

Characterization of Spam Advertised Website Hosting Strategy

Chun Wei
University of Alabama at
Birmingham
1300 University Blvd.
Birmingham, AL, USA
weic@cis.uab.edu

Alan Sprague
University of Alabama at
Birmingham
1300 University Blvd.
Birmingham, AL, USA
sprague@cis.uab.edu

Gary Warner
University of Alabama at
Birmingham
1300 University Blvd.
Birmingham, AL, USA
gar@cis.uab.edu

Anthony Skjellum
University of Alabama at
Birmingham
1300 University Blvd.
Birmingham, AL, USA
tony@cis.uab.edu

ABSTRACT

This paper surveys three months of spam data and investigates the hosting strategy of spam domains that are used to sell pharmaceutical, luxury goods and sexual enhancement tools. Thousands of domains have been found and most of them use wildcard DNS records to support non-existing machine names. The hosting IP addresses are much fewer than the number of domains, with a large number of domains hosted on a limited number of hosts. The majority of these heavily-used hosts reside in networks outside the USA. These hosts are stable and have good connectivity and availability. As a result, many domains on these hosts are alive for the entire three-month investigation period. The hosting IP addresses began to move in late March. The new IP addresses, however, are still in the same network range as the old IP addresses. The result suggests that further investigation on spam web hosting will be beneficial in disrupting the spamming network.

1. INTRODUCTION

The reason spam has become popular is that it is profitable and almost risk-free. According to a survey by New Research from Marshal London, 29% of internet users have purchased goods from web sites advertised by spam emails[9]. According to the Microsoft SIR report [8], 49% of all spam emails are advertisements for pharmaceutical products. Researchers at UCSD 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 [6]. However, spam emails are a means to an end. The end is the web site that will complete online transactions with buyers. To establish a spam-advertised site, a spammer needs to find a hosting place and register a domain name. Because of anti-spam efforts, a spam domain is likely to be detected quickly and reported to a blacklist [13]. To maintain the site availability, spammers usually have to register many domain names that will resolve to the same site. If some domains get blocked by spam filters, then new ones will be added. The hosting servers should have good connection and high availability just as a regular online store website and are unlikely to reside on a bot. Therefore, we predict the hosts should be on more stable networks and are probably owned and maintained by spammers.

2. RELATED WORK

So far many spam researches have been focused on spam messages, such as spam filtering and categorization [3, 7] and its

distribution channel— the botnets [2, 4], but much less effort on the hosting and C&C servers that are actually owned by spammers. Compared to bots, we believe the hosting structure should be more stable and thus easier to trace. The Spamsscatter project [1] probed the linkage in spam emails and clustered the destination web sites to explore the scam hosting infrastructure. They found most scam hosts were steady and had high availability and good connectivity. Our research both in previous work and in this paper is focused on targeting the origin by analyzing the spam emails. We began with clustering the email headers and subjects [10], but that revealed little about where emails came from. Tracking the sender's IP addresses can lead to the botnets, but the botnets are actually intended to protect the spammers from being traced [11]. To trace back to the C&C server, one needs to analyze the communication traffic of a bot and any suspicious computer on the Internet. After spammers adopted Peer-to-Peer C&C infrastructure [5], the tracking became more difficult. Therefore, we started to explore the derived information from the spam-email-embedded URLs, such as web hosts information, domain registrar information and web content. They provide more useful clues than the email itself in tracing the spammers. This paper shows the result of our first attempt to analyze the spam hosting infrastructure in a systematic way. Our findings support our prediction that spam-advertised websites selling products and services are less likely to be hosted on botnets, which are more transient. In contrast to the Spamsscatter findings, we discovered many significant hosts (hosting vast number of spam domains) reside outside the US and more than half of the domains are hosted in more than 1 physical IP address.

3. METHODOLOGY

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. The data consist of approximately 1.2 million email messages received from January 2009 to March 2009.

