THE JOURNAL OF
DIGITAL FORENSICS,
SECURITY AND LAW

Journal of Digital Forensics,
Security and Law

Volume 5 | Number 1 Article 2

2010

Clustering Spam Domains and Destination Websites: Digital Forensics with Data Mining

Chun Wei
University of Alabama, Birmingham

Alan Sprague
University of Alabama, Birmingham

Gary Warner
University of Alabama, Birmingham

Anthony Skjellum
University of Alabama, Birmingham

Recommended Citation

Wei, Chun; Sprague, Alan; Warner, Gary; and Skjellum, Anthony (2010) "Clustering Spam Domains and Destination Websites: Digital Forensics with Data Mining," *Journal of Digital Forensics, Security and Law*: Vol. 5 : No. 1 , Article 2.
DOI: https://doi.org/10.15394/jdfsl.2010.1070
Available at: https://commons.erau.edu/jdfsl/vol5/iss1/2

Clustering Spam Domains and Destination Websites: Digital Forensics with Data Mining

Chun Wei
Dept. of Computer and Information Sciences
Univ. of Alabama at Birmingham
USA
weic@cis.uab.edu

Alan Sprague
Dept. of Computer and Information Sciences
Univ. of Alabama at Birmingham
USA
Sprague@cis.uab.edu

Gary Warner
Dept. of Computer and Information Sciences
Univ. of Alabama at Birmingham
USA
gar@cis.uab.edu

Anthony Skjellum
Dept. of Computer and Information Sciences
Univ. of Alabama at Birmingham
USA
tony@cis.uab.edu

ABSTRACT

Spam related cyber crimes have become a serious threat to society. Current spam research mainly aims to detect spam more effectively. We believe the identification and disruption of the supporting infrastructure used by spammers is a more effective way of stopping spam than filtering. The termination of spam hosts will greatly reduce the profit a spammer can generate and thwart his ability to send more spam. This research proposes an algorithm for clustering spam domains extracted from spam emails based on the hosting IP addresses and tracing the IP addresses over a period of time. The results show that many seemingly unrelated spam campaigns are actually related if the domain names in the URLs are investigated; spammers have a sophisticated mechanism for combating URL blacklisting by registering many new domain names every day and flushing out old domains; the domains are hosted at different IP addresses across several networks, mostly in China where legislation is not as tight as in the United States; old IP addresses are replaced by new ones from time to time, but still show strong correlation among them. This paper demonstrates an effective use of data mining to relate spam emails for the purpose of identifying the supporting infrastructure used for spamming and other cyber criminal activities.

Keywords: Digital Forensics, spam email, cyber crime, clustering, data mining.

1. INTRODUCTION

According to the McAfee threat report (McAfee Avert Labs, 2009), there were 153 billion spam messages per day in 2008 and over 90% emails were spam. Most spam emails are sent by botnets, infected computers controlled by commanding servers. In the first quarter of 2009, nearly twelve million new IP addresses were detected as bots, an increase of almost 50% from the last quarter of 2008 (McAfee Avert Labs, 2009).

Spam has been used to spread malware to recruit more bots; to trick people to phishing sites and steal vital information; to lure people into false transactions by exploiting human greed, such as promising lottery winnings, overseas inheritances, or easy work-at-home jobs with great salaries; and to advertise counterfeit products and services, such as pharmaceuticals, luxury jewelry and watches, sexual-enhancement products and pirated software.

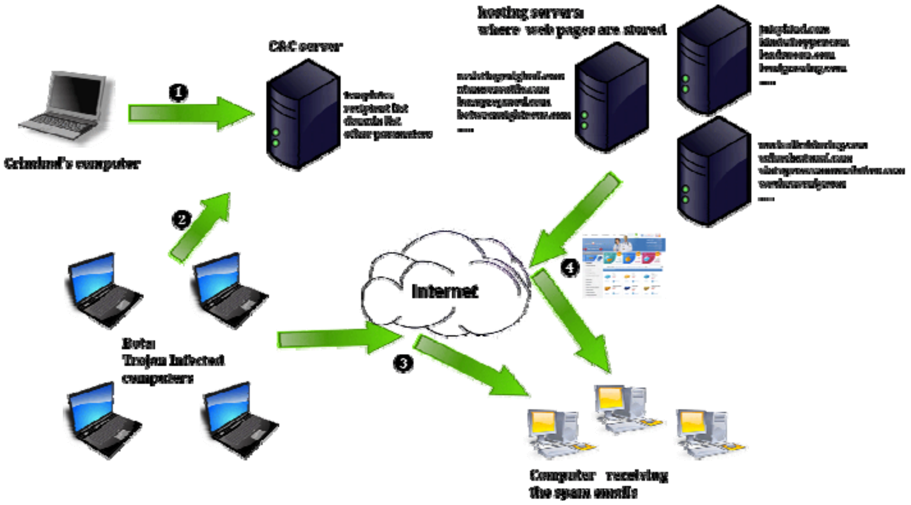

Figure 1: Information flow on spamming network

Figure 1 shows how a spammer operates its network in order to protect his identity. The spammer controls the bots, infected computers, through a command & control server (C&C), for example, the notorious rogue hosting provider McColo terminated in late 2008 (Clayton, 2009). He updates information on the

C&C server and each of his bots contact the server to receive new commands, new spam templates and email address lists. Then the bots send out the spam emails with URLs pointing to websites. The spammer also maintains the websites on various web-hosts, as well as maintaining the corresponding DNS entries on name servers.

Much anti-spam research seeks to create better spam filters to prevent spam from reaching email recipients. This ignores the well-established concept of deterrence, "the inhibiting effect of sanctions on the criminal activity of people other than the sanctioned offender." (Blumstein et al., 1978, p.3). When society believes, and sees through repeated example, that criminals are punished for their action, fewer people are likely to become offenders themselves. Filtering of spam fails to deliver this potential deterrent effect that would be a predicted outcome of regularly prosecuting spammers. Spam can be more effectively stopped by disrupting its source, such as C&C and hosting servers shown in Figure 1. This research targets the hosting servers because it is not necessary for the email recipient to be able to find the origin of a spam email in order to process the message, but it is essential to the delivery of the spammer's end goal, the sale of a product or service, for an actual location of the advertised website to be reachable to the email recipient. If the recipient cannot reach the point-of-sale website, no transaction can occur. Besides, occasionally the hosting servers and commanding servers are at the same location.

The hosting servers can be traced by investigating the URLs in spam emails. According to previous research by Pu and Webb (2006), the majority of current spam contains URLs. The URLs point to websites where the vital actions take place for spammers to make a profit, for example, a web site selling counterfeit pharmaceutical products. Researchers at University of California at San Diego (Kanich et al. 2008) studying the Storm Worm projected that the pharmaceutical spam portion of the Storm Worm activities may have generated as much as $350 Million for the botnet controllers. Whoever created these websites on the hosting servers is obviously responsible, either directly or as part of the same criminal conspiracy, for the spam emails that lead to those websites.

In order to protect the websites from termination, spammers developed a new way to combat domain blacklisting: registering a large number of new domains every day. Even though it costs more for spammers to register so many domains, St Sauver (2008) summarized several major benefits for spammers to do that: (1) to reduce the chance of spam being blocked by SURBL/URIBL filtering (two popular spam "black lists" used by spam filtering solutions) because new domains are less likely to be on the blacklist; (2) to reduce the risk of being prosecuted by law enforcement. Because the large volume of spam has been distributed among many different domain names, each will appear to be a small-volume spamming group, thus reducing the chance of catching law enforcement's attention; (3) to balance the traffic and increase the chance of survivability. In order to shut down the spam, one has to take down all of the domains or all of the back-end servers.

However, the spam domains tend to cluster in a limited number of hosting IP addresses. According to the research by Wei et al. (2009), several IP blocks were found to host more than ten thousand domains in the first quarter of 2009, meaning that spammers have their favorite hosting places. Therefore, the clustering of domains based on their IP locations appears to be a very promising way of grouping spam emails that likely stem from the same spamming organization. This also provides the law enforcement personnel a more effective way of terminating spam domains by shutting down the hosting servers.

