The Harvester, the Botmaster, and the Spammer: On the Relations Between the Different Actors in the Spam Landscape

Gianluca Stringhini, Oliver Hohlfeld†, Christopher Kruegel, and Giovanni Vigna
Department of Computer Science, UC Santa Barbara
†RWTH Aachen University
{gianluca, chris, vigna}@cs.ucsb.edu, oliver@comsys.rwth-aachen.de

ABSTRACT

A spammer needs three elements to run a spam operation: a list of victim email addresses, content to be sent, and a botnet to send it. Each of these three elements are critical for the success of the spam operation: a good email list should be composed of valid email addresses, a good email content should be both convincing to the reader and evades anti-spam filters, and a good botnet should efficiently sent spam. Given how critical these three elements are, figures specialized on each of these elements have emerged in the spam ecosystem. Email harvesters crawl the web and compile email lists, botmasters infect victim computers and maintain efficient botnets for spam dissemination, and spammers rent botnets and buy email lists to run spam campaigns.

Previous research suggested that email harvesters and botmasters sell their services to spammers in a prosperous underground economy. No rigorous research has been performed, however, on understanding the relations between these three actors. This paper aims to shed some light on the relations between harvesters, botmasters, and spammers. By disseminating email addresses on the Internet, fingerprinting the botnets that contact these addresses, and looking at the content of these emails, we can infer the relations between the actors involved in the spam ecosystem. Our observations can be used by researchers to develop more effective anti-spam systems.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Abuse and crime involving computers

Keywords

Botnets; Spam; Cybercrime; Underground Economy

ASIACCS'14, June 2-3, 2014, Kyoto, Japan.

http://dx.doi.org/10.1145/2590296.2590302.

1. INTRODUCTION

Email spam is a wealthy business for cyber criminals. Recent research showed that a successful spam campaign can generate revenues between $400,000 and $1,000,000 [11, 12]. Given this profitability, a whole economy has emerged around email spam. Similar to legitimate economic ecosystems, many different parties are involved in a spam campaign. On the one side, the spammer needs to have a good list of target email addresses for the campaign to be effective, as well as a botnet able to efficiently send emails to them [21]. On the other side, spammers need an effective infrastructure to sell the illicit goods that they advertise. This infrastructure includes the websites that sell the goods, the shipping facilities, and the payment processors [15]. Thus, there can be three main parties involved in the spam ecosystem: the email harvester, the botmaster, and the spammer.

Studying the relationship among these different parties involved in the spam ecosystem deepens the understanding of the spam underground economy and can pave the way for new spam mitigation techniques. In this way, it first helps to estimate the magnitude of the spam problem and can reveal new trends. Second, it allows to identify bottlenecks and critical points in the spamming pipeline; these critical points can be used to develop mitigation techniques to fight such threats. For these reasons, previous work analyzed *individual* aspects of the parties involved in the process. In particular, researchers studied the harvesting process of email addresses on the web [9, 20], the structure and operation of spamming botnets [5, 18, 21], or the email templates used by spammers [13, 14]. Other work focused on studying the financial conversion of spam [11, 12] or the workflow that goes from when an illicit good is purchased to when it gets delivered [15]. These recent advances in the understanding of individual parties now open the question on their *relationship*.

To the best of our knowledge, no research has focused on the operational relations and the interactions among the different parties in the spam ecosystem. Some work addressed *economic* interactions of some of the different players on the underground market [21]. This work suggested that spammers buy email lists from email harvesters, rent botnets from botmasters, and then use them to send spam. However, the question on their *operational* relation and the interactions among *multiple* parties has not been answered yet. Thus, a complete understanding of relations and interactions of players in the spam ecosystem is still missing. Open research questions include the following aspects. Do some spammers

harvest email addresses by themselves? Do some spammers rent multiple botnets to send the same type of spam? And if they do, do they use the same email list across different botnets?

This paper presents the first analysis of the relations among email harvesters, botmasters, and spammers. In an attempt to contribute to answering these questions, we run a multi-step experiment. In the first step, we set up a large number of email addresses, each pointing to a mailserver under our control and advertise them on web pages. Then, we record the accesses to those web pages, to fingerprint the email harvesters. We then log the connections that we receive on our mailserver. Since the email addresses that we disseminated on the web are not used for legitimate purposes, we assume that any connection that we receive is generated by a botnet (or by a mailserver operated by spammers). We then apply a technique known as *SMTP dialects* [22] to assess which botnet or mailserver generated each connection. As a last step, we analyze the content of the spam emails that we receive, and group them into *spam campaigns*. Our assumption is that a single spammer will be responsible for each spam campaign. Such assumption has been confirmed by previous work [21]. After having logged this information, we compare the different datasets, checking whether the same spammer has rented multiple botnets, and whether multiple spammers share the same email list or botnet.

The results of this study give us new insights into how spammers operate. In particular, our findings suggest that spammers typically rent a single botnet and that a fraction of them set up their own mail transfer agents (MTAs) to spread spam. Another interesting discovery is that spammers tend to stick with a single list of email addresses for long periods of time, even years.

In summary, this paper makes the following contributions:

- We perform a large-scale experiment that tracks how email addresses are harvested, which botnets are contacting the harvested addresses, and what type of spam they are sending.
- We provide detailed statistics about the email harvesters, the spamming botnets, and the spam campaigns that we observed. We make novel observations on the way email harvesters operate and on the geographic distribution of the bots in large botnets.
- We analyze the relations between email harvesters, botmasters, and spammers. We show that, with rare exceptions, spammers purchase their email lists from professional email harvesters and rent a single botnet to send spam.
- We discuss how our observations can aid researchers in spam mitigation. We argue that the consistent habits of spammers in terms of the email lists and botnets they use can be leveraged for detection.

2. METHODOLOGY & DATA COLLECTION

The analysis infrastructure that we used in our measurement is composed of three parts:

1. **Email Harvester Logging.** In this part, we first advertise a large number of unique email addresses on the web. Each email address is only displayed once and points to a mailserver under our control. To be able to correlate received spam with harvesting information, we log web page retrieval logs for each issued email address. This approach thus allows us to identify which harvester fetched a certain email address.
2. **SMTP Dialect Fingerprinting.** Previous research showed that each botnet, email client, and mail transfer agent (MTA) uses a different implementation of the SMTP protocol [22]. This allows us to fingerprint the email engine of a host that is talking to a mailserver. We leverage this technique to fingerprint the different botnets and MTAs that send emails to the harvested email addresses.
3. **Spam Campaign Analysis.** To avoid easy detection, spammers slightly alter the content of their spam emails over time. However, previous research showed that it is possible to group spam campaigns by looking at the domains of the URLs that are advertised in the email body [27]. We leverage a similar technique, and identify spam campaigns that are carried out by the same spammer.

