

Identifying New Spam Domains by Hosting IPs: Improving Domain Blacklisting

Chun Wei, Alan Sprague, Gary Warner, Anthony Skjellum

Dept. of Computer and Information Sciences,
Univ. of Alabama at Birmingham
1300 University Blvd.
Birmingham, AL, USA
{weic, sprague, gar, tony}@cis.uab.edu

ABSTRACT

This paper studies the possibility of using hosting IP addresses to identify potential spam domains. Current domain blacklisting may not be effective if spammers keep replacing blacklisted domains with newly registered domains. In this study, we cluster spam domains based on their hosting IP addresses and associated email subjects. We found some hosting IP addresses were heavily used by spammers to host a large number of domains and persisted for much longer period of time than related domains. Our results show that hosting IP blacklisting should be effective against many point-of-sale spam campaigns, such as pharmaceutical, sexual enhancement and luxury good spam, which mainly use static IP addresses to host their websites. The IP addresses remain active from several days to even a couple of months before replaced by a set of new IPs. Therefore, even when new spam domains appear from time to time, they can be immediately detected as spam domains by looking up the hosting IP address. The reported IP addresses are also useful for law enforcement investigators to identify ISPs that provide bullet-proof hosting services to spammers. The detection and termination of spam domains and their hosts will severely impede spammers' capability to generate revenue from spam.

1. INTRODUCTION

Spam has become a major threat to Internet security today. According to the McAfee threat report [8], there were 153 billion spam messages per day in 2008 and over 90% emails were spam. The reason spam is so hard to kill is that more spam messages are sent from bots, infected computers controlled by command and control servers and people are buying from spam sites because of the relatively cheaper price [7]. The revenue allows the spammers to develop new technologies and send out more spam messages. Spam is also used for phishing, spreading viruses and online fraud, causing more damage to society. One way to combat spam is domain blacklisting, for example, SURBL/URIBL filtering (two popular spam "black lists" used by spam filtering solutions) [19, 20]. The URLs within the spam emails are analyzed and reported to the blacklist. Further incoming emails containing blacklisted domains will be blocked. Measures can also be taken to shut down the domains. The effectiveness of domain blacklisting is challenged by new techniques adopted by spammers. In order to protect the spam websites from termination, spammers keep registering a large number of new

domains every day. St Sauver [18] summarized three major benefits for spammers to do that despite the increase of cost: (1) to reduce the chance of spam being blocked by SURBL/URIBL filtering and increase the chance of survivability 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. In order to shut down the spam, one has to take down all the domains or all the back-end servers. Sheng et al. [15] analyzed the effectiveness of phishing URL blacklisting and found 63% of the phishing campaigns only last for less than 2 hours. By contrast, the majority of the phishing URLs took almost 12 hours to appear on blacklists. Therefore, before a phishing URL was reported to a blacklist, it might already been replaced by a new one. Wei et al. [24] found several IP addresses hosted more than 1000 domains in the first quarter of 2009, showing that the domains owned by the same spammer tend to cluster on certain hosting IP addresses. If some existing domains hosted at an IP address were found to be spam domains, the chance of new domains hosted there also being spam domains is very high. This paper investigated the effectiveness of using hosting IP addresses to detect new spam domains.

2. RELATED WORK

Most research on IP blacklisting focuses on detecting spam-sending machines based on sender's IP addresses [4,11] or the behaviors of the suspected bots [13,14]. Ramachandran and Feamster [11] studied the network-level behaviors of spammers and found the majority of spam messages were sent from a few concentrated portions of IP address space. However, the top 3 networks on their list were Korean Internet Exchange, China Telecom and Sprint. Obviously, we cannot just block all emails from Sprint or China Telecom. Because a large portion of spam emails are sent by botnets [2], Ramachandran et al. [13] tried to detect possible bots by observing the lookups to Domain Name System-based Blackhole Lists (DNSBLs), which are lists of IP addresses that originate spam. Later they [14] analyzed the behavioral patterns of sending machines to discover botnet membership. But the detection of botnets is still a passive countermeasure against spam. The prevention of computers from infection and disruption of the command and control servers would be more effective against botnets than blocking the sending

IP addresses [3]. Ramachandran et al. [12] did a preliminary study on the effectiveness of DNSBLs and found only 5% of bot IP addresses were ever listed at Spamhaus PBL [17].