2. RELATED WORK

In recent years, there has been research on spam characterization and categorization, which provides deeper insight on spamming strategies and scam host infrastructure.

Tom (2008) clustered spam emails if any of the following three attributes are identical: sending IP address, message body and email subject. The biggest cluster reported contains 85% of all emails in a 9-day period of time in Dec 2007. The emails were related to replica watches, gambling, porn and sexual enhancement. However, the consideration of only identical subjects and message bodies will fail to find email messages with similar subjects or message bodies that are generated by templates, which is very commonly seen in today's spam messages. Moreover, emails with common subjects like "Re:" and "Fwd" may not necessarily have any connections between them.

Calais et al. (2008) used four attributes (language, message type, message layout and URLs) to cluster spam campaigns. Emails that share common frequent features will be grouped. Some big spam groups reported consist of more than 100,000 spam emails, which were collected by honeypots in several Brazilian networks. The paper also investigated the network patterns of the sending machines (abuse of HTTP, SOCKS proxies and open relays). The paper did not further investigate the URLs, such as fetching the web pages, finding hosting IP addresses or WHOIS information.

Pu and Webb (2006) observed trends in spam email message construction, especially obfuscation methods in HTML-based spam emails. They then built a Webb Spam Corpus, which consists of nearly 350,000 web pages that are obtained from URLs in the HTML-based spam emails (Webb et al. 2006). They also found that the web hosts in their Corpus were tightly connected to each other by web links. But the graph was too heavily clustered to see any detailed information of how the hosts were actually connected. Using the Webb Spam Corpus, they categorized the web pages into five categories: Ad Farms, Parked Domains, Advertisements, Pornography and Redirection (Webb et al. 2007). They found web spam pages tend to have more duplicates and redirections than normal web pages. They also identified Top 10 hosting IP addresses with the most web page count and two IP ranges account for 84% of the hosting IP addresses. However, they did not indicate whether these IP addresses were related or not.

The Spamscatter project (Anderson et al. 2007) also fetched web pages using the links in spam emails and clustered the web pages based on screen shot similarity. They categorized spam campaigns based on the content of the web sites. However, ten largest virtual-hosted scam categories they listed contained three “watches” categories, two “pharmacy” categories and two “software” categories and there was no indication whether they were related or not. They traced domains for about two weeks and found that multiple virtual hosts (different domains served by the same server) and multiple physical hosts (different IP addresses) are infrequent. This might not be true anymore since the largest cluster we found contain many domains, each hosted by the same set of IP addresses. They also investigated the lifetime of scam hosts and found the majority of them were short-lived. However, a spammer can point a website to a different IP address by changing DNS entries and creating new domain names to delete replace old ones that are blacklisted. Therefore, the termination of a host or domain name does not necessarily mean a spam campaign has ended. In our study, the largest cluster lasts for the entire experiment period, while new domain names are introduced every day and hosting IP addresses shifted from time to time.

Our research combines the email subject with information from URLs in the emails to cluster spam messages. Because our collected emails contain emails with different body format, some with only text, some html coded, some with image attachments, some with all of the above, we decided not to compare email bodies but only the email subjects. However, we applied a fuzzy matching algorithm so that similar subjects resembling a pattern can be discovered. Instead of fetching the web pages for each URL, we extracted the domain names from URLs and fetched hosting IP addresses, because we observed many phantom host names that were created by attaching a random string before a real domain name. It is not efficient to fetch all the phantom host names when they actually point to the same wildcard DNS record. Also some URLs in emails are used for obfuscation, disguising the spam emails as legitimate emails. Fetching the web pages for those URLs will be misleading.

Since new spam emails appear each day, it is not efficient to rebuild the clusters if new data arrives. We develop an algorithm that can group new emails as they arrive and link new clusters to old ones. The design is motivated by research in data streams. The nature of data streams demands three critical requirements for clustering algorithms (Barbara, 2002): (1) Compressed representation of data; (2) Fast, incremental processing of newly arriving data points; (3) Identification of outliers. In data stream research, data compression is achieved by Clustering Feature (CF) (Zhang et al. 1996). Subsequent research (Aggarwal et al. 2003;

Cao et al. 2006) developed similar data structure based on CF tree. Spam data is different from other data streams because the traditionally useful attributes are usually numerical attributes while most of our attributes are nonnumeric. Although it is possible to find distance between two spam emails based on similarity, it is impossible to define a centroid for a spam cluster using coordinates, which are used to define micro-clusters in data stream papers. Therefore, we cannot use the same CF of other data stream papers for spam data. Instead, we cluster spam emails on a daily basis, and the daily clusters serve as micro-clusters. We then compare daily clusters according to similarity on global attributes. The outliers are usually ignored because we are interested in leading clusters with a great number of emails and domain names. Within a cluster, we are also interested in the hosting IP addresses that host many domain names.

3. DATA COLLECTION AND PREPARATION

3.1 Data Collection

For this study, we gathered spam emails from a number of domains controlled by our researchers, including the "catch-all" email accounts for these domains. Traditionally, emails sent to non-existent users are rejected by a mailserver, but in the case of a catch-all address, all emails are accepted, but emails intended for non-existent addresses are sent to a single default account regardless of its original destination. Emails sent to these catch-all domains which do not correspond to real users are spam by nature and they form a large percentage of our data set. From our dataset of more than 7 million emails, we chose to focus on a two-month period containing 638,678 email messages received during the months of June and July of 2009.

3.2 Extracting URLs from Spam Emails

Some attributes extracted from spam email messages suggest relationship among different emails.

Attributes that can be directly extracted from an email header and content are called inherent attributes. Extracted inherent attributes include email subject, sender's name, sender's email address, sender's IP address, date received, embedded URLs, email attachment. Among them the URL is the most interesting, because it leads to the hosting spam websites. In our dataset, over 90% of spam emails contain URLs in the email text. Some spam use image attachment and have URLs embedded in the image, therefore we are actively working to incorporate OCR into our system in order to detect the URLs in the image. In this paper, we only include the emails with URLs in text format.

Derived attributes are information derived from inherent attributes. The URL can be used to fetch the websites, the hosting IP addresses and WHOIS information. The derived attributes provide more useful information leading to the spam origin than inherent attributes. On the other hand, some URLs may point to websites which are no longer available, such as the Ad farm and parked domain pages found by Webb et al. (2007). Therefore, the information is harder to retrieve.

3.2.1 Extracting wildcard domain name

During the process of extracting the domain name portion of the URL, we observed that many spam domains use wildcard DNS records.

A wildcard DNS record is a DNS record that will resolve requests for non-existent host names having a matched domain suffix (Wikipedia, 2009). It is specified by using a "*" as the left most part of a host name, e.g. `*.domain.com`. Therefore, if a user requests a domain name ending with “domain.com” that does not have a corresponding entry in the DNS records, the wildcard record will be used to resolve the request.

To test a wildcard domain, we first extract the domain name portion from the host name, e.g. the domain name for “zhpt.tarecahol.cn” would be “tarecahol.cn”. Then we create our own phantom host name by attaching a random string to the domain name. If the new host name can still be resolved, and provides the same data as the original, it proves the domain is using wildcard DNS records. Then it is very likely all other host names ending with the same domain name should also resolve to the same site. This strategy greatly reduces the number of host names that need to be fetched.

3.2.2 Probing the hosting IP addresses

The UNIX “dig +short [hostname]” command is used to check the IP address of the advertised hostname. We save the domain-IP pair in a database table. Since a domain can be hosted on more than one IP address and an IP address can host many domains, there is a many-to-many relationship between domain and IP and each domain-IP pair is a unique entry. We also record the date when the domain is first observed in spam emails and the last time it is observed. The WHOIS information for each IP is also retrieved using the “dig” command, and we store the network block, organization name, country code and ASN number in another table. The two tables are linked by IP index.

4. METHODOLOGY

This section describes the methods we use to cluster spam emails based on the domain names, hosting IP addresses and email subjects. The emails in the same cluster are considered to be originated from the same spam organization.