Site	Description	# of Addresses	# of Emails
A	Private blog	1	72
B	Gaming web site	274	1,511
C	Mail archive	187	387
D	Private web page	4	74
E	Spamtrap page	71	153

Table 1: Number of advertised email addresses (ids) and number of received spam emails per web site. As it can be seen, the gaming web site is the one whose addresses received the most spam.

In the remainder of this section, we describe the three parts of our data analysis infrastructure in detail.

2.1 Email Harvester Logging

We identify email harvesters by using a methodology that relies on issuing unique spamtrap email addresses via the web [9]. As the addresses are uniquely generated for each page request, their usage can be directly mapped to a specific page request once the first spam email is received. To allow for this mapping, we log basic information such as the requesting IP addresses, timestamps, and HTTP header information for all page requests. This per-request information allows us to analyze further properties such as the user agent strings submitted by the harvester bots.

The generated addresses are embedded into nine low-profile web pages of various type. Table 1 provides a description of the websites and statistics about the advertised email addresses. This methodology is implemented in web sites by including a server-side dynamic script that generates unique email addresses for each page request and logs information about the visitors. Each web site advertises six different spamtrap addresses, each being displayed with one of the following presentation and obfuscation techniques: *i*) a *mailto:* link, *ii*) non-linked, plain-text address, *iii*) address obfuscated in the form of *user [at] domain [dot] tld*, *iv*) address obfuscated using JavaScript code, *v*) address included in a hidden data field of a web form, and *vi*) plain-text address inside an HTML comment. All of the above described addresses consist of random strings of 10 characters each (*RND*

IDs, e.g., "jdi4gj8bzx"). We use random strings as they are difficult to guess and, therefore, we can be confident that a spammer who targets those addresses obtained them from the harvesters, and did not randomly guess them. In addition to random strings, we issue realistic looking addresses containing random combinations of first and last names generated from phone book records (e.g., "john.doe"). Compared to random strings, the assumption is that realistic looking addresses are harder to identify as fake addresses, but are also easier to guess.

Email addresses are advertised by appending different domains and Top Level Domains (TLDs). Our email domains are handled by several mail exchange servers located in different networks. These servers provide us with the unfiltered email feed via IMAP. We consider any email sent to those addresses as spam. As our SMTP dialect classification (see next section) relies on detailed SMTP transaction logs, we needed to capture detailed traffic traces at each mail exchange server. Unfortunately, this was only possible at one server due to administrative restrictions in other networks.

The data collection started on December 14, 2012, and ended on May 15, 2013. In this period, the system received 2,197 spam emails sent to 613 unique spamtrap addresses. The summary of the number of emails received per web site is reported in Table 1.

During the measurement period, the mail exchange server also received 1,299 emails sent to 75 email addresses that were not advertised by the system. Out of this set, 115 emails were addressed to `admins@foo.tld`, `info@foo.tld`, `contact@foo.tld`, `contactus@foo.tld`, and `sales@foo.tld`, where `foo.tld` denotes the domain advertised on the web and assigned to the mail exchange server. The remaining 1,184 emails were addressed to external email domains, including `163.com`, `gmail.com`, and `yahoo.com.tw`. As none of our mailservers is configured as an open relay, these mails were declined.

Despite the size small size of the dataset, it enabled us to observe interesting interactions between the different actors in the spam landscape. Our results therefore provide a first step in getting an understanding on how different parties involved in the spam process cooperate, which ultimately aims at gaining a better understanding of the online underground economy.

2.2 SMTP Dialect Fingerprinting

In our previous work, we showed that it is possible to reliably fingerprint an SMTP client by analyzing the SMTP messages that it exchanges with the mailserver. More precisely, each SMTP implementation, both in legitimate programs and in malware, shows differences in the way it implements the SMTP protocol [22]. We call these variations of the SMTP protocol *SMTP dialects* [22]. To identify different botnets, we extract the SMTP dialect spoken by each client trying to send an email to the mailserver.

The SMTP protocol is defined as an alternating dialogue between a client and a server. The messages sent by the client are called *SMTP commands*, while the messages sent by the server are called *SMTP replies* [1]. The client first specifies the sender of the email (in the form of an email address), one or more recipients, and then asks for permission to send the actual email content (with a `DATA` command). If the server grants this permission, the client starts transmitting the content of the email, otherwise the connection is aborted (typically because of an error). To study the SMTP dialect spoken by a client, we are only interested in the sequence of commands and replies until the `DATA` command is issued, or an error is reported.

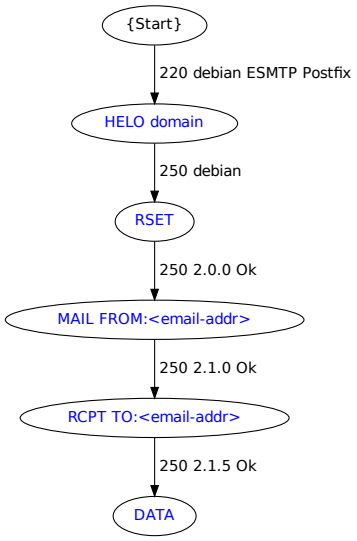

Figure 1: An example of an SMTP dialect. The transitions are labeled with server replies, while the states are labeled with client commands. Reaching the `DATA` command represents and email correctly being sent.

In our previous work, we defined an SMTP dialect **D** as a state machine

$$\mathbf{D} =< \Sigma, S, s_0, T, F_g, F_b >,$$

where Σ is the input alphabet (composed of server replies), S is a set of states, each labeled with a client command, s_0 is an initial state, and T is a set of transitions. $F_g \subset S$ is a set of good states, representing an email being processed correctly, while $F_b \subset S$ is a set of bad states, representing an error in the email sending process. Since SMTP commands include variable fields (such as email addresses), we abstract specific fields of the commands into regular expressions. In particular, we substitute fully-qualified domains, IP addresses, host names, and email addresses with generic identifiers. This operation is described in detail in our previous paper [22].