Most spam filters use email content to decide if an email is spam. The content of a spam email can tell more about the spammer, for example, the URLs in the email. The URLs point to websites where the vital actions take place for spammers to make a profit. A spammer can create fake sending email address and sender's name because it is not necessary for the email recipient to be able to find the true origin of a spam email in order to process the message. However, the URL is usually true because 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 and the spam email becomes useless. Same for a phishing website, if the user cannot open the website, no harm will be done. Therefore, the detection of IP addresses of spam domains appearing in spam messages will be another effective anti-spam measure by disrupting its end goal. If a hosting IP is shut down, all related domain names will be ineffective before they can be moved to a new host. New spam domains can be easily detected by checking the hosting IP addresses if they are still alive. The IP addresses can also reveal the networks that are patronized by spammers. The shutdown of those networks can be a major blow to the spammers. But most current Domain Block Lists (DBL) do not provide hosting IP blacklisting, for example, the Spamhaus DBL is actually a domain URI block list and does not support IP lookup [16].

Several spam research papers have studied the URLs in spam messages. Pu and Webb [9] started by observing trends in spam email message construction, especially obfuscation methods used in HTML-based spam emails to evade filtering. They then built a Webb Spam Corpus, which consists of nearly 350,000 web pages that were obtained from URLs in the HTML-based spam emails [21]. Using the Webb Spam Corpus, they categorized the web pages into five categories: Ad Farms, Parked Domains, Advertisements, Pornography and Redirection [22]. The categorization provided a good survey of spam websites, but was not very helpful for law enforcement personnel, who would want to know more details about the spam campaigns. Their research also identified Top 10 hosting IP addresses with the most web page count and two IP ranges that account for 84% of the hosting IP addresses. However, they did not indicate what the websites looked like and if they were related or not. Later, they used hosting IP address as a feature in clustering web spam because they found the hosting IP range of spam hosts were quite different from the IP range of legitimate hosts [23]. However, on IP range over 204.*, the distinction was not so clear. Our research targets individual spam hosting IP address because we want to identify significant IPs that are heavily used for hosting spam websites and account for a large number of spam messages. These IP addresses pose more threat than other spam IPs that account for fewer spam domains and emails. The detected IPs can be reported to law enforcement agency in order to shut them down. The list can also be use for blocking further spam domains if they cannot be terminated immediately.

The Spamscluster project [1] also fetched web pages using the links in spam emails and clustered the web pages based on

screenshot similarity. They categorized spam campaigns based on the content of the web sites. 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) were infrequent. This might not be true anymore since spammers are improving their hosting infrastructures to protect the servers. 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 replace old ones that are blacklisted. Therefore, the termination of a host or domain name does not necessarily mean a spam campaign has ended.

Konte et al. [5] studied the spam domains hosted using fast-flux service network (FFSN). They collected about 3000 domain names from over 115,000 emails in 2007 and found many point-of-sale domains were hosted at distributed machines with each IP address serving for a short period of time before being replaced by a new IP. Our data set is much larger, averaging about 900,000 per day over a week period of time. Our results showed that static IP addresses were still used to the host majority of the point-of-sale spam websites. One IP was used to host more than 9000 domains pointing to Canadian Pharmacy spam. The IP query always returned the same IP addresses at different times. The IP address remained active for a couple of months, before being replaced by a new IP address.

3. METHODOLOGY

This research aims to cluster spam domains based on their hosting IP addresses and identify the most dominant hosts. Our approach is different from common content-based spam detection research, for example, the ALPACAS [25], which focuses on measuring the similarity between emails. Our study targets spam domains that appear in spam emails and use the hosting IP addresses to group related domains. We believe the shutdown of spam domains and their hosting servers will deliver a heavier blow to spamming infrastructure than filtering spam at the receiving end.

