Spamming Chains: A New Way of Understanding Spammer Behavior

Pedro H. Calais Guerra
Federal University of Minas Gerais (UFMG)
pcalais@dcc.ufmg.br

Dorgival Guedes
Federal University of Minas Gerais (UFMG)
dorgival@dcc.ufmg.br

Wagner Meira Jr.
Federal University of Minas Gerais (UFMG)
meira@dcc.ufmg.br

Cristine Hoepers
Brazilian Network Information Center (NIC.br)
cristine@cert.br

Marcelo H. P. C. Chaves
Brazilian Network Information Center (NIC.br)
mhp@cert.br

Klaus Steding-Jessen
Brazilian Network Information Center (NIC.br)
jessen@cert.br

ABSTRACT

In the effort of keeping their identities hidden, spammers rely on many weapons, such as the use of open proxies, open relays and compromised machines to conceal the spam origin before they deliver messages through SMTP. In this work, we study how today's sophisticated spammers combine such techniques, chaining machines along the network to deliver their messages anonymously. Our analysis was based on the observation of HTTP and SMTP traffic from connections established by spammers to a set of low-interaction honeypots. The main contribution of this paper is to show how the understanding of such chains can unveil information beyond that obtained from previous spam analysis techniques, often characterized by focusing on a single point of the spam dissemination process. In particular, we show that honeypots that emulate open proxies and open relays allow the detection of end-user compromised machines, because of the chains established by spammers linking open proxies to those machines before delivering the message to a legitimate SMTP server. We also show that spammers reach open proxies and then spread their abuses to several open relays and compromised machines at the same time, creating chains that behave similarly to botnets. Comparing spam traffic collected from 2006 to 2009, we concluded that open proxies still must be considered a threat, despite the claims from other works which have argued that most spam traffic nowadays is solely due to compromised user machines and botnets.

1. INTRODUCTION

One of the major concerns of spammers is to conceal their identities, including in terms of network location. This happens for two reasons: first, spamming is considered an abuse in most networks, thus spammers need to stay anonymous while disseminating their unsolicited messages. Second, if spammers sent spams directly from their (known) machines to the victims' mailboxes, they would be easily blocked by mail servers [3].

Actually, direct spamming was the preferred strategy adopted by spammers on the early spamming days, opening an SMTP connection to the SMTP server of the intended recipient. Since legitimate SMTP servers tend to log the IP address of the machine sending each message, and have also become more restrictive on their conditions to accept incoming messages from unknown origins, other techniques have been added to the spammer's "bag of tricks" in order to hide from the final SMTP server. That can be done by using servers along the network that might be improperly configured, behaving as open proxies that are willing to forward connections to other hosts in the network, or open mail relays, accepting all messages handed to them for delivery, oblivious of their relation to the messages' actual source and destination addresses [15]. Additionally, spammers can also exploit user machines compromised by some kind of malware [22] that can be able to perform one (or both) of those functions (open proxy/mail relay). Machines infected that way join a botnet or just remain available to whomever happens to find them in the network.

In the on-going *spam arms race*, today's sophisticated spammers tend to combine different identity concealing techniques, creating *chains* of machines. For example, a spammer can chain a connection through multiple open proxies and then abuse an open relay before delivering the message to the recipient's SMTP, turning the tracking of his or her origin almost impossible. Other possible chains are represented on Figure 1.

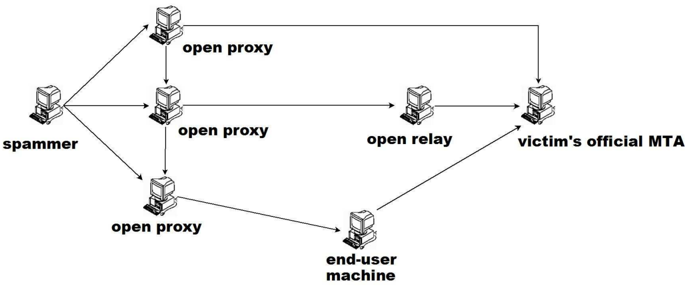

Figure 1: different chains established by spammers to disseminate spams

Although some of those chains have been reported by the

CEAS 2009 - Sixth Conference on Email and Anti-Spam July 16-17, 2009, Mountain View, California USA

spam research community as technically possible [3, 2] and are discussed informally on security mailing lists and forums, the scientific characterization of chains of machines for dissemination of spam is still limited. In general, researchers focus on only one specific hop of the path to delivery (usually the logs of mail servers or spam traps). Not only that, but the importance of some kinds of vulnerabilities, like open relays, has been considered of minor importance recently, due to the widespread of botnets, perceived as a greater threat as a spam source.

In this paper, we investigate spammer behavior in terms of the chain of machines they use to deliver their messages. For that, we take a vantage point inside the network, observing intermediary links of those chains. The understanding of the different paths followed by spams in the Internet network infrastructure can open new directions on many anti-spam research topics, such as estimation of spam campaigns size, estimations of infrastructure size and development of reputation-based anti-spam techniques inside the network.