To build an SMTP dialect, we capture the SMTP conversations between a client and the mailserver and iteratively build the state machine, as described in [22]. Figure 1 shows an example of an SMTP dialect. The transitions represent server replies, while the states represent client commands. In this particular case, the client first announces itself and its domain (`HELO` command), then sends a `RSET` command, specifies the sender and the recipient email addresses (with the `MAIL` and `RCPT` commands, respectively), and starts sending the content of the email.

2.2.1 Learning the Dialects Spoken by Botnets

As we said, we are interested in understanding what botnet or MTA generated the emails that have been sent to the mailserver. To this end, we need previous knowledge about which SMTP dialects are spoken by different botnets and MTAs. To accomplish this task, we analyzed the SMTP conversations generated by the malware samples submitted to *Anubis*, a popular sandbox malware analysis system [2]. For each malware sample, we first learned the SMTP dialect spoken by it. Then, we grouped together all the samples that speak the same dialect.

We analyzed the malware samples submitted to Anubis for the period between January 1, 2013 and May 15, 2013. In total, 18,849 submitted samples connected to a mailserver and started an SMTP conversation. These samples include spamming botnets, self-propagating worms that send a copy of themselves over email, and generic malware that sends an email to their botmaster as a heartbeat signal. The sandbox environment did not allow any SMTP connection to the outside, but redirected them to a local mailserver. This way, we were able to log the SMTP conversations generated by the malware samples. In total, the malware samples that we analyzed speak 72 different SMTP dialects. As we will show in Section 3.2, these dialects cover the vast majority of the hosts that contacted the email addresses that we disseminated on the web.

To assign a name to each dialect, we proceeded as follows. First, we grouped all the malware samples that speak the same dialect together. Then, we collected the name given to those malware samples by popular antivirus programs, by using `Virustotal` [25]. Finally, we assigned as a name for the malware family the most common label assigned to the samples that speak that dialect by the antivirus programs in `Virustotal`. This approach has been already used in previous work, and it proved to be reliable [22].

Not all spam is sent by botnets though [24]. Some spammers set up their own mailservers and use them to send spam, instead of setting up botnets. To identify this type of spam, we need to learn the dialects spoken by popular Mail Transfer Agents (MTAs). To this end, we set up a number of virtual machines, and installed a different MTA on each of them. More precisely, we used Virtualbox as our virtualization environment — we set up Ubuntu Linux 11.10 virtual machines, and set up one of the Exim, Postfix, Qmail, and Sendmail MTAs on each of them; we also set up a Windows Server 2008 virtual machine, and ran Microsoft Exchange 2010 on it. On each virtual machine, we set up a script that sent emails automatically to a mailserver under our control. We then leveraged the SMTP conversations generated by the virtual machines to learn the dialects spoken by each MTA. As we already noted, each MTA speaks a different dialect from the others [22].

2.3 Spam Campaign Analysis

We call a spam operation a *spam campaign*. Broadly speaking, a spam campaign is composed by a set of emails that advertise the same product. Typical goods advertised in spam emails are pharmaceutical products, counterfeit goods, dating sites, and others [21]. In this paper, we consider a spam campaign as being indicative of a single spammer. This assumption is motivated by the fact that previous work showed that each spammer sets up his/her email templates when sending spam [21]. In principle however, the same spammer might run multiple spam campaigns at the same time. We perform a more comprehensive analysis of these assumptions in Section 3.4.

Even if they advertise the same type of goods, the spam emails belonging to the same campaign are not identical. Spammers add some variations in the content, in the subject lines, and in the advertised URLs to avoid easy detection by template-based anti-spam techniques [19, 27]. For this reason, we have to adopt more advanced similarity techniques to group spam emails into spam campaigns. We extract four features that characterize an email; in the following, we describe them in detail. We consider two emails as belonging to the same campaign if they match any of these four criteria.

Subject Line. Subject lines are important in spam emails, because they are the first piece of information seen by a victim, and might lure them into opening the full email. Spammers need to make their subject lines captivating, but also vary them enough to avoid easy detection. We consider two emails as belonging to the same spam campaign if their subject lines are either identical or share four or more words.

URL domain. Miscreants change the domains that they use to host their malicious web pages quite often. Since purchasing domains has a non-negligible cost, however, they use each of them for multiple emails. For this reason, we consider two emails as belonging to the same spam campaign if they advertise a URL that shares the same domain. A similar technique has been leveraged by previous research [27].

Mailer. Spammers often use fake mail user agent strings in the email headers (mailer). This mailer is often the same for the emails belonging to the same campaign [21]. Therefore, we group together emails that share the same mailer.

Sender email address. From our observations, we noticed that most spammers set the `From` address in their emails. We also observed that this from address is often shared by multiple emails in the same campaign. For this reason, we group together emails that share the same sender address.

2.4 Assumptions & Limitations

The methodology that we follow in this paper is based on a set of assumptions. In the following, we discuss these assumptions in detail.

Different actors in the spam chain engage in a market economy. We assume the presence of three different roles, i.e., the harvester, the botmaster, and the spammer. As suggested by previous work [21], these roles are likely to be represented by different entities that engage in a market economy. For example, email addresses are collected by the harvester and then sold to spammers who rent botnets for the spam dissemination. In other cases, the spammer and the harvester might represent the same entity.

Different usage pattern in the harvesting and spamming process are indicative for the different actors. While we cannot observe the different actors directly, we assume that they can be fingerprinted by different observable usage pattern. For instance, short turnaround times in which spam arrives almost immediately after harvesting [9, 20] can be *indicative* for cases in which the harvesting is performed by the spammer. In turn, long turnaround times of up to multiple years, however, can be *indicative* of cases in which email addresses were sold on the market. Conversely, spammer and harvester are likely to be two in-

dependent entities. We note that this first step provides *indications* rather than strict evidence.

Email addresses are being harvested from the web. In this paper, we only focus on e-mail addresses that were harvested from public web pages by using crawlers. We remark that there are other ways to harvest email addresses, such as malicious software locally running on compromised machines [14]. These further harvesting techniques are, however, out-of-scope of this paper.

SMTP dialects are indicative for different botnets. To identify different botnets, we rely on SMTP dialect fingerprinting. We assume that this technique can reliably distinguish between malware families that send spam or between misused MTAs. This assumption is supported by our previous work, in which we show that all the SMTP implementations that we encountered in the wild are different [22].

