Peeking into Spammer Behavior from a Unique Vantage Point

Abhinav Pathak[1], Y. Charlie Hu[1], Z. Morley Mao[2]
[1]Purdue University and [2]University of Michigan

ABSTRACT

Understanding the spammer behavior is a critical step in the long-lasting battle against email spams. Previous studies have focused on setting up honeypots or email sinkholes containing destination mailboxes for spam collection. A spam trace collected this way offers the limited viewpoint from a single organizational domain and hence is short of reflecting the global behavior of spammers. In this paper, we present a spam analysis study using sinkholes based on open relays. A relay sinkhole offers a unique vantage point in spam collection: it has the broader view of spam originated from multiple spam origins destined to mailboxes belonging to multiple organizational domains.

The trace collected using this methodology opens the door to study spammer behaviors that were difficult to do using spam collected from a single organization. Seeing the aggregate behavior of spammers allows us to systematically separate High-Volume Spammers (HVS, e.g. direct spammers) from Low-Volume Spammers (LVS, e.g. low-volume bots in a botnet). Such a separation in turn gives rise to the notion of "spam campaigns", which reveals how LVS appear to coordinate with each other to share the spamming workload among themselves. A detailed spam campaign analysis holds the promise of finally reverse engineering the workload distribution strategies by the LVS coordinator.

1. INTRODUCTION

The battle against unsolicited emails, or spam, has been on-ongoing for over a decade, with both spammers and filter providers developing increasingly sophisticated solutions [14]. As with any battle, increasing our understanding of the enemy, i.e., the spammer behavior, plays a critical role in the long-lasting battle against spam as it directly assists in the development of counter-measures that target or exploit the weakness of the spammers.

Towards this goal, a natural and effective approach is to set up honeypots or mail sinkholes to attract a large amount of spam and perform off line analysis. Many studies [24, 1, 25, 13, 32] have pursued this approach and made progress towards revealing several aspects of spammer behavior. For example, Ramachandran and Feamster [24] used data from mail sinkholes from a few domains to study the network level properties of spammers [24] and Anderson et al. [1] used data from a single domain sinkhole to study the properties of scam infrastructures.

However, the spam traces collected by the existing approaches analyzing spammer behavior based on honeypots or email sinkholes offer the limited viewpoint from a single organizational domain at a time. Hence, they can not be easily used to retrieve and analyze the global behavior of spammers which typically spam far more than just a single organization.

In this paper, we present spam analysis based on spam data collection at a mail relay sinkhole that overcomes the above limitation of conventional spam sinkholes. In particular, we use open relays (also known as "proxy pots") as a form of sinkholes to attract and collect spam. Such an open relay sinkhole offers a unique vantage point in spam collection: it has the broader view of spam originated from multiple spam origins going to mailboxes belonging to multiple organizational domains.

The trace collected using this methodology opens the door to study spammer behaviors that are difficult to do using spam collected from conventional sinkholes which mimic individual organizational domains. Using a spam trace collected using an open relay over a period of three months consisting of 40 million spam deliveries originating from about 200,000 unique IP addresses destined to 24 million mailboxes, we present several case studies of these spammer behaviors. We identify two classes of spamming hosts based our observation of the data from the sinkhole. The first set consists of dedicated spam sources, which are brute force spammers, each spamming in an enormous number every day. We call this set High-Volume Spammers (HVS). The second set consists of a large number of hosts (mostly compromised machines) working under a central provision, each typically spamming with a low volume. We call the second set Low-Volume Spammers (LVS). Due to the sheer number of LVS, spam due to them amount to a major percentage of the total spam worldwide [33, 28]. But the "stealth" spamming behavior of individual hosts makes them much harder enemies to identify and defeat. Open relay data offers a much broader view of the aggregate behavior of spammers which allows us to separate HVS from LVS with a much

higher confidence level than when observing spammers from a single domain.

Second, the separation of LVS from HVS reveals many global properties of LVS that enhance our understanding of their coordination and workload distribution. In particular, it exposes the clear notion of "spam campaigns" used by LVS, which reveals how hosts appear to coordinate with each other to share the spamming workload among themselves. A detailed spam campaign analysis which is a difficult problem on its own holds the promise of finally reverse engineering the workload distribution strategies by the coordinator of LVS.

The main contributions of the paper are: (1) we describe in detail the methodology of spam collection through open relays, (2) we present a methodology to separate the two major spamming sources: HVS from LVS, and (3) we draw several implications of such a separation which points to a promising direction to study the internal workload distribution among LVS hosts.

2. SPAMMING ORIGINS

There are primarily two types of spammers in the Internet: A *direct spammer* is a dedicated host that leases a connection from a "spam-friendly" ISP [24] and spams continuously. Such spammers make repeated connections to a mailserver to deliver spam to different mailboxes at the domain. Hosts involved in this kind of spamming are frequently seen to spam a particular domain or several domains. A *botnet* consists of a large number of compromised hosts, called bots, to carry out spamming activities on its behalf [6, 24, 31, 33]. Each machine in the botnet typically sends only a few spam[1] to a domain every day to avoid detection. The bots are coordinated by a botmaster, who owns and operates the bot army. Bots amount for a high percentage of spam in the Internet. Several studies [33, 28] have reported about 85 - 95% of the Internet spam are generated by bots.