3.1 Data Collection

For this study, the spam emails collected in the first week of January were used for clustering and identified IP addresses were traced over the entire months of January and February to find how long these IP addresses persist. Spam emails were gathered from a volunteer ISP with more than 10,000 email accounts, which are no longer active, implying that most of email messages sent to these accounts would be spam. The dataset contains 35 million email messages in January and 17 million in February. For unknown reasons, there was a decline in the number of emails in February, which was only half of that of January. About 95% of the emails contain URLs. About 175,254 domain names were extracted from the URLs. For example, the domain name portion of the URL "yruz.dixuxigob.cn/index.php" will be "dixuxigob.cn". Then we clustered the domain names based on IP address similarity and email subject similarity. The IP addresses were fetched using UNIX "dig +short" command every 15 minutes until the domain names disappeared from the spam emails.

3.2 IP address similarity

A domain name can be resolved to several IP addresses as a way of load balancing and improving search results. 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 sets.

The Kulczynski coefficient on sets A and B is defined by:

$$Kulczynski(A, B) = (|A \cap B|/|A| + |A \cap B|/|B|) / 2$$

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

It yields a value between 0 and 1.

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, $|A| \leq |B|$, we match each IP address in A to all IP addresses in B and choose the maximum matching score S_i . In order to generalize the Kulczynski coefficient to allow partial matches, we replace $|A \cap B|$ in the formula by $\sum_{i \in A} S_i$ ($|A| = n$, $|A| \leq |B|$)

Since a hosting company may host many domains for small business, there is the possibility that domains hosted at the same IP address may not relate to each other at all. However, if two domains are resolved to the same set of IP addresses of size four or more, the chance that they belong to different businesses but are hosted at the same place is very small. Therefore, a coefficient is added to adjust the IP similarity score based on the size of the IP sets being compared.

According to the statistics of our dataset, over 99% domains resolve to less than 5 IP addresses. Therefore, we set the maximum size to 4. If the average size of two IP sets being compared is 5 or more, the coefficient is set to 1.

The IP similarity score will be:

$$S(A, B) = C * Kulczynski(A, B),$$

where

$$C = \sqrt{\min\left(\frac{|A| + |B|}{2 * MaxSize}, 1\right)} = \sqrt{\min\left(\frac{|A| + |B|}{8}, 1\right)}$$

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}

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

3.3 Subject Similarity

If two domains are hosted at the same subnet and their associated email subjects resemble, they may also be related. To strengthen the relationship between two domains that are partially matched in IP addresses, we also compare the email subjects that are associated with the two domains. 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 blanks. A subject will be regarded as a sequence (or string) of tokens. The number of tokens will be defined as the subject length, similar to the number of characters for the string length. Spammers sometimes use templates to create customized spam messages. In order to match similar subjects that are generated by a pattern, we developed a fuzzy matching algorithm for subject matching based on Levenshtein distance [6].

3.3.1 Inverse Levenshtein Distance

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 e _ _ c t i o n _
String t: s e d u c _ i _ n g
 $ILD(s, t) = 5$

3.3.2 String similarity

In section 3.2, the Kulczynski coefficient on two sets A and B is defined.

We want to define a Kulczynski coefficient for strings in a way analogous to sets. Having the number of matches from the alignment using ILD, 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 $Kulczynski(\text{"section"}, \text{"seducing"}) = (5/7 + 5/8) / 2 = 0.67$.

3.3.3 Subject similarity score based on partial token matching

Since a subject is a string of tokens, we can apply the string similarity on subjects: The similarity of subjects a and b is computed as $Kulczynski(a, b)$, a and b are matched as two strings, each token in a and b is treated like a character in a string.

Each token is actually a string of characters. We observed some tokens were generated by a pattern to produce variation in email subjects. For example,

Personal 72% off
Personal 73% 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 in p and q , compare p_i with q_i . Hence $match(p, p) = 1$. Thus the matching score for the above example is $(2.667/3 + 2.667/3) / 2 = 0.89$, because 72% is partially matched to 73%, yielding a score of 0.667.

