Mining Spam Email to Identify Common Origins for Forensic Application

Chun Wei
Dept. of Computer and Information Sciences, Univ. of Alabama at Birmingham
1300 University Blvd.
Birmingham, AL, USA
1-205-934-8669
weic@cis.uab.edu

Alan Sprague
Dept. of Computer and Information Sciences, Univ. of Alabama at Birmingham
1300 University Blvd.
Birmingham, AL, USA
1-205-934-8513
sprague@cis.uab.edu

Gary Warner
Dept. of Computer and Information Sciences, Univ. of Alabama at Birmingham
1300 University Blvd.
Birmingham, AL, USA
1-205-934-8620
gar@cis.uab.edu

Anthony Skjellum
Dept. of Computer and Information Sciences, Univ. of Alabama at Birmingham
1300 University Blvd.
Birmingham, AL, USA
1-205-934-2213
tony@cis.uab.edu

ABSTRACT
In recent years, spam email has become a major tool for criminals to conduct illegal business on the Internet. Therefore, in this paper we describe a new research approach that uses data mining techniques to study spam emails with the focus on law enforcement forensic analysis. After we retrieve useful attributes from spam emails, we use a connected components clustering algorithm to form relationships between messages. These initial clusters are then refined by using a weighted edges model where membership in the cluster requires the weight to exceed a chosen threshold. The results of the cluster membership are validated by WHOIS data, by the IP address of the computer hosting the advertised sites, and through comparison of graphical images of website fetches. This technique has been successful in identifying relationships between spam campaigns that were not identified by human researchers, enabling additional data to be brought into a single investigation.

Categories and Subject Descriptors
H.4.3 [**Information Systems Applications**]: Communications Applications – Electronic mail

K.4.1 [**Computers and Society**]: Public Policy Issues – Abuse and crime involving computers

General Terms
Algorithms, Experimentation, Security.

Keywords
Electronic Mail, Spam, Data Mining, Forensic Analysis, Cyber Crime

SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

1. INTRODUCTION
In recent years, Spam email has become a major problem for society not only because the number of spam emails is astonishingly massive and growing but also because more and more spam emails are related to cyber crimes. For example, phishing emails are sent to people to trick them to log into phishing sites that will steal their personal information, some spam emails provide false information about an unknown company in order to lure people to buy its stocks, other spam emails are selling pirated software, illegal drugs, or promoting online gambling. Even though these kinds of spam emails have violated laws and caused damage, it is difficult for law enforcement personnel to stop them for the following reasons: 1) the daunting volume of spam emails has made it virtually impossible for human to collect evidence from it; 2) criminals who create and distribute spam emails are using various techniques to disguise their true identities and make it hard to track them down: for example, many spam emails are delivered from "zombies" (computers that are affected by Trojan viruses), and thus will receive orders, in this case sending spam emails, from a commanding computer, which remains much harder to identify.

On the other hand, most of the academic research thus far has been focused on spam filtering, which separates the "bad" email form the "good" email. Various techniques have been applied, such as Naïve Bayesian [7], Support Vector Machines [4], k-Nearest Neighbor [9], Neural Network [3], Genetic Algorithm [8], and Rough Set Theory [13]. In addition, there is also research on text categorization [12], authorship identification [11] and monitoring of email user behaviors (EMT) [10]. But none of these has put focus on the advanced analysis of spam emails, especially the ones related to cyber crimes. ScamSlam[1] did further clustering analysis on scam, which they defined as intelligently designed spam messages for illegal purposes, and reported that half of the scam messages were produced by 20 individuals or collaborating groups. However, the scam messages they tested are limited to one topic category and their approach is based on the content of the emails. It is not surprising that different techniques have been applied to spam emails to obfuscate their content in order to trick the spam filters. For example, some spam emails put their real messages in the graphic attachment while the content of the email is just an unrelated

paragraph from a novel. Therefore, it is necessary to consider additional attributes of the email, not only looking at the content.