Our data collection architecture is based on the deployment of low-interaction honeypots emulating open proxies and open mail relays in Brazilian networks [20]. It has already been argued that techniques for colleting spam data usually provide only a sample of all the spams disseminated [1]; we argue that the issue is even more complicated: because of machine chains and their ramifications, from a single vantage point the perceived behavior of *each* spammer is also a sample. The challenge is, then, to understand how spammers act even though data collected by a limited set of honeypots provide an incomplete view of the spammer behavior.

By observing the origins of the connections established to the honeypots, the next steps attempted by the spammers in the process of chaining machines, and the identification of all messages associated with each spam campaign, we were able to get information about the sequences of machines exploited by spammers, building a clearer picture of the process. For our purpose, we define a chain as the sequence of connections that may be used to forward the content of a group of spam messages until they are delivered over an SMTP connection, whether that connection happens to be to the mail server of the final intended recipient or not. Our decision is based on the fact that, once a spam message is delivered to a properly working SMTP server, it will follow the same path of all other mail, guided by the information of DNS MX records, mostly.

The main contributions of the paper are: (1) we demonstrate how the study of machine chains for dissemination of spams can unveil previously undocumented spammer behaviors, (2) we demonstrate behaviors previously mentioned by security specialists but not yet demonstrated in a scientific work, (3) we show that, because of machine chains created by spammers, the value of honeypots emulating open proxies and open relays are not limited to studies of these types of abuse, but can also help characterize the dissemination of malware infected compromised machines, for example, and (4) we show that spammers that are able to establish chains to a larger set of machines and avoid blacklisting usually send higher volume of messages.

2. RELATED WORK

There are many works that characterize how spammers abuse network resources. However, most of those collect data from a specific spamming dissemination strategy, such as botnets [9, 10], spam traps [5] and open relays [14]. Because of that, they only focus on a specific step of the path traversed by the messages. In case of researches that analyze mail servers logs [11, 6], only the last hop abused by the spammer before reaching the server is analyzed. Spams collected that way do not allow the study of machine chains, since SMTP headers can be easily forged by spammers, and they do not record any TCP connection chaining through proxies.

There are previous works that analyze connections established by low-interaction honeypots, but they focus on the analysis of the characteristics of the abuses targeting the honeypots, such as CC (Internet Country Code) of origin and IP address distribution of incoming connections [4, 20]. Our approach is different because we consider both the origin and destination of the connections established with the honeypots, as well as information about the composition of campaigns, what allow us to improve our knowledge about the different paths spam messages follow. Our technique to identify campaigns is based on a previous work [4]. It is based on the extraction of relevant characteristics from spam messages (such as URL and subject fragments, message layout and encoding type) and insertion of these features on a tree structure (Frequent Pattern Tree) that identify the invariant parts among spam messages that define spam campaigns. Our technique differs from other approaches proposed in the literature [1, 8, 24] by being more adaptive to changes in spam obfuscation techniques since it do not use pre-defined patterns for classification.

Some works mention the creation of chains of machines for sending spam as something possible [3, 2, 13], but they do not effectively characterize and demonstrate such behaviors. Our paper aims to fill that gap.

3. CHARACTERIZATION METHODOLOGY

Our methodology to analyze chaining of machines for spam dissemination is comprised by three well defined steps. First, spam data is collected through the deployment of low-interaction honeypots emulating open proxies and open relays. Next, we show evidences that chains established between open proxies and open relays/compromised machines are a frequent behavior in our dataset. Finally, we characterize and quantify how each spam campaign makes use of the different SMTP targets in terms of number of connections, average number of connections per host and volume and duration of abuses.

3.1 Data Collection

Our data collection architecture comprises a set of sensors based on low-interaction honeypots [16] to study the spam problem, in particular the abuse of open proxies and open relays. A proxy is a server that acts as an intermediary, making connections on behalf of other clients. An open proxy allows connections to be made from any origin to any destination IP address or port, and is traditionally abused for sending spam. Examples of common proxy protocols are HTTP and SOCKS. Misconfigured SMTP servers, usually acting as open relays, allow the delivery of messages from any source to any recipient and are also abused by spammers.

We deployed 10 honeypots in 5 Brazilian broadband ISPs networks (both cable and ADSL), that captured approximately 525 million spams over 15 months. These spams came from 216,888 different IP addresses, allocated to 165 different countries (Country Codes (CC), as defined in ISO 3166) and would have been delivered to 4.8 billion recipients [20].

These honeypots did not collect data at the final spam destination, like in spamtrap accounts or in mail servers. Instead, we measured the abuse of proxies and relays by spammers, and captured the spam at that stage, before reaching its final destination. This allowed us to analyze the different destinations of the spams.