3.3.4 Adjusted score based on subject length

Some subjects are longer than others, containing more tokens. The chance of two long subjects matching each other 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

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

3.3.5 Subject similarity score between two domains

In the above section, we define the similarity score between two email subjects. Because each domain is linked to a set of subjects, we also use Kulczynski coefficient to measure the subject similarity between two domains.

For subject sets A and B , $|A| \leq |B|$, we match each subject in A to all subjects in B and choose the maximum matching score S_i . The similarity score is then calculated using the Kulczynski coefficient, where $|A \cap B|$ in the formula is replaced by $\sum_{i \in A} S_i$ ($|A| = n$, $|A| \leq |B|$).

3.4 Overall Similarity Score

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

When two domain names have a perfect IP or subject similarity score, 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. RESULTS

We clustered the spam domains from January 1 to 8. Due to the volume of spam which averages about 1,300,000 per day, we clustered on an hourly basis. Each hour there were about 1000 clusters, most of them being small clusters with less than 100 emails. We suspect those are newsletters and promotional emails sent by advertising companies. The clusters having more than

100 emails were used for further investigation of hosting IP addresses. There are about 30 – 60 clusters in that category per hour. Those clusters usually accounted for over 70% of total emails in a particular hour. The email and domain count was then used to identify the most dominant spam clusters and corresponding hosting IP addresses.

4.1 Canadian Pharmacy Spam

The most significant cluster is the Canadian Pharmacy (CP) spam (identified by fetched websites). It contains domain names that are primarily hosted at 60.172.229.102 and 61.235.117.75, with a few hosted at 58.218.199.97. The campaign became dominant at the 7th hour of Jan 3, and account for about 6 thousand emails for that hour. Figure 1 shows the hourly email count starting from the 7th hour of Jan 3 until the end of Jan 8 for the identified campaign comparing to the total email count. There were about 50,000 – 60,000 emails per hour during that period of time. At some hours, the CP spam accounted for over half of the total amount of emails. Even at hours when the volume of the identified campaign is low, it still accounted for about 10% of the total emails. Surprisingly, when the CP spam volume went up and down, the total spam volume remains stable, meaning the volume of other spam campaigns increased when CP spam volume declined.

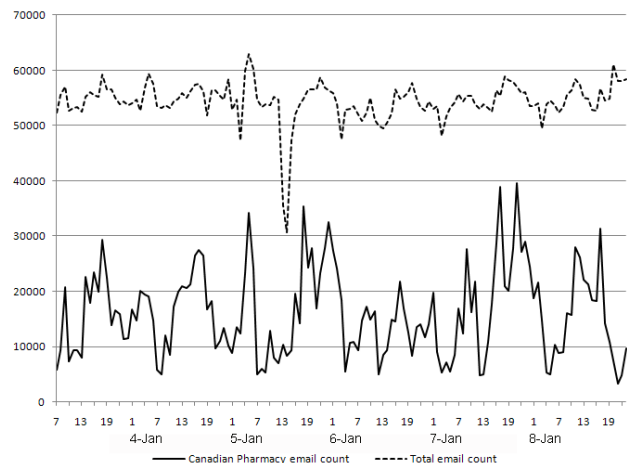


Figure 1: Hourly email count of Canadian Pharmacy spam comparing to total email count during Jan 3-8, 2010

In most days, our clustering algorithm separated the cluster into three sub-clusters because the spammers use three different subject patterns (the words in bracket are customized):

Notification to [username] special 80% OFF of Pfizer

Special 80% discount for customer [name] on all Pfizer

Valued customer [email address] 80% OFF on Pfizer.

Since there is only one IP address, the adjusted coefficient will generate an IP similarity score of 0.5. Different subject patterns will force the domains into different clusters. Human investigation on the subject patterns and fetched web pages confirm they are one campaign. The emails were probably generated by botnets using different templates.