3. SPAM COLLECTION USING OPEN RELAYS

In this section, we first present our new spam collection methodology using open relays. We then give detailed statistics of our spam collection.

3.1 Open Relay Sinkholes

Open relay provides a unique vantage point for observing Internet spam traffic. Since spammers typically spam mailboxes in many organizational domains, a conventional sinkhole which pretends to be a normal mail server at an organization only observes the spam traffic to that single organizational domain. Such a sinkhole therefore only observes a portion of the spam originated from the spammers. In contrast, a spam sinkhole that masquerades as a normal open relay has a much broader view point of the spam traffic. Figure 1 depicts the position of an open relay in the spam cycle. An open relay on one side sees a plethora of origin spammers that attempt to relay mails through it and on the other side sees all the final destinations of the mails. Such a broader view point of the spam traffic potentially reveals the global behavior of spammers. For example, in the case of HVS, it allows us to study how the spammers schedule spams destined to different destination domains, and in the case of LVS, it potentially allows us to analyze the coordination of hundreds of thousands of LVS in spamming all the destination mailboxes and domains [23, 30].

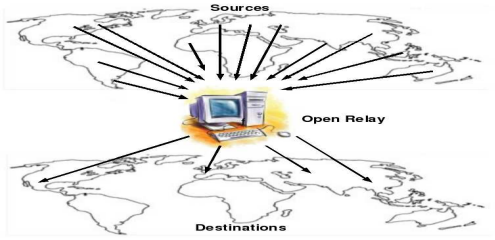

Figure 1: Position of open relay in the Internet.

We note a potential limitation of open relay sinkholes is that an individual open relay may not capture all the spam traffic going to a domain, as a HVS or LVS army may employ multiple open relays or directly spam destination mailboxes. The trace collected at our relay sinkhole effectively provides a sampling of the spam traffic from multiple spam origins to multiple destination domains.

Open Relay Scanners. Spammers use relay testing softwares [26] to scan the Internet for open relays that could be exploited by them for spamming. To detect open relays, they first scan the hosts that have mail servers running on port 25 (SMTP). The hosts that are detected to accept port 25 connections are then checked for relay. A spammer tries to relay a test mail to its own email address through the detected host. Typically the subject or the body of such a mail contains the IP address of the host being tested. Once the test mail is successfully received, the IP address of the host is extracted from the body and the host is confirmed to relay mails.

Below is one such testing mail that we intercepted at our open relay.

```
From s2ui0d5g4b0d1@yahoo.com
    Wed Dec  5 00:55:41 2007
Return-Path: <s2ui0d5g4b0d1@yahoo.com>
Received: from --XX-- (219.84.177.81)
    by --XX-- with SMTP;
    for service168tw@yahoo.com.tw;
    Wed, 05 Dec 2007 00:55:40 -0500 (EST)
    (envelope-from s2ui0d5g4b0d1@yahoo.com)
X-Avenger: version=0.7.7; receiver=--XX--;
    client-ip=219.84.177.81;
Subject: Super webscan open relay check
    succeded, hostname = --XX--
```

In this example, the spammer tries to deliver a mail to an email account, service168tw@yahoo.com.tw. The subject of

[1]We note researchers have observed in some recent botnets, such as the Storm Worm Botnet, each bot spams in relatively high volume.

the mail contains IP address of our host (anonymized as "–XX–"). Upon receiving such a mail, the spammer confirms the detection of an open relay at our IP address.

How to sustain spam at open relays. Once an open relay is detected, multiple spammers start exploiting the host to relay spam through it. The relay testers periodically (about once a week observed by our relay) checks whether the hosts is still relaying the mail using the technique above. We observed that if the host stops responding to relay testers at any time, spamming through the relay is stopped within a few days.

To sustain spam collection through the relay without actually relaying all the spam mails to the final destination mailboxes, which can result in our open relay soon be blacklisted by DNSBLs, we carefully configured our open relay to only relay the mails that are doing the relay testing. In this way, the relay testers are given continuous false assurance that the relay continuous to relay all the mails whereas in reality only the testing mails are relayed and all others are stored and not forwarded.

An important step here is to identify which mails are for testing the relays and which are actual spam messages. Most of the relay testers could be trivially identified as they contained the IP address of our relay server in either the mail body or in the subject lines. Some of them also contained words like "relay", "test", "successful", etc. So any mail that contained either the relay's IP address or these keywords were let through. An important point to note here is that relay testing done by many DNSBL(s), for blacklisting purposes, also contain the IP address of our relay in their mail bodies. We denied any mail that contained words like "dnsbl", "ordb", "sorbs", etc. from passing through. We note that the relay tester behavior is based on observations by our relay and hence our mechanism for detecting relay testers is not necessarily general[2].

3.2 Data Collection

We set up an open relay by configuring the Mail Avenger MTA [16] to selectively relay mails, i.e., only relaying relay testing mails, and store all the through traffic, as described in Section 3.1. In addition to logging the mail body sent in each connection, we also configured Mail avenger to record various information about the connecting hosts such as TCP SYN fingerprints from which we can derive the OS running at the spammer hosts, DNSBL status of the spammer IP in five blacklists (cbl [8], sbl-xbl [29], dsbl [11], dnsbl.sorbs [10] and spamcop [5]), and traceroute to the spammer host at the time of receiving mails.