4.1 Fuzzy String Matching

First we define several terms that will be used to calculate similarity between two strings.

4.1.1 Inverse Levenshtein Distance

The most common way to measure disagreement between strings is through edit distance, also referred as Levenshtein distance (Levenshtein, 1966). Because we want to measure the similarity rather than distance, we use dynamic programming to find the alignment between a pair of strings *s* and *t* that maximizes the number of matches. The resulting number of matches between strings *s* and *t* is called their inverse Levenshtein distance, written as ILD(*s*,*t*). For example,

String s:	S	a	t	u	r	d	a	y
String t:	S	_	_	u	n	d	a	y

ILD(*s*, *t*) = 5

4.1.2 String Similarity

We prefer the measure of similarity between a pair of strings to be always between 0 and 1. We want it to express the portion of the strings that match. The Kulczynski coefficient accomplishes this but is defined for sets instead of strings. The Kulczynski coefficient on sets A and B is defined by:

Kulczynski (A, B) = (|A∩B|/|A| + |A∩B|/|B|) / 2

where |A| and |B| are the size of set A and B.

It yields a value between 0 and 1.

We want to define a Kulczynski coefficient for strings in a way analogous to sets. Having the number of matches from the alignment, we define the Kulczynski coefficient for strings s and t by:

Kulczynski (*s*,*t*) = (ILD(*s*,*t*)/|*s*| + ILD(*s*,*t*)/|*t*|) / 2

where |*s*| and |*t*| are the length of strings *s* and *t*.

Therefore Kulczynski(“Saturday”, “Sunday”) = (5/8 + 5/6) / 2 = 0.59.

4.1.3 Subject Similarity

We next describe the matching algorithm to compute similarity of email subjects, which contain multiple tokens. We define: a *token* is a sequence of nonblank characters in a subject; tokens are separated by spaces. A subject will be regarded as a sequence (or string) of tokens. The number of tokens will be defined as the subject length, analogous to the string length as the number of characters in the string.

4.1.3.1 Subject similarity score based on partial token matching

Since a subject is a string of tokens, we can compute similarity of subjects as described in section 4.1.2: the similarity of subjects a and b is computed as Kulczynski(a, b), a and b are matched as two strings, where each token in a and b is treated like a character in a string.

However, each token is actually a string of characters. We observed some tokens could partially match each other because they were generated by a pattern to produce variation in email subjects. For example, the discount amount in the following two subjects:

February 70% OFF

February 75% OFF

Therefore, when matching a pair of tokens, we allow tokens to partially match each other if they have the same length. In particular, if two tokens *p* and *q* have the same number of characters, say *n* characters: length*(p)* = length*(q)* = *n*, we define match(*p*, *q*) = *m*/*n* where *m* is the number of matching characters. The matching is done like this: for each character $(p_1, p_2, ..., p_n)$ in *p* and $(q_1, q_2, ..., q_n)$ in *q*, compare p_i with q_i. Hence match(*p*, *p*) = 1. Thus the matching score for the above example is 2.667/3 = 0.89, because 70% is partially matched to 71%, yielding a score of 0.667.

4.1.3.2 Adjusted similarity score based on subject length

Some subjects are longer than others, containing more tokens. The chance of two long subjects matching each another is much less than that of two short subjects matching each other, while yielding approximately the same similarity score. Therefore, a coefficient is introduced to adjust the subject similarity score based on the subject length. The purpose of the coefficient is to decrease the credit given to short subjects that match each other.

According to the statistics of our dataset, about 60% of all subjects have 5 or fewer tokens. We consider 5 to be the critical length: if the average subject length of two subjects being compared is 5 or more, the coefficient will be 1, but if their average subject length is less than 5, the coefficient will be less than 1, decreasing the credit for matching. The similarity score for subjects a and b will be:

*S(a,b) = C * Kulczynski*(a,*b*),

where $C = \sqrt{\min(\frac{|a|+|b|}{2 \times MaxLength}, 1)} = \sqrt{\min(\frac{|a|+|b|}{10}, 1)}$

4.2 Clustering Spam Domains on a Daily Basis

Since new emails are added to the database every day, we need an on-going clustering method for spam emails. It is not effective to re-cluster the entire data set each time we receive new emails. We want to not only cluster new emails as soon as they arrive but also identify relationships between the new clusters and the previous clusters. Therefore, we use a daily clustering strategy and then link clusters in two adjacent days if they resemble based on interesting email attributes. In doing so, we can find what the clusters look like in the most recent days as well as tracing them back and find out what they look like historically.

The purpose of the daily clustering is to sort spam domains into different groups based on where they are hosted and the subject line of the emails where they are found. (The subject line is often a useful indicator of the content of the emails, but even when the subject is misleading, it can be used to show relationships between non-identical URLs.) In order to do that, we need to define similarity for host IP addresses and email subjects. Because each domain may correspond to several IP addresses (multiple DNS entries) and several email subject lines, similarity coefficients will be used to compute similarity between two sets. We also have to define how to match two subjects and two host IP with some fuzziness.

4.2.1 Hosting IP similarity between two domains

A domain name can be resolved to several IP addresses as a way of load balancing and improving search results. The nameserver will direct requests to different IP addresses based on the order they arrive. If the domain has three IP entries, usually the *n*th request will go to (n%3)th IP address. Apparently spammers are taking advantage of this to increase their site availability. Therefore, the comparison of IP addresses between two domain names becomes a set operation. We use Kulczynski coefficient to measure the similarity between two IP address sets.

When matching two IP addresses, we allow two IP addresses be partially matched if they belong to the same subnet, which is recognized by matching the first three octets. For example, 1.2.3.4 will partially match 1.2.3.5; we assign a score of 0.5 in this case.

For IP sets A and B, |A| <= |B|, we match each IP address in A to all IP addresses in B and choose the maximum matching score S_i. The sum of S_i is $\sum_1^n S_i$ (|A| = n, |A| <=|B|), which replaces the |A∩B| in the Kulczynski coefficient formula.

Some domains have many hosting IP addresses, while others have fewer. Considering the size of IP set, the chance of two sets of size four matching to each another is much less than that of two sets of size one matching each other. If a pair of domains each corresponding to four IP addresses and has perfect match, it is very unlikely that this happens by chance. Therefore, a coefficient is added to adjust the IP similarity score based on the size of an IP set.

According to the statistics of our dataset, only 10% of all domains have resolved to more than 4 IP addresses. Therefore, the maximum size is set to 4. If the average size of two IP sets being compared is larger than 4, the coefficient is set to 1.

The IP similarity score will be:

$S(A,B) = C * Kulczynski(\mathrm{A, B})$,

$$\text{where } C = \sqrt{\min(\frac{|A|+|B|}{2 \times MaxSize},1)} = \sqrt{\min(\frac{|A|+|B|}{8},1)}$$

For example, if domain A has IP set {1.2.3.4, 4.5.6.8, 3.5.6.1} and domain B has IP set {1.2.3.4, 3.5.6.2}

S(A,B) = 0.79*(1.5/3 + 1.5/2)/2= 0.49

4.2.2 Subject similarity between two domains

We also retrieve the email subject from emails that reference a certain domain. Each domain is linked to a set of subjects. The subject similarity between two domains is calculated in the same way as IP similarity using Kulczynski coefficient. The subject similarity score is used to strengthen the relationship between domains that partially match to each other in IP addresses.

As we observed some spam subjects are generated using patterns, for example "Coupon ID ####", the only difference is in the ID number. No common subjects will be found between these two sets of subjects using exact match, but we know they are related. Taking this into account, we substitute the exact string matching with the fuzzy matching algorithm described in section 4.1. By using fuzzy matching, the comparison between two subjects yields a score between 0 and 1, instead of a "yes" or "no" answer.

For subject set A and B, |A| <= |B|, we match each subject in A to all subjects in B and choose the maximum matching score S_i. The sum of S_i replaces the $|A \cap B|$ in the Kulczynski coefficient formula.

The similarity score is then calculated using the Kulczynski coefficient.