Figure 2 shows the transition from IP both 60.172.229.102 to 61.235.117.75. Before Jan 6, all domains in that cluster were hosted at 60.172.229.102. On Jan 6, the domains were hosted at both 60.172.229.102 and 61.235.117.75. On Jan 7, more domains were hosted on 61.235.117.75 than 60.172.229.102. On Jan 8, no domains names were found to be hosted at 60.172.229.102 in our database. The IP 61.235.117.75 was traced all the way to the first week of March in our database before being replaced by a new IP address. Both IP addresses are located in China.

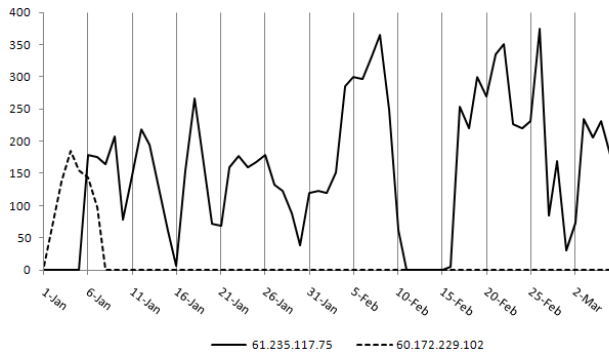


Figure 2: The number of domains hosted at IP addresses 61.235.117.75 and 60.172.229.102 from Jan 1 to Mar 6

Starting from Jan 7, there were 9,921 domain names hosted at 61.235.117.75 that are related to Canadian Pharmacy spam. The domains have either “.cn” or “.ru” as the top-level-domain, indicating they were either registered in China or Russia. The two dominant patterns in the domain names are: (1) concatenation of two English words, such as “senseleast.ru”; (2) alternation between constants and vowels, such as “quzixenov.cn”. The “.cn” or “.ru” domains were redirected to a “.com” domain. For example, “cottonwe.ru” was redirected to “pillsgreatenter.com”. Both domains are hosted at the same IP addresses, with most current DSN records pointing to 202.111.175.31 and 218.93.201.53. The WHOIS information shows that a person named “Zhaohua”, a very common Chinese name, registered “pillsgreatenter.com” on Jan 8. There have been only 2 changes on NS records and 3 changes on IP A records since then. The same person owns 725 other domains. The reverse IP records show that 2063 other domains are also hosted at the same server. This means that the IP record change is not frequent and the same IP address can be used to detect many other spam domains hosted there.

Among the 9,921 domains we discovered, 5,722 domains showed up in spam emails for only a single day, implying that if they were ever reported to domain blacklists as spam domains by human investigators, one would probably never see them again in spam emails. However, if the hosting IP address is used to blacklist the domains, those 9,000+ domains can be easily detected as spam domains.

4.2 Ultimate Replica Watches Spam

Another interesting spam campaign is the Ultimate Replica Watches Spam. Those domains were hosted at 116.127.27.188, which is located in South Korea. However, some other domains hosted at the same IP address pointed to sexual enhancement websites. In some hours of a particular day, the clustering results

show 2 to 3 clusters all hosted at that IP address. The clusters were differentiated by the email subjects. For example, on 9th hour of Jan 3, one cluster was about replica watches, another about penis enlargement, the third was DrMaxman (also a sexual enhancement product).

The IP address 116.127.27.188 was traced till Feb 16 (Figure 3). This cluster is much smaller than the Canadian Pharmacy, averaging about several hundred spam emails per hour and 20 new domain names per day. However, on Jan 18, there was a burst of domain names on that IP address. Reverse WHOIS information of sampled domains showed that the same guy, under the name “Zhang Cheng”, registered 1000+ domains.