Deploying a set of honeypots, and not just one, brought some advantages for characterization of machine chains. First, we could observe chains of machines involving more than one honeypot. They also allowed us to measure how spammers abused more than one open proxy/open relay.

We used `Honeyd` [16] and its SMTP and HTTP server emulation subsystems to capture spam. A SOCKS proxy emulator was developed to complement the existing emulators [20]. The SMTP emulator stores each message received, along with information about the host originating the SMTP connection. The HTTP proxy emulator recorded the IP address of the machine that contacted it, along with the identification of the machine and port targeted through the proxy (IP address, destination port number, and the machine name, when available from the command sequence or through DNS reverse mapping). The SOCKS emulator did not record the destination machine name, unless it was provided in the command sequence, and that only for version 4.0 of the protocol.

Messages were stored locally and never delivered to recipients. The only exceptions were e-mail probes sent by spammers to test whether the proxy/relay was actually working (delivering messages). Such probes were identified early in the configuration of the honeypots and added to the processing routines.

On this work, we have analyzed connections to HTTP and SMTP ports of the honeypots. We did not consider SOCKS connections because most of them used earlier versions of the protocol, which did not register the hostnames which were targeted by the connections. As discussed next, that information is required in our methodology for identification of abuse types. This meant that 36.8% of all messages recorded were not considered, but we believe the remaining volume is still representative.

4. UNDERSTANDING SPAM MACHINE CHAINS

On this section we present our results after applying our methodology to our data. Table 1 provides a general view of the dataset. Data was collected in two distinct periods (July 2006–June 2007 and October 2008–April 2009). During the period of almost 18 months, over 260 million messages were delivered by the spammers to the honeypots's fake open HTTP proxies, over 97 million distinct connections (an average of 2.7 messages delivered by each connection). Those connections originated from 93,757 unique IP addresses and targeted a larger number of different destination addresses (459,218). Based on the analysis of the messages we identified the campaigns using the Frequent Pattern Tree proposed on [4] and the addresses of intended recipients (and their mail domains). In the discussion that follow we highlight our major findings.

Table 1: Overview of the data associated with connections established to the honeypots' HTTP proxy emulators

messages	262,121,899
spam campaigns	45,121
connections	97,136,321
unique source IP addresses	93,757
unique recipients	3.2×10^9
unique mail domains	6,710,121
unique target IP addresses	459,218

4.1 Evidences of Spamming Chains

Our basic approach to identify chains of abused machines is to analyze the relation between the **recipient mail domains** and the **target hosts** of the HTTP connections spammers established with the honeypots. To illustrate our strategy, Table 2 shows a (real) sample extracted from our dataset, for a given source IP. This IP address sent 6 messages aimed at different mail domains. The chains started with an abuse to a honeypot's HTTP port and then the spammer tried to connect to different target hosts. The first message was delivered to the MX server associated with the mail domain found on the spam message; in this case, the spammer tried to deliver the message from an open proxy to the victim's MTA. The same behavior was observed for the third message send by this spammer. Messages 2, 4, 5 and 6, however, were sent by open proxy + compromised machine chain established by the spammer (according to the spammer's impression, since the honeypots do not deliver any spam). In the case of message 4, 5 and 6, the host `<IP-NUMBER>.HINET-IP.NET` (IP address supressed) was responsible to forward the message to other host on the network or to deliver the message to the recipient's MTA. We believe those end-user machines are not part of botnets, but just machines infected by some malware that opened their port 25 for spammers. This is because we verified that open relays – misconfigured mail servers – are also abused by spammers as the target of their HTTP connections, on our dataset, what indicates that spammers are just probing port 25 to find machines and not really coordinating an attack through botnets.

msg.	mail domain	target host
1	`hotmail.com`	`mx1.hotmail.com`
2	`yahoo.com.tw`	`<IPnumber>.veloxzone.com.br`
3	`ms29.hinet.net`	`ms29a.hinet.net`
4	`ms29.hinet.net`	`<IPnumber>.HINET-IP.net`
5	`ms29.hinet.net`	`<IPnumber>.HINET-IP.net`
6	`ms12.hinet.net`	`<IPnumber>.HINET-IP.net`

Table 2: Sample of connections attempted by a spammer

Looking to the diversity of mail domains and target hosts on the whole dataset, we found that the HTTP proxy connections attempted by spammers to one of our honeypots were targeted to almost 460 thousand distinct hosts. On the other hand, more than 6.7 million unique mail domains were targeted by spammers (Table 1). Since the number of mail domains is almost 15 times higher than the number of hosts targeted by the connections, it suggests that most of the chains do not end at the final mail server: spammers do not deliver all spams to the recipient's MTA after reaching an open proxy, but try to insert more intermediaries on the chain to increase their anonimity.