Using our relay set up, we collected spam traffic starting October 1, 2007 for three months. All the mails received by our open relays were spams, as all the mails received were to be relayed while legitimate mail servers do not use mail relays without authorization. Table 1 gives a summary of the three-month spam collection at the relay.

[2]And it is likely that the spammers will mutate their testing methods after reading this paper.

Table 1: Trace statistics.

Collected at a relay sinkhole Oct-Dec 2007.	in Millions
Number of outgoing mails asked to relay:	39.7
Number of SMTP connections:	2.3
Number of unique IP addresses:	0.19
Number of unique recipients:	24.7
Number of destination domains:	0.27

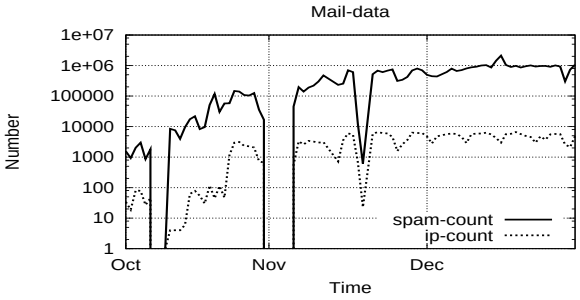

Figure 2: Number of mails supposed to be relayed and number of unique spammer IPs per day.

Source statistics. Figure 2 shows the number of mails that our relay received and the number of unique IP addresses that made the SMTP connections every day during the three-month period. We see that the spam through the relay was initially low in volume but later ramped up. Once the relay became popular, the number of spam remained constant at about one million per day originated from a few thousands of IP addresses. We notice three drop points in spam numbers. While the drop in mid November was due to maintenance at our mail relay, the first two drops seems to be due to decisions taken at the spammers' end. The source IPs of the spammers connecting to our relay fall into many regions of the IP address space. Figure 3 shows the CDF of spammer distribution across the IP address space. The spammers originated from 150 countries with a majority of them situated in India, Argentina, Brazil and China.

Destination statistics. The mails that we received had mail addresses corresponding to about 264,000 unique domains. Figure 4 plots the number of mails that were destined to each domain, in increasing order. We see that 10294 domains received more than 100 spam. The four domains that received the most spam include hinet.net, yahoo.com.tw, msn.com, and gmail.com.

4. SEPARATING SPAM ORIGINS

From the sinkhole data we observe the prevalence of two sets of spamming hosts. The first set contains a large number of hosts that spam in low volume which appear to be highly coordinated, and the second set contains a small number of hosts that spam in high volume and do not appear to be coordinated. Based on these characteristics, we term the first set of hosts as Low-Volume Spammers (LVS) and the second set of hosts as High-Volume Spammers (HVS). We conjecture

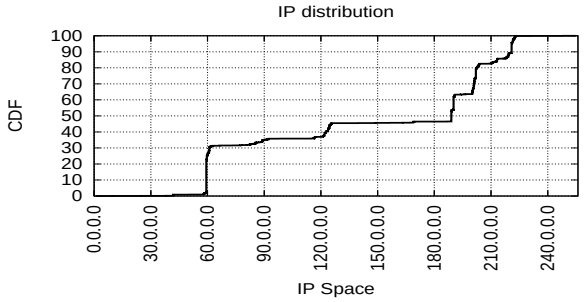

Figure 3: **CDF across IP address space of spammer IP addresses that originated mail to our relay.**

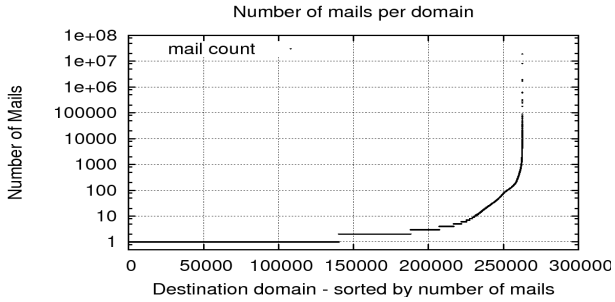

Figure 4: **Mail count to different destinations, sorted by the mail count per destination.**

that HVS observed in our trace correspond to direct spammers, and LVS could correspond to the bots in some botnets where the bots send low-volume spam.

In this section, we show how the trace collected at the relay sinkhole enables us to identify and isolate the spam origins into these two sets. Separating the two types of spammers enables us to perform further in-depth analysis of either type of spammers' spamming behavior.

Our separation technique is based on the following observation. Once a HVS discovered our open relay, it is likely to divert a significant fraction of its spamming traffic to the relay, disregarding the number of destination domains such traffic are destined to. On the other hand, each individual host in LVS using our open relay is likely to send the usual, low volume traffic to the relay. Such different spamming behaviors between HVS and LVS are much more easily observed by our relay than by a conventional sinkhole which only observes the spam traffic from a spammer to individual domains.