4.2.3 Overall similarity score between two domains

An overall similarity score is calculated by taking the average of the hosting IP and subject similarity scores.

The weight and threshold is assigned by forensic investigators based on empirical experiences. When two domain names have the perfect IP or subject similarity scores, we are confident these two domain names are related. Therefore, we set the threshold to be 0.5, which will cover the scenarios when IP score is perfect regardless of what the subject score is or when subject score is perfect regardless of what the IP score is. When the IP and subject scores are not perfect, the average score is a linear function: $x + y \geq 1$, all the points above the line $x + y = 1$ will be accepted. We also tried quadratic function: $x^2 + y^2 \geq 1$ and found the result was almost the same for leading clusters, because the domain names usually have both hosting IP addresses and subjects in common.

4.2.4 Bi-connected components

Using the domain name as vertex and similarity score as edge, we can build a graph. Initially each connected component is considered a cluster. We then use the bi-connected component algorithm to find if the domain names in a cluster are well-connected. According to the definition of bi-connected components (Baase, 1988), a graph is bi-connected if and only if it contains no articulation point, also called a cut vertex. The removal of an articulation point will cause the graph to be disconnected. We try to find if there is any domain name in a cluster that acts as an articulation point. Such domain names may be popular domain names being referenced in a spam email. Therefore, we build the graph by connecting two domain names if their similarity score passes the threshold, and apply the bi-connected component algorithm to detect any articulation points. We will ignore an articulation point if it separates a single domain vertex from the whole graph because it has trivial impact. But if it connects sub-components with a considerable size, we will break up the graph into several bi-components.

4.2.5 Labeling emails based on domain clusters

Once the domains are grouped, we label the emails accordingly. However, there is a problem of conflict if an email references several domains that point to different hosting IP addresses. This situation usually happens if a spam email references common websites, for example, "yahoo.com" or "pctools.com", etc. To deal with this, we come up with a heuristic rule. Because a spam host is likely to host many spam domains at a time for various reasons, a spam domain are more likely be connected with other spam domains. But a referenced URL is unlikely to be grouped with other domain names, for example, "yahoo.com" and "pctools.com" will probably stand by themselves. Knowing this, we assign an email to the IP group that contains the most number of domain names if a conflict occurs. Therefore, an email is more likely to be assigned to the spam group rather than the referenced domain name group. The rule might not work for newsletters, but we are not interested in investigating those emails, which usually form small clusters in our experiment.

4.3 Linking Daily Clusters

Because most leading spam campaigns will last for a long period of time, it will be worthwhile to observe the evolution of a campaign through a period of time. Pharmaceutical spam is a primary example of this, with several campaigns which spanned the entire dataset for this study.

Daily clustering provides a summary of daily spam campaigns. Next, clusters from two different days can be compared based on cluster features – attributes of emails that suggest relationship between clusters.

The method is as follows: when clusters of the current day are produced, we try to match them to the clusters of the previous day. We may want to focus on the leading clusters, those that account for most the spam emails that day. However, some spam may subside for several days and come back again. Therefore, if a cluster of current day cannot be matched any cluster of the previous day, we'd like to keep tracing back for at least a week before we declare it as a new clusters with no predecessors.

4.3.1 Similarity between two clusters

Two clusters are matched to each other based on the same two attributes used in daily clustering: subject and hosting IP addresses. Each cluster includes a group of emails, which contain domains names. The emails are associated with a set of subjects and domain names with a set of hosting IP addresses. We use Kulczynski coefficient to compute the similarity between subject and IP sets.

4.3.1.1 Host IP Similarity between two clusters

Consider the following two real clusters:

Cluster A from day 1

ip_address	domain count
60.191.221.126	327
220.248.186.101	327

Cluster B from day 2

ip_address	domain count
60.191.221.126	348
60.191.221.135	1
64.182.91.176	1
68.183.244.105	1
72.32.79.195	1
72.51.27.51	1
219.152.120.12	1
220.248.172.37	1
220.248.186.101	348

A daily cluster may contain many domains that are hosted at different IP addresses. Some IP addresses may host more domains than other IP addresses. In the above example, the IP addresses 60.191.221.126 and 220.248.186.101 are dominant in Cluster A and B, hosting 99% of domains in both clusters. The other IP addresses are obvious outliers, hosting only 1 domain each. It may caused by falsified IP information or wrong inclusion of domain names in daily clustering. Simple set comparison will find poor IP overlap between the two clusters. Therefore, the domain count needs to be taken into account.

Two IP addresses will still be matched in the same way as in section 4.2.1. Two identical IP addresses will have a matching score of 1. If they reside on the same subnet (the first three octets match), a score of 0.5 is assigned.

For IP sets A and B, |A| <= |B|, we match each IP address in A to all IP addresses in B and choose the maximum matching score S_i, then each matching score will be multiplied by the square root of the smaller domain count of the two IP addresses. The sum of adjusted S_i will replace |A∩B| in Kulczynski coefficient formula.

|A∩B| = $\sum_{1}^{n} C_i S_i$ (|A| = n, |A| <=|B|), where $C_i = \sqrt{\min(a_i, b_k)}$

Here, a_i and b_k are the domain count of two matching IP addresses yielding score S_i. In the above case, the perfect matching on 60.191.221.126 and 220.248.186.101 will be counted as $\sqrt{327} \times 1$.

|A∩B| = $\sqrt{327} + \sqrt{327}$

The set size will also be the sum of square roots of all the domain counts. If a_i is the domain count for an IP address in cluster A and b_i is the domain count for an IP address in cluster B, then

|A| = $\sum_{i=1}^{m} \sqrt{a_i} = \sqrt{327} \times 2$

|B| = $\sum_{i=1}^{n} \sqrt{b_i} = \sqrt{348} \times 2 + 7$

S(A,B) = Kulczynski(A, B) = (|A∩B|/|A| + |A∩B|/|B|)/2

= (1+36.17/44.31)/2=0.9

4.3.2.1 Subject Similarity between two clusters

The subject similarity between two clusters is computed in the same way as the subject similarity between two domain names in daily clustering (See 4.1 and 4.2.2). The cluster with fewer subjects is matched to the other cluster using fuzzy string matching. The best match is found in the larger cluster for each subject in the smaller cluster, the summation is then taken as the intersection. The Kulczynski coefficient is used to capture the subject similarity of the two clusters.

4.3.2 Linking two clusters

The average of subject and IP similarity scores between two clusters is used to decide whether the two clusters are related. Because the two clusters are from different days, we want to relax the threshold a little bit. However, we are not sure how much is appropriate because it is hard to predict when the spammer will make major changes to his spamming strategies. Therefore, we will store all the similarity scores in the database as long as they are not zero. The investigator has the choice to set a threshold to select the scores that interest him. For the experiment, we set the threshold to be 0.4 for clusters from adjacent days, a littler lower than the threshold for daily clustering.

5. RESULTS

From the 638,678 emails collected from June and July 2009, 350,394 emails were used for clustering. The remaining emails were excluded either because the parsing program did not find a URL in the emails or the domain name extracted from the URLs could not resolve to an IP address, indicating that the advertised website was unavailable.

We extracted 16,348 domains from the emails, and most of them have used wild-card DNS entries. The ratio between the number of host names and the number of domain names is over 100: 1. The host name here is a sub-domain that is created from an existing domain by attaching a string before the domain name. For example, “live.com” is a domain name, and “ghl234.live.com” is a host name. This indicates that by studying domains instead of URLs in emails, we effectively compress the data while not losing valuable information.

5.1 Daily Clustering Result

Most daily clusters are very small, containing at most six emails and at most two domain names. The largest daily cluster usually has more than 1000 emails and more than 100 domain names. Figure 2 shows the number of emails in the top 5 daily clusters compared to the total number of emails which are used in clustering. The emails in top 5 daily clusters account for about 83% of total emails.

The leading clusters are most interesting to us and probably also to law enforcement personnel. Therefore, we further examine the large clusters to validate if the domains and emails in those clusters are really related. For example, the largest cluster on July 30 has 2617 emails and 155 domains, which account for almost 48% of the emails included in clustering that day. This shows how dominant the leading clusters are in our dataset.