3.1 Extracting Domain Name

We extract URLs from email messages and then the domain name portion of the URL. During the process, 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 machine names having a matched domain suffix [12]. It is specified by using a “*” as the left most part of a machine 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.

Therefore, we create our own phantom machine by attaching a random string to a domain name under test. If the phantom machine can be resolved, it proves to be using wildcard DNS record. Then we believe others machine names end with the same domain name would resolve to the same site. This strategy greatly reduces the number of machines that need to be fetched.

3.2 Clustering Domains on IP Group

The Unix “dig” command is used to check hosting IP information for domains. There is a many-to-many relationship between domain and IP. Once the IP addresses of the domains are fetched, we cluster the domains on a daily basis depending on the IP group they belong to. An IP group is one or several IP addresses that simultaneously host the same domain. Since many domains have the same IP combination in DNS record, the clustering result reveals the most heavily hosted IP groups, which is interesting to spam investigators. We are also able to trace a particular IP group by monitoring the domains hosted there each day.

4. RESULTS

Over 95% of the spam in our dataset contained URLs, among which we found 42,703 domains that were using wildcard DNS entries and which accounted for 1,050,180 host machine names appearing in the email messages.

4.1 Top Hosting Networks

There are over 4000 hosting IP addresses found in the three-month period of time. The hosting IP addresses are distributed among 1391 network blocks residing in 61 countries. Table 1 shows the Top 6 network organizations that host the most number of domains (NW is abbreviation for Network). Note if a domain is hosted on multiple network blocks, so it will be counted multiple times in this table. 5 of them are in China and the other one is in Russia.

Table 1: Top 10 hosting organizations

Organization	Network block	IP count	Domain count
China Unicom Hunan NW	220.248.160.0/19	6	22343
	110.52.0.0/15	3	2991
China Jinhua Telecom Co.Ltd	60.191.192.0/18	4	10454
China Qingdao Cnecgroup NW	203.93.208.0/21	1	5948
China United Telecom Corp.	211.91.224.0/20	1	5268
CHINANET Chongqing NW	219.152.0.0/17	1	2104
	222.176.0.0/14	1	977
Russian Federation Moscow	87.242.64.0/18	2	1395

4.2 IP Shifting

Starting from around Mar 18, some old IP groups began to stop receiving new domains and corresponding new IP groups started to emerge for the same spam cluster. The chart (Figure 1) shows

the number of domains on two IP groups per day from Jan 01 to March 31, 2009. Both IP groups are found to host counterfeit “Canadian Pharmacy” websites. Therefore, the spammer apparently moved the hosting IP addresses by replacing old IP addresses with new ones in the nameserver entries. We then checked the domains on old IP groups and found they also changed to the same IP groups. However, the new IP addresses are still in the same network range.

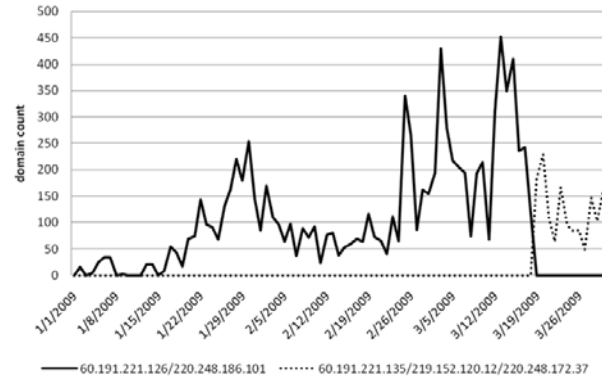


Figure 1: Websites Domain count per day on two IP groups hosting "Canadian Pharmacy" websites, Jan-Mar 2009

5. CONCLUSIONS AND FUTURE WORK

This paper analyzed the IP hosting of web links appearing in spam emails. Many spam domains are found to be using wildcard DNS records, resulting in large number of phantom machines in spam emails. The results confirmed our speculation that hosting IPs are more stable and easier to trace than botnets. The shifting of hosting IP addresses does occur but much less frequent than the rate of domains. Most domains showed up in our spam email collection for only one day and ceased to appear again, even though many of them are still alive. We suspect the cause is a counter-strategy of spammers against domain blacklisting.