A single campaign is indicative for a single spammer. Spam campaigns advertise goods, point to scam web pages, or disseminate malware. We assume that such campaigns are run by an entity that we refer to as a spammer. The spammer himself remains invisible to us and only the launched campaigns can be observed. Previous work showed that spammers use individual templates for composing spam mails [21] and thus supports our hypothesis. In case of a joint campaign run by multiple spammers, individual entities cannot be differentiated by our approach and are merged into distinct observable campaigns. Thus, we cannot derive exact figures on the number of spammers observed.

The assumed behavior of the harvester, the botmaster, and the spammer are based on indirect measurements of both the harvesting and the spamming process. The nature of the applied indirect measurements yields correlations among the behavior of different actors in the spam chain. However, our approach does not allow us to derive *exact* figures and causal relationships for the involved actors. Despite these limitations, this work provides a novel first step in understanding these relationships.

3. ANALYSIS OF THE COLLECTED DATA

In the following, we first analyze the harvesters that fetched our email addresses. Then, we analyze the botnets (or MTAs) that sent the emails and we study the spam campaigns that targeted our email addresses. Finally, we discuss the relations among the three actors involved in the spam delivery process (harvesters, botnets, and spammers).

3.1 Analysis of the Harvesters

In total, the spamtrap email addresses in our dataset were harvested by 75 unique IP addresses. Note that the number of IP addresses does not map to the number of harvesters. In distributed infrastructures, for instance, multiple IP addresses can belong to the same harvesting entity. A distributed infrastructure helps in two ways: it is more efficient, and it is stealthier. From what we observed in our experiments, however, the main activity concentrates on a small set of IPs. In particular, four IP addresses harvested 70% of the email addresses, which ended up receiving 74% of the total spam.

To classify the harvesters that crawled the web sites, we manually analyzed the dataset. We consider two IP addresses as belonging to the same harvester if *i)* they are located in the same autonomous system or if *ii)* they announce the same user agent string or share patterns in the HTTP

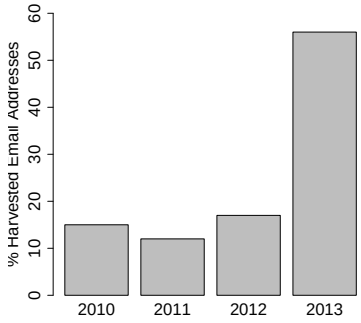

Figure 2: Harvesting year. As it can be seen, most addresses have been harvested during 2013.

headers. While several distinct harvester bots can run in the same AS, we consider the submitted user agent string pattern to be one characteristic of the harvesting software being used.

In total, we observed nine distinct harvesters, which are summarized in Table 4. Five harvesters used a single IP address, while the others were distributed. The largest distributed harvester (C) was observed on 56 distinct IP addresses located in a residential DSL access network in Germany. One unique characteristic of this harvester are random-looking user agent strings. The total number of machines used by the harvester is likely to be different from the number of IP addresses, as residential users in this network get new dynamic IP addresses assigned every 24 hours. While the 56 IP addresses could in theory belong to the same physical machine, we observed several parallel crawling activities from multiple IP addresses. This indeed suggests a distributed harvesting infrastructure.

Interestingly, the size of the harvester, estimated by the number of IP addresses, does not correlate with harvesting activity and spam volume received by the harvested email address: The largest number of email addresses was collected by a harvester (D) that was using only three IP addresses.

We assume user agent strings to be a main characteristic of a harvesting software. Table 4 also shows example user agents for the email harvesters that we observed. Some of the user agent strings mimic legitimate browsers, while others identify software libraries or consist of random strings. In some cases, such as harvester I, the user agent does not change at subsequent requests. For some harvesters, the user agent is constantly changed, arguably to avoid easy detection. At the extreme is harvester C, which sets a different random string as user agent after each request.

We observe two interesting patterns in the harvesters that contacted us. Firstly, the Java user agent used by harvester D correlated with high harvesting activity and caused large spam volumes. Earlier observations of this user agent in the context of harvesting [9, 20] suggest that this harvesting software has existed for at least eight years and thus to be a stable pattern in the harvesting landscape.

As we observed earlier [9], in some cases addresses that we returned only to harvesters presenting the user agent of the Google bot received spam. This denotes the case of harvester F that not only presents a legitimate user agent, but

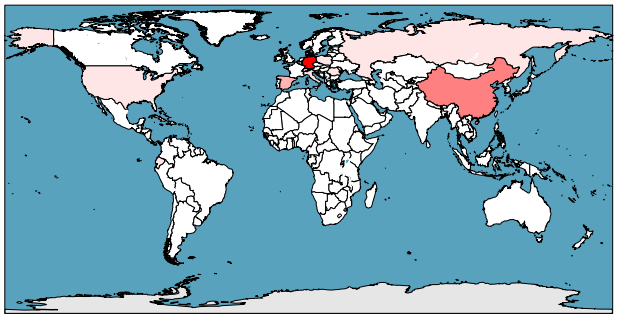

Figure 3: Country distribution of the email harvesters. 73% of the harvester IP addresses were located in Germany, while 9% of them came from China.

also originates from the Google AS. It suggests that harvesters use search engines as a proxy to either *i)* hide their own identity or *ii)* optimize the harvesting process itself. We identified harvesting software that offers the functionality of querying search engines. For example, the advertisement for ECrawl v2.63 [17] states: "Access to the Google cache (VERY fast harvesting)," while the description of the Fast Email Harvester 1.2 reports that the "collector supports all major search engines, such as Google, Yahoo, MSN" [6].

It is interesting to look at where the harvesters are located. By using the Maxmind geolocation database [16], we map harvester IP addresses to geolocation. We additionally map harvester IP addresses to AS numbers. We observe harvesters from ten countries (eleven Autonomous Systems), respectively. Figure 3 shows the country distribution of the IP addresses that harvested the email addresses that we disseminated. A darker color in the map represents a larger fraction of IP addresses from that country. As it can be seen, Germany is the country hosting the largest number of harvesters (73% of the total), followed by China (9%), and Spain (5%). The prevalence of German IP addresses can be explained with the fact that the IP addresses of harvester C, which is composed of 56 IP addresses, are all located in Germany. On the other hand, the IP addresses of harvester D, which was the one that was able to gather the highest number of IP addresses, are located in the Netherlands.