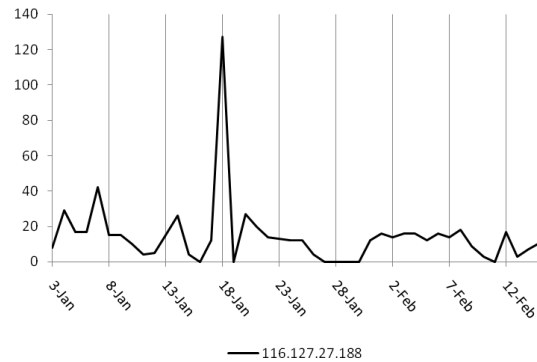


Figure 3: The number of domains hosted at IP address 116.127.27.188 from Jan 3 to Feb 16

4.3 Tracking a Phishing Campaign

Phishing spam is quite different from pharmaceutical spam. Each domain can resolve to a large number of IP addresses, probably by using FFSN. From our clustering results, we initially identified six outstanding IP addresses 112.165.14.77, 190.213.196.98, 190.241.253.188, 200.169.71.144, 210.106.80.90 and 218.153.64.25 that were tightly connected to each other. The DNS lookup returned all six IP addresses. By tracing the clusters containing any of these six IP addresses, we found the emergence of a phishing campaign.

Table 1: The number of IP addresses used by the phishing campaign

Time stamp	Jan 6, 9am	Jan 7, 8am	Jan 8, 8am	Jan 8, 12pm
# of IPs	21	198	455	476
# of emails	233	264	618	1032

Table 1 shows the time stamp and the corresponding number of resolved IP addresses. Starting from Jan 6, 9am, there were only 21 IP addresses identified. At Jan 7, 8am, the number jumped to 198. At Jan 8, 8am, it burst to 455, then at 12pm, to 476. The email count was also on the rise, but not as dramatically as the IP count.

The following is a sample email message:

Dear user of the email.com mailing service!

We are informing you that because of the security upgrade of the mailing service your mailbox (username@email.com) settings were changed. In order to apply the new set of settings click on the following link:

http://email.com/owa/service_directory/settings.php?email=username@email.com&from=email.com&fromname=username

Best regards, email.com Technical Support.

Letter_ID#B2S602QQE9P3X

The actual URL (shown below) points to a domain “okqwac.com.pl” instead of “email.com”.

http://username.email.com.okqwac.com.pl/owa/service_directory/settings.php?email=username@email.com&from=email.com&fromname=username

With the URL already disabled, we are not sure if it was used to steal personal information or to spread a virus. (Note: every user received a customized email, we substituted the real username with “username” and real domain name with “email.com”).

4.4 Other IP addresses

Apart from the IP addresses described in the previous sections, we also discovered several additional IP addresses that are worth noting. Table 2 shows some of the other significant IP addresses identified during the first week of January. These IP addresses did not persist as long as the ones used by Canadian Pharmacy spam. The spam messages came sporadically, but during their peak times, they could account for a significant number of domain names and emails. At some hours, the total related spam emails surpassed 10,000.

We also checked WHOIS information of sampled domains in each of the cluster and found the same scenario: a single identify is registering many domains that will expire in one year and the name and contact information looks dubious.

Table 2: Summary of other significant hosting IP addresses

IP addresses	# of domains	Active period	Products
124.61.222.223	711	Jan 3 – 8	Watches and sexual enhancement
58.218.250.107	255	Jan 1 –28	Drugs
202.111.175.126	68	Jan 3 –5	Sexual enhancement
116.123.221.91	61	Jan 3 – 7	Drugs Casino

5. CONCLUSIONS

Current domain blacklists are maintained manually by collecting domain names from spam emails. Many domain names are