The grim situation of crime-related spam emails has prompted the need for research of spam emails with focus on the requirements of forensic analysis. The findings of ScamSlam points to a fact that by sending out spam emails, a spammer inevitably reveals some information about himself/herself in the emails, just as a criminal will leave telltale traces in the crime scene. And the information can be retrieved and studied. Therefore, by applying data mining techniques to the deep analysis of spam email, we plan to utilize computer power to aid humans in finding clues among spam emails and make the job of law enforcement personnel who are fighting spam emails more manageable.

2. EXTRACTING EMAIL ATTRIBUTES

Previous research works on spam email usually start with building a word corpus based on the email content or studying email traffic, such as the domain name portion of the senders' email [6]. However, the evolving obfuscation techniques used by spam email senders makes it inadequate to study only a limited number of attributes of the spam emails. The email content approach is likely to fail on spam emails with no content, but only an attachment. In fact, our email collection shows that most spam emails with attachment have no body content. And the sender's domain approach may not work on spam emails sent from "zombies." Many spam emails contain a fake "From" header, so the sender's email address does not really exist. Therefore, there is no simple solution, and it is necessary to extract as many attributes from the emails as possible. In our research, eleven attributes have been successfully parsed from the messages : "message_id", "sender_IP_address", "sender_email", "subject", "body_length", "word_count", "attachment_filename", "attachment_MD5", "attachment_size", "body_URL", "body_URL_domain". Some attributes are broken down into two sub-attributes, for example, "body_URL" into "machine_name" and "path". Some attributes are useful for global clustering because most email message have a non-null value in that attribute, such as email subject or sender's IP address. But these attributes may be weak evidence that do not prove two emails are related. Two emails with common subject, such as "Re:" and "Fwd", may actually come from different spammers. Other attributes are good for clustering a specific sub-group of spam emails, such as "body_URL_domain", which only works with spam email with URLs. But a domain name, especially a spam domain, is very strong evidence showing relationship between two emails if they both point to the same domain.

Apart from the inherent attributes that can be directly retrieved from the email, such as the eleven attributes just mentioned, derived attributes, those that cannot be directly acquired from emails but can be derived from inherent attributes by looking them up in additional sources, are also important. Some examples are the WHOIS data from the Domain Name Registrar and screenshots of web pages of domains. The derived attributes provide further evidence of relationship between spam emails or spammers. For example, if two different domains point to the same IP address, then they are related; and if two IP addresses host the same web pages, then the two IP addresses are related. Derived attributes are very useful in finding non-obvious relationships and validating initial clusters built from inherent attributes.

3. CLUSTERING METHODS

Two clustering methods have been used in our experiments thus far. The agglomerative hierarchical algorithm is used for the global clustering of the entire dataset. When this clustering method is applied, the largest cluster contained too many emails, indicating the assertion of relationships which were not present. Next, the connected component with weighted edges algorithm is used to overcome this false positive situation. If a cluster resulting from the first method is found to be weak, the second clustering algorithm is applied, which is designed to require stronger evidence for clustering.

3.1 Agglomerative Hierarchical Clustering Based on Common Attributes

An agglomerative clustering method [5] is used for global clustering to group spam emails based on common values of email attributes. In the beginning, each email message by itself is a single cluster. Then clusters that share a common attribute are merged. Each time a new attribute is introduced, clusters from the previous iteration will be merged based on the common values in the new attribute. The old clustering results are backed up in case the process needs to be reversed due to false positives.

D(i, j) is defined as the distance between cluster i and j. D(i, j) = 0 if cluster i and j share a common value in an attribute and D(i, j) = 1 if not. Two clusters are merged if distance is 0. A common attribute value means exact string matching.

In our experiment, 'subject' is used in the first iteration of global clustering. Therefore, two clusters are merged if they share a common subject. 'Subject' is used because most emails contain a subject and two emails with the same subject are presumed to be more likely to be originated from the same source. There are of course exceptions, subjects that are blank or common phrases cause false-positives in the result, which brings up a counter-measure method that will be discussed in the next section.

