Botnet Spam Campaigns Can Be Long Lasting: Evidence, Implications, and Analysis

Abhinav Pathak[1], Feng Qian[2], Y. Charlie Hu[1], Z. Morley Mao[2], and Supranamaya Ranjan[3]

[1]Purdue University, West Lafayette, IN 47907
[2]University of Michigan, Ann Arbor, MI 48109
[3]Narus Inc., Mountain View, CA 94043

ABSTRACT

Accurately identifying spam campaigns launched by a large number of bots in a botnet allows for accurate spam campaign signature generation and hence is critical to defeating spamming botnets. The straight-forward approach of clustering all spam containing the same label such as an URL into a campaign can be easily defeated by techniques such as simple obfuscations of URLs. In this paper, we perform a comprehensive study of content-agnostic characteristics of spam campaigns, *e.g.,* duration and source-network distribution of spammers, in order to ascertain whether and how they can assist the simple label-based clustering methods in identifying campaigns and generating campaign signatures. In particular, from a five-month trace collected by a relay sinkhole, we manually identified and then analyzed seven URL-based botnet spam campaigns consisting of 52 million spam messages sent over 2.09 million SMTP connections originated from over 150,000 non-proxy spamming hosts and destined to about 200,000 end domains. Our analysis shows that the spam campaigns, when observed from large destination domains, exhibit durations far longer than the five-day period as reported in a recent study. We analyze the implications of this finding on spam campaign signature generation. We further study other characteristics of these long-lasting campaigns. Our analysis reveals several new findings regarding workload distribution, sending patterns, and coordination among the spamming machines.

Categories and Subject Descriptors: C.2.3 Computer Communication Networks: Network Operations–network management; C.2.0 Computer Communication Networks: General–security and protection

General Terms: Measurement, Security

Keywords: Spam campaign, botnet, burstiness, distributedness, open relay

1. INTRODUCTION

Ever since spam first became a major problem, spamming techniques have escalated in complexity in response to the increasing sophistication of spam filtering techniques. Due to the fundamental weakness of content-based spam filtering techniques, *i.e.,* there are simply too many ways to obfuscate the content, researchers have developed IP-address-based techniques. Such techniques maintain a blacklist of IP addresses that are known to have originated spam in the past, and offer a simple interface such as DNS for convenient and efficient lookups by mail servers in the future. In reaction to DNS blacklisting, spammers resorted to employing *botnets*, each of which consists of a large number of compromised machines, typically operated under a central Command-and-Control (CnC) to originate spam. The low volume of spam originated from each individual bot significantly adds to the difficulty of accurate and timely blacklisting of these bot machines. Because of the sheer size of botnets, spam due to bots amount to a major percentage of the total spam worldwide [24, 20].

The difficulty with identifying individual bots due to their stealthy spamming behavior suggests that an effective approach to identifying and defeating botnets has to resort to identifying the *collective behavior*, *i.e.,* the spam campaign that the bots in a botnet launch collectively for pushing the same spam (including all of its obfuscated forms) to millions of mailboxes. Under such an approach, the new challenge in the battle against botnet spam is shifted to the ability to accurately cluster spam belonging to the same campaign and generate signatures characterizing a campaign immediately upon its onset. Such spam campaign signatures can then be used to identify and filter future spam belonging to the same campaign.

A key observation about spam campaigns is that spam belonging to the same campaign typically share the same spam label, such as a URL or a phone number, which is needed to carry out the spamming purpose, for example, to tell the recipients how to buy the medication advertised in the spam. However, clustering spam into campaigns solely based on the labels embedded in the message such as URLs is insufficient as it is easily defeated by techniques such as simple obfuscations of URLs, using HTTP features such as URL redirection, and including additional legitimate URLs in the mail body. A natural question then is whether the label-based spam campaign clustering approach can benefit from exploiting some defining content-agnostic characteristics inherent in a campaign?

In this paper, we perform a trace-driven analysis that searches for such definitive characteristics of spam campaigns. In particular, we focus on characterizing the *burstiness*, *i.e.,* how long a spam campaign lasts, and *distributedness*, *i.e.,* how widespread the sources of a spam campaign are. To facilitate our investigation, we leverage the technique previously developed for peeking into spammers' behavior from relay sinkholes [15] which provides the unique and broad view of numerous, possibly concurrent, spam campaigns hitting many diverse end domains (including spam to major domains such as Hotmail, Yahoo! mail, Gmail). In particular, from a five-

SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.
Copyright 2009 ACM 978-1-60558-511-6/09/06 ...$5.00.

month trace collected by a relay sinkhole, we manually identified seven URL-based botnet spam campaigns which consist of 52 million spam messages sent over 2.09 million SMTP connections originated from over 150,000 non-relay, non-proxy IP addresses and destined to about 200,000 end domains.

We observe that nearly all spam campaigns, when observed from large destination domains, exhibit a burst duration far longer than the five-day period used as the campaign burstiness threshold in a recent study [22]. One immediate implication of this finding is that one can not simply cluster all the spam from a large number of spamming sources that contain similar URLs and the delivery of which is finished within a short period of time as a single spam campaign, as burstiness is not a definitive characteristic for many spam campaigns. In other words, while burstiness plus distributedness may be a sufficient condition for identifying a spam campaign, it is not a necessary condition. Hence, using it as the definitive characteristic can result in a high false negative ratio. Our analysis shows that the complete URL signatures generated using a five-day burstiness cutoff and a 20-AS distributedness cutoff on our seven spam campaign trace result in a false negative ratio of 98.21% (considering only the spam that satisfy the 20-AS distributedness cutoff). In contrast, the study in [22] never reported false negative results while evaluating their spam campaign signature generation technique.

A second important contribution of the paper is an in-depth analysis of these long-duration botnet spam campaigns identified in our trace. Unlike previous analysis of spam campaigns [22, 23], the unique vantage point of our relay sinkhole allows us to study the spamming patterns of individual spammers to multiple destination domains, as well as the coordination of the sending patterns of individual spammers on behalf of different spam campaigns. Overall, the major findings of our study include:

- Many spam campaigns are *not* bursty in nature, whether observed from the relay's point of view or from an end domain's point of view; they continue on for months.
- Though a spam campaign as a whole may not be bursty in nature, the bots carrying out the work can in fact be bursty and stealthy within the campaign. In many cases, bots complete their entire workload for a particular spam campaign within the first one hour of their arrival into the campaign, as observed by the relay.
- There exist many common spamming IPs across multiple spam campaigns.
- The bots that appear in multiple spam campaigns typically spam for different campaigns in close-by time instances. Further, they appear to spam nearly the same workload (number of spam emails) but to distinct recipients across the multiple campaigns.
- An individual spamming host's involvement in a campaign is related to its upload link bandwidth. The higher the upload bandwidth, the more spam was observed to originate from it.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 briefly reviews the methodology and advantages of using relay sinkholes to collect spam. In Section 4, we discuss the challenges with spam campaign identification and how we identified seven URL-based spam campaigns manually. We characterize the burstiness and distributedness of botnet spam campaigns in Section 5, and perform an in-depth analysis of individual spammer behavior and the coordination among spammers in Sections 6 and 7. Finally, we conclude in Section 8.

2. RELATED WORK

The major shift of spammers from using high-volume spamming hosts to low-volume but coordinated bots has significantly increased the difficulty in identifying individual spamming sources. In reaction, a number of recent studies [2, 23, 10, 22, 11] have focused on learning and identifying bots' collective behavior, *i.e.,* the spam campaign they launched, with the goal of deriving spam campaign signatures which can then be used to filter future spam.

One approach is to perform string extraction on the entire mail content in order to identify common "shingles" that characterize a campaign. For example, Zhuang *et al.* [23] applied the shingling algorithm [4] to spam mail bodies to separate different spam campaigns in spam traces collected from Hotmail.

A less resource-intensive approach than string extraction from the email body is to look for labels that characterize the intention behind the message, *e.g.,* the advertised URLs. For instance, Anderson *et al.* [2] used spam traces collected from a top-level four-letter domain to study the spam infrastructure. They used the simple notion of a spam campaign to consist of all emails that contain the same URL (after all redirections). However, as exhibited by our trace, it is possible for two URLs that do not render the same page to belong to the same campaign. One way to work around this is to compare the page displayed for different URLs as images [2] and cluster them based on similarity of the rendered pages using image-shingling. However, this method can be quite resource-intensive.

In contrast, in this paper we take a different approach by first manually identifying spam campaigns and then analyzing content-agnostic characteristics of a campaign that can assist with campaign identification. In this regard, similar to our approach, Konte *et al.* [11] also first manually identified campaigns and classified spam in their trace containing URLs belonging to 3,360 domain names into 21 distinct campaigns. However, their goals are different than ours in that they focus on how fast-flux service networks are used to host the online scams advertised by the spam messages.

Most recently, Xie *et al.* [22] proposed to use *burstiness* and *distributedness* as two definitive characteristics of botnet spam campaigns to assist in identifying URL-based botnet spam campaigns. In particular, in extracting clusters of same-URL-containing spam from an email trace sampled at 1:25,000 from a large mail service provider, the authors exercised a burstiness filter of active period less than five days, and a distributedness filter of spammers originating from over 20 ASes. The URLs contained in such clusters are then used in a regular-expression-based automated campaign signature generation process for URL-based spam campaigns.

Related to spam campaign analysis, Kanich *et al.* [10] studied the conversion rate of spam – the probability that an unsolicited mail will ultimately elicit a sale – by infiltrating CnC channels of "The Storm" botnet and injecting three spam campaigns that spammed 500 million recipients.

Finally, several studies have focused on analyzing the network-level properties of spammers. As an example, Ramachandran and Feamster [17] analyzed the network-level behavior of spam originating from botnets and discovered a new spamming technique, called BGP spectrum agility, that uses hijacked prefixes to send spam. Ramachandran *et al.* [18] focuses on analyzing the similarity in the sending patterns of individual spammers such as the temporal and spatial locality (*e.g.,* in destination domains) in the spam they generated. They used email traces from 115 domains to develop a *behavioral blacklist* based on the sending behavior (the set of target domains that a particular IP address sends spam to) of spammers rather than a fixed identity such as an IP address. Such an approach requires sharing of spam collected from across multiple end domains.

3. METHODOLOGY

In our study, we used the methodology described in [15] for collecting a large spam trace from the unique vantage point of an open relay sinkhole. Such a unique vantage point provides a snapshot view of many spam campaigns involving a large number of coordinated hosts and many destination domains. In the following, we briefly review the methodology and discuss its advantages and limitations, and summarize the trace we collected.

3.1 Spam Collection using an Open Relay

Open relay sinkhole. An open relay is an MTA (Mail Transfer Agent) that forwards emails from any client to any destination. In general, spamming through an open relay is lucrative for a spammer since they can go undetected, as the final mail receiver sees only the mail relay as the spamming source. While the bots in a botnet generally send spam in low volume and hence are less likely to be detected, using a relay whenever possible remains lucrative as long as the relay is not blacklisted.

Spammers use relay testing software [19] 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 if they also relay. A spammer tries to relay a test email to its own email address through the detected host. Typically, the subject or the body of such an email contains the IP address of the host being tested. Once the test email is successfully received, the IP address of the host is extracted from the body and the host is confirmed to relay emails.

Once an open relay is detected, the spamming hosts start exploiting the host to relay spam. The relay testers periodically (about once a week as observed by the relay) check whether the host is still relaying the email 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 compromising it, *i.e.,* the relay being blacklisted by DNSBLs, the open relay is carefully configured to forward only the emails that are involved in relay testing. In this way, the relay testers are given continuous false assurance that the relay continues to forward all emails whereas in reality only the testing emails are relayed and all others are stored (and not forwarded). An important step here is to identify which emails are for testing the relays. Most of the relay testers could be trivially identified as they contain the IP address (in plain text/hex) of the relay server in either the mail body or in the subject lines. Some of them also contain words like “relay”, “test”, “successful”, *etc.* So any email that contains either the relay’s IP address or these keywords are let through. An important point to note here is that relay testing is also done by DNSBL(s) for purposes of blacklisting and these test emails also contain the IP address of the relay in the mail body. Hence, any email that contains words like “dnsbl”, “ordb”, “sorbs”, *etc.*, are denied from passing through to prevent our relay from being blacklisted. We note that this relay testing behavior is based on observations from our relay and hence the mechanism for detecting relay testers may not necessarily be general.

Advantages and Limitations of a Relay Sinkhole Trace. Upon being recruited by spammers to relay spam, an 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 on one side sees a plethora of origin spammers that attempt to relay spam through it, and on the other side sees all the final destinations of the spam. Such a broader view point of the spam traffic potentially reveals the global behavior of spammers. In particular, it potentially captures snapshots of many spam campaigns, each of which targeting many destination domains.

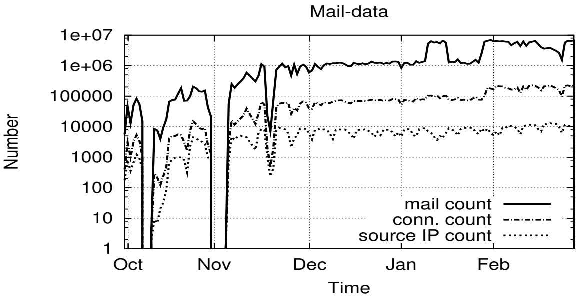

Figure 1: Number of emails supposed to be relayed, number of incoming connections and number of unique spammer IPs per day.

In principle, trace collection at a single domain can potentially cover 100% of all the spam destined to that domain. In practice, however, due to the sheer volume of spam at such domains, studies using them typically use sampling techniques. For example, in [22], the trace used by the authors are sampled at the ratio of 1:25,000. The spam trace collected at the relay sinkhole can also be viewed as a form of sampling, *i.e.,* of the spam campaigns going to each destination domain. Since the uniformity of this sampling of our relay trace is arguable, *i.e.,* the trace may not see all the spammers that are part of a campaign, in this paper, we will only extract properties of spam campaigns that are *lower-bounded* by what is seen in the trace. For example, if the duration of a spam campaign seen in the trace is X days, then we know the actual duration *can only* be longer, but not shorter. In other words, we will analyze the trace in a way that avoids its limitation.

3.2 Trace Statistics

Using an open relay, we collected a spam trace for a period of five months from September 30, 2007 to February 28, 2008. Figure 1 plots the number of emails that our relay was asked to relay per day and the number of unique IPs that generated these requests per day. We observe that the spam through the relay was initially low in volume but ramped up to one million per day in the second month until mid-January. From mid-January we observe an even higher volume of about 10 million spam per day. Except for the first month, the spam each day originated from about 10,000 spamming hosts. We observe that spam through the relay drops to zero in two time periods (in early October and early November). These two drops were not a result of data collection issues, thus we surmise that they were likely due to a modified decision at the spamming sources with respect to our relay.

Table 1 summarizes the statistics of the collected spam. In total, on the source side, over half a million unique IP addresses originated about 11 million SMTP connections to the relay. These IP addresses are spread across 2,243 ASes and 168 countries as determined by consulting the routing table database published by the RouteViews project [1][1]. On the destination side, our relay was

[1] Routeviews publishes routing table updates every 15 minutes. An IP is resolved to the AS that announced the longest prefix covering the IP.

Table 1: **Statistics of the whole trace, collected at a relay sinkhole from Sep. 30, 2007 to Feb. 28, 2008.**

Source side statistics	
# incoming SMTP conn.:	11,269,081
# unique IP addresses:	543,828
Avg. # recipients per conn.:	25.32
# network prefixes:	9,222
# ASes:	2,243
# spam originating countries:	168
Destination side statistics	
# mails to be delivered:	285,422,502
# unique recipients:	53,950,777
# destination domains:	628,092

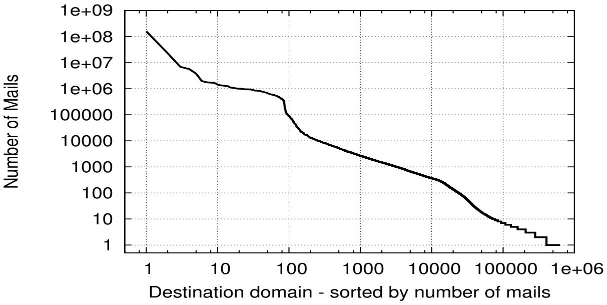

Figure 2: **Emails received per destination domain.**

asked to relay about 285 million messages in the five month period to about 54 million unique mailboxes distributed across 628,092 end domains. We found that about 75% of the hosts were already blacklisted in at least one of the five DNSBLs [5, 21, 9, 8, 3] which were queried by the relay at the time of receiving the spam. Figure 2 shows the number of emails that were destined to each domain, where the domains are sorted by the number of emails. We observe that a few providers, *e.g.,* Yahoo!, Gmail, Hotmail and Hinet, are the target of a lot of spam (89.39% of 285 million spam), indicating that the stakes are higher for them in the arms race against spam. These large end domains could also serve as a good spam information source for potentially helping others via a coordinated spam sharing mechanism (in a similar manner as Google's safe browsing API). We also observe in our trace that spam sources that generate high volumes, spam almost all domains, and conversely, domains that receive the most spam, receive from nearly all the spamming hosts.

In the rest the paper, we focus on the number of SMTP connections as opposed to the total number of spam messages to be received by all the destination mailboxes, for the following two reasons. First, in a single SMTP connection, a spamming source typically delivers the same spam message to multiple destination mailboxes in the same destination domain. This is achieved by transferring one spam message body and multiple destination mailboxes to the receiving mail server. Hence the number of SMTP connections is a more relevant metric for measuring the workload executed (as opposed to assigned) by a spammer, compared to the total number of spam messages finally delivered (*e.g.,* by the receiving mail servers) to all the destination mailboxes. Second, an SMTP connection carries the spamming source information and the time when the spam was delivered from the source, and hence is more specific than the multiple mailboxes being spammed in that connection when we analyze the duration and distributedness of spam campaigns.

Table 2: **Breakdown of the 11 million connections by spam label type.**

SC Label Type	% of Spam	# of distinct spam labels
URLs	58.63[a] (6.50 million)	2,501
Phone Number	15.02	10
Skype ID	9.45	2
Mail ID	1.43	1
No Subject/Body	5.31	1
Others	9.90	-

[a] We removed 0.26% of the connections which contain invalid URLs or URLs of the form `www.dom.tld/location-OOOO` where OOOO denotes per-connection obfuscated string as each such URL appears in spam sent in only one connection.

4. SPAM CAMPAIGN IDENTIFICATION

In this section, we formally define spam campaigns (SCs), and discuss challenges with identifying URL-based SCs and how we manually identified seven URL-based SCs accounting for 52 million spam messages and over 2 million connections in our trace.

4.1 Definitions

We formally define a *spam campaign* to be the set of email spam that are meant to achieve the same *spamming purpose*, for example, in trying to sell a certain medication. To evade the increasingly sophisticated content-based spam filters, the content of a spam campaign is typically obfuscated (in more and more sophisticated manners). However, no matter how much the content is obfuscated, spam belonging to the same campaign have to contain the same underlying "contact information" to execute the purpose, *e.g.,* to "sell the medication." Such "contact information" comes in various forms. The first and still most widely used format is a URL that points to the web page that in turn contains detailed information for executing the spamming purpose. More recent "contact information" types include phone numbers and Skype IDs.

We denote such "contact information" as the *spam label*, as in principle each of them uniquely identifies the spam that belong to a campaign. In practice, realizing that the static "contact information" contained in the messages for a campaign can be used by a content-based spam filter to detect spam, the spammers try to obfuscate the "contact information" to the extent they can. For example, this is easy to do for URLs by using standard features of HTTP protocols such as redirection (using simple HTTP or javascript).

We denote different types of "contact information" as different *spam label types*. Table 2 gives a breakdown of the 11 million connections in our spam trace according to different label types. We observe that 58.63% of the connections are for spam containing URLs (6.50 million connections), 15% contain phone numbers, 9.45% contain Skype IDs, 1.43% contain Mail IDs (email addresses in spam messages for recipients to contact the spam originators) as the "contact information". We denote the above label types as *simple label types*. Interestingly, 5.31% of the connections have completely empty subject field and mail body. The purpose of such spam remains unclear and we suspect they could be due to bugs in the spam template of the spammers. Finally, 9.9% of the connections do not belong to any of the above simple signature types. They do not contain URLs or specific contact information. These spam emails seek to influence the behavior of the recipient without listing a means for contacting the spam originator. Pump and Dump stock spams have these characteristics though there are several other kinds in our trace.

Table 3: Example URLs belonging to the same spam campaign.

URL	Spam Campaign
http://dhnaXXXX/hljtehnahaj	SC-Software1
http://dhnaXXXX/paepyhaeot	SC-Software1
http://dhnaXXXX/654j6d4jj	SC-Adult2
http://dhnaXXXX/gfjxh985034	SC-Adult2
http://dhnaXXXX/tgg3w3rq4324ty345	SC-Adult2
http://www.988.idv.XXX	SC-Book
http://www.ohinet.net/XXX/magic/	SC-Book
http://www.myweb-gmail.com/XX/composition_phrase/	SC-Book

4.2 Challenges

Since different label types are simply alternate ways of providing the "contact information", there is little incentive to provide redundant "contact information", as confirmed by examining our spam trace. For spam containing label types of phone numbers, Skype IDs, or Mail IDs, we observe that each of them contains only a single occurrence of a single label type, i.e., it contains either a single phone number, a single mail ID, or a single SkypeID. Further, labels of these types are never obfuscated. The situation with clustering spam containing URLs as the spam labels is much more complicated, as URLs can be easily obfuscated. Hence, we focus on URL-based spam in the rest of the paper. The most common type of URL obfuscation we observed was HTTP redirection using standard redirectors such as those provided by Yahoo! and AOL. It is straight-forward to extract final URLs from these redirectors (almost all spam filters extract the final URL before generating a spam score).

Automatically identifying spam campaigns purely based on examining the URLs contained in the message poses several challenges. First, spam containing URLs that share the same domain name can belong to different spam campaigns. An example of this is given by the first and second groups of spam in Table 3. Second, conversely, spam containing URLs that differ in the domain name can belong to the same spam campaign. An example of this is given by the third group of spam in Table 3. Third, spammers can insert legitimate URLs along with the spam URLs into the message to confuse URL-based identification schemes.

4.3 Manual Identification of URL-based Spam Campaigns

In light of the above challenges in automated identification of spam campaigns, to enable our characterization study of spam campaigns, we manually classify the 6.50 million connections (52 million spam) into seven major botnet initiated campaigns as follows. First, we grouped all spam that have in common the same URL into a separate cluster. Note that a spam email containing multiple URLs will appear in multiple clusters. There are 2,501 distinct URLs in total, and we end up with 2,501 clusters. Second, we randomly pick one spam from each cluster, and gave the resulting spam to a human who manually clustered 2,042 of these 2,501 URLs into seven distinct spam campaigns, corresponding to 4.21 million connections. Finally, we removed connections from open proxies and open relays. Spammers use several indirect sources such as open relays and open proxies to relay their spam emails in order to hide the infected machines. In fact, spam emails that are to be forwarded through our relay make one such example. Since our goal is to study spam originated from individual bots, ideally we would like to replace such open proxies and open relays in our trace with the bots behind them. However, it is actually non-trivial to determine the actual source behind a proxy that originated the spam. Instead, we simply drop all spam sources that are listed as open proxies or open relays in the Spamhaus policy blacklist [16] which was recently integrated with NJABL [14]. We note that this is effectively a sampling of our trace. This sampling does not affect our analysis, however, as long as we only extract properties of spam campaigns that are *lower-bounded* by what is seen in the (sampled) trace. After removing spam due to about 40,000 different open proxies and relays, we obtained the same number of URLs, 2,042 while the number of connections was reduced in half, from 4.21 million to 2.09 million. In the rest of the paper, we study these 2.09 million connections containing 2,042 URLs belonging to the seven manually identified spam campaigns.

We note that in sharp contrast to past reports that the amount of spam originating from open proxies/relays has reduced substantially, we found evidence to the contrary - about half of the 4.21 million connections for the seven manually classified URL-based spam campaigns are for such emails. This is indicative of open proxies and open relays being used as the "front line" by spammers to hide their bots.

Table 4 summarizes the nature and statistics of the seven major SCs identified. Each of the seven campaigns is identified by a set of URLs that were contained in the spam in that campaign. For example, SC-Book corresponds to a set of 1,567 URLs, all of which are related to selling books and ultimately pointed to an online bookstore. Note that each campaign exhibits diversity in terms of multiple distinct URLs and in five out of the seven campaigns, even the URL domain names are different. This underscores the difficulty in obtaining spam campaign signatures in an automated fashion. In the rest of the paper, we study various content-agnostic characteristics of campaigns to explore whether they can assist with automated campaign identification.

5. SPAM CAMPAIGN CHARACTERISTICS

In this section, we characterize the duration and distributedness of the seven major URL-based spam campaigns identified from our relay trace. We further study the implications of these characterizations on the automated campaign signature generation technique proposed in [22].

5.1 Campaign Duration and Distributedness

We first analyze the duration of the spam campaigns in our relay trace. Table 4 shows that the seven URL-based spam campaigns identified in our trace have duration between 1 to 99 days, and the spam sources originate from 8 to 1,173 ASes.

We further analyze the duration of these spam campaigns when observed from top four destination domains in our trace, Yahoo!, Gmail, Hotmail, and Hinet. The duration of these campaigns again last between 1 to 99 days. As discussed in Section 3.2, since our relay trace may not capture all the spam belonging to these seven campaigns headed to those individual destination domains, the actual duration and distributedness of the spam campaigns may be even longer and wider.

5.2 Per-URL Duration and Distributedness

Since the spam campaign signature generation scheme in [22] starts by identifying clusters of spam containing the same URL that satisfy the burstiness and distributedness criteria, we also analyze the duration and distributedness of such individual clusters in our trace. As explained before, the seven manually identified campaigns contain 2,042 unique URLs. Since our five-month trace collection could potentially start from and end in the middle of the campaign containing a particular URL, we removed all the spam containing an URL that ends in the first month or starts in the last month of the five-month period. The resulting trace contains 1,774 URLs and 1.3 million connections. This filtering also removed SC-

Table 4: **Summary of the seven manually identified major URL-based spam campaigns.**

SC Name	Ad Type	Distinct URLs	Distinct URL Domains	# of Source IPs	# of Source ASes	# of SMTP Connections	# of Destination Domains	Duration of SC (Days)	Start Date
Book	Book Store	1,567	31	8,555	8	287,705	94,466	71	Dec 19
Adult1	Adult Drug	38	12	92,441	1,173	720,076	80,034	92	Nov 28
Adult2	Adult Site	306	12	62,117	1,055	419,750	54,622	99	Nov 19
Adult3	Adult Tool	24	1	228	80	4,611	143	1	Jan 28
Shopping	Shopping	5	1	20,375	592	107,934	26,917	36	Nov 28
Software1	Software	54	4	28,502	702	265,476	45,308	12	Feb 12
Software2	Software	48	3	36,178	856	279,799	44,472	63	Dec 22

Table 5: **Distribution of URLs and SMTP connections in the six campaigns over varying burstiness and distributedness ranges – all spam (1,774 URLs and 1.3 million connections).**

Duration (days)	Distributedness (# of ASes)					Total
	1-10	10-20	20-100	100-500	≥ 500	
0-5	3.66 (6.32)	0.96 (0.02)	0.00 (0.00)	0.45 (0.96)	0.00 (0.00)	5.07 (7.30)
6-10	1.58 (6.04)	0.00 (0.00)	0.00 (0.00)	1.52 (1.26)	0.00 (0.00)	3.1 (7.30)
11-20	1.52 (10.74)	0.00 (0.00)	0.00 (0.00)	5.13 (19.40)	0.00 (0.00)	6.65 (30.14)
21-40	2.76 (17.89)	0.00 (0.00)	0.00 (0.00)	1.80 (17.70)	0.00 (0.00)	4.56 (35.59)
>40	79.71 (40.03)	0.00 (0.00)	0.28 (0.07)	0.00 (0.00)	0.06 (10.67)	80.05 (50.77)
Total	89.23 (81.02)	0.96 (0.02)	0.28 (0.07)	8.90 (39.32)	0.06 (10.67)	

Table 6: **Distribution of URLs and connections in the six campaigns over varying burstiness and distributedness ranges - only spam destined to the Yahoo! domain (1,774 URLs and 0.9 million connections).**

Duration (days)	Distributedness (# of ASes)					Total
	1-10	10-20	20-100	100-500	≥ 500	
0-5	3.72 (7.53)	0.90 (0.03)	0.06 (0.04)	0.45 (1.00)	0.00 (0.00)	5.13 (8.6)
6-10	1.80 (5.93)	0.00 (0.00)	0.00 (0.00)	1.52 (1.24)	0.00 (0.00)	3.32 (7.17)
11-20	1.30 (8.52)	0.00 (0.00)	0.00 (0.00)	5.13 (21.27)	0.00 (0.00)	6.43 (29.79)
21-40	2.82 (16.14)	0.00 (0.00)	0.00 (0.00)	1.80 (17.96)	0.00 (0.00)	4.62 (34.1)
>40	79.59 (37.94)	0.23 (0.03)	0.00 (0.00)	0.00 (0.00)	0.06 (10.80)	79.88 (48.77)
Total	89.23 (76.06)	1.13 (0.06)	0.06 (0.04)	8.90 (41.47)	0.06 (10.80)	

Table 7: **Distribution of URLs and connections in the six campaigns over varying burstiness and distributedness ranges – only spam from non-blacklisted IPs and destined to the Yahoo! domain (1,748 URLs and 200,000 connections).**

Duration (days)	Distributedness					Total
	1-10	10-20	20-100	100-500	≥ 500	
0-5	6.01 (27.17)	0.06 (0.05)	0.46 (0.93)	0.00 (0.00)	0.00 (0.00)	6.53 (28.15)
6-10	3.38 (7.59)	0.00 (0.00)	1.54 (0.98)	0.00 (0.00)	0.00 (0.00)	4.92 (8.57)
11-20	4.29 (9.43)	0.00 (0.00)	2.80 (11.94)	2.57 (20.80)	0.00 (0.00)	9.66 (42.17)
21-40	21.11 (22.85)	0.00 (0.00)	0.00 (0.00)	1.66 (14.49)	0.00 (0.00)	22.77 (37.34)
>40	53.83 (19.01)	0.00 (0.00)	0.00 (0.00)	0.06 (10.91)	0.00 (0.00)	53.89 (29.92)
Total	88.62 (86.05)	0.06 (0.05)	4.8 (13.85)	4.29 (46.2)	0.00 (0.00)	

Software1, all URLs for which started in the last month of the five-month period. A breakdown analysis of distributedness and burstiness of the 1,774 URLs for the remaining six campaigns is given in Tables 5, 6 and 7. Each cell in the tables provides two numbers, the percentage of URLs and the percentage of connections that are characterized by a certain range of distributedness and burstiness. Note that the percent values for URLs across a row or column in the tables add up to 100% but this may not be true for percent values for connections. This is because a spam message may contain multiple URLs where each URL may be characterized differently, *i.e.,* in different cells in the table. Hence, an SMTP connection that sent such a message would get counted multiple times, once for each of the URLs that it corresponds to.

Table 5 shows the distribution of these 1,774 URLs (1.3 million connections) in the six campaigns over varying burstiness and distributedness ranges. We observe that only 0.45% of the URLs (0.96% of SMTP connections) have duration of five days or shorter and are distributed over more than 20 ASes. If we ignore distributedness, relatively few connections (7.30%) and URLs (5.07%) occur over short duration (five or fewer days) whereas a majority of the connections (50.77%) and URLs (80.05%) advertise URLs that lasted a long duration (more than 40 days). If we ignore duration, a majority of connections (81.02%) and URLs (89.23%) originate from a small number of ASes (10 or fewer) while a relatively smaller number (49.99% connections and 8.96% URLs) originate from a larger number of ASes (100 or more).

We further analyze whether the above distribution will change significantly when a campaign is viewed from an individual destination domain. Such an analysis is useful as spam filtering techniques are often deployed at individual domains. Table 6 shows the distribution of the 1,774 URLs corresponding to over 0.9 million connections in the six campaigns that were destined to Yahoo! only. We chose Yahoo! as the destination domain in this analysis as about 70% of the spam in the six campaigns were destined to it. We observe the distributions are very similar to those in Table 5; a significant percentage of spammed URLs exhibit persistence and are advertised via a small number of ASes.

Finally, we analyze whether the above distribution will change when the spam trace is examined after filtering out the spam originated from blacklisted source IPs. We perform this analysis as the trace in [22] was processed this way and hence it helps us to isolate factors that potentially contribute to the different observations. After removing the spam originated from blacklisted source IPs (over 75% spam sources were blacklisted in at least one of the five blacklists queried), there are 1,748 URLs and about 200,000 connections in the six spam campaigns that were destined to Yahoo! only. Table 7 shows the distribution of these URLs and connections. Compared to the distribution before removing spam due to blacklisted IPs, there is a significant reduction in distributedness, but only a relatively small reduction in duration. In particular, the URLs that have duration of five days or shorter and distributed over more than

20 ASes remain at 0.46% (and the connections at 0.93%). If we ignore distributedness, 6.53% of the URLs have duration of five days or shorter, but the URLs with duration over 40 days remains substantial, at 53.89%. If we ignore duration, 88.62% of the URLs orignate from 10 or fewer ASes, while 4.29% of URLs originate from 100 or more ASes.

The above finding on the duration of per-URL spam clusters from our trace is in sharp contrast to the previous study in [22] where the authors observed that the vast majority of URLs are actively advertised for a short burst of less than five days. This finding has significant implications on the design of signature extraction algorithms for detecting spam, as we show next.

5.3 Implications on Signature Generation

To estimate the impact of this finding on the efficacy of campaign signature generation, we re-implemented the spam signature generation technique proposed in [22]. Specifically, we analyzed the six manually identified campaigns as observed by Yahoo! after filtering out spam from blacklisted source IPs. The URL/connection distributions of this trace are in Table 7. We first used the same at-least-20-ASes cutoff for the "distributedness" property as in [22] to filter out the spam containing URLs that are not distributed enough. For the remaining spam, we then used the burstiness duration of five days as in [22], and generated Complete URL (CU) signatures. We focused on the CUs as the authors of [22] reported a majority (70.3-79.6%) of the spam campaigns identified belonged to the CU category. Using the signatures generated, we calculated the number of connections that would be blocked by those signatures, as a percentage of all the spam mails containing all the URLs originated from 20 or more ASes. This gives us the "false negative ratio" in using the CU signatures to filter the spam. The false negative ratio is calculated to be 98.21%. We repeated the scheme using a 20-day burstiness cutoff, and the false negative ratio is reduced to 79.21%.

To summarize, while the authors in [22] considered burstiness as a necessary criteria in narrowing down the candidate pool of URLs on which they apply signature extraction, we argue otherwise. Using a short duration of five days as the burstiness criteria can lead to very high false negative ratios. Our empirical observation highlights the challenges in accurately extracting spam signatures, since to reduce false negatives, it would be necessary to apply keyword extraction and signature generation algorithm across a very large set of spam that encompass duration as large as several months. This suggests that relying on burstiness as a criteria for signature extraction may prove counter-effective as it may be too late to react and block the spam.

6. INDIVIDUAL SPAMMER BEHAVIOR

In this and the next section, we analyze botnet spam campaigns in detail to gain insights into the workload distribution and coordination amongst the members of a botnet that originate a spam campaign. Since we no longer analyze the spam for individual URLs, we make use of the entire spam trace for all seven manually identified campaigns of 2.09 million connections and 2,042 URLs. Due to space restrictions, in this section we focus on three spam campaigns identified in our trace, based on their distinctive features: SC-Book has a large number of distinct URLs, SC-Adult1 has a large number of source IPs, and SC-Software1 is relatively short in duration but is highly distributed in terms of the number of ASes corresponding to the spam sources. As explained in Section 3.2, we use the number of SMTP connections to characterize the workload.

Figure 3 shows the number of SMTP connections made per day to our spam relay and number of unique IP addresses per day that contact our relay for the three selected spam campaigns. We observe similarly as in Figure 1, that all three campaigns have a fairly steady total workload per day over the 2.5-month period. But does this stability in aggregate behavior for a spam campaign imply uniformity in terms of workload per spammer as well? In particular, we seek to answer the following two questions related to the workload distribution amongst the spammers of one campaign:

- Are the spammers belonging to the same campaign assigned an equal amount of workload, *i.e.,* connections made? If not, what causes the uneven workload?

- How does each spammer accomplish its workload over time?

6.1 Workload Distribution

Figure 4 shows the workload distribution per spammer IP for the three spam campaigns. Note that despite the relative stability in terms of the total number of connections per spam campaign over time, the workload of each spammer for a given campaign varies widely from 1 to 1,000 SMTP connections per spammer IP address.

A natural question that arises is whether this uneven workload distribution is correlated with, and hence explained by, the arrival time of spammers joining a spam campaign. In other words, do spammers that join a campaign earlier than others send more spam for a given campaign, potentially due to their confirmed ability to deliver spam?

While it is possible that a spammer joined a campaign long before it initiated its first SMTP session to our relay, there is no easy way to establish the ground truth about this. Therefore, we use the time a spammer initiated its first SMTP session for a particular campaign to our relay as its arrival time for that campaign.

Figure 5 plots the spamming activities of spammers sorted in increasing order of their arrival time, for the three spam campaigns. We see a fairly even spread in arrival times for spammers, and a majority of spammers continue spamming for a campaign for long periods of time. In particular, for SC-Adult1, a few spammers were seen to be actively spamming for almost 2.5 months.

Next, Figure 6 plots the number of spam connections per IP address for the three campaigns, sorted in increasing order of their arrival time, as in Figure 5. From Figure 6, it is hard to observe a trend for SC-Book, SC-Adult1, and SC-Software1. Hence we also plot a Simple Moving Average (SMA) curve for these campaigns. For every IP address on the y-axis, we average the values for the 250 IP addresses above it and 250 below it, and join these points per IP through a curve which we call SMA. Now a trend can be observed for SC-Book, with the latter arrivals sending less spam than the earlier arrivals. Interestingly, no such trends are present for SC-Software1 and SC-Adult1, with each IP address making a similar 7-10 connections during its lifetime of the campaign.

6.2 Workload Spread Over Time

Next, we study how an individual spammer in a campaign accomplishes its workload over time.

Short-duration spammers. Figure 7 depicts a percentage-percentage plot, where the y-axis is the percentage of spam sources and the x-axis is the percentage of spam sent by them, during the first 1, 8 and 24 hours of their arrival into the three campaigns, Book, Adult1, and Software1, as seen by the relay. A point (x, y) on the curve for t hours for a particular campaign means y% of spam sources spam at least x% of their workload in the first t hours of their arrival.

For SC-Book, Figure 7(a) shows that 20% of sources complete their entire workload (100% on the x-axis) within the first hour of their arrival. However, only 60% of the sources send 10% or more of their total workload within the first hour of spamming. Furthermore, note that the percentage of spam sent by a spammer during

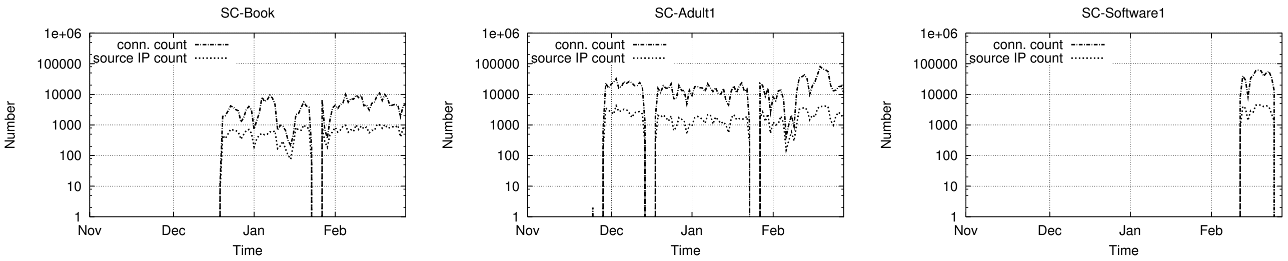

Figure 3: **Number of mails and number of source IPs per day for SC-Book, SC-Adult1, and SC-Software1.**

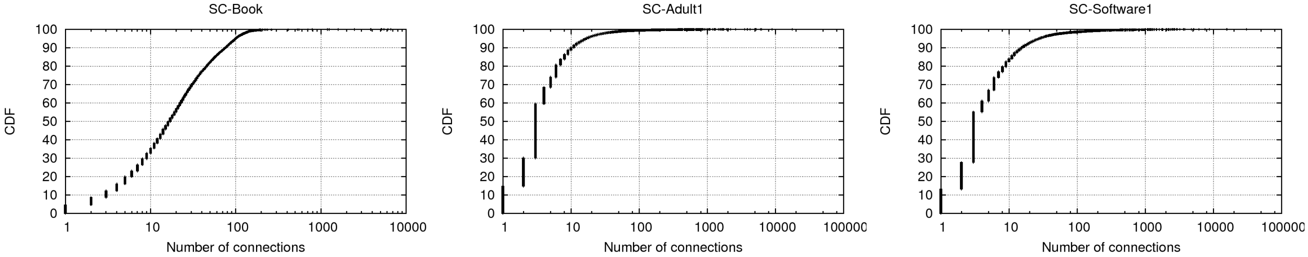

Figure 4: **CDF of workload per source IP for three spam campaigns.**

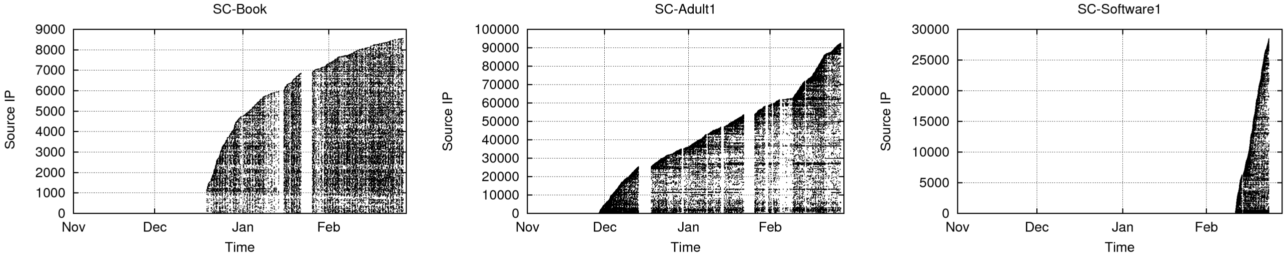

Figure 5: **Correlation between spammer IP and its spamming activities, sorted in increasing order of arrival time, for the three spam campaigns.**

the first 1, 8 and 24 hours of its arrival in the campaign remains nearly the same. These facts imply that some spammers (20%) finish their workload as soon as they first arrive in the campaign, while others make multiple visits across several days before they deliver 100% of their workload.

For SC-Adult1 and SC-Software1, we see a completely different trend when compared to SC-Book. We see that these two campaigns observe different workload behavior during the first 1, 8, 24 hours. More than 90% of the spam sources complete all their work within 8 hours after their arrival, and about 75% complete their entire workload during the first hour.

Bursty spamming behavior. The above analysis of SC-Adult1 and SC-Software1 indicates the "spam and move on" behavior by *individual* spam sources. We emphasize that this is not to be confused with the "burstiness" of spam campaigns as defined in Section 5. SC-Book, on the other hand, shows reuse of a spammer for spamming since a spammer does not complete its workload on the first day (Figure 7(a)) of its appearance. We next focus on SC-Book to investigate how often spam sources revisit our relay and the characteristics of their spamming activities such as the number of connections and the duration of spamming at each visit.

We define a "burst" event by a spam source as the time duration that starts when we observe the first spam from an IP and ends at a point after which there is no spamming activity from it for at least a pre-chosen period of time which we denote as the "numb" period. Within a burst, the time gap between any two consecutive spam emails would be less than the "numb" period. Using this definition, the spam history of an IP follows the pattern of oscillating periods of burst and numb, *i.e.,* burst - numb - burst - numb, and so on. We select a "numb" period of one hour since Figure 7(a) shows for SC-Book the spam workload distribution for the first hour nearly equals the workload distribution for the first 8 and 24 hours.

We now study the burst characteristics for SC-Book. Figure 8 plots the CDF of the number of bursts per IP. We see that 20% of IP addresses participate in the campaign with only one burst. The fact that spammers revisit the relay to continue SC-Book is vindicated by observing that about 18% of spam sources participate in the campaign for more than 10 bursts. Figure 9 plots the CDF of the average time duration of a burst per IP address. We see that more than 80% of IPs have average burst lengths longer than 100 seconds and nearly all of them finish in less than 1,000 seconds. This indicates the stealthy behavior of spammers where they finish their workload per burst within 100-1,000 seconds. Figure 10 plots the CDF of the average workload per burst per IP address. 90% of the IPs for this campaign have an average workload per burst of fewer than 10 connections. We also plot the CDF of the spamming workload per IP address for the first hour. We observe that the first burst by an IP typically delivers fewer spam emails compared to the average workload per burst. In summary, SC-Book shows a stealthy behavior, where spammers visit the relay multiple times, with each visit of "bursty" short duration on the order of 100-1,000 seconds, and comprising of a low volume of 10 or fewer connections.

6.3 Workload and Access Link Capacity

We now turn back to the question raised in Section 6.1 regarding the cause for the uneven workload distribution among spammers and examine possible correlation between the workload and the network access link capacity of different spammers.

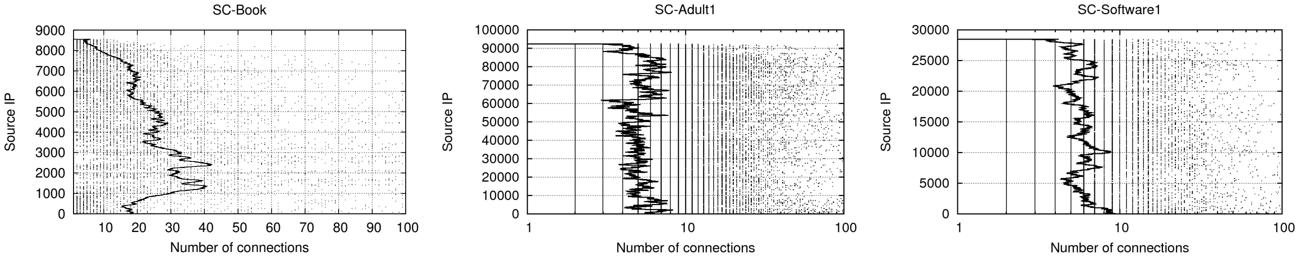

Figure 6: Number of connections per source IP for three spam campaigns, sorted in increasing order of their arrival time. (note: the x-axis has linear scale for SC-Book and log scale for other two graphs.)

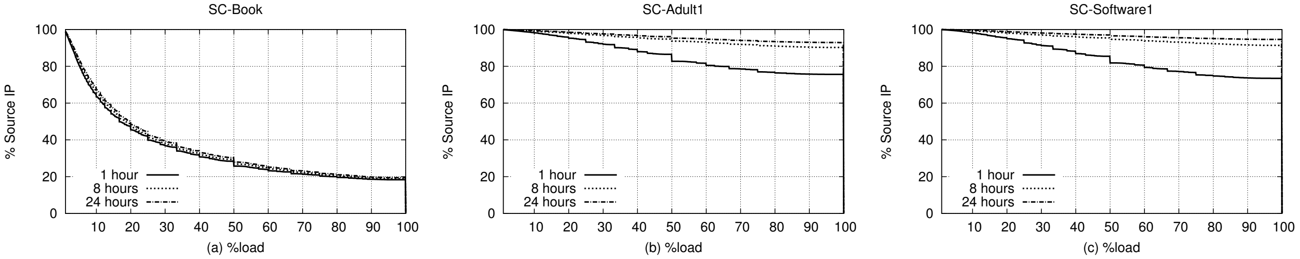

Figure 7: Distribution of workload (in connections) accomplished by spammers in the first 1/8/24 hours for SC-Book, SC-Adult1, and SC-Software1.

Measuring the access link capacity of a link remotely is a difficult problem [7]. We used the tool proposed in [7] to measure the access link capacity of a remote spamming host as follows. For each five-hour duration, we collected all the new IP addresses that spammed our relay. At the end of it, if there were more than 800 IPs collected, we randomly sampled 800 of them (otherwise we chose all of them)[2] and fed them to four machines, each of which used the tool described in [7] to measure the link capacity of 200 hosts. The cap of 800 was so that the measurement could finish in five hours using the four machines. We repeated this procedure for about one month and measured the link capacity for a total of 92,000 spam sources. Unfortunately, only 2,231 hosts responded to the packet trains sent by the prober.

Since we had the capacity information for only 2,231 hosts, instead of performing a per spam campaign analysis for them, we performed an aggregate study for these sources. In particular, for each of these 2,231 IP sources, we found out the average burst workload, average burst length, and number of bursts across all the campaigns combined, with the "numb" period being one hour. Figures 11(a)-(c) show the scatter plots between these metrics for each of these IPs, versus its upload capacity. We also plot the Simple Moving Average (SMA) for all three scatter plots by taking an average of 50 points above and 50 below a particular point.

Figure 11(a) shows that the higher the spammer upload capacity, the larger the workload per burst to the relay from that spammer (note that the x-axis is in log scale). The SMA of the average burst workload in fact shows an exponential increase as the upload capacity increases. Figure 11(b) suggests that the higher the upload link capacity (of a spamming host), the longer the burst duration. The SMA of the average burst time increases from about 30 minutes for low link capacity spammers to about three hours for high link capacity spammers. Figure 11(c) shows a similar trend in the number of bursts from an IP address when compared with its upload capacity.

We conclude from above that the higher a spammer's capacity to spam, the heavier workload it tends to deliver. This suggests that the controllers exploit the upload capacities of their spamming bots. The simplest way to achieve this appears to be one where a controller divides the workload into equally sized chunks, and as a bot finishes a piece, it retrieves another. This would enable the controller to exploit the different capacities of bots, as well as deal reasonably well with bots that get turned off unexpectedly.

7. COORDINATION AMONGST SPAMMERS

In Section 6, we made the observation that spam campaigns are fairly stable over time in terms of the number of SMTP connections and the number of source IPs per campaign (see Figure 3). We also saw that individual spammers could be bursty, with spammers for an example campaign (SC-Book) sending multiple short bursts, each lasting 100-1,000 seconds. Finally, we saw that individual spammers are assigned a diverse workload per campaign ranging from 1 to 1,000 SMTP connections per campaign (see Figure 4). That still leaves the following question unanswered: despite the disparity in workload across spammers, why does the aggregate behavior of a campaign still appear relatively stable over time? In this section, we provide a breakdown of a spam campaign in terms of the geographical distribution of spammers in an attempt to answer this question. Further, we also study the coordination amongst the spammers that participate in the same campaign and those that participate across different campaigns. In Sections 7.1 and 7.2, we use SC-Adult1 which involves over 90,000 spammers from all over the world.

7.1 Steady Aggregate Behavior of a Campaign

To better understand the disparity between the steady aggregate behavior of a campaign and the diverse individual behavior of spammers, we break down the spammers belonging to a campaign by the domains targeted and by the geographic regions from which the spammers originate. We obtain the geographic region for an IP using Maxmind [13].

We study one week of spam activity in SC-Adult1 from February 10^{th} to 17^{th}. We start by looking at all the spamming activities destined to all the domains as seen at the relay. In this one week period, our relay observed a total of about 0.15 million connections containing spam messages destined to 90,000 domains. Figure 12(a) shows the number of SMTP connections per hour made

[2] We note that an IP address seen by our relay could be a NAT. We do not make a distinction among spam sources in this regard.

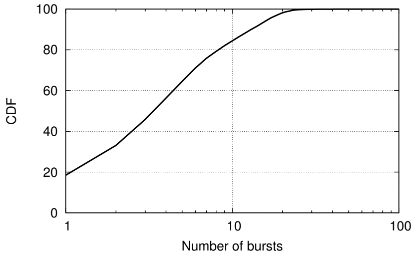

Figure 8: **CDF of number of bursts per IP for SC-Book.**

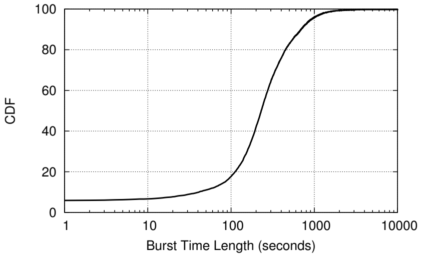

Figure 9: **CDF of average length of bursts per IP address for SC-Book.**

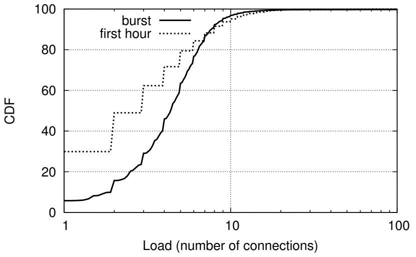

Figure 10: **CDF of average workload per burst per IP, and of workload in the first hour of spamming per IP, over all IPs, for SC-Book.**

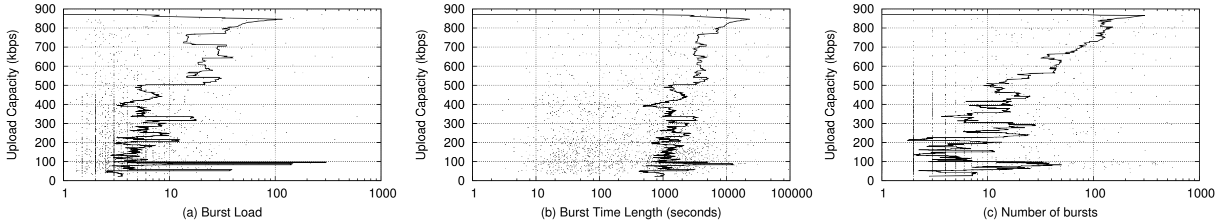

Figure 11: **Scatter plots of the upload capacity of an IP vs. its burst characteristics (across all campaigns spammed by the IP), (a) average burst workload, (b) average burst duration, and (c) number of bursts, for 2,231 spamming IPs. Also shown are the SMA (simple moving average) across 100 neighboring IPs.**

to our relay over the one week time period. Similar to what was seen from Figure 3, the aggregate behavior in terms of the number of connections is stable over time.

Per-day activities to different domains. We next separate spam for SC-Adult1 by the different domains targeted. We identify that this campaign targets users in four major domains: Yahoo!, Gmail, Hotmail, and Hinet. Interestingly, the number of connections made to our relay for this campaign that is heading to any of these domains is still stable over time (figures not shown).

Per-day activities from different regions. We next separate spam for SC-Adult1 according to where the spammer is geographically located: India, Argentina, Brazil, and China, as shown in Figure 12(b)-(e). Note that the time on the x-axis for all figures is in EST. Interestingly, the number of connections made from each geographic region follows a diurnal variation, peaking during the local timezone's mid-day. For instance, for sources originating from India, the leftmost point on the x-axis corresponds to 1:30 pm IST, which is when the peak activity occurs. Similar behavior is seen for Argentina and Brazil. The number of connections from China does not exhibit as pronounced diurnal variations and we suspect this could be due to the country straddling many longitudes which could contribute to a statistically multiplexed behavior. The fact that bots from each country seem to generate the maximum spam during their local mid-day could be explained by previous reports about the working of spam botnets. For instance, the authors in [12] reported that Storm Bots typically finish majority of their spam workload within four hours of the machine being booted up.

Interestingly, the geographically distributed nature of botnets combined with the different timezones explains the reason why campaigns such as SC-Adult1 (and all other campaigns observed by our relay) exhibit a stable behavior over time.

Per-hour activities per region to a domain. Finally, we study the number of connections made by spammers from different countries to a single destination domain, Yahoo! mail. The resulting graphs, not shown due to space limitation, appear very similar to those in Figure 12(b)-(e). This is because about 60% of the recipients for SC-Adult1 are in the Yahoo! domain. Plots for other domains such as Hotmail, Gmail and Hinet also show similar behavior. The main observation we derive from these results is that there is no destination-specific scheduling of spam by spammers, with each of the large domains exhibiting similar per-origin-country pattern as the aggregate spam campaign.

7.2 Intra-Campaign Coordination

Next, we explore the coordination amongst spammers that participate in the same campaign. This coordination is reflected in a very interesting way in the mailboxes being spammed. First, we observe that each SMTP connection delivers spam mails to alphabetically close recipients (an observation similar to [6, 15]). To further investigate the workload balancing strategy used across the spamming sources, we analyze the correlation between the spam source IPs and the corresponding mailboxes. In Figure 13, we plot a scatter plot of spam source IPs versus the recipients to whom they spam in SC-Adult1, where all the recipients in the campaign, *i.e.,* the email addresses, are sorted alphabetically, and the spam source IPs are sorted in increasing order of their arrival time in the campaign. For clarity, we only show a portion of the whole campaign - 10,000 contiguous spam sources labeled from 10,000 to 20,000.

We observe a few "slanted" lines in Figure 13. Based on this we conjecture that either the Botnet controller or a "job dispatcher" is maintaining the list of recipients in an alphabetically sorted manner. Two IPs that contact the controller one after another are given out recipients from the sorted list in a first-in first-out order, and hence a slanted line joins two IPs close in time with recipient lists that are also close in the alphabetical order.

Interestingly, there are several such slanted lines in the figure. One possible explanation for this could be due to the existence of multiple "job dispatchers" (which could be located on a single machine or on several machines in an overlay as observed in the Storm Botnet [12]), where each dispatcher has a sorted recipient list of its own. Thus, IPs in a botnet contact their respective job dispatchers, and the parallel slanted lines correspond to jobs obtained from different dispatchers.

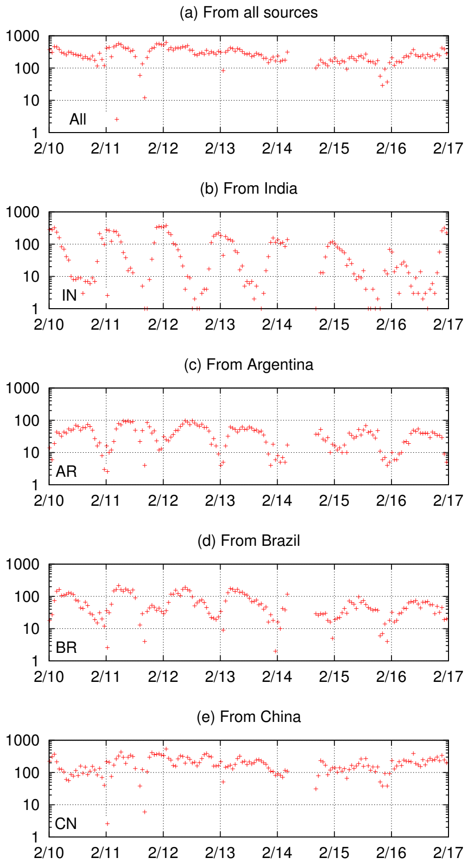

Figure 12: Connection count per IP per hour for SC-Adult1 according to all sources, sources from India, Argentina, Brazil and China.

7.3 Cross-Campaign Interactions

Next, we study cross-campaign interactions from the point of view of spam sources. Specifically, we analyze the overlap between spammers across different spam campaigns.

Do spam sources participate in different campaigns? We first find out whether there is an overlap amongst spammer IPs that participate in different campaigns. For this purpose, we define a metric, *source overlap*, to quantify the overlap in terms of spam sources across two campaigns. If set A and set B contain IP addresses spamming campaigns A and B respectively, then we define source overlap $O_{A,B}$ between these two sets as

$$O_{A,B} = \frac{\parallel A \cap B \parallel}{min(\parallel A \parallel, \parallel B \parallel)} \times 100 \qquad (1)$$

Figure 14 plots the source overlap metric for each pair of the eight spam campaigns that were identified in Table 4. From the figure we see that SC-Book does not have much overlap with other SCs, but there is a considerable overlap amongst the other SCs.

We observe that SC-Adult1 and SC-Software1 have about 48% source overlap. Earlier we had observed that nearly all the spam sources for these two campaigns finish their spam workload in the first one hour of spamming (Figure 7). We next explore what these spam sources do once they are done with sending spam for one campaign. In particular, do they immediately start sending spam for the other campaign?

In the following analysis, we carefully construct a set of IP addresses (described next) and study their properties across these two campaigns, Adult1 and Software1. Since SC-Software1 is observed for about 12 days, starting mid-February, we first prune all the spam mails in SC-Adult1 that lie outside the 12-day period (February 12 - 25) of SC-Software1. This leaves us with 22,567 total spam sources and 228,000 total SMTP connections for SC-Adult1. Second, since we are interested in spam sources that spam most of their workload in the first one hour, we further prune sources that spam more than one burst from both campaigns. Third, we extract the spam sources left after step two that were observed spamming in both campaigns. We call this set the “common-set”. We now study the properties of the 6,550 spam sources belonging to this common-set.

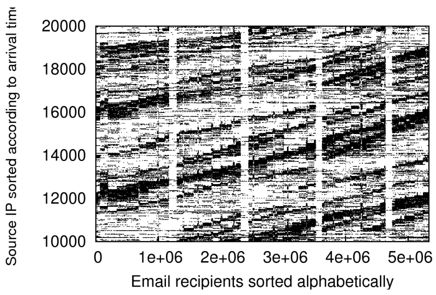

Figure 13: Scatter plot demonstrating the breakdown of workload in terms of spam recipients across IPs for SC-Adult1. Each point in the plot depicts a particular IP on y-axis spamming a recipient on x-axis.

	Book	Adult1	Adult2	Adult3	Shopping	Software2	Software1
Software1							100
Software2						100	5.00
Shopping					100	6.64	3.56
Adult3				100	23.87	76.63	15.32
Adult2			100	72.61	38.78	37.87	32.10
Adult1		100	69.24	40.20	39.99	29.77	47.79
Book	100	0.11	0.06	0	0.02	0	0

Figure 14: Pairwise “Source IP Overlap” among the seven campaigns.

How close in time do the sources spam for different campaigns? For the spam sources in the common-set, we calculated the time difference when they started spamming for SC-Adult1 and SC-Software1. Figure 15 plots the CDF of the absolute time difference between the starting time of the bursts by the common IPs across the two campaigns. We observe that about 70% of the spam sources spam the two campaigns within a period of one hour. About 85% of sources spam the two campaigns within a period of two hours. This clearly suggests that spam sources switch to the next campaign as soon as they are done with their workload in one campaign.

Do the sources spam the same amount of workload across campaigns? For the spam sources in the common-set, we calculated the number of SMTP connections (workload) they spammed for each of the campaigns. Specifically, we are interested in finding whether they spammed different amount of workload in the two campaigns. Figure 16 plots the CDF of the absolute difference of per-IP workload in the two campaigns, for every IP in the common-set. From the figure we see that 40% of IPs belonging to the common-set had a workload difference of at most one connection. About 80% of IPs had a workload difference of less than four connections. This suggests that the spamming workload of a source is a property dependent on the characteristics of the spam host and not much dependent on the campaign that it spams for.

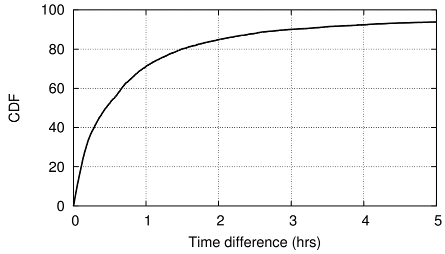

Figure 15: **CDF of absolute time difference between first spam by an IP to SC-Adult1 and first spam to SC-Software1, for all the IPs in the common-set.**

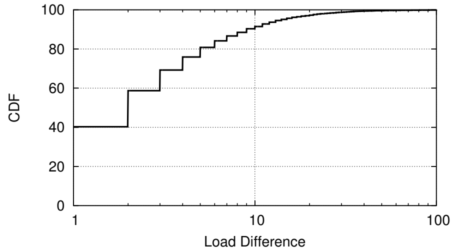

Figure 16: **CDF of absolute difference between workload for all the IPs in the common-set for SC-Adult1 and SC-Software1.**

Do spam sources spam the same recipients across campaigns? We next analyze whether the spam sources belonging to the common-set spam the same set of recipients. The 6,550 IPs belonging to the common-set made about 87,294 and 96,387 connections to our relay to deliver spam to SC-Adult1 and SC-Software1, respectively. They attempted to relay spam to 207,538 and 209,566 recipients, respectively. Out of 6,550 IPs, only three had at least one common recipient across the two campaigns. This suggests though the common IP addresses spam in close time intervals across campaigns and with the same workload, they do not spam the same mailboxes.

8. CONCLUSIONS

In this paper, we presented a trace-driven analysis that characterizes the burstiness and distributedness of botnet spam campaigns. To enable our study, we manually identified seven major URL-based botnet spam campaigns containing 2,042 distinct URLs and 2.09 million SMTP connections from a five-month trace captured by a spam relay. Our study shows that URL-based campaigns can be prolonged, sometimes lasting for periods of 99 days, suggesting that burstiness can not be used as a necessary criteria to assist clustering URL-based spam campaigns. Indeed, we showed burstiness-based campaign generation with a cutoff of five days led to a high false negative ratio (98.21%). Our finding suggests that despite recent advances, the problem of campaign identification with reasonable false negatives is harder than previously thought.

We further studied the characteristics of three botnet campaigns. Though campaigns are found to be long lasting, individual bots can exhibit a bursty behavior, with bots for one campaign (Book) contacting the relay in multiple bursts (18% of bots arrive in 10 or more bursts) with each burst lasting a short duration of 100-1,000 seconds. We also studied the coordination across spammers in a campaign and for one campaign (Adult1), we identified a pattern where the spammers that arrived at the relay in contiguous time were found spamming recipients that were in alphabetically sorted order, suggesting FIFO job dispatching by the spam coordinators. Finally, we studied the interactions among the bots that spam on behalf of multiple campaigns and found that such common bots generate spam for each campaign in close-by time, with the same workload per campaign, but to distinct recipients across campaigns.

In future work, we plan to study the interactions between spam campaign identification schemes and IP blacklisting. On one hand, bots identified in botnet spam campaigns can help to enhance the freshness of IP blacklists. On the other hand, the IPs already in the IP blacklists can potentially help to speed up the detection of new or evolving campaigns. Finally, we plan to develop automated schemes that exploit the spam content in addition to spam labels for accurate and timely campaign identification and signature generation.

Acknowledgments: We thank the anonymous reviewers for their helpful comments. We are especially grateful to our shepherd, Martin Arlitt, whose detailed comments drastically improved the presentation of this paper. This work was supported in part by NSF grant CNS-0738419.

9. REFERENCES

[1] Route Views Project Page. http://www.routeviews.org.

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

[3] Bl: Spamcop blocking list. http://bl.spamcop.net.

[4] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. *Comput. Netw. ISDN Syst.*, 29(8-13):1157–1166, 1997.

[5] CBL: Composite Blocking List. http://cbl.abuseat.org/.

[6] R. Clayton. Do zebras get more spam than aardvarks? In *Proc. of CEAS*, 2008.

[7] M. Dischinger, A. Haeberlen, K. P. Gummadi, and S. Saroiu. Characterizing residential broadband networks. In *Proc. of ACM SIGCOMM IMC*, 2007.

[8] SORBS: Spam and Open-Relay Blocking System. http://dnsbl.sorbs.net.

[9] DSBL: Distributed Sender Blackhole List. http://list.dsbl.org.

[10] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: an empirical analysis of spam marketing conversion. In *Proc. of ACM CCS*, 2008.

[11] M. Konte, N. Feamster, and J. Jung. Dynamics of online scam hosting infrastructure. In *Proc. of PAM*, 2009.

[12] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the spam campaign trail. In *Proc. of USENIX LEET*, 2008.

[13] Maxmind - ip geolocation and online fraud prevention. http://www.maxmind.com/.

[14] Njabl: Spam blocking blacklist. http://www.njabl.org/.

[15] A. Pathak, Y. C. Hu, and Z. M. Mao. Peeking into spammer behavior from a unique vantage point. In *Proc. of USENIX LEET*, 2008.

[16] Pbl: The policy block list. http://www.spamhaus.org/pbl/.

[17] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *Proc. of ACM SIGCOMM*, 2006.

[18] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proc. of ACM CCS*, 2007.

[19] Super webscan. http://www.sharewareconnection.com/super-webscan.htm.

[20] Joe st sauver: Evolving methods for sending spam and malware. http://www.ftc.gov/bcp/workshops/spamsummit/presentations/Evolving-Methods.pdf.

[21] The spamhaus project. sbl-xbl.spamhaus.org.

[22] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: signatures and characteristics. In *Proc. of ACM SIGCOMM*, 2008.

[23] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, I. Osipkov, G. Hulten, and J. Tygar. Characterizing botnets from email spam records. In *Proc. of USENIX LEET*, 2008.

[24] 2006 spam trends report: Year of the zombies. http://www.commtouch.com/downloads/Commtouch_2006_Spam_Trends_Year_of_the_Zombies.pdf.