, and Phil Chan. Learning Patterns from Unix Process Execution Traces for Intrusion Detection AAAI Workshop: AI Approaches to Fraud Detection and Risk Management, July 1997
7. Matthew G. Schultz, Eleazar Eskin, and Salvatore J. Stolfo. Malicious Email Filter - A UNIX Mail Filter that Detects Malicious Windows Executables. Proceedings of USENIX Annual Technical Conference - FREENIX Track. Boston, MA: June 2001.
8. Damashek, M. Gauging Similarity with n-grams: language independent categorization of text Science, 267 (5199), 843-848, 1995.
9. Mitchell, T. Machine Learning, McGraw-Hill, 1997, pg., 180-183.
10. Hogg, R.V. Introduction to Mathematical Statistics, Prentice Hall, 1994.

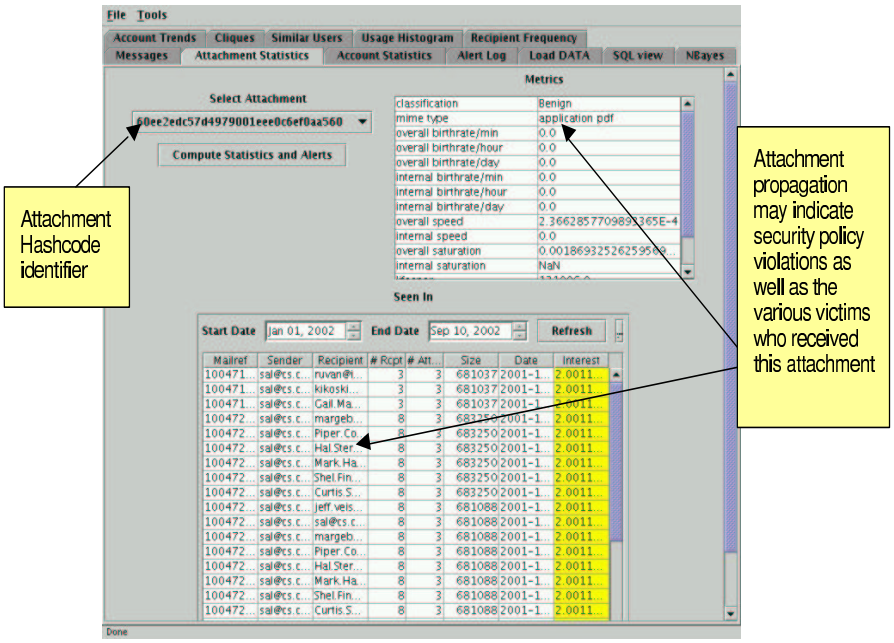

Figure 3 – Attachment Statistics

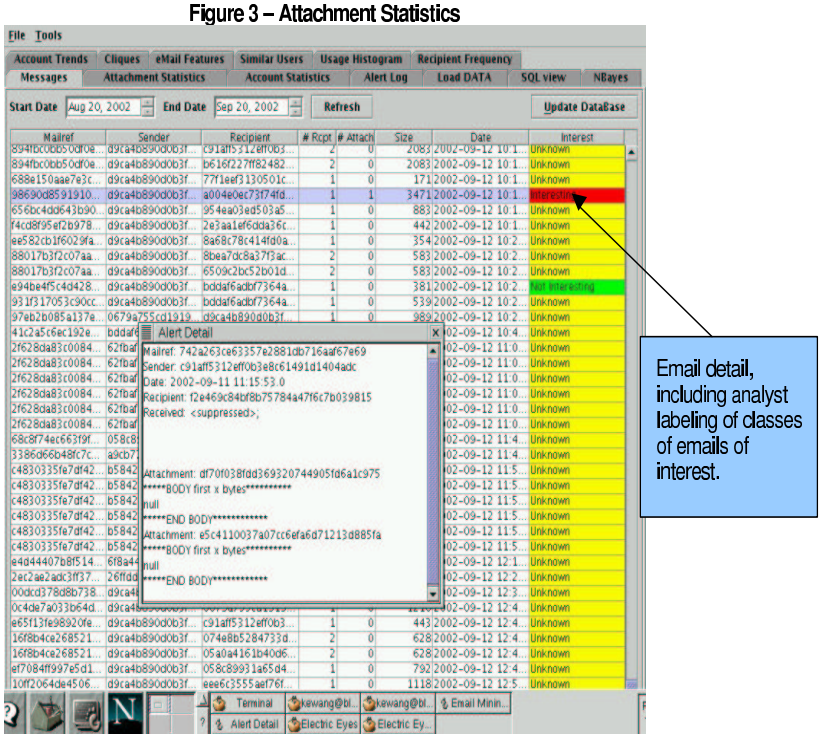

Figure 4. Main analyst window to sort and inspect specific emails.