Figure 2 confirms those differences for the majority of the source IPs that tried to abuse one of our honeypots' HTTP ports. The scatter plot relates the number of mail domains targeted by each source IP address with the number of unique IP addresses targeted by it. The majority of the source IPs are plotted below the $y{=}x$ line and thus those IPs fit on the case discussed on the previous paragraph: they contact less IP numbers than expected if they always targeted the Mail Exchange servers associated to each of their victims' mail domains.

In our datasets, more than 50% of the target IP addresses received messages directed to more than two distinct mail domains; more than 10% received messages addressed to more than 10 domains and some IPs received messages to more than 100 mail domains, indicating that those hosts are not the final destination of the messages, but just intermediaries that would be responsible for further distributing messages to their final destination.

About 15% of the source IPs, however, tried to abuse significantly *more* IP numbers than mail domains after reaching open proxies (in the case, our honeypots). Some IPs, for example, target only 10 mail domains but they deliver messages to more than 100 unique IP numbers. Those cases also indicate establishment of longer machine chains for spam delivery. In summary, evidence of the establishment of chains are given by the green dots plotted far from the $y{=}x$ line. If a spammer never establishes chains to open relays or compromised machines, his behavior will be represented by a dot near the $y{=}x$ line or a little above it, since some domains mail deilver messages to more than on MX server (e.g., messages targeted to `yahoo.com.tw` can be delivered to `mta-v1.mail.vip.tp2.yahoo.com` or `mta-v2.mail.vip.tp2.yahoo.com`).

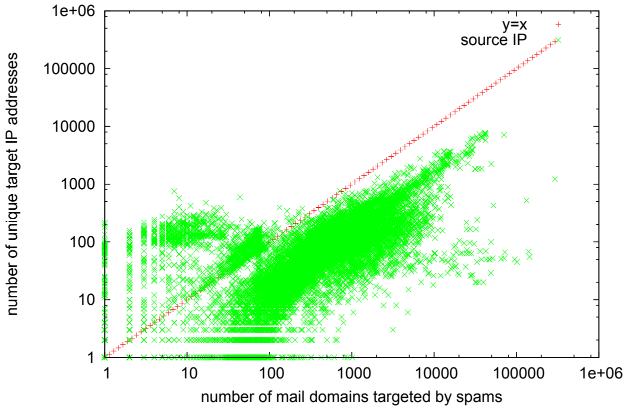

Figure 2: scatter plot showing number of mail domains and target IP address for each source IP number

Seeking more evidences of the establishment of chains between open proxies and open relays/infected machines, we analyzed how the mail domains and IP numbers targeted by each source IP varied across time. Figures 3 and 4 shows the variability of the mail domains (green dots) and target IP addresses (red dots) targeted by two distinct source IP addresses over time. Mail domains were mapped to sequential numeric ids, represented on the y axis. Whenever a new mail domain was targeted by the source IP, a new, greater, id was assigned to it. The same was done for the target IP addresses. For example, the sequence of mail domains on Table 2 would be {1, 2, 3, 3, 3, 4}, while the sequences of IP addresses (hosts) would be {1, 2, 3, 4, 4, 4}.

Figure 3 illustrates a spammer that abused only one mail domain during September 2008. On the other hand, more than 500 different target IPs were abused over this period. It is interesting to observe that most of the IP addresses were abused only once and never were abused again, what can be noticed by the dominant red curve with increasing ids and few dots with repeating ids (which indicate that the spammer abused a host already abused in the past). The different aspects of the red and green curve confirms the establishment of open proxy + compromised machine/open relay chains for this spammer.

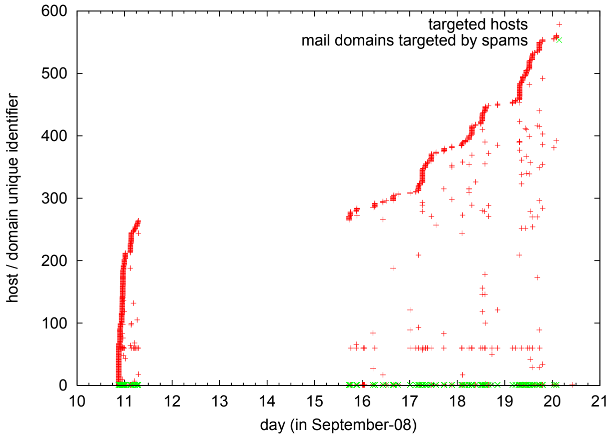

Figure 3: spammer abusing a unique mail domain (green dots) establishing chains with different hosts over time (red dots)

Figure 4 illustrates another case which also indicates the establishment of chains. This spammer, in January 2009, disseminates a campaign to various distinct (700+) mail domains and it can be observed that about 200 distinct mail domains are abused all over the campaign duration and about 500 mail domains are abused only once. However, when we look to the list of 250 abused hosts (in red dots), very little repetition is observed; almost every target host is abused only once. The difference on the distribution of red and green dots are the indication that the abuses of target hosts is not related to the mail domains contained on spam targets.