Domain name is used as the attribute for the second iteration. A domain name (*.e.g.,* yahoo.com) is the part of a URL that is the human readable representation of an IP address. Two clusters are merged if they contain emails which point to the same domain.

The agglomerative clustering method is desirable because only in the first iteration, the runtime of the algorithm is a function of the number of emails, but starting from the second iteration, the runtime is a function of the number of previous clusters, which is constantly reducing. The weakness of the method is that coincidence, common phrases and sheer luck can cause untrustworthy relationships to be introduced since our logic is that two emails are linked as long as they share at least one common attribute. In our experiment, we stop after two iterations because we have encountered a false-positive problem: the biggest cluster contains more than 67% of the emails with URLs. To counter false-positives, a connected component with weighted edge method is introduced in the next section to break the biggest cluster into smaller clusters.

3.2 Connected Components with Weighted Edges

To eliminate chance conjoining of unrelated spam campaigns into the same cluster, the concept of "connected component of weighted edges" was applied.

A *connected component* [2] in an (undirected) graph is a set *S* of vertices such that for every vertex *v* of *S*, the set of vertices reachable (by paths) from *v* is precisely *S*. The weight of an edge shows the strength of the connection between the two vertices. The goal is to find connected components of this graph, considering only edges with weight above a threshold. This goal stems from the following reasoning: Suppose a spammer owns 10 domains and has a list of 10 subjects, and he sends out emails by randomly picking a subject and a domain. There are totally 100 possible combinations. If he sends out enough emails and we have enough collection of his emails, we should see examples of all 100 combinations. So if domains are assigned as vertices and subjects as edges, we will evidently find that the ten domains are tightly connected to each other with strong edges. On the other hand, if two domains are owned by two different spammers and they are connected to each other by chance because the two spammers share a common subject, the connection between domains, in this case, will be weak since the probability of two spammers picking the same subject is relatively lower. If a group of domains in the biggest cluster are tightly connected to each other, they are very likely to be owned by the same spammer.

Therefore, all domains from the biggest cluster are retrieved and assigned as vertices. The edges connecting them will be any common subject and the weight of the edge is the number of common subjects shared by two domains. A threshold is then selected and all edges with weight below that threshold will be dropped. The remaining connected components should be tightly related.

The algorithm is designed to allow the threshold be adjusted to produce a more favorable result. By applying the algorithm to a cluster that has false positives, the cluster is divided into smaller clusters that are more tightly related. If the result still shows too many false positives in our sub-clusters, the threshold will be incremented. Or if the result shows too many tiny clusters, the threshold will be decremented. In the experiment, thresholds 2, 3 and 5 are used and the result turned out to be most accurate with threshold 3, which will be explained in more detail in section 4.4.

4. EXPERIMENTAL RESULTS

4.1 Data Set

The dataset consists of three months of email submitted by a single researcher that has been manually identified by that researcher as spam. This researcher collects a high volume of spam through the use of "catch all" email addresses. A "catch all" configuration accepts mail for all possible addresses at a given domain. One common technique spammers use to "harvest" new target addresses is to send emails to randomly generated userids at well-known domains. Mail which does not "bounce" or reject is assumed by the spammer to have been delivered. Because a "catch all" address configuration accepts ALL mail, spammers treat all tested addresses as valid for this researcher's domains. In addition to numerous catch-all addresses, spam emails sent to the researcher's true email addresses have also been included. This researcher contributed 211,000 emails during the months of June, July and August 2007. Neither this researcher nor his ISP does any form of spam filtering prior to our data extraction.

4.2 Results of Agglomerative Hierarchical Clustering

In the beginning, each of the 211,000 messages of our dataset was a cluster of size 1. In the first iteration, emails with matching subject are brought together, so that the number of clusters became equal to the number of subjects, 72,160, with the largest cluster containing 9,380 emails sharing a common subject. After experimenting with other possible attributes, it was decided that the next attribute for clustering would be the domain portion of the URLs contained within the bodies of the spam email messages. For the purposes of this experiment, focus was placed on the 33,993 clusters which contained some emails with at least one URL. Future work will address other types of spam emails.