inactive by the time they are blacklisted. The IP blacklist only focuses on the sending IP addresses of spam emails, which is not effective against botnet-generated spam. This paper studies the possibility of using hosting IP blacklist to detect new spam domains. The domain names were extracted from spam emails and hosting IP addresses were fetched. Then domain names were clustered based on IP similarity and subject similarity. All these processes can be done automatically within 2-3 hours after the email arrives. The original goal of the research is to identify spam hosts that are heavily used by spammers to host many spam domains and account for a large number of spam messages. The identified IP addresses and related spam domains be used as forensic leads to trace the identity of the owners by using WHOIS information and IP history records. The information can also be used by law enforcement personnel as evidence against certain ISPs who provide bullet-proof-hosting services to spammers. If the hosts cannot be terminated immediately, the IP list can be used for checking spam domains in newly arrived spam emails. Since our goal is not filtering spam domains from legitimate domains, we did not measure our method against legitimate data. However, since over 90% emails are spam nowadays [8], the leading domain clusters we identified are confirmed to be spam domains. A legitimate business is not likely to register hundreds of domains for their websites. The legitimate domains are not likely to cluster with each other. Therefore, we believe the domains and IP addresses listed in our results are solely for spam purposes and their termination can cripple some major spam campaigns.

Our results show that this method will be effective against pharmaceutical, sexual enhancement and luxury good spam, which mainly use static IP addresses to host the websites. The number of domain names clearly outnumbers that of the hosting IP addresses. The IP addresses remain alive from several days to even a couple of months. The spammers keep registering new domains to replace old domains. But because the hosting IP addresses did not change, new domains can be easily linked to old domains. Occasionally, the spammer will change the hosting IP addresses. But there was a short period of time when both old IP and new IP addresses were used, allowing us to link the new IP address to the old IP address. The IP addresses detected are located either in China or South Korea. We suspect the ISPs there provide bullet-proof hosting services (no intervention on what are hosted there as long as fees are paid) due to the lack of legislation in these countries [10].

The method is not effective against phishing spam, where the number of IP addresses outnumbers that of domain names. Hacked domains, blog spaces and infected computers can be used to redirect a user to the real phishing site. The reason is probably because phishing sites are vigorously pursued by investigators. On the other hand, the sites that sell drugs and sexual enhancement products are seldom touched, making it unnecessary for spammers to change the hosting IP addresses frequently. They only need to flush out old domain names that are already blacklisted by spam filters. Therefore, blocking hosting IP addresses will still be effective against point-of-sale spam websites and associated spam emails. The pressure from the investigators may push the spammers to use more advanced techniques, such as the FFSN, to protect their hosts. Nevertheless, this will raise the operation cost and reduce the effectiveness of spam.

6. FUTURE WORK

The goal of our spam research is to identify spamming infrastructure that belongs to the same spamming group so that they can be put to an end. Currently we focus on the hosting servers of spam domains. We want to trace a lasting spam campaign and detect any changes in the hosting infrastructure. We also want to detect emerging spam campaigns, such as a new phishing spam. We hope to build IP blacklist and domain blacklist that can be reported to law enforcement agents. Currently, it is done by SQL queries. We hope to automate this process by having a computer program query the database and generate reports.

The tracing of a spam campaign by using hosting IP address may not be sufficient. We will lose the trace of a spam campaign if the identified IP address ceases to work before a new IP is introduced. While the campaign continues by switching to a new IP address, our algorithm will regard it as a new spam campaign. Therefore, it will be useful to collect and analyze more attributes, such as subject, sender's email address, attachment type and sample email text, which can be used to relate a new cluster with an existing one.

The fetched websites are also useful to relate similar spam campaigns and find out what a cluster is really about. The fetching has to be timely when the site is still alive. In our experiment, the data was more than one month old, many websites were no longer alive, thus the content of those websites remains a mystery. However, it may not be necessary to fetch all the domain names in a cluster. A small sample may be sufficient to deduce that the rest of the domains will point to the same website.