We next focus on harvesting activity periods. The harvesting time of email addresses that received spam during the measurement dates back to 2010. We show the number of addresses harvested per year in Figure 2. We split the harvesting activity in activity periods. We define activity periods as periods of time in which the harvester connected to our pages at least once in each of two consecutive hours. The number of activity periods with their respective start and end time is shown in Table 2. The harvesting behavior that we observed during an activity period was typically intense, with a larger number of crawled web sites in a small time period of several seconds to minutes.

We investigated what happens to e-mail addresses after they have been harvested by focusing on the usage of harvested addresses. Concretely, we denote the time between the address being harvested and the first spam email received at that address as the *turnaround time* and show their median value in Table 3. Harvester D and G show

Harvester	Activity Periods	Start	Stop
A	2	2010-07-02	2010-07-02
B	1	2010-10-30	2010-10-30
C	41	2010-02-10	2010-10-20
D	3	2011-01-08	2013-04-06
E	1	2011-01-15	2011-01-15
F	1	2013-03-29	2013-03-29
G	1	2012-08-30	2012-08-30
H	4	2012-10-19	2012-12-05
I	4	2011-08-27	2012-02-16

Table 2: Activity periods per harvester. As it can be seen, some harvesters were active in bursts for short periods of time, while others were constantly observed for very long periods.

Harvester	Median Turnaround Time
A	26 days
B	15 days
C	343 days
D	5 days
E	548 days
F	35 days
G	5 days
H	42 days

Table 3: Email turnaround times. This table shows after how many days from the harvesting the first spam email was received.

the fastest turnaround of 5 days, while medium turnaround times of less than two months are observed for harvester A, B, F and H. Harvester C and E show the longest turnaround time of more than one year. Long turnaround times suggest emails being sold on the market and used by entities other than the harvester. As previous research showed [21], the purchase of email lists on underground sites is heavily influenced by the reputation of the creator of the list. A reason for longer turnaround times might be that the reputation of the harvester is not yet established, and therefore spammers are less likely to purchase his lists.

3.2 Analysis of the SMTP Dialects

We learned the SMTP dialects for all the clients that sent emails to the mailserver, as described in Section 2.2. We discarded any SMTP conversation that generated an error, or for which the client abruptly closed the connection. Therefore, each SMTP conversation analyzed in this section corresponds to an email being delivered to the mailserver. We logged 2,024 correctly sent emails in total.

Our system identified seven different dialects among the clients that sent emails to the mailserver. A summary of these results is shown in Table 5. As it can be seen, the mailserver was targeted by three of the largest active spamming botnets (*Cutwail*, *Lethic*, and *Kelihos*). It was also targeted by *MyDoom*, which is a generic name used by antivirus companies to refer to email-spreading worms. Our mailserver was also contacted by two types of MTAs, set up by miscreants to send spam (*Postfix* and *Sendmail*). Note that, in principle, multiple spammers might use the same botnet, or set up the same MTA to send spam. We discuss this possibility in Section 3.4.

The *modus operandi* of the spammers using each botnet, worm, and MTA setup is rather different. For instance, the spammers using Lethic did not send emails only to harvested email address, but also to generic ones (e.g., `info`, `admin`). All other setups leveraged the email addresses harvested by

Harvester	# of IPs	# of Email Addresses	Sample User Agent
A	1	2	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; Alcohol Search;)
B	1	1	Mozilla/4.0 (compatible; Synapse)
C	56	60	r5wRofjnmtbqqrea5igfhmjisyqjikweoepo
D	3	415	Java/1.6.0_04
E	1	2	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0;)
F	1	2	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
G	1	20	libwww-perl/6.04
H	2	13	Mozilla/5.0 (Windows; U; Windows NT 6.1; rv:11.0)
I	9	20	Mozilla/4.0(compatible; MSIE 5.0; Windows 98; DigExt)

Table 4: Summary of the observed email harvesters. For each email harvester, we include the number of IP addresses used, as well as the number of email addresses collected by the harvested, and a "sample" user agent. For those harvesters that vary their user agent, we included a random one among the ones used.

Botnet or MTA	Harvesters	Campaigns Sent	# of Email Addresses	# of Emails	# of IPs
Cutwail	A,B	A,B	3	244	243
Kelihos	C	C	5	8	5
Lethic	D,None	D	29	533	101
Mydoom	C,D,E,F,G	*	431	1,191	52
Postfix	C,D,H	E,F,G	41	43	5
Sendmail	B,C	H	5	5	4

Table 5: Summary of the SMTP dialects, the email lists, and the spam campaigns observed. The horizontal line separates dialects that belong to botnets or generic malware from those that belong to regular Mail Transfer Agents (MTAs). "None" means that the botnet sent emails to email addresses that were not harvested, but that are popular default addresses (e.g., `admin`). We omitted the list of campaigns spammed by Mydoom, for simplicity.

one or more harvesters. The way in which these addresses are used varies according to the type of setup too. In general, botnets tend to send several emails to each harvested address, while MTAs limit themselves to one email per address, on average.

Table 5 also shows the number of IP addresses that we observed belonging to each botnet and MTA. As it can be seen, botmasters use different strategies while managing their bots: Each email sent by Cutwail came from a different IP address (i.e., bot), while Lethic bots are reusing their email addresses. The time at which the different botnets or MTAs contacted the mailserver also varies. In particular, the Cutwail botnet constantly sent spam emails to our harvested addresses during the observation period, sending a small number of emails per day, while Lethic started contacting the mail server in April 2013, sending a higher number of emails. Mydoom, on the other hand, focused its activity during March 2013. A summary of the activity of the various botnets and MTAs that contacted us is pictured in Figure 4. The size of the circles is proportional to the number of emails sent by each botnet or MTA during that day.

As a last element, the different botnets show a very different country distribution of their bots: Figure 5 shows the country distribution for the Cutwail botnet. As it can be seen, most Cutwail bots are located in Korea (18% of the total), India (13%), and Serbia (9%). Lethic, on the other hand, has 92% of its bots located in South Korea (see Figure 6). This country distribution is not necessarily representative of all Cutwail and Lethic instances: as previous research showed, spammers rent single instances of command and control servers and buy their malware installations separately [3,21]; spammers can go as far as selecting the countries in which they want their bots to be located. Therefore, our observation suggests that the botnet users that sent spam to us purchased their bots in a small number of countries. Other instances (and customers) of the same botnet might show very different country distributions. The fact that each spammer uses bots located in different countries is consistent with previous work, which showed that the physical location of a bot does not influence the overall spamming performance of the botnet [10].