Figure 5 shows the number of times each IP address made a connection to our relay to deliver spam mails. The graph is sorted by the number of connections made by each IP. We see that most of the hosts made few connections to our relay during the three-month period. About 25% of the hosts connected to our relay just once, and more than 75% connected fewer than 10 times. About 0.9% (nearly 1700) of the hosts made more than 100 connections (more than 3 per day on average) and were responsible for about 59% of the total spam. About 0.1% (nearly 190) of the hosts made more than 1000

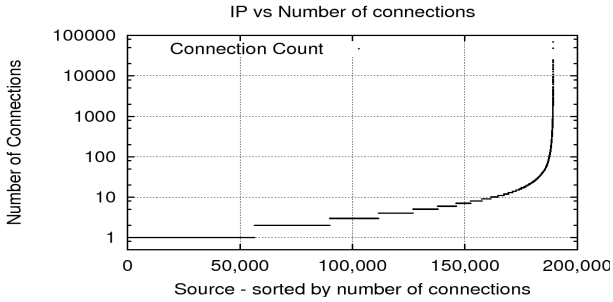

Figure 5: **Number of connections made by each unique spammer IP, sorted by the number of connections made.**

connections during the same duration which were responsible for 43% of the spam. We see that there are two distinguishing sets of hosts originating spam to the relay. The first set of hosts keep a low profile by sending only a few spam each, whereas the second set of hosts send a large volume of spam each. Based on these observations, we conjecture that the first set of hosts are part of botnet, whereas the second set of hosts are dedicated spam servers. In the following, we show how to derive some heuristics to separate these two sets robustly.

Why spammers use open relays for spamming. One interesting question that arises in our spam collection is why spammers use open relays for spamming? During our spam collection, for every host that connected to our relay, we performed DNSBL lookups for its IP to five popular IP based blacklists. We found that 75% of the hosts were already blacklisted in at least one of the five DNSBLs we queried at the time of receiving the spam (51% were blacklisted in at least 2 blacklists, and 1.5% were blacklisted in all the 5 blacklists). We speculate that there are two reasons that spammers use relays. First, the hosts that were already blacklisted because of their previous spamming activity use relays so their spam will not be filtered by the destination mail servers that use DNSBLs. Second, the hosts that were not blacklisted yet use relays to hide their identities to reduce their chance of getting blacklisted.

4.1 The Notion of Chunks

Before we present the heuristics for separating LVS and HVS, we introduce the notion of "chunks" which is used to assist the analysis of the spam collected by the relay.

The notion of chunks is motivated by observing the granularity of mailboxes that LVS appear to spam at. Individual members in the LVS set appear to be coordinated. The Coordinator maintains a list of recipients to spam. It breaks down the list and assigns each member a part of the list. After receiving the mail text and its part of the recipient list, each host in LVS starts spamming the end hosts. The Coordinator also provides each LVS with information about which open relays/proxies to use. Anecdotal evidence [3, 4] and our analysis of the spam collected at our mail relay suggest that the Coordinator appears to sort the list of recipients al-

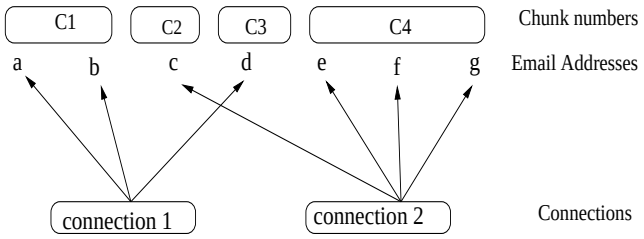

Figure 6: Defining chunks from alphabetical sorting of email addresses

phabetically, then break the list down into small segments, and finally distribute the segments to the individual LVS as individual LVS continuously request for them. The individual LVS typically spam consecutive recipient mailboxes in a received segment using a single SMTP session with multiple `RCPT TO`.

However, the trace collected at our relay sinkhole contains spam from both HVS and LVS which are potentially intermixed in the list of all the recipients sorted alphabetically. We now define the notion of chunks in this mixed list and show later how it is used to assist the separation of the two types of spammers as well as the validation of the separation.

Given the alphabetically sorted list of spam collected by our relay sinkhole, we define a chunk as a set of consecutive recipients in this list that were delivered the same spam in one connection by a single spammer (source IP). Note a spammer may have originated several chunks that are separated apart in the list. For example, a host may spam to mail addresses starting with "a" and later to mail addresses starting with "c", while another spams mail addresses starting with "b" and "d".

Figure 6 illustrates an example of how we define chunks and assign chunk numbers. Suppose "connection 1" from an end host delivers mails to recipients with email ids "a", "b" and "d". These recipients would get the same mail text as the mail was delivered in one connection. Now if "connection 2" from another (or may be the same) end host delivers mails to recipients with email ids "c", "e", "f" and "g", after sorting the list of recipients we define chunk "c1" to be consisting of recipients "a" and "b", chunk "c2" consisting of recipient "c", and so on. Figure 6 shows the sorted list of recipients and the corresponding chunk numbers. After applying the chunk definition to our trace, we find that a typical chunk contains about 5 to 10 recipients.

4.2 Separating Heuristics

Using the definition of chunks, we now present a few heuristics for separating HVS and LVS. These are based on the fundamental observation that HVS spam in high volume, whereas individual LVS usually spam in low volume to evade detection. An additional observation exploited by our heuristics is that LVS act under a common coordinator, which implies that they are coordinated to share the workload among themselves, whereas HVS are not coordinated.

Connection Count. LVS differ from HVS as they try to evade detection and spam in low volume. From our trace we see that LVS generally tend to make one or two connections per day to an end host to deliver spam. So in the three-month time frame, an LVS would make on the order of 100 connections. This leads to the first heuristic we use for separating these two sets of spammers, i.e., by using a cutoff threshold on the number of connections a spammer has made to our relay. Figure 5 shows a cutoff at 100 connections would classify 99.1% of the spammers as LVS.