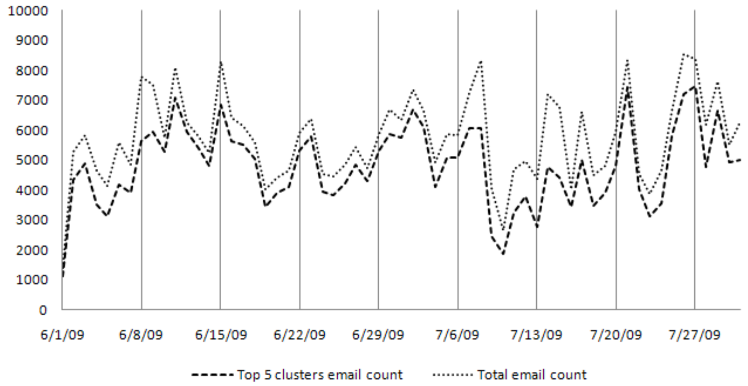

Figure 2: The number of emails in top 5 clusters compared to total emails

Figure 3 shows interconnectivity of domain names, hosting IP addresses and the country of the network containing the IP addresses in the largest cluster on July 30. The domain names are connected to the IP addresses and IP addresses to the countries. The 155 domains are divided into three subgroups based on hosting IP addresses. The biggest sub-group contains 140 domains, which were all hosted at four dominant IP addresses. The second sub-group contains 13 domains, which were hosted at several other IP addresses in addition to the four main IP addresses. The third sub-group contains only two domains: one is hosted at five IP addresses, out of which four are common to the ones in sub-group 1, the other is hosted at 159.226.7.162. Three of the four dominant IP addresses reside in China and the other in Russia. It is unlikely for an investigator to relate an IP in Russia to IP addresses in China unless there is sufficient evidence to support that.

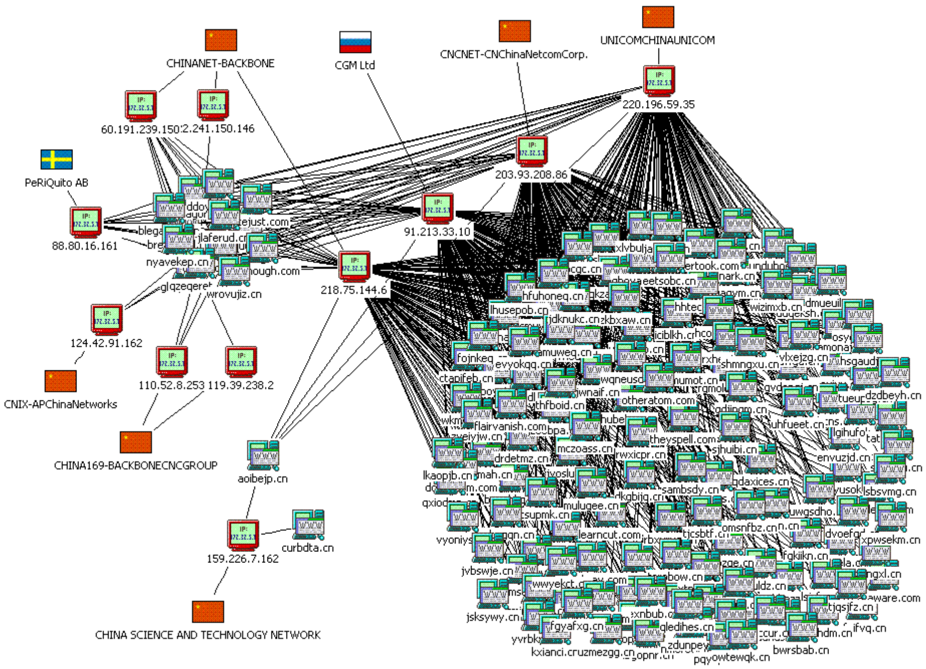

Figure 3: Domain names and related IP addresses in July 30 largest cluster

We then pulled email samples for several domains in each of the sub-groups. Figure 4 shows the connection between some sample emails, domain names and hosting IP addresses. Sample domain names are taken from each subgroup from Figure 3 and put into the middle column. The first two domains are sampled from the second sub-group, the last two from the third sub-group and rest from the largest sub-group. Sample email screenshots are taken for the 10 domain names and put into the left column. The associated hosting IP addresses are put into the right column. The links show that they are all related to each other: they either share the same host IP addresses or are referred to in emails with the same template. Subgroup 2 is linked to subgroup 1 by the common hosting IP addresses. Subgroup 3 is linked to subgroup 1 by common emails. We can see at least four different email templates that are substantially different from each other in appearance. A human may still able to link sample email #3 with #4, but is not likely to link #1 with #2, and #5 together. There are several more email templates from the largest cluster not illustrated here.

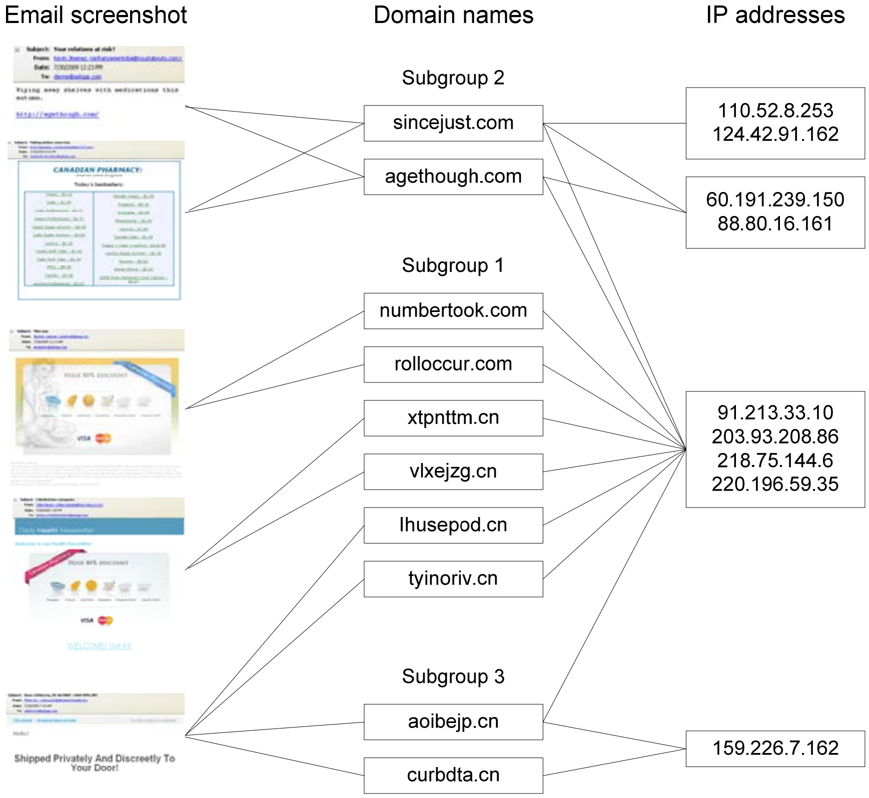

Figure 4: Connection between sample emails, domain names and hosting IPs from largest cluster on July 30

We checked sample domains in each of the three subgroups and found they were all "Canadian Pharmacy" scam websites. We believe the remaining domain names are likely also "Canadian Pharmacy" scam. The fetched web pages may also group these domain names together, but the process is time-consuming and the fetched web content may not be correct. For example, we found some hosts have counter-measurements that will ban an IP if it tries to probe the server repeatedly, thus we will always get a time-out response. Another concern is the business model of affiliate program spammers. Many large affiliate programs, for example the GlavMed program, which owns the illegal "Canadian Pharmacy" content, pay individuals for creating traffic which results in purchases of their products. On one level, all of the “Canadian Pharmacy” spammers are related, because they are all spamming members of the GlavMed affiliate network. However, it is more valuable to identify the spammers by their individual organizations. A familiar example may be the franchise program for a large fast-food restaurant such as McDonalds. Some McDonald's franchisees own only one restaurant while others own several dozen. But it would be incorrect to say that all McDonald's restaurants are owned by the same company. They are affiliated. In addition, a restaurant franchisee may own many kinds of restaurants, not just McDonald's. In the same way, a spammer may spam for several different programs, one may send spam for pills and watches, while another sends spam for pills and pornography. By concentrating on what spam is sent, and where the spammed websites are hosted, we believe we are identifying the "franchisee", rather than making the error of grouping together all spammers who belong to the same affiliate program. In clusters from other days, we see websites such as fake Rolex watches, Canadian pharmacy and Bank of America phishing mingled together.