After the second iteration of the algorithm, clustering by Subject x Domain, the number of clusters under consideration was reduced from 33,993 clusters to 3,247 clusters. Each of the newly formed clusters was formed by linking existing clusters which shared at least one common domain. 89.7% of all of the email fell into 42 unique clusters when this process was applied. Many smaller clusters existed as well, but this paper, and the anticipated user of this system, will focus on the larger clusters, as the desire is to identify the greatest nexus of criminal spamming activity.

4.3 Validation of Results

It is necessary to determine whether the resulting clusters are valid for purposes of cybercrime investigation. For instance, will an investigator gain a clear knowledge that messages in a single cluster are related in a valid way? Clusters were evaluated using a visual inspection method at this time. Because the clusters in this experiment were conjoined by the presence of a common "domain" portion of their URL, a routine was developed to fetch and save a graphical image, or thumbnail, of the appearance of each destination website. Where the resultant collection of website images from a single cluster was visually confirmed to be the same by sorting of the resultant webpage images, a high confidence was placed upon the integrity of the cluster. Where the resultant collection of website contained divergent images, a second level of validity checking was required.

For second level validity checking, a list of the Internet domains contained in a given cluster is checked using a "WHOIS" command which returns information about where the domain is hosted (what IP address), what registrant information is associated the domain, and what nameservers are providing services for the domain.

4.3.1 First Level Validation: Website Image Comparison

When first level validity checking is run, the smaller clusters report as being "highly trustworthy" based on the identical or nearly identical images which are returned when the corresponding webpages are retrieved. The largest seven clusters are the following (Table 1). This is out of a total of 42 clusters contained more than 100 messages each.

Cluster	Number of emails	Number of subjects	Number of Domains (Theme)
A	105,848	16,125	10,845 (many themes)
B	3,810	112	20 (Downloadable Software)
C	1,284	48	37 (Elite Herbal)
D	851	13	62 (Downloadable Software)
E	744	224	157 (several themes)
F	584	125	207 (ED Pill Store)
G	554	88	9 (Diamond Replicas)

Table 1: Largest 7 clusters

Clusters B, C, D, F, and G each contained exactly one website image pattern, indicating that all of the spam messages in each of these clusters had as their purpose to drive recipients to these common websites. This was considered a high level of validity.

4.3.2 Second Level Validation: WHOIS and Host Data

Cluster A and E contained multiple images patterns, indicating that messages identified as related by the system were being used for different spam campaigns. It was necessary to do secondary validation to determine if these messages were indeed related. The secondary validation technique involves running a process to perform a "WHOIS" and Host resolution on each of the domains in question. The result of the Host resolution is an IP address of the computer running the web server for this domain. Two spammed URL domains which each resolve to the same IP address have a strong correlation. The WHOIS data results were checked to see if two domains in the same cluster were using the same NameServer, or were registered to the same Owner. These commonalities were judged by researchers to determine if the correlation was strong. For instance, two domains which use the nameserver "ns1.yahoo.com" may have very weak correlation, but two domains which each use the nameserver "ns1.strawpusnips.com" may be considered to have a stronger correlation as there are very few domains which use that nameserver.

Cluster E was found to contain 100 domains which were still resolving at the time of our validation. There were 22 unique image patterns which were found among these 100 domains. Analysis of Host webserver IP address, Owner Registration data, and Nameserver data showed that all 100 domains were using only 6 IP addresses to host their traffic. Only 2 unique sets of Owner Registration data were found, and only 3 sets of Nameservers were found.

On further investigation, it was determined that examples of each image pattern could be found on each of the IP addresses, and from domains registered by each owner, and on domains using all three nameservers. The evidence that examples of all 22 patterns could be found on each of the IP addresses in cluster E, but not on any address in other clusters, lead the researchers to rate this as a strongly related cluster.