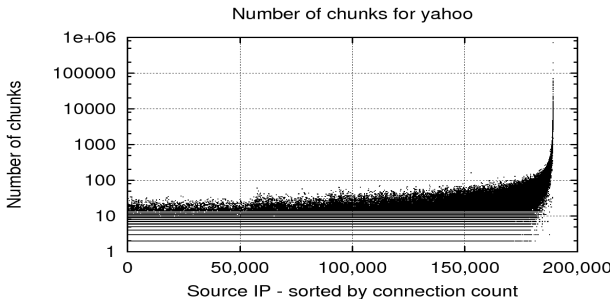

Figure 7: Number of chunks by each unique spammer IP, sorted by the number of connections made by host.

Number of Chunks. Our second heuristics is based on the number of chunks delivered to a single large domain by individual spammers[3]. For chunk-based analysis hereafter, we use only the mails delivered to Yahoo. There are about 120,000 IP addresses that spammed this domain. Figure 7 plots the number of chunks each spamming source spammed Yahoo, sorted by the total number of connections made by each source (as in Figure 5). We observe that most of the sources (more than 95%) deliver less than 100 chunks. Also, 387 sources (< 0.5%) deliver more than 1000 chunks.

Average Chunk Gap. We define the Inter Chunk Gap as the number of recipients between two consecutive chunks originated by same spammer in the sorted list of recipients collected at the relay. If a host originates more than one chunk in the sorted list, we define the Average Chunk Gap (ACG) as the average inter chunk gap between all of the adjacent pairs of chunks originated by that host in the list. For example, in Figure 6, either of the two connections has two chunks, with an ACG value of one. For spammers that originate only one chunk, we define the average chunk gap for that spammer as 1. Most LVS deliver only one chunk. An LVS that delivers multiple chunks usually has the chunks spread apart in the alphabetical listing; the chunks could be requested from the Coordinator at different time. This implies that ACG for LVS will be usually high. In contrast, HVS do not usually spam in chunks, i.e., they tend to spam randomly chosen mailboxes (i.e., not consecutive) in their list of mailboxes using the same connection. In addi-

[3]In this paper, for simplicity, we analyze only spam destined to Yahoo. All plots hereafter concerning chunks will depict hosts that spam this domain only. Other hosts are removed. We leave analysis of correlating spam to different domains as future work.

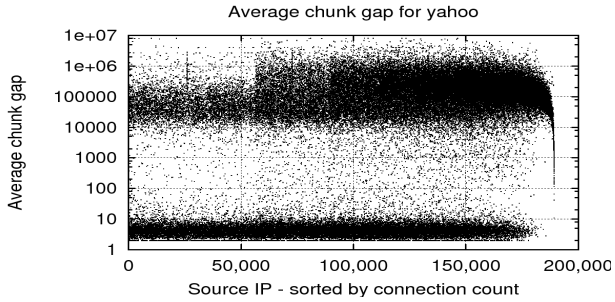

Figure 8: Average chunk gap for individual spamming hosts sorted by number of connections made by each host.

tion, HVS typically spam in large numbers. As a result, the chunks from their spam are often small and inter-mixed with those due to LVS in the total sorted recipient list. Therefore their average inter-chunk gaps are expected to be low.

Figure 8 plots the ACG for all the hosts that spammed the Yahoo domain, sorted by their connection counts. We observe that hosts in the initial part of the graph fall in two categories - one that has low ACG and one that has exceptionally high ACG. Towards the heavy spamming zone, i.e., far right on the graph, we see a sharp decline in the ACG value. This indicates that hosts with high numbers of connections tend to have low ACG.

4.3 Separation Rules

Based on the three heuristics that we have defined, we now present the *separation rules* for separating LVS and HVS. We define a separation rule (SR) as a tuple (cc, nc, acg) that separates the hosts into two sets, LVS and HVS, as follows

```
Separation Rule: (cc, nc, acg)
foreach spamming host h in the relay trace
   if( connection count by h > cc &&
       number of chunks by h > nc &&
       ACG of h < acg)
         h is a HVS;
   else
         h is a LVS;
```

The SR tuple (cc, nc, acg) splits the spamming hosts into LVS and HVS. The values of cc, nc and acg determine the degree and effectiveness of the separation. A good tuple is decided based on heuristics. Table 2 gives several separation rules, the corresponding number of hosts that qualify as HVS due to the cut offs, and the percentage of mails generated by them. We see that as the cutoff values are lowered, the number of HVS classified increases rapidly but the percentage contribution of spam generated by them do not increase so profusely.

Selecting cutoff thresholds. The choice of cutoff thresholds used in the separation rule determines the accuracy of the separation. We devise a method that provides feedback on the separation quality for different cutoff thresholds. Since HVS spam a large number of recipients randomly through the list, the recipients they generate are likely to be inter-mixed with that from LVS in the total sorted recipient list,

Table 2: Exploring separation rules parameters.