5.2 Tracing Clusters over the Experiment Period

In this experiment, we traced clusters from adjacent days for a period of two months. A threshold of 0.4 is used: if the average of IP similarity score and subject similarity score passes that threshold, the two clusters are considered related. The biggest cluster is traced from the beginning of June to the end of July, with average IP score of 0.89 and average subject score of 0.28.

Figure 5 shows the number of emails and new domain names belonging to the biggest cluster for the experiment period. Here new domain names means the domain names have never been seen in our database prior to the current date. Therefore a domain name will only be counted at the date when it first appeared no matter how long it lasts. The total number of emails is 221,654, compared to 7,386 domain names. There were no new domain names found on July 16, even though the spam emails kept coming; maybe the spammers took a day off.

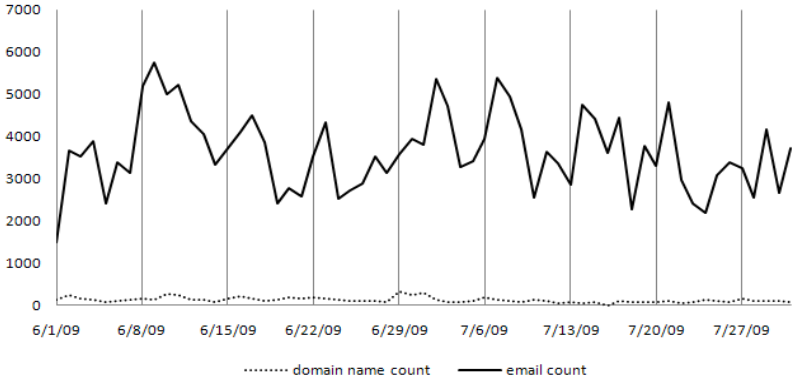

Figure 5: Daily email and domain count of the largest cluster

Most domain names last for a very short period of time, 59% lasting for one day and 39% lasting for two days. If we further break down the domain names by their top-level domain, most of the domain names end with ".cn", followed by ".com", with the ".com" count being only 10% of the ".cn" count. In Table 1, the second row shows the number of domain names for each top-level domain in our largest cluster; the third row shows the number of domain names which last for only one day for each top-level domain; the fourth row is the percentage of row 3 in proportion to row 2; the last row is the time period in which each top-level domain appeared. The ".com" domain names usually live longer than ".cn".

Table 1: domain name count of top level domain in largest cluster

Top-level domain	cn	com	ru	com.cn	net
# of domains	9107	1029	303	26	2
# of domains lasting for one day	5538	400	229	2	1
Percentage of domains lasting for one day	61%	39%	75%	8%	50%
Period seen	6/1 – 7/31	6/1 – 7/31	6/1 – 6/11	6/4 – 7/15	7/22 – 7/29

The biggest cluster contains 42 IP addresses, 14 of which have more than 1000 associated domain names. Some IP addresses actually appeared in our database as early as in late May and some of them are still alive in August. Table 2 shows some of the top IP addresses (associated with most domain names). They are located on different networks in China. Some IP addresses are used for a short period of time, but IP 203.93.208.86 is used throughout the experiment period. IP 58.17.3.41 is used in the first half, stopping at June 21 and IP 218.75.144.6 picks up in the second half, from June 20 to July 31.

Table 2: Top hosting IP addresses of the largest cluster

IP address	Host owner	Country	Active period	Domain count
58.17.3.41	China Beijing Superman Internet Cafe	China	5/27 – 6/21	3427
60.191.221.123	Jinhua Telecom Co.,Ltd	China	6/10 – 6/21	1965
60.191.239.150	Jinhua Telecom Co.,Ltd	China	7/1 – 7/26	1947
60.191.239.153	Jinhua Telecom Co.,Ltd	China	6/20 – 6/28	1008
61.191.191.241	Hefei Chinanet Anhui Province Network	China	6/10 – 6/30	2779
119.39.238.2	Cnc Group Hunan Yueyang Network	China	6/20 – 7/5	1965
203.93.208.86	Qingdao China Unicom IP Network	China	5/22 – 7/31	7600
218.75.144.6	Changsha Chinanet-hn Changde Node Network	China	6/20 – 7/31	3861
222.241.150.146	Changsha Chinanet-hn Hengyang Node Network	China	6/29 – 7/5	1051

An interesting thing is the correlation between different IP addresses on the number of associated domain names. The reason is that when a new spam domain appears, it usually points to several IP addresses. As a result, we see high correlation on domain name count among IP addresses over a period of time. In this case, only domain names never before seen will be counted. Figure 6 shows the correlation between 58.17.3.41 and 203.93.208.86 from June 1 to June 19, and the correlation between 218.75.144.6 and 203.93.208.86 from June 22 to the end of July. June 19 to June 22 appears to be the transition period when the DNS entries are being updated. There were correlations between short-lived IP addresses as well. Figure 7 shows the correlation between 218.75.144.6 and two other IP addresses during the second half of the experiment. 218.75.144.6 is perfectly correlated with 119.39.238.2 from June 20 to July 5, and perfectly correlated with 60.191.239.150 from July 8 to July 22. Even though the spammers are moving domains among different IP addresses, some IP addresses are more consistent. In addition, some domains will still point to old IP addresses during the transition period before they disappeared. We were able to find partial IP overlap between clusters of adjacent days during the transition period.

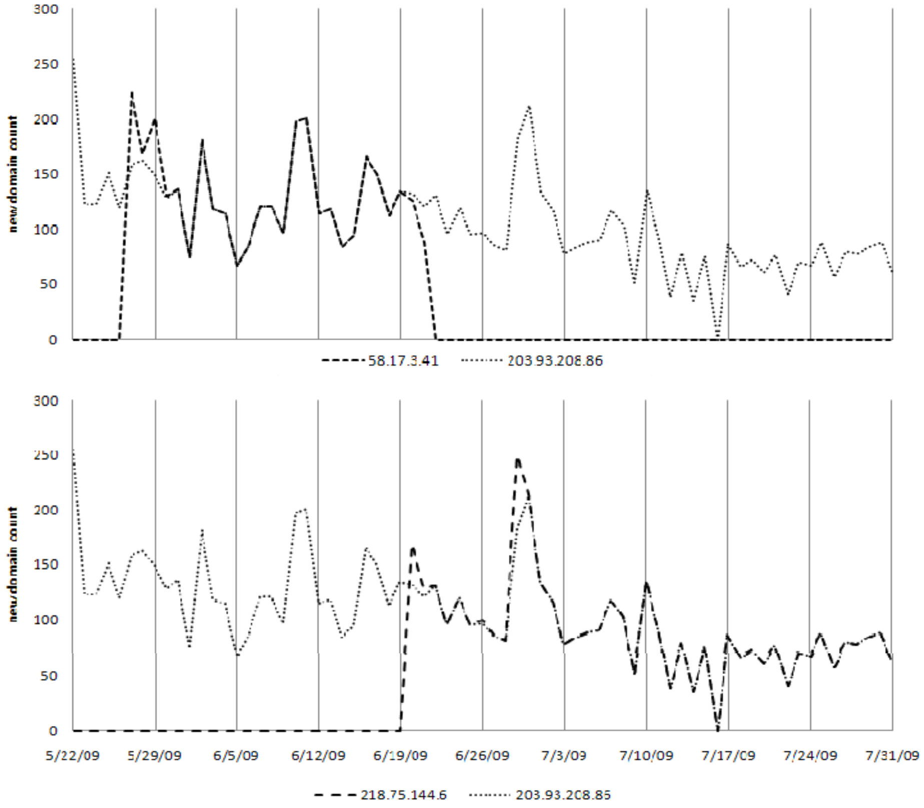

Figure 6: The number of new domain names hosted on IP addresses 58.17.3.41, 218.75.144.6 and 203.93.208.86