On the other hand, the Mydoom worm has most of its victims in Poland (19% of the total IP addresses) and in the United States (17%) (see Figure 7). We omit Kelihos because this botnet sent a very small amount of spam, and the map would not be meaningful.

We also wanted to understand in which countries spammers set up their mailservers to send spam. Figure 8 shows the country distribution of the Postfix and Sendmail installations that contacted the mailserver. Two servers were located in Russia, while the United States and Canada hosted one server each. The United Kingdom and Spain also hosted one server.

3.3 Analysis of the Spam Campaigns

We applied the clustering technique described in Section 2.3 to the emails that we received. In total, we obtained 63 spam campaigns. Table 6 reports a summary of some of these campaigns. We omitted the 55 spam campaigns carried out by Mydoom. We suspect that Mydoom might be a generic label that antivirus companies give to unknown malware samples, and therefore analyzing the different spammers grouped under this label is not very meaningful. For this reason, we did not analyze this botnet any further.

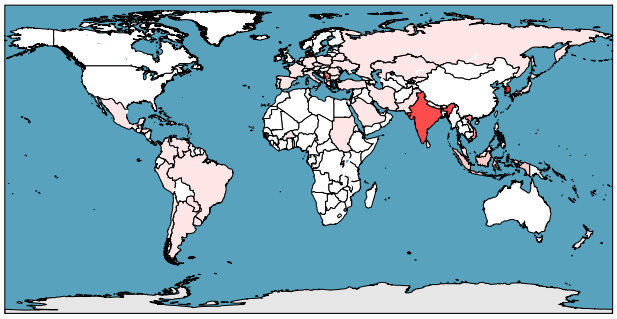

Figure 5: Country distribution of the Cutwail botnet. Most bots are in South Korea (18%), followed by India (13%).

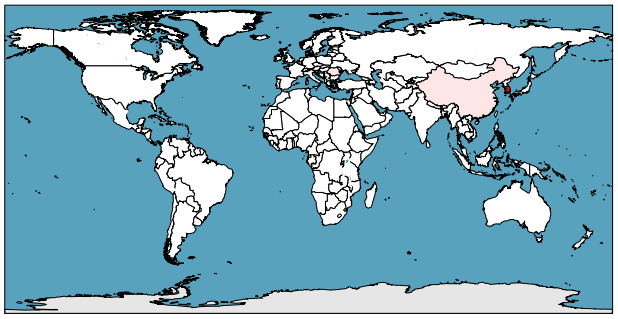

Figure 6: Country distribution of the Lethic botnet. The vast majority of the bots are located in South Korea (92%).

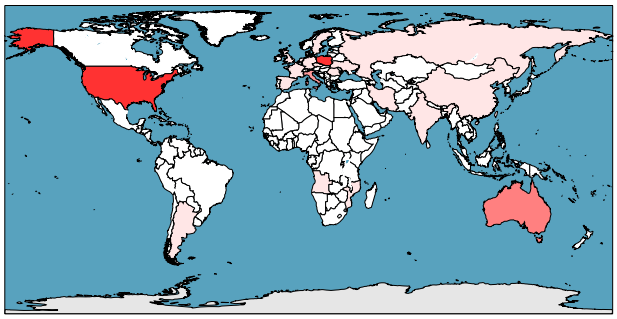

Figure 7: Country distribution of the Mydoom worm. The countries with most bots are Poland (19%) and the United States (17%).

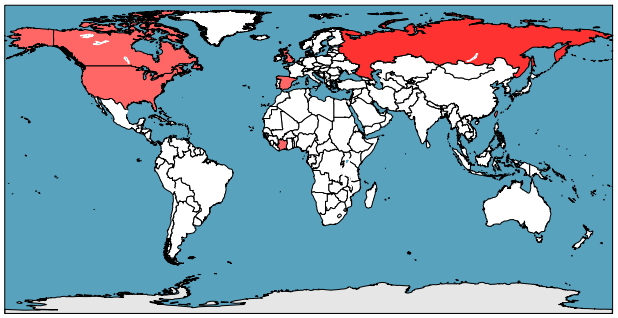

Figure 8: Country distribution of the MTAs used to send spam (Postfix and Sendmail). Interestingly, there are a number of rogue mailservers located in Western Europe and North America.

Campaign	# of Emails	Topic
A	64	Counterfeit goods
B	180	Online dating
C	8	Financial scam
D	533	Search Engine Optimization
E	7	Email marketing
F	6	Phishing scam
G	30	Phishing scam
H	5	Phishing scam

Table 6: Summary of the observed spam campaigns. The topic of the campaign was selected by manual analysis of the spam emails. We omitted the campaigns carried out by Mydoom for simplicity.

As it can be seen in Table 6, the spam campaigns that we logged cover a large variety of goods and services. Surprisingly, we did not observe any email advertising pharmaceutical products, which has been the focus of the majority of the underground economy research in the past [11]. While this might arguably be an artifact in our dataset, it might also suggest that spammers are moving on to exploring other ways of generating revenue, such as blackhat Search Engine Optimization (SEO). We leave a detailed investigation of changes in spamming trends for future work.

3.4 Relations Between the Actors

In this section, we discuss the relations between the different parties involved in the spam delivery process: email harvesters, spammers, and botmasters.

First, we try to understand how many spammers contacted the spamtrap system. The basic relations are reported in Table 5. As we mentioned, we consider a specific spam campaign as being indicative of a single spammer. However, a spammer might perform multiple spam campaigns, either at the same time, or at two different points in time. We consider two campaigns as being carried out by the same spammer if both the botnets or MTAs and the email lists used to carry out the two campaigns are the same.

Interestingly, all the botnets that we observed were used by a single spammer each. Lethic and Kelihos carried out a single spam campaign, while Cutwail carried out two different campaigns, at two distinct points in time. However, both campaigns used the email lists A and B. Therefore, we associate them to the same spammer. On the other hand, we identified three different installations of Postfix, each carrying out a different campaign, and each one using a different email list. We consider these three servers as being managed by three different spammers.