No.	nc	cc	acg	# ds	ds mail %
Rule 0	0	0	0	0	0
Rule 1	5000	5000	10000	39	36.25
Rule 2	1000	1000	10000	183	49.19
Rule 3	180	180	10000	475	55.65
Rule 4	50	50	10000	477	55.66

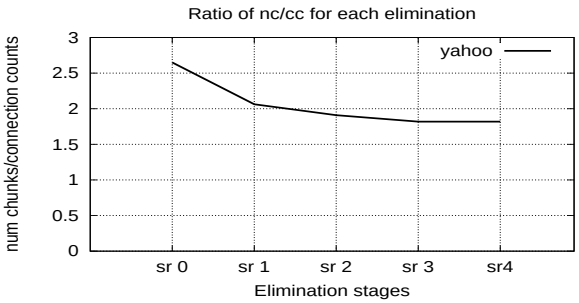

Figure 9: Ratio of the number of chunks over the connection count for various separation rules.

and hence inflate the number of chunks originated by each host. After removing HVS, we expect the average number of chunks (NC) delivered by a LVS in an SMTP connection to converge to some constant. Hence we can use the observation of such a convergence as a validation that we have removed all or most HVS. Figure 9 plots the ratio of the total number of chunks delivered by all LVS combined to the total number of connections by the LVS, *after* applying separation rules as described in Table 2 and removing the chunks due to the identified HVS. We see that the ratio begins at 2.7 under Rule 0, which means no separation, and gradually drops down to about 1.8 for subsequent rules. After Rule 3, the ratio remains nearly steady for the subsequent separation rule, indicating we have achieved the separation and further lowering the cutoff values may classify aggressive LVS as HVS. Deciding the cut off rules in the algorithm depends on the frequency of spam and the domains being spammed to, and is currently done manually. We plan to study automating this process in our future work.

Validation. Though we do not have ground truth to validate our separation results, we tried to verify the blacklisting status of the HVS identified. We found most of the HVS found by our algorithm were blacklisted as open proxies in an open proxy database [19]. Out of the 39 HVS identified by Rule 1, 30 were blacklisted as open proxies. Out of the 477 HVS identified by Rule 4, 349 were blacklisted as open proxies. The increased number of identified HVS blacklisted as open proxies suggests Rule 4 is a more accurate separation rule than Rule 1.

Limitation. Our separation heuristic described above may be affected by the presence of NATs in the Internet. A NAT may have many LVS behind it, and hence its connection count and number of chunks delivered can be very high. Such a NAT has a high probability of being classified as

HVS. We plan to address this issue in our future work [7].

5. SEPARATION IMPLICATIONS

Separating HVS from LVS have many implications. Separating the two types of spammers enables us to perform further in-depth analysis of either types' spamming behavior. Spam from HVS can be easily mitigated by using a DNS based blacklist. Blacklisting them would result in stopping about 50-60% of spam reaching user inboxes (see Table 2 for spam contributed by HVS identified). For the remaining 40-50% of spam for which LVS are responsible, using a DNSBL might not be effective as a large number of hosts are involved. Separation of LVS and HVS allows us to study LVS behaviors in isolation, for example, their workload sharing model and the dynamics of their aggregate spamming behavior such as spam campaigns. In the following, we discuss several interesting LVS spamming behavior observed in the trace collected at our relay sinkhole.

5.1 Strength: Destination repetition

How much time do LVS take to spam the same email addresses twice? Answers to this question have implications to estimating the strength of the LVS in terms of its size. We define the interarrival time for a destination mailbox as the time between when the relay received two consecutive spam mails (from the same source or different sources) destined to the destination mailbox. For destinations that receive only a single spam in the three-month period, we assign the interarrival time as more than 90 days. Figure 10 plots the CDF of the average interarrival time for destination email addresses. We make three observations from the plot. First, about 55% of destinations through our relay received only a single spam in the three-month period, i.e., CDF stops at 45%. We note that these recipients could have received other spam through other relays/open proxies. Second, most of the remaining recipients received more than one spam within a period of 30 days. Third, few recipients (nearly 6%) received spam twice almost instantaneously (in less than 10 minutes). We observe that these recipients received spam from two different sources (mostly spread apart in location). We conjecture that this could arise due to the loose synchronization among LVS as two LVS may take up the same job of delivering the same spam to the same chunk. A detailed investigation is left as future work.

5.2 Dynamics: Spam Campaigns

Separating HVS from LVS also exposes interesting dynamics of LVS spamming behavior. Figure 11(a) plots a graph of chunk vs. time for spam from both HVS and LVS, during a week-long period starting December 16, 2007. For each chunk in the trace, we assign a chunk number that is equal to the number of chunks preceding it in the sorted list of recipients (or chunks). We then plot the chunk number and the time the chunk was delivered to the relay.

Figure 11(b) plots the same chunk vs. time graph but after removing all the spam from HVS identified using the separation rule 4 and reassigning consecutive chunk numbers. We observe that the spam from HVS contribute to the vertical lines in Figure 11(a), i.e., reflecting that they spam in a brute force manner during particular time periods, and they do not work in much coordination as LVS do. After removing these vertical lines due to HVS, we observe many streaks of slanted lines made up of small squares. This interesting pattern of slanted lines, which we call "streaks", which begin in lower chunk numbers and ends in higher chunk numbers, reflects the dynamics of spamming by LVS. Our preliminary, manual, investigation of mails belonging to a few of these individual streaks shows that nearly all the mails in them bear a common aim. This could be in the form of a common URL embedded[4] in all the mails belonging to one such streak, or some common medical treatment giving the same Skype/email address for consultation, and so on. The spam that contain one common aim (e.g. URL, skype id) could belong to many such streaks. We term the collection of all the streaks that have the same common aim as a "spam campaign".