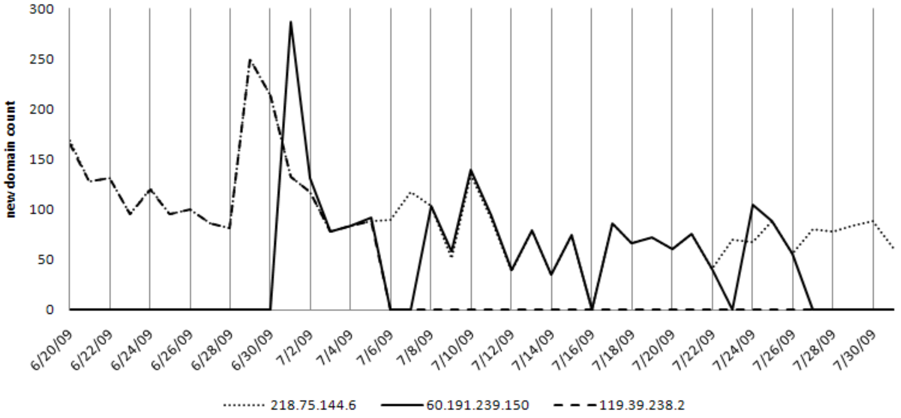

Figure 7: The number of new domain names hosted on IP addresses 218.75.144.6, 60.191.239.150 and 119.39.238.2

We also checked IP addresses of the sending machine, located in the "Received" records of email headers. In the largest cluster, the number of sending IP addresses is about 70% of the number of emails. The number of sending IP addresses increased and decreased along with the number of emails (Figure 8). The sending IP addresses are evenly distributed among different IP ranges, thus the spam emails are coming from all over the world. When the number of spam emails increased on some days, it was because more machines were sending spam, not because some machines were sending more emails. The large number of sending IP addresses suggest that the spam in the largest cluster is probably sent via botnets. Therefore, the spammer who created the web sites is likely either responsible for spreading Trojan viruses and turning computers into bots, or does business with the botnet creator.

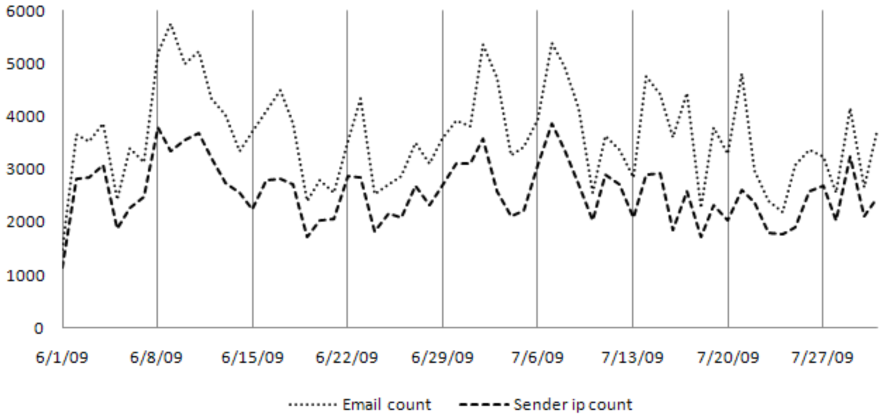

Figure 8: The number of emails and sending IP addresses in the largest cluster

6. CONCLUSION

Starting with spam, we investigated the domain names appearing in emails and their hosting IP addresses, combined with email subjects. We are able to link spam campaigns that are seemingly unrelated based on observation of their inherent attributes by human investigators, as shown in Figure 4. Our biggest clusters account for from one-third up to half of daily spam emails. Based on human observation, the spam is mostly pharmaceutical spam.

The results of this experiment confirms our expectations of spammers registering a large number of domain names to combat domain blacklisting. The largest spam group we found was associated with close to a hundred new domain names each day. We fetched the registrar records of some domain names in our cluster and found that in many cases a single identity was used to register hundreds of domain names in a short period of time and the identity was obviously a disguise. For example, a Canadian Pharmacy spam domain was registered by a Chinese interior remodeling company, who also registered over 100 other domains. It is hard to imagine that an individual company will register so many domain names for legitimate purposes and what was hosted there had nothing to do with the company's business. Many domains were registered in China, ending with ".cn". We checked the destination web sites and found many ".cn" domains used in spam were actually redirected to ".com" domains when the website was visited. Therefore, the short-lived domains are used to protect the real destination domains which never appeared in the spam emails we collected. We suspect the ".cn" domains are probably easier to register and cheaper to buy and the registrar does not care what is actually hosted there.

The 7,000 plus domain names found in our largest cluster discovered through the 2-month experiment period were linked to 221,654 emails. Even though many emails had different appearances in email bodies, as shown in Figure 4, we believe they are the work of one spam group because the destination web sites have the same look and feel. Just using inherent attributes from emails, such as email content and email header, would fail to group them together.

The spammers also exploit wild-card DNS records to create many phantom host names from a single domain name. This suggests domain name blacklisting will be more effective than URL blacklisting if we can confirm that a domain is registered solely for spam usage. It also explains why Webb et al. (2007) found many duplicate web spam pages in their corpus. If so many host names are created from a relatively smaller set of domain names that are actually hosted at the same place, it is not surprising the fetched web pages will be identical. Some of the domains in our cluster were associated with more than 10,000 hostnames. Therefore, fetching the web pages for all of them would not be efficient considering the volume of spam today.

By monitoring the hosting IP addresses, we discovered several networks that are heavily used by spammers, mostly residing in China. The lack of adequate regulation and legislation (Qi et al. 2009) in that country is probably the main cause that these networks are exploited by spammers. Fourteen IP addresses have been found to host more than 1000 domain names. The spammers register many domain names and point each of them to several IP addresses as a way of load balancing. Domain names created during the same period of time will have high correlation on hosting IP addresses. From time to time, the spammers will redistribute domain names to new set of IP addresses. However, some IP addresses remain active for a longer period of time, allowing us to link new IP addresses to old IP addresses. The results suggest that new spam domains can be more effectively detected by checking their hosting IP addresses and significant hosting IP addresses can be reported to law enforcement personnel for termination.

7. FORENSIC APPLICATIONS

Some Internet Service Providers (ISPs) are favored by spammers because they provide “bullet-proof hosting” services that ignore all abuse complaints. Other rogue ISPs are using complex “data center shifting” techniques to hide their criminal activity by shifting between many data centers, claiming to have terminated the offending customer, while really moving him to a new location. This paper has demonstrated a technique to identify “bullet-proof hosting” centers, but more importantly, it can also identify spam clusters migrating from one IP or network to another which can be used to track “data center shifting”. For example, the “Pricewert/3FN” organization, which was described by the Federal Trade Commission of the United States as a “Rogue Internet Service Provider”, was terminated after researchers were able to demonstrate the many types of cybercrime all sharing a common infrastructure (Federal Trade Commission, 2009). Identifying and terminating criminal spamming infrastructure has been shown to provide significant decreases in the world-wide spam volume, for example, after the McColo shutdown spam volumes were reduced by 36% in the United States, and by as high as 73% in other regions of the world (DiBenedetto et al. 2009, Mori et al. 2009). Our research will help forensic investigators to identify those ISPs that are assisting spamming and other cyber criminal activities. An individual investigator may see a dozen spam domains every day, but to relate thousands of domains together will not be easy. Traditional law enforcement technology has not scaled well in cases involving millions of data elements. This paper demonstrates an effective use of data mining to respond to this challenge. Once a network has been discovered to host a large number of spam domains over considerable period of time, a case can be built and measures taken to take the ISP down. It is very likely that the same network is also used by spammers to engage in other kinds of cyber-crimes, such as phishing and online fraud.