Primary image validation of Cluster A revealed several main image patterns, including "Canadian Pharmacy", "ED Pill Store", "Elite Herbal", "Herbal King", "International Legal RX", "My Canadian Pharmacy", "Penis Enhancement Patch", and "US Drugs". There were also a small number of outlier images, some of which represented false spam reporting by the researcher.

Cluster A was subjected to 2nd level Validation. This largest cluster contained many divergent images, which also did not show a high correlation through second level validity. It was believed that False Positives were causing Cluster A to bring unrelated messages together, forcing 50.1% of our entire spam sample into one cluster.

4.3.3 False Positive Identification

The Clustering Algorithm is desirable because of its very fast nature, and the non-exponential manner in which additional attributes may be added to the data clustering process. The weakness of the algorithm is that coincidence, common phrases, and sheer luck can cause untrustworthy relationships to be introduced.

A great number of spammers evidently choose to use either a blank subject line, or a subject line consisting of simple words such as: "Re:", "Hi", and "Hello". Other simple phrases are commonly used by spammers who choose the phrase for its likelihood to intrigue a reader into opening the message. "Was this from you?" or "Alert!" or "Thank you" may be chosen by unrelated spammers because of this fact. As the researchers noticed this trend in previous experimentation, certain phrases, such as "Re:" are ignored when clustering by Subject.

4.4 Results of Weighted Edges

To markedly reduce the chance conjoining of unrelated spam campaigns, the concept of "Weighted Edges" was applied. With this model, comparisons are made between the vertices and their edges, but rather than a single "zero distance" edge being sufficient to force a Cluster Label change, the discovery of a zero distance edge causes a counter to be incremented towards a threshold value. The algorithm is designed to allow the threshold to be adjusted, and the validation processes repeated to determine whether the new threshold delivers a more favorable result. Beginning with a Cluster which has failed to show strong trustworthiness, the weighted edges algorithm is applied. Vertices which have a relationship which exceeds the threshold value are now related to one another in "subclusters".

Through experimentation with threshold values of 2, 3, 5, it was determined that for our current email population, 3 was an ideal threshold value for achieving a trustworthier group of subclusters.

Beginning with Cluster A, with a population of 10,845 domains, the Weighted Edges Algorithm was applied to create 26 significantly sized SubClusters, and many smaller clusters and singletons.

SubClusters 1 through 26 were then validated using the methods described in 4.3.1

SubClusters 2 through 26 each showed a very high correlation visually, isolating related spam messages into distinct families of spam.

In the case of Cluster A, the application of Weighted Edges was able to identify a number of incidental common subjects which had caused emails to cluster together which were advertising unrelated websites. While this may be the result of chance, as

mentioned above, it is also possible that this is the result of a successful spammer serving more than one criminal enterprise.

SubCluster 1, the largest remaining group after applying Weighted Edges, still had a number of distinct visual patterns present. Some of these distinct patterns, such as "Herbal King" and "Elite Herbal" were shown through secondary validation to be tightly linked. 125 domains were found which were all registered to "Danny Lee" of "Health Worldwide, Inc" in Kowloon, Hong Kong. Each of the live domains from this group used the Nameserver "ns1.chongdns99.com", and each was hosted on the IP address "210.14.128.34". 65 additional domains were found to be registered to "Sammy Lee" of "Liquid Ventures, Inc", also of Kowloon, Hong Kong. The live domains from this group also used the Nameserver "ns1.chongdns99.com", and each was hosted on the IP address "210.14.128.34". Sample domains from each of these two groups were found to represent each of the image patterns. Yet "Herbal King" and "Elite Herbal" showed no commonalities in secondary validation with the other image patterns in this SubCluster.