To validate our assumptions about the establishment of chains involving open proxies and other machines on the network, we considered 619,782 messages delivered during the period from 2009-03-10 to 2009-03-20. We mapped the mail domains found on each spam to its list of Mail Exchange (MX) servers (using `dig`), and then verified if at least one of these servers were present as a target of the SMTP connections spammers established. The results demonstrated that only 90,657 (14,6%) of the mail domains targeted by spam-

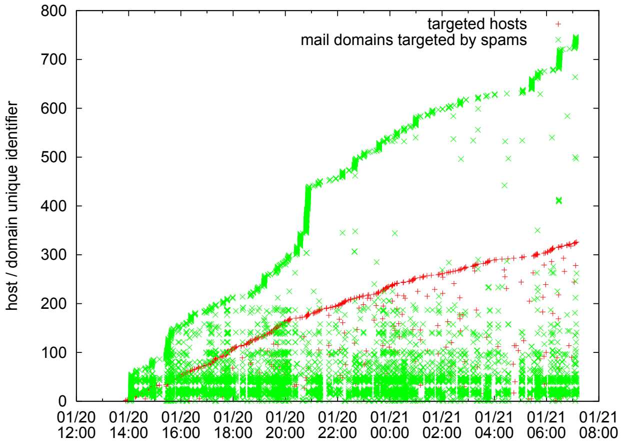

Figure 4: spammer abusing a list of abused hosts that rarely repeat; on the the other hand, the recipient list constantly abuses a list of 200 distinct mail domains

mers had a corresponding MX server on the list of hosts abused by spammers as the target of their HTTP connections, which we consider a strong evidence that, in fact, open proxy and open relay/compromised machines chain are a common behavior in our dataset. To the best of our knowledge, we are the first to demonstrate those chains on a scientific work. On the next sections, we derive spamming behaviors that can be determined by analyzing those chains.

4.2 Chains allow the observation of samples of each campaign

Machine chains make the measurement of the full behavior of a spammer very difficult and researchers have to keep in mind that all observations are just a sample of the spammer's acts (see Figure 5). For example, on Figure 3 presented on Section 4.1, we cannot guarantee that the spammer did not perform any spamming activity from 09/11 to 09/15; he or she might just have decided to abuse other open proxies in the network, out of the range of our HTTP emulators.

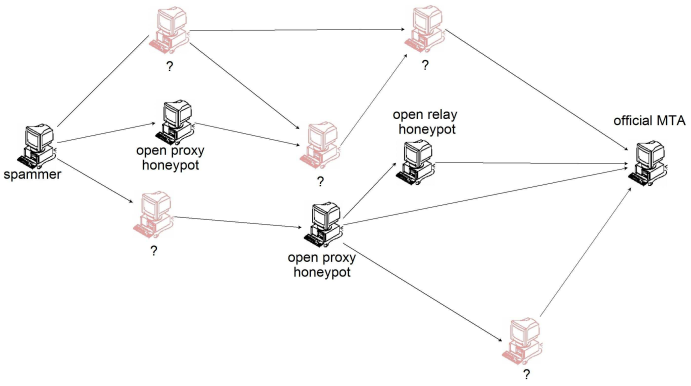

Figure 5: because of chains, honeypots allow the observation of only a sample of each spammer's behavior

Only chains that included at least one of the honeypots got observed, and this might explain why, on average, the campaigns we have identified are small (90% of the campaigns sent less than 5,000 messages), when it is been widely said that spam campaigns are much larger than that, reaching millions of recipients. To verify that, we checked the number of messages each spam campaign sent to each of our honeypots' open proxy ports, in terms of total number of messages and the average number of messages per honeypot. That is shown in Figure 6, where we plotted only campaigns that abused more than one honeypot. It can be observed that spammers explicitly and intentionally sent a small volume of messages to each open proxy. As many campaigns send less than 1,000 messages to each honeypot, they may actually have exploited hundreds of other open proxies available over the Internet. A future work is to investigate temporal patterns that could indicate gaps of time during a campaign, when spammers would be exploiting other open proxies rather than those of the honeypots.

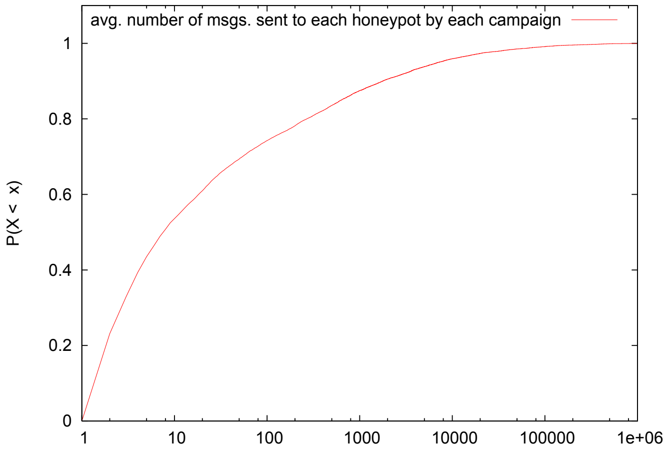

Figure 6: Spammers spread abuses among open proxies

To analyze how campaign sizes can vary depending on how they are observed, we investigated the correlation between average campaign sizes with the dispersion of the origins of the abuses that originated each one of the campaigns. In other words, we verified how campaign sizes varied when campaigns were observed, from the honeypots' viewpoint, from different number of sources, either abusing the honeypots' HTTP or SMTP ports. Figures 7 and 8 illustrate each one of these cases.