8. FUTURE WORK

This research is part of the Spam Data Mining for Law Enforcement project at University of Alabama at Birmingham. We have established a large database of spam emails on grid computing, and are expecting to receive large amount of spam from several sources this year. By that time, the daily volume will reach one million. The focus is to use computing power to process spam emails and extract useful evidence for law enforcement usage. In this research, we used domain names and hosting IP addresses to cluster spam emails, supplemented by email subjects, and produced promising result. In the future, we plan to extract more information from the domain names and IP addresses, such as the nameservers, registrar information and hosting history. With more attributes, we can improve the cluster quality by reducing false-positives and false negatives. We shall also able to discover more relationships among different clusters and produce a more detailed report of the clusters. However, there are some difficulties in retrieving and parsing the registrar information. Each registrar will return the records in different formats, making the parsing very difficult. Before we can process the information automatically with minimum errors, the registrar information can only used for human validation of the clusters. However, we are able to parse the nameserver from the registrar information because of the unified format so it will be the next attribute included in our clustering algorithm.

As the size of our spam corpus increases, we are currently changing the "daily clustering" represented in this paper by clustering of increasingly smaller time intervals. "Emerging clusters" may then be evaluated on a frequency based on the time interval encountered. In our current research new patterns may be identified daily, but we are moving to report hourly on new threats, or eventually every ten minutes, or in even smaller time intervals.

Fetching of the websites is also useful to find out what the products are and relate spam in the same product category. It is also helpful to identify multiple spam campaigns operated by the same spammer. However, we found some hosts have counter-measurements that will ban an IP if it tries to probe the server repeatedly. Therefore, after probably one hundred tries, we will always get time-out feedback. We need to find a way around this preventive mechanism. The existing clustering results will reduce the number of websites we need to fetch: in a well-connected cluster, we just need to fetch some sample domains and the rest will probably be the same. This will also enhance the efforts designed to protect consumers from malicious websites by scanning these websites because the scope of scans for malicious activity can be narrowed only to the "emerging clusters". New websites belonging to existing clusters which are known to be malicious can be categorized as potentially harmful without the need to investigate each website individually.

In this research, the thresholds are pre-set by human based on empirical observation. We plan to add some artificial intelligence in deciding the splitting boundaries, for example, a self-learnt decision tree. However, one concern is that we are dealing with criminals, not machines. A spammer may decide to change his spamming strategies entirely if the old one is not effective or he senses threats. Therefore, the old training data may become totally obsolete. On the other hand, sudden change also means more effort and cost for the spammers too. He has to develop new codes and install new servers. Basically, it is an arms race and we are trying to gain the edge over them.

This research only covers the portion of the spam that contains domain names having a static hosting place. There are spam domains that are hosted at a blog personal space, hacked servers or bots. In these cases, the investigation of the host may not reveal any interesting results. Therefore, we have to analyze those spam messages using a different approach.

9. REFERENCES

Aggarwal, C. C., Han J., Wang, J. and Yu, P. S. (2003). 'A framework for clustering evolving data stream'. The 29th International Conference on Very Large Data Bases. Sept. 9-12, 2003. Berlin, Germany.

Anderson, D. S., Fleizach, C., Savage, S., and Voelker, G. M. (2007). 'Spamscatter: Characterizing internet scam hosting infrastructure'. The 16th USENIX Security Symposium. Aug. 6-10, 2007. Boston, MA.

Baase, S. (1988). 'Graphs and digraphs', in Computer Algorithms: Introduction to Design and Analysis, (2nd ed.). Addison-Wesley, Boston, MA.

Barbara, D. (2002). 'Requirements for clustering data streams'. ACM SIGKDD Explorations Newsletter, 3(2), 23 – 27.

Blumstein, A., Cohen, J. and Nagin, D. (Eds.) (1978). Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates. National Academy of Sciences, Washington, DC.

Calais, P. H., Pires, D. E. V., Guedes, D. O., Meira, W. Jr., Hoepers, C. and Steding-Jessen, K. (2008). 'A Campaign-based Characterization of Spamming Strategies'. The 5th Conference on Email and Anti-Spam. Aug. 21-22, 2008. Mountain View, CA.

Clayton, R. (2009). 'How much did shutting down McColo help?' The 6th Conference on Email and Anti-Spam. Jul. 16-17, 2009. Mountain View, CA.

Cao, F., Ester, M., Qian, W. and Zhou, A. (2006). 'Density-Based Clustering over an Evolving Data Stream with Noise'. The 6th SIAM International Conference on Data Mining. Apr. 20-22, 2006. Bethesda, MD.

DiBenedetto, S., Massey, D., Papdopoulos, C. and Walsh P. J. (2009). 'Analyzing the aftermath of the McColo shutdown'. The 9th Annual International Symposium on Applications and the Internet. Jul. 20-24, 2009. Seattle, WA.

Federal Trade Commission. (2009). 'FTC Shuts Down Notorious Rogue Internet Service Provider, 3FN Service Specializes in Hosting Spam-Spewing Botnets, Phishing Web sites, Child Pornography, and Other Illegal, Malicious Web Content', http://www.ftc.gov/opa/2009/06/3fn.shtm, retrieved on Oct 20, 2009.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G., Paxson, V. and Savage, S. (2008). 'Spamalytics: An empirical analysis of spam marketing conversion'. The 15th ACM Conference on Computer and Communication Security. Oct. 27-31, Alexandria, VA.

Levenshtein, V. I. (1966). 'Binary codes capable of correcting insertion and reversals'. Soviet Physics - Doklady, 10, 707 – 710.

McAfee Avert Labs. (2009). 'McAfee threats report: first quarter 2009'. http://img.en25.com/Web/McAfee/5395rpt_avert_quarterly-threat_0409_v3.pdf, retrieved on Sept 15, 2009.

Mori, T., Esquivel. H., Akella. A., Shimoda, A. and Goto. S (2009). 'Understanding the World's Worst Spamming Botnet'. ftp://ftp.cs.wisc.edu/pub/techreports/2009/TR1660.pdf, retrieved on Oct 17, 2009.

Pu, C., and Webb, S. (2006). 'Observed trends in spam construction techniques: A case study of spam evolution'. The 3rd Conference on Email and Anti-Spam. Jul. 27-28, 2006. Mountain View, CA.

Qi, M., Wang, Y. and Xu, R. (2009) 'Fighting cybercrime: legislation in China'. International Journal of Electronic Security and Digital Forensics, 2, (2). 219-227.

St Sauver, J. (2008). 'Spam, domain names and registrars'[PDF document]. MAAWG 12th General Meeting. Feb. 18-20, 2008. San Francisco, CA. http://www.uoregon.edu/~joe/maawg12/domains-talk.pdf, retrieved on Aug. 15, 2009.

Tom. P. (2008). 'Latent botnet discovery via spam clustering'. The Expanded MIT Spam Conference 2008. Mar. 27-28, 2008. Boston, MA.

Webb, S., Caverlee, J. and Pu, C. (2006). 'Introducing the Webb Spam Corpus: Using email spam to identify web spam automatically'. The 3rd Conference on Email and Anti-Spam. Jul. 27-28, 2006. Mountain View, CA.

Webb, S., Caverlee, J. and Pu, C. (2007). 'Characterizing Web Spam Using Content and HTTP Session Analysis'. The 4th Conference on Email and Anti-Spam. Aug. 2-3, 2007. Mountain View, CA.

Wei, C, Sprague, A., Warner, G and Skjellum, A. (2009). 'Characterization of spam advertised website hosting strategy'. The 6th Conference on Email and Anti-Spam. Jul. 16-17, 2009. Mountain View, CA.

WikiPedia. (2009). 'Wildcard DNS Record'. http://en.wikipedia.org/wiki/Wildcard_DNS_record, Retrieved on Jun. 10, 2009.

Zhang, T., Ramakrishnan R. and Livny, M. (1996). 'BIRCH: An Efficient Data Clustering Method for Very Large Databases'. The 1996 ACM SIGMOD International Conference on Management of Data. Jun. 4-6, 1996. Montreal, Canada.