We next investigated whether email harvesters collect email addresses to sell them, or whether spammers are doing the harvesting themselves. The intuition here is that if the email

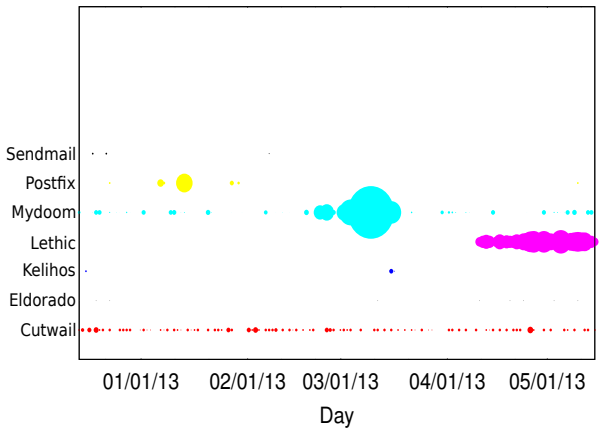

Figure 4: Email activity of the different botnets / MTAs per day. Each line represents a different dialect, and the diameter of the circles is a function of the number of emails sent by clients speaking that dialect during a given day.

addresses collected by a harvester are used by multiple spammers, they have likely been purchased on the black market. On the other hand, if a spammer is using a set of email addresses exclusively, he might have harvested them himself. Our findings lead us to speculate that harvesters C and D are likely to be ran by professional harvesters, who are selling their email lists on the black market. This is supported by the fact that multiple spammers used them for their spamming operations: the email list harvested by harvester C was used by spammers that rented Kelihos, Mydoom, Postfix, and Sendmail. Similarly, the email list sold by harvester D was used by Lethic, Mydoom, and Postfix users. The fact that multiple spammers purchased these email lists suggests that these harvesters are successfully collecting email addresses and advertising their email list. This is further supported by the fact that harvester C is the one with the largest infrastructure among the logged one (56 IP addresses), and that harvester D was the most successful one, with 415 harvested email addresses. It is also interesting that the first spam email that was received by one of the email addresses harvested by D arrived after five days from the harvesting (see Table 3). This shows that the operators of D are very efficient in harvesting email addresses and selling them on the black market. The operators of harvester B probably also sold their email list, since it has been used by both Cutwail and Postfix.

The emails harvested by two harvesters were used only by a single spammer (E, H). This might suggest that in these cases the spammers and the harvesters are the same entity, and are using the email addresses that they harvested exclusively for launching their spam campaigns. This intuition is supported by the fact that one of the Postfix installations was located in the same country as harvester H, which collected the email addresses used by that server (Spain). Interestingly, the email addresses collected by harvester I were not used by any spam setup.

4. DISCUSSION

Previous research showed that reputation is key for the success of a cybercriminal [21]. For example, an email harvester is more likely to sell his email lists if he is a respected member of the underground community, and a botmaster is more likely to rent his botnet to spammers if he has proven that his botnet works well in delivering spam. In the experiments performed for this paper, we found evidence of this behavior. Spammers seem to stick with the same email harvesters, as well as with the same botnets, for long periods of time. This suggests that spammers establish some sort of customer loyalty with harvesters and botmasters, and that this relationship hardly breaks (in the absence of major events, such as botnet takedowns).

The fact that spammers are using the same botnet and email list for long periods of time can be leveraged by security researchers for detection. For example, a system that was recently proposed observes bots as they contact different email servers, in practice fingerprinting the email list that the bots are using [23]. By observing a small number of bots that are known to belong to a botnet, this system is able to find more bots that use the same email list. Since spammers use the same email list for long periods of time, this approach could prove to be a very effective method to track spambots in the wild. Similarly, spammers that keep using the same botnet can be effectively detected by techniques that fingerprint the email engine used by such botnets [22]. Moreover, since spammers seem to rely on a single botnet at a time, taking down the botnet that they are using can have significant effects on their business. This observation makes techniques that identify command and control servers particularly important [4, 7, 26].

The fact that spammers concentrate their bots in a small number of countries could further help in the detection. Previous research showed that the geographical distance between spamming bots and their victims is higher on average than the distance between legitimate email senders and recipients [8]. Similarly, having spambots that are concentrated in few far countries can be used as a strong indicator for a mailserver performing spam detection.

As we said previously, we did not observe any pharmaceutical spam sent to the email addresses that were harvested. This is somewhat surprising, since much of previous research focused on studying spam schemes that advertise pharmaceuticals [11, 12]. Besides being caused by measurement artifacts, another reason for this discrepancy might be that pharmaceutical spam has been steadily declining over the last two years, as recent reports noted [24]. Alternatively, it could be that large pharmaceutical affiliate programs harvest their own email addresses, and that they directly provide them to their affiliates, who do not have to look for email lists on the black market.

Note that our findings are based on *correlations* among the observed behavior of different actors in the spam chain—the harvester, the botmaster, and the spammer. Despite the correlative nature of our analysis and the limited dataset size, the collected data enables us to observe interesting interactions between the different actors in the spam landscape. Thus, we see our work as a promising first step toward understanding how different parties involved in the spam process cooperate, which ultimately improves our understanding of the online underground economy. We expect additional data to deepen this understanding and plan to expand this pre-

liminary study in future work. The next steps include going beyond our correlation analysis to establishing causal relationships among the different spam actors.

5. RELATED WORK

A wealth of research has been conducted on email spam. By studying the underground economy surrounding spam, as well as the challenges that spammers and botmasters face, researchers can develop new mitigation techniques that attackers cannot easily evade. Previous research falls in two main fields: *Studying the conversion of spam* and *Studying the spam delivery infrastructure*.

Studying the conversion of spam. Studying to what extent the goods advertised in spam emails are purchased helps in dimensioning the spam ecosystem, and in understanding how much money spammers can make. Kanich et al. infiltrated the Storm botnet, and modified the spam emails sent by the botnet to point to their fake pharmaceutical site [11]. This way, they were able to track the number of users that would have purchased the counterfeit goods. In a follow-up work, Kanich et al. studied a large rogue pharmaceutical website [12]. By leveraging a vulnerability in the website that allowed to enumerate any order that was made, they estimated the size of the whole spam operation. Levchenko et al. studied the workflow of spam-advertised goods, from when a good is purchased, to when it is delivered to the customer [15]. They were able to identify the financial institutions involved in transactions related to the spam business. Although studying the economic conversion of spam is very important, it goes beyond the scope of this paper. Instead, we look at how the different entities operating in the spam delivery business operate and are related.