5. IMPACT

This initial experiment showed interesting results as significant clusters of emails were found which through the two phase verification technique were shown to be tightly related, regardless of the disparity of the Subject, Contents, or Header information. The result is not perfect as we are still exploring and improving our methods. But we believe it is a promising research area that worth further pursuit. The findings have received positive response from members of the law enforcement and anti-spam communities who have indicated that these methods have generated clusters that they feel are worth criminal investigation.

6. SUMMARY AND FUTURE WORK

This paper has proposed a new approach to analyze spam emails with a focus on the needs of law enforcement personnel. Initial results show that the data mining technique creates clusters of related emails which can easily be assessed for their validity. The resulting clusters have been primarily related to spam messages which are trying to encourage the purchase of a product or service. Clusters of spam used for spreading viruses through attachments, or spam which sends visitors to hacked websites for purposes of phishing or other fraud were not readily identified using the current method.

The next stage of the research is to introduce more attributes into analysis, especially derived attributes. Some of them have already been used in validation, such as domain WHOIS information and web page screenshots. If these derived attributes are stored in the database, they can also be useful to for finding relationships between currently unrelated clusters. It would also make it possible to automate the current manual verification process. We also would like to investigate the spam emails that do not contain a URL in the content. Proper attributes need to be identified to do this; some candidates include attachment related attributes, which are useful for emails with an attachment but no content.

The next issue is the scalability of the system. Work is already underway to parallelize the task of initial email parsing to allow us to work with much larger spam collections, with a goal of receiving real-time spam feeds from major spam recipients. Additional research in parallelizing the analysis tasks is anticipated as we consider the possibilities of much larger spam collections, and the needs to use different attributes to identify clusters conforming to different spam use cases.

7. REFERENCES

[1] Airoldi, E. and Malin, B. *ScamSlam: An Architecture for Learning the Criminal Relations Behind Scam Spam.* Carnegie Mellon University, School of Computer Science, Technical Report CMU-ISRI-04-121. Pittsburgh: May 2004.

[2] Baase, S. Computer Algorithms: Introduction to Design and Analysis. (2nd ed.). Addison-Wesley, 1988.

[3] Clark, J., Koprinska, I. and Poon, J. *A neural network based approach to automated e-mail classification.* In Proceedings of IEEE/WIC International Conference on Web Intelligence, 13, 17, (Oct. 2003), 702 – 705.

[4] Drucker, H., Wu, D. and Vapnik, V.N. *Support vector machines for spam categorization.* IEEE Transactions on Neural Networks, 10, 5, (Sep 1999), 1048 – 1054.

[5] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques. (2nd ed.).* Morgan Kaufmann, San Francisco, CA, 2006.

[6] Jung, J. and Sit, E. *An empirical study of spam traffic and the use of DNS black lists.* In Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. (Oct. 2004) 370 – 375.

[7] Sahami, M., Dumais S., Heckerman, D. and Horvitz, E. *A Bayesian approach to filtering junk email.* AAAI Workshop on Learning for Text Categorization, AAAI Technical Report WS-98-05. Madison, Wisconsin. July 1998. 55 – 62.

[8] Sanpakdee, U., Walairacht, A. and Walairacht, S. *Adaptive spam mail filtering using genetic algorithm.* In Proceedings of the 8th International Conference on Advanced Communication Technology. (Feb. 2006). 441 - 445.

[9] Soucy. P and Mineau, G. W. *A simple KNN algorithm for text categorization.* In Proceedings of 2001 IEEE International Conference on Data Mining, (Nov – Dec 2001) 647-648.

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

[11] Vel, O. D., Anderson, A., Corney, M. and Mohay, G. *Mining Email Content for Author Identification Forensics.* SIGMOD: Special Section on Data Mining for Intrusion Detection and Threat Analysis, 30, 4, (Dec. 2001) 55 - 64.

[12] Yang, Y. and Liu, X. *A Re-examination of text categorization methods.* In Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (Aug. 1999). 42-49.

[13] Zhao, W. and Zhang, Z. *An email classification model based on rough set theory.* In Proceedings of the 2005 International Conference on Active Media Technology. (May 2005). 403 – 40.