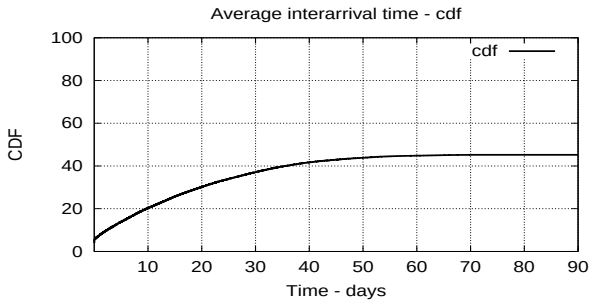

Figure 10: Interarrival time for destinations spammed through LVS.

Streaks in "spam campaigns" reveal how LVS appear to coordinate with each other to share the spamming workload among themselves. A detailed spam campaign analysis which is a difficult problem on its own holds the promise of reverse engineering the workload distribution strategies by the coordinator.

6. RELATED WORK

The phenomenal increase of email spam in the recent past has generated much interest from the research community. Much effort have been put into developing mitigation schemes [12, 17, 20, 27, 18]. Closely related to developing mitigation schemes are the numerous studies that aim to understand spammer behaviors. Several studies have used email sinkholes to study spammer properties. In [24], the authors used a mail sinkhole to analyze the network-level properties of spammers, such as their geographical and network distribution. They also documented that some spammers used prefix hijacking for spamming which they termed as BGP spectrum agility. Anderson et al. in [1] used a sinkhole to

[4]The final URL after all redirections.

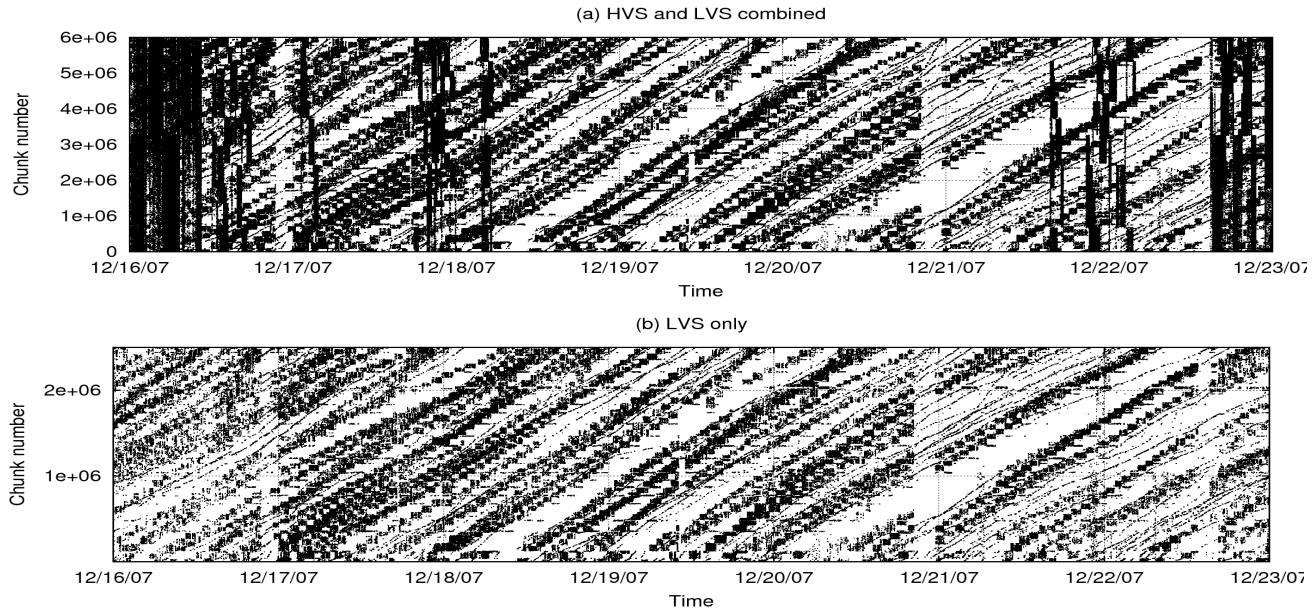

Figure 11: Chunk vs time, before and after removing spam due to HVS. Note: chunk numbers are recalculated in (b) after removing chunks from HVS

study the scam infrastructure that spammers use. They used the notion of spam campaigns as mails that correspond to a particular scam. In [9, 13, 32], the authors also used a well established domain to study spam and spammer behavior. In [21, 22], the authors set up experiments to trap bots, i.e., the spamming sources, for behavior analysis. In [25], the authors used traces from 115 domains to develop a new blacklisting technique, called behavioral blacklisting, that takes into account the spamming behavior of a host rather than its IP address. While spam traces from several domains give a broader view of spamming pattern than from a single domain, it is administratively very difficult to obtain diverse traces in this way. Getting the message bodies for each individual spam delivered seems an elusive dream in such scenarios as this data is bound by privacy policies. Our study analyzed spammer behavior from the unique view point of a man-in-the-middle of the spamming cycle, which overcomes the above difficulties.