Although the USA has the highest number of hosting IP addresses, the most heavily used IP addresses reside outside USA, beyond the reach of US jurisdiction. Some IP addresses in China are found to host more than 8000 real domains in a period of three months. The number of domains created for spam is astonishing, suggesting strong organized criminal behavior.

In the future, we want to make several improvements. We need to check for redirection URLs and record the destination domains and their corresponding IP addresses. We also need to check the domains more regularly to see if the hosting IP addresses change. We suspect the spammers have multiple IP addresses in several different networks and shift the hosts back and forth from time to time. Tracking the nameserver of the hosting IPs should also be useful because it is where the spammers update the DNS records. The registrar information may also be useful in relating different spam domains. Our goal is to find all the hosts controlled by the same spammers, what kinds of websites are hosted there, where are the nameservers. The results should catch the attention of law enforcement and effort will be taken to put down the spam hosts. The spam emails will become useless as hosts are gone. We believe it will be an effective way to reduce spam eventually.

6. REFERENCES

- [1] Anderson, D. S., Fleizach, C., Savage, S., & Voelker, G. M. (2007). Spamscatter: Characterizing internet scam hosting infrastructure. *In Proc. of 16th USENIX Security Symposium*, pp.125-148.
- [2] Bacher, P., Holz, T., Kotter, M. and Wicherski, G (2005). "Know your enemy: Tracking botnets". *The Honeynet Project and Research Alliance*, <http://www.honeynet.org>
- [3] Calais, P. H., Pires, D. E. V., Guedes, D. O., Meira, W. Jr., Hoepers, C. and Klaus, S. (2007). A campaign-based characterization of spamming strategies. *In Proc. of the Fifth Conference on Email and Anti-Spam*.
- [4] Cooke, E., Jahanian, F. and McPherson, D. (2005). The zombie roundup: Understanding, detecting and disrupting botnets. *Workshop on Steps to Reducing Unwanted Traffic on the Internet*, pp. 39-44.
- [5] Grizzard, J., Sharma, V. and Dagon, D. (2007). Peer-to-peer botnets: overview and case study. *HotBots '07: Workshop on Hot Topics in Understanding Botnets*.
- [6] Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G., Paxson, V. and Savage, S. (2008). Spamalytics: An empirical analysis of spam marketing conversion. *In Proc. of 15th ACM Conference on Computer and Communication Security*.
- [7] Li, F. and Hsieh, M. H. (2006). An empirical study of clustering behavior of spammers and group-based anti-spam strategies. *In Proc. of the Third Conference on Email and Anti-Spam*.
- [8] Microsoft Security Intelligence Report, Volume 6, 2H08. <http://www.microsoft.com/sir/>
- [9] Sex, Drugs and Software Lead Spam Purchase Growth <http://www.marshall.com/pages/newsitem.asp?article=748>
- [10] Wei, C., Sprague, A., Warner, G. and Skjellum, A. (2008). Mining spam email to identify common origins for forensic application. *In Proc. of 2008 ACM Symposium on Applied Computing*, pp. 1433-1437.
- [11] Wei, C., Sprague, A. and Warner, G. (2008). Detection of Network Blocks Used by the Storm Worm Botnet. *In Proc. of 46th ACM Southeast Conference*.
- [12] Wildcard DNS Record: http://en.wikipedia.org/wiki/Wildcard_DNS_record
- [13] Zhang, J., Porras, P. and Ullrich, J. (2008). Highly Predictive Blacklisting. *In Proc. of the 17 Conference on Security Symposium*. pp. 107-122.