7. REFERENCES

- [1] Anderson, D. S., Fleizach, C., Savage, S., & Voelker, G. M. Spamscluster: Characterizing internet scam hosting infrastructure. In *Proc. of 16th USENIX Security Symposium*, 125-148, 2007.
- [2] Bacher, P., Holz, T., Kotter, M. and Wicherski, G. Know your enemy: Tracking botnets. *The Honeynet Project and Research Alliance*, 2005. <http://www.honeynet.org/papers/bots/>
- [3] Cooke, E., Jahanian, F. and McPherson, D. The zombie roundup: Understanding, detecting and disrupting botnets. *Workshop on Steps to Reducing Unwanted Traffic on the Internet*, 39-44, 2005.
- [4] Dietrich, C. and Rossow, C. Empirical research on IP blacklisting. *ISSE 2008 Securing Electronic Business Processes*, 163, 2009.
- [5] Konte, M., Feamster, N. and Jung, J. Dynamics of online scam hosting infrastructure, In *Passive and Active Measurement Conference (PAM)*. 2009.
- [6] Levenshtein, V. I. Binary codes capable of correcting insertion and reversals. *Soviet Physics - Doklady*, 10, 707 – 710, 1966.
- [7] M86 Security. Sex, drugs and software lead spam purchase growth. <http://www.marshall.com/pages/newsitem.asp?article=748>, 2008.
- [8] McAfee Avert Labs. McAfee threats report: second quarter 2009. http://www.mcafee.com/us/local_content/reports/6623rpt_avert_threat_0709.pdf
- [9] Pu, C., and Webb, S. Observed trends in spam construction techniques: A case study of spam evolution. *The 3rd Conference on Email and Anti-Spam*. Mountain View, CA. 2006.
- [10] Qi, M., Wang, Y. and Xu, R. Fighting cybercrime: legislation in China. *International Journal of Electronic Security and Digital Forensics*, 2, (2). 219-227. 2009.
- [11] Ramachandran, A. and Feamster, N. Understanding the network-level behavior of spammers. *The 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. Pisa, Italy, 2006.
- [12] Ramachandran, A., Dagon, D. and Feamster, N. Can DNS-based blacklists keep up with bots. *The 3rd Conference on Email and Anti-Spam*. Mountain View, CA. 2006.
- [13] Ramachandran, A., Feamster, N. and Dagon, D. Revealing botnet membership using DNSBL counter-intelligence. In *Proc. of the second conference on Steps to Reducing Unwanted Traffic on the Internet*. San Jose, CA. 2006.
- [14] Ramachandran, A., Feamster, N. and Vempala, S. Filtering spam with behavioral blacklisting. In *Proc. of the fourteenth ACM Conference on Computer and Communications Security*, Alexandria, VA, 2007.
- [15] Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J. and Zhang, C. An empirical analysis of phishing blacklists. *The Sixth Conference on Email and Anti-Spam*. Mountain View, CA. 2009.
- [16] Spamhaus DBL. <http://www.spamhaus.org/dbl/>
- [17] Spamhaus PBL. <http://www.spamhaus.org/pbl/>
- [18] St Sauver, J. Spam, domain names and registrars. *MAAWG 12th General Meeting*. San Francisco, CA. 2008. <http://www.uoregon.edu/~joe/maawg12/domains-talk.pdf>
- [19] SURBL. <http://www.surbl.org>
- [20] URIBL. <http://www.uribl.com>
- [21] Webb, S., Caverlee, J. and Pu, C. Introducing the Webb Spam Corpus: Using email spam to identify web spam automatically. *The 3rd Conference on Email and Anti-Spam*. Mountain View, CA. 2006.
- [22] Webb, S., Caverlee, J. and Pu, C. Characterizing Web Spam Using Content and HTTP Session Analysis. *The 4th Conference on Email and Anti-Spam*. Mountain View, CA. 2007.
- [23] Webb, S., Caverlee, J. and Pu, C. Predicting web spam with HTTP session information. In *Proc. of the 7th Conference on Information and Knowledge Management (CIKM 2008)*. Napa Valley, CA. 2008.

[24] Wei, C, Sprague, A., Warner, G and Skjellum, A.
Characterization of spam advertised website hosting strategy.
The sixth Conference on Email and Anti-Spam. Mountain
View, CA. 2009.

[25] Zhong, Z., Ramaswamy, L. and Li, K. ALPACAS: A large-
scale privacy-aware collaborative anti-spam system. In *Proc.*
of the 27th International Conference on Computer
Communication (INFOCOM 2008). Phoenix, AZ. 2008.