In [2], the authors described setting up open relays and open proxies using tools described in [15]. They configured open relay to relay just the first message as they suggested that the first message is usually for relay testing. Though our study builds on similar ideas, we observed that testing messages are sent repeatedly and hence such messages need to be relayed continuously. We also presented a mechanism for detecting and relaying them without relaying the actual spam. Further, we pointed out that the relay should not forward testing messages from DNSBL(s) to evade being blacklisted.

7. CONCLUSION

In this paper, we presented a new methodology for spam trace collection using an open relay which offers a unique vantage point in the spamming cycle. The broader view of the spamming cycle, i.e., from a diverse set of spam origins to a diverse set of destination domains, allows us to separate HVS from LVS in a systematic way and we presented an algorithm for separating these two categories of spam origins. This separation allows us to isolate spam due to LVS and analyze the coordination among LVS. Our study exposed the interesting dynamics of "spam campaigns" by LVS.

The battle against spam is ongoing and to win the battle we need to have a good understanding of the spammer behavior which appears to evolve continuously. In our future work, we plan to analyze the properties of spam campaigns by LVS in the hope of reverse engineering LVS workload distribution strategies and designing new anti-spam techniques.

Acknowledgment

This work was supported in part by NSF grants CAREER-0238379 and CAREER-0643612. We thank our shepherd Thorsten Holz and the anonymous reviewers for their helpful comments.

8. REFERENCES

[1] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In *USENIX Security*, 2007.

[2] M. Andreolini, A. Bulgarelli, M. Colajanni, and F. Mazzoni. Honeyspam: honeypots fighting spam at the source. In *SRUTI'05: Proceedings of the Steps to Reducing Unwanted Traffic on the Internet Workshop*, pages 11–11, Berkeley, CA, USA, 2005. USENIX Association.

[3] Silicon valley north - unsolicited e-mail (spam) answers. http://www.svn.net/helpdesk/spam.html#address3.
[4] Junkbusters guide to staying off junk email lists. http://www.junkbusters.com/harvesting.html.
[5] Bl: Spamcop blocking list. http://bl.spamcop.net.
[6] Zdnet security news. most spam generated by botnets, expert says. http://news.zdnet.co.uk.
[7] M. Casado and M. J. Freedman. Peering through the shroud: The effect of edge opacity on IP-based client identification. In *Proc. 4th Symposium on Networked Systems Design and Implementation (NSDI 07)*, Cambridge, MA, Apr. 2007.
[8] Cbl: Composite blocking list. http://cbl.abuseat.org/.
[9] R. Clayton. Email traffic: a quantitative snapshot. In *Proc. of CEAS*, 2007.
[10] Sorbs: Spam and open-relay blocking system. http://dnsbl.sorbs.net.
[11] Dsbl: Distributed sender blackhole list. list.dsbl.org.
[12] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazieres, and H. Yu. Re: Reliable email. In *Proc. of NSDI*, 2006.
[13] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida, and J. Wagner Meira. Workload models of spam and legitimate e-mails. *Perform. Eval.*, 64(7-8):690–714, 2007.
[14] J. Goodman, G. V. Cormack, and D. Heckerman. Spam and the ongoing battle for the inbox. *Commun. ACM*, 50(2):24–33, 2007.
[15] Fighting spammers with honeypots. Laurent Oudot 2003-11-26, http://www.securityfocus.com/infocus/1747.
[16] Mailavenger. http://www.mailavenger.org.
[17] T. H. Isidore Rigoutsos. Chung-kwei: a pattern-discovery-based system for the automatic identification of unsolicited e-mail messages (spam). In *Proc. of CEAS*, 2004.
[18] B. Medlock. A language model approach to spam filtering.
[19] Njabl: Spam blocking blacklist. http://www.njabl.org/.
[20] T. Oda and T. White. Developing an immunity to spam.
[21] H. Project. Know your enemy: Tracking botnets. Published on the Web.
[22] M. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. A multifaceted approach to understanding the botnet phenomenon. In *Proc. of IMC*, 2006.
[23] M. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. My botnet is bigger than yours (maybe, better than yours). In *Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets (Hotbot)*, 2007.
[24] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *Proc. of SIGCOMM*, 2006.
[25] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proc. of 14th ACM Conference on Computer and Communications Security (CCS)*, 2007.
[26] Super webscan. http://www.sharewareconnection.com/super-webscan.htm.
[27] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk E-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.
[28] Joe st sauver: Evolving methods for sending spam and malware. http://www.ftc.gov/bcp/workshops/spamsummit/presentations/Evolving-Methods.pdf.
[29] The spamhaus project. sbl-xbl.spamhaus.org.
[30] Srizbi now leads the spam pack. http://www.marshal.com/trace/traceitem.asp?article=567.
[31] Cnn: Expert: Botnets no. 1 emerging internet threat. http://www.cnn.com/2006/TECH/internet/01/31/furst/.
[32] S. Venkataraman, S. Sen, O. Spatscheck, P. Haffner, and D. Song. Exploiting network structure for proactive spam mitigation. In *Proc. of Usenix Security*, 2007.
[33] 2006 spam trends report: Year of the zombies. http://www.commtouch.com/downloads/ Commtouch_2006_Spam_Trends_Year_of_the_Zombies.pdf.