We can see that campaigns in which abuses to the honeypot's HTTP ports originate from 1 to 10 CCs were usually big and sent tens of thousands of spams (Figure 7). They also tended to abuse larger numbers of honeypots. On the other hand, campaigns on which open proxy abuses originated from a significant number of CCs (greater than 40) were very small and abused no more than 2 honeypots. It is interesting that, despite their varied origins, those campaigns were target at only one or two honeypots each, suggesting a high level of coordination among them.

Actually, less can mean more: those short campaigns originating from many CCs may, actually, be much larger than those which originate from fewer CCs. Because more powerful spammers spread their abuses, each machine being

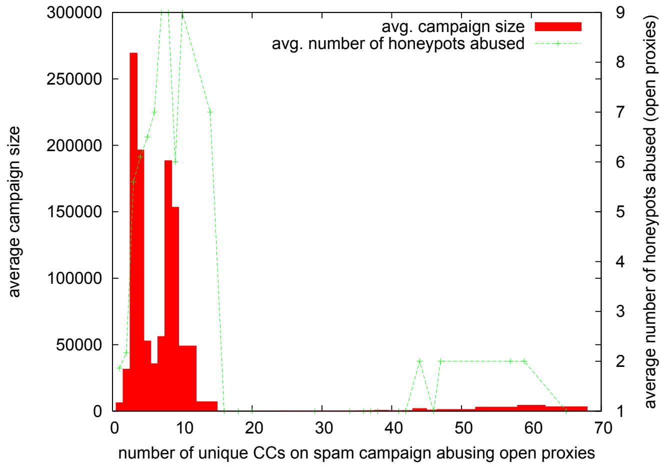

Figure 7: Avg. campaign size and avg. number of honeypots abused as open proxies per number of CCs of origin

abused has the impression that the campaign being disseminated is small.

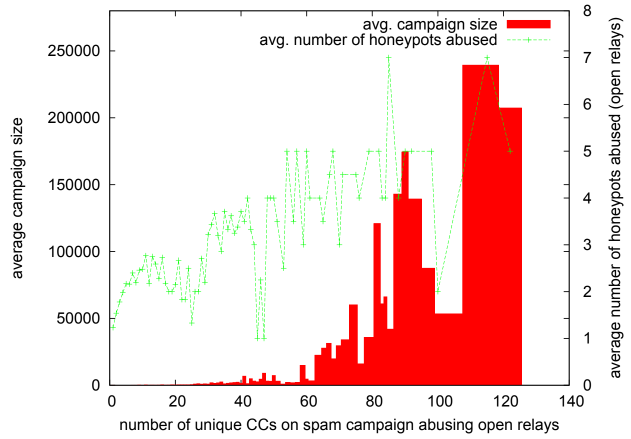

Figure 8: Avg. campaign size and avg. number of honeypots abused as open relays per number of CCs of origin

When we look at the same relationships for the abuses to the honeypots' open relays, the pattern is significantly different (Figure 8). Now, more honeypots are abused and campaign sizes increase, on average, as the origin of the open relay abuses become less concentrated in general.

As the origin of the abuses become more disperse, intensity of abuses decrease when honeypots are abused as proxies and increase when they are abused as SMTP relays. We believe such difference is due to the fact that abuses to open relays and compromised machines occur on the last hop of the chain and that all open proxies which are disseminating a given spam campaign target the same open relays and compromised machines. These would explain the behavior on Figure 8. On the case of Figure 7, we are observing an intermediary path of the chain and the more distributed the spammer activity is on this step, less messages are observed on each honeypot participating on the chain.

These results indicate how spam's figures and statistics can vary depending on the vantage point. For example, if we look solely to the abuse of HTTP and SOCKS ports of our honeypots, we will conclude that over 70% of spam are originated from Asian countries. On the other hand, if we look at the abuses to the honeypots' SMTP port, we will see the abuses spreaded all over the world (see [4]). In this sense, understanding of chains in the context of spam campaigns can provide useful insights that clarify these contradictions.

When we group our data using the identified campaigns, we see that the majority of the observed spam campaigns (91%) abused our honeypots as open proxies, only. The remaining 9% exploited the honeypots both as open proxies *and open relays*, what allowed us to improve our understanding of the chains established by spammers. For those campaigns, abuses to our honeypots as open proxies and open relays occurred during the same time frames for 96.5% of the campaigns. This means that both forms of abuse resulted from a common effort of the spammer. We have, then, two views of the chains from different angles: honeypots being abused as open proxies and abusing open relays/compromised machines, and honeypots being abused as open relays, by machines spreaded over the world which are probably open proxies.

4.3 Chains of open proxies with end-user compromised machines

A manual inspection of the hostnames of the machines targeted by spammers through HTTP connections established with the honeypots revealed that a significant number of machines were, in fact, end-user machines, probably compromised by some type of malware that instructed them to disseminate spam. This observation agrees with the results presented on Section 4.2, that indicated that many HTTP connections established by spammers are not targeted to a legitimate MX server associated to the recipient's mail domains, but to other machines such as open relays and end-user compromised machines.

To quantify the abuse to end-user machines, we applied a simple heuristic based on the hostnames targeted by the HTTP connections. We use the fact that ISPs often assign names for user machines (when they do so) that combine fixed tokens with a variable string that differentiates each host, usually a numeric id or the IP address which has been assigned to the host. For example, clients of the American ISP Verizon are usually named using the format `static-<IP>.<LOCATION>.dsl-w.verizon.net`. Machines under responsibility of HINET (in Taiwan) are identified by the format `<IP>.HINET.-IP.hinet.net`. On the other hand, legitimate mail servers usually have well defined names, such as `mail.ufmg.br`.

Our technique to differentiate between chains to mail servers and end-user machines was based on that observation. First, we broke the hostnames of those targeted by HTTP proxy connections into tokens for each level in the DNS hierarchy, preserving the information about their level. Next, the tokens were inserted on a data structure known as *Frequent-Pattern Tree* (FP-Tree) [21, 4]. On that kind of tree, insertions are performed in such a way that tokens from the same hostname define a path on the tree and the most frequent tokens are found at the higher levels and infrequent or random ones are inserted closer to the bottom of the tree, near the leaves. We also register, for each token, how many connections/messages used that name. This way, exploited end-user machines hosted on large ISPs share most of their paths on the root of the tree, because of the fixed parts in the format of their hostnames. Such hostnames differ only by tokens which correspond to their unique identifiers, often (part of) their IP addresses. As those features are less frequent than the fixed fragments, the hostnames belonging to the same ISP end up forming a sub-tree with a large number of siblings at the leaves and a heavily used common path. Our approach is not exact and can lead to false positives and false negatives; our intention was to detect some clear groups of end-user machines to demonstrate our hypothesis that, in fact, end-user machines are being chained with open proxies over the Internet to disseminate spam.

After applying this heuristic, we have identified 94,480 hosts that represent end-user machines and that are not mail servers. Based on that we can say they are compromised machines, either poorly configured or infected by any sort of malware that makes them behave as open mail relays. Those hosts were distributed among 894 groups (ISPs). Table 3 shows the top 10 countries hosting infected machines, in terms of unique IP addresses. It is not surprising that the U.S. appears on top of the list; previous reports from security companies have pointed the country as the world leader in number of compromised machines.

CC	number of unique IPs (ISPs)	%
US	59800 (351)	36.6
TW	38925 (61)	23.8
CN	24708 (19)	15.1
HK	6880 (28)	4.2
GB	6564 (59)	4.0
KR	5925 (8)	3.6
JP	5631 (48)	3.5
DE	5627 (50)	3.4
BR	5049 (37)	3.1
CA	3958 (35)	2.4

Table 3: Top 10 countries in number of compromised machines

The FP-Tree also allowed us to group them by their domains. Based on that we can say they are compromised machines, either poorly configured or infected by any sort of malware that makes them behave as open mail relays. Those hosts were distributed among 894 groups (ISPs). Table 4 shows the top 11 groups identified, in terms of unique IP addresses. According to the CC information associated with each network, we can see infected machines can be widely distributed.

It is interesting to note that, based on their domain names, some groups of infected machines were found in domains of dedicated hosting services and datacenter providers (e.g., `ev1servers.net`). We were not able to determine whether that was due to their servers being poorly configured, infected by malware, or even intentionally configured that way by a client (the spammer).

These results indicate that, although widely reported that the majority of spam are sent from compromised machines and that open proxies are not common anymore [18], open proxies are still used as a technique for spam distribution. The underestimation of the impact of open proxies may be due the fact that observations from mail server logs will end up identifying the compromised machines from the countries listed on Table 3 as the last real `received:` line of the message headers. However, open proxies are a common mechanism for identity concealment and, because of that, they are still used on the different chains discussed on this paper. Thus, fighting open proxies is still an important way of fighting different spam dissemination strategies which include open proxies in their routes, including when botnets and other kinds of compromised machines appear to be the last originators of spam.

Table 4: Main groups of compromised user machines

Group	CC	Unique IP addrs.
$< IP >$.HINET-IP.hinet.net	TW	15.045
$< IP >$.ev1servers.net	US	1.417
rrcs-$< IP >$.central.biz.rr.com	US	1.228
$< IP >$.static.isl.net.tw	TW	1.191
Red-$< IP >$.staticIP.rima-tde.net	ES	1.022
$< IP >$.seed.net.tw	TW	966
$< IP >$.ptr.us.xo.net	US	882
$< IP >$.dsl.scrm01.pacbell.net	US	877
ip-$< IP >$.ip.secureserver.net	US	849
$< IP >$.dynamic.hinet.net	TW	746
c-$< IP >$.hsd1.nj.comcast.net	US	735

By understanding machines chains we can characterize infrastructures which are a step before or after the machines where we effectively collected the data. In this example, we are investigating the geographical distribution of compromised machines by observing connections would be directed to them through the mediation of our (fake) open proxies. Although the honeypots emulate only open proxies and open relays, the chains established by spammers allow us to measure other types of abuse. The honeypots emulating open proxies and open relays can be deployed as a early warning system to issue alerts to ISPs with the most recently machines on their network which have been abused by spammers, for example.

4.4 Impact of Chains

In this section, we analyzed how the the number of machines abused and the intensity of the abuse to each machine affects the volume of messages the spammer deliver and for how long they persist their abuses.

Figure 9, in log-log scale, verifies the correlation, for each source IP, between the number of unique target machines contacted by the spammer and the volume of messages sent by it. Although there is a considerable scattering of the dots, the correlation coefficient is significative (72%). We can note that only spammers that count with lists of target machines greater than 10,000 elements managed to send more than 1 million spams.

We also correlated the number of target IP addresses abused by each source IP with the duration of the abuse (in number of days). The result can be seen on Figure 10. It is clear than only spammers that count with infrastructure to abuse thousands of IP addresses can send messages for many months.

Figure 11 shows that spammers that manage to send messages for many months are the same that establish, on average, few connections to each machine they abuse. This

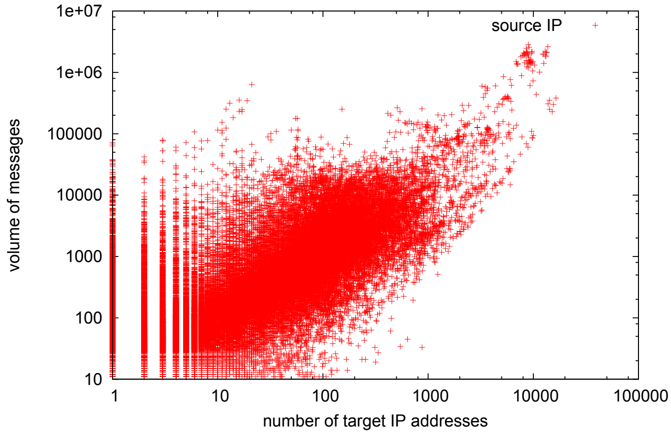

Figure 9: number of target IP addresses x volume of messages

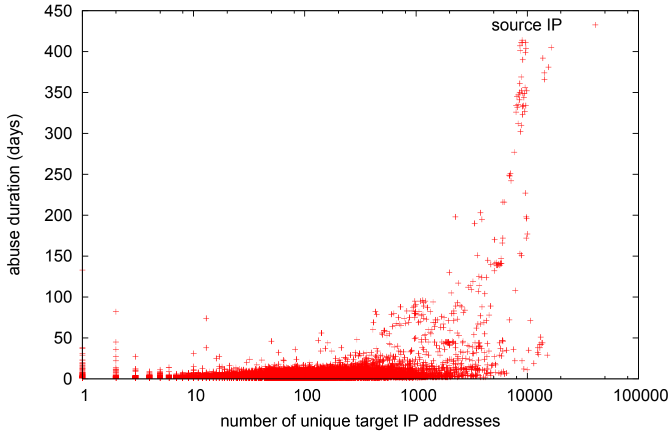

Figure 10: number of target IP address abused by each source IP x number of days

observation indicates that the most successfull spammers are the ones which are able to spread their abuses and, thus, remain unnoticed. What seems to limit the volume of messages a spammer deliver does not seem to be the bandwith to which they have acess, but the ability they have to chain their connections to many different intermedaries at the same time. This characteristic has been identified on the literature as a remarking characteristic of botnets [19] and raise many challenges for spam fighting [12].

We noticed that each spammer that connects to a honeypot HTTP port establish chains with multiple open relays and compromised machines, because they will be the machines which will appear on the last `received:` line of the SMTP header. Each of these machines are targeted few times by each spammer. But how is the traffic pattern when the honeypots themselves are abused as open relays (and then forwarded to the recipient's MTA – according to the spammer's belief)?