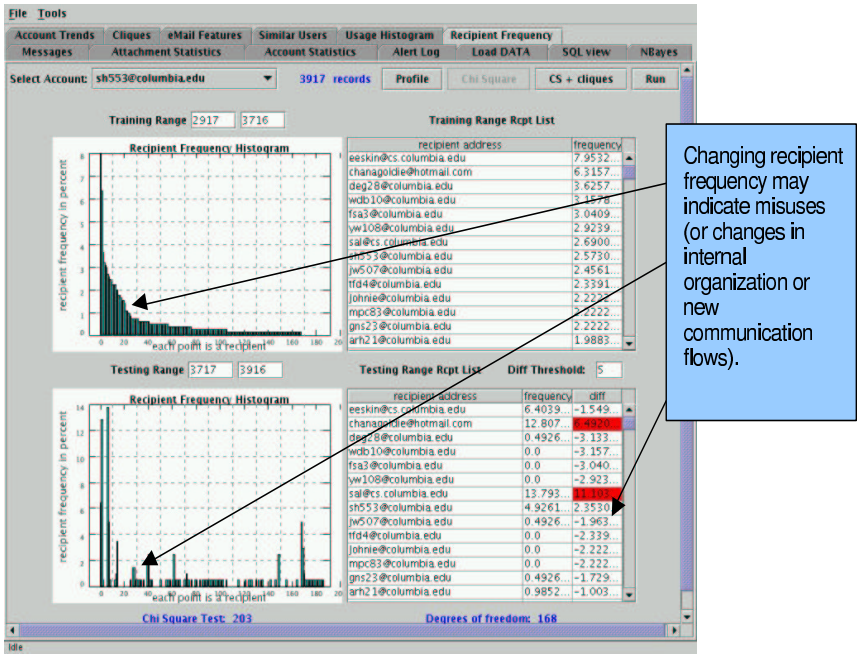

Figure 5 - Chi Square Test of recipient frequency

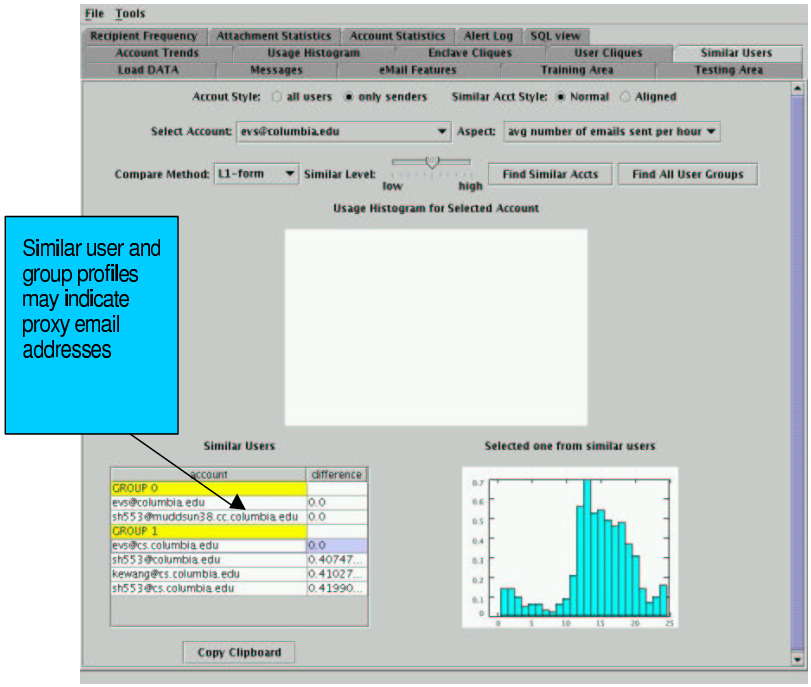

Figure 6 – Histogram Comparison to Detect Similar users

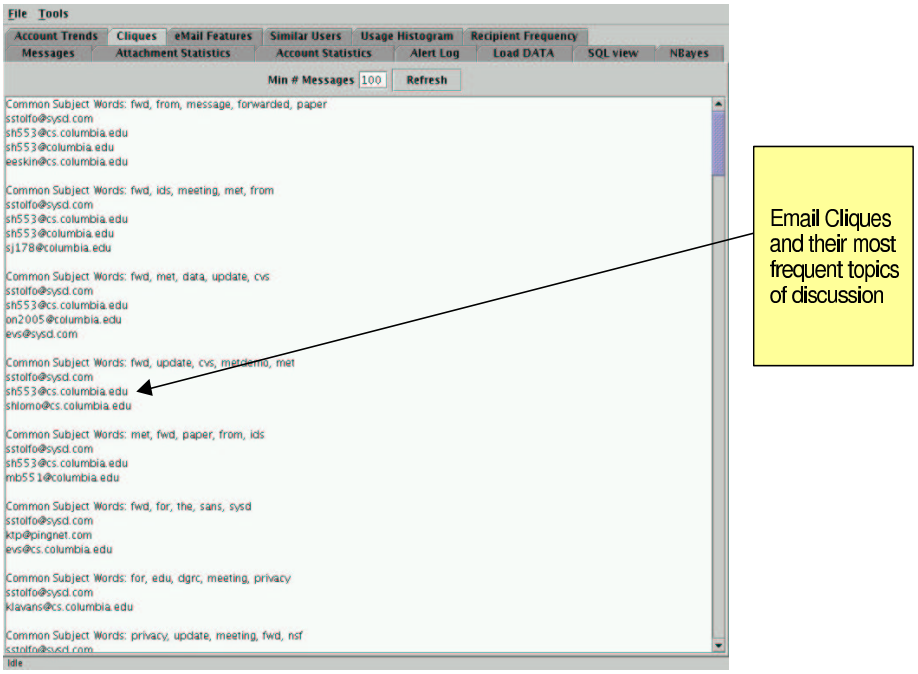

Figure 7 - Clique generation for 100 messages

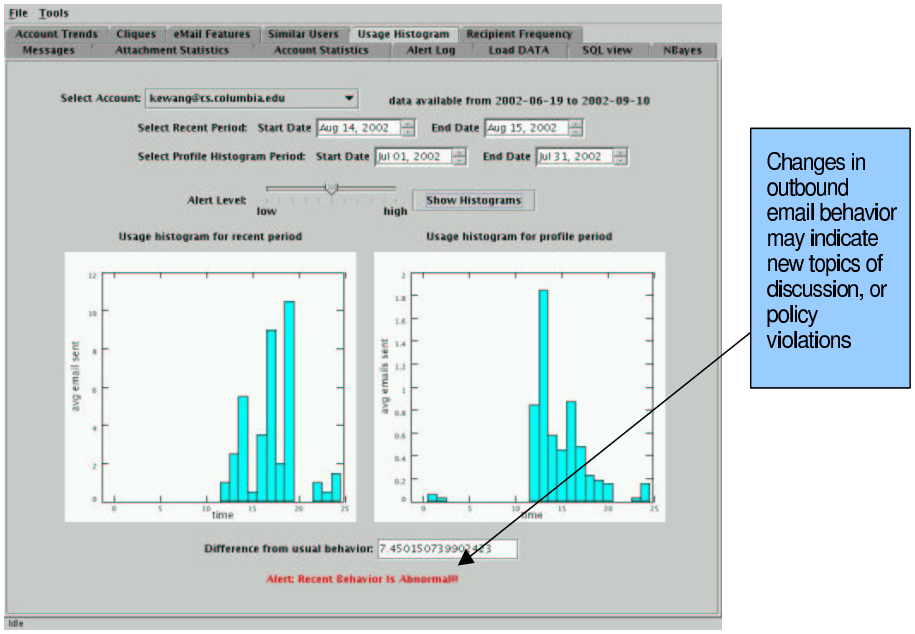

Figure 8 – Anomalous user behavior detected by histogram comparison