Studying the spam delivery infrastructure. When studying the spam delivery infrastructure, it is important to understand how the different parties involved in it operate. Previous work studied how miscreants collect email addresses on the web [9,20]. This research showed how automated harvesters operated, and suggested some simple obfuscation techniques that can prevent them to read email address, while still making them intelligible to humans.

A wealth of research has been conducted in studying the command and control (C&C) infrastructure of spamming botnets, and how botmasters manage their bots. Stone-Gross et al. analyzed a number of C&C servers from the Cutwail botnet, studying how spammers used them, and the challenges that they had to face [21]. Cho et al. infiltrated the MegaD botnet, providing interesting insights on how a large-scale botnet operates. A number of infiltration operations have been performed against peer-to-peer botnets, showing how these botnets work, and the type of spam that they send [13, 18].

Another interesting aspect is understanding how miscreants purchase infected machines. Caballero et al. [3] showed that there are complex schemes, made of malware-delivery networks that can download the payload that is required by customers on a large number of infected machines.

Although previous research suggested that there is a rich underground economy trading all the components required to set up a successful spam campaign (email lists, botnets, and malware installations) [21], no work actually studied these dynamics. Our paper provides a first look at this phenomenon, focusing on the relations between email harvesters, botnets, and spammers.

6. CONCLUSIONS

In this paper, we analyzed how the different actors involved in the spam delivery process cooperate, and what type of resources are shared among them. Our preliminary study suggests that spammers typically rely on a number of professional email harvesters to populate their email lists. Also, our findings suggest that spammers typically rent a single botnet, instead of using multiple ones at the same time. This work is a first step in understanding how spammers operate, and how the underground economy landscape look like. We hope that the insights provided in this paper will help researchers in finding the weak points in the spam delivery chain, and in developing better mitigation techniques.

7. ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (ONR) under grant N000140911042, the Army Research Office (ARO) under grant W911NF0910553, and Secure Business Austria.

8. REFERENCES

[1] RFC 821: Simple Mail Transfer Protocol. `http://tools.ietf.org/html/rfc821`.

[2] U. Bayer, A. Moser, C. Kruegel, and E. Kirda. Dynamic analysis of malicious code. *Journal in Computer Virology*, 2(1):67–77, 2006.

[3] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring Pay-per-Install: The Commoditization of Malware Distribution. In *USENIX Security Symposium*, 2011.

[4] J. Caballero, P. Poosankam, C. Kreibich, and D. Song. Dispatcher: Enabling Active Botnet Infiltration Using Automatic Protocol Reverse-Engineering. In *ACM Conference on Computer and Communications Security (CCS)*, 2009.

[5] C. Cho, J. Caballero, C. Grier, V. Paxson, and D. Song. Insights from the Inside: A View of Botnet Management from Infiltration. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2010.

[6] eMarkSofts. Fast email harvester 1.2. `http://fast-email-harvester.smartcode.com/info.html`, 2009.

[7] G. Gu, R. Perdisci, J. Zhang, and W. Lee. BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-independent Botnet Detection. In *USENIX Security Symposium*, 2008.

[8] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser. Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine. In *USENIX Security Symposium*, 2009.

[9] O. Hohlfeld, T. Graf, and F. Ciucu. Longtime Behavior of Harvesting Spam Bots. In *ACM SIGCOMM Conference on Internet Measurement*, 2012.

[10] J. Iedemska, G. Stringhini, R. Kemmerer, C. Kruegel, and G. Vigna. The Tricks of the Trade: What Makes Spam Campaigns Successful? In *International Workshop on Cyber Crime*, 2014.

[11] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamalytics:

An Empirical Analysis of Spam Marketing Conversion. In *ACM Conference on Computer and Communications Security (CCS)*, 2008.

[12] C. Kanich, N. Weaver, D. McCoy, T. Halvorson, C. Kreibich, K. Levchenko, V. Paxson, G. Voelker, and S. Savage. Show Me the Money: Characterizing Spam-advertised Revenue. *USENIX Security Symposium*, 2011.

[13] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the Spam Campaign Trail. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2008.

[14] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamcraft: An Inside Look at Spam Campaign Orchestration. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2009.

[15] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, et al. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE Symposium on Security and Privacy*, 2011.

[16] MaxMind. GeoIP. `http://www.maxmind.com/app/ip-location`.

[17] Northworks Solutions Ltd. Ecrawl v2.63. `http://www.northworks.biz/software.html`, 2012.

[18] C. Nunnery, G. Sinclair, and B. B. Kang. Tumbling Down the Rabbit Hole: Exploring the Idiosyncrasies of Botmaster Systems in a Multi-Tier Botnet Infrastructure. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2010.

[19] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. M. Voelker, V. Paxson, N. Weaver, and S. Savage. botnet Judo: Fighting Spam with Itself. In *Symposium on Network and Distributed System Security (NDSS)*, 2010.

[20] M. Prince, B. Dahl, L. Holloway, A. Keller, and E. Langheinrich. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2005.

[21] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2011.

[22] G. Stringhini, M. Egele, A. Zarras, T. Holz, C. Kruegel, and G. Vigna. B@BEL: Leveraging Email Delivery for Spam Mitigation. In *USENIX Security Symposium*, 2012.

[23] G. Stringhini, T. Holz, B. Stone-Gross, C. Kruegel, and G. Vigna. BotMagnifier: Locating Spammers on the Internet. In *USENIX Security Symposium*, 2011.

[24] Symantec Corp. State of spam & phishing report. `http://www.symantec.com/content/en/us/enterprise/other_resources/b-intelligence_report_02-2013.en-us.pdf?om_ext_cid=biz_socmed_AR_pv_180313_scom_socialmedia_SIRFeb13`, 2013.

[25] VirusTotal. Free Online Virus, Malware and URL Scanner. `https://www.virustotal.com/`.

[26] P. Wurzinger, L. Bilge, T. Holz, J. Goebel, C. Kruegel, and E. Kirda. Automatically Generating Models for Botnet Detection. In *European Symposium on Research in Computer Security (ESORICS)*, 2009.

[27] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming Botnets: Signatures and Characteristics. *SIGCOMM Comput. Commun. Rev.*, 38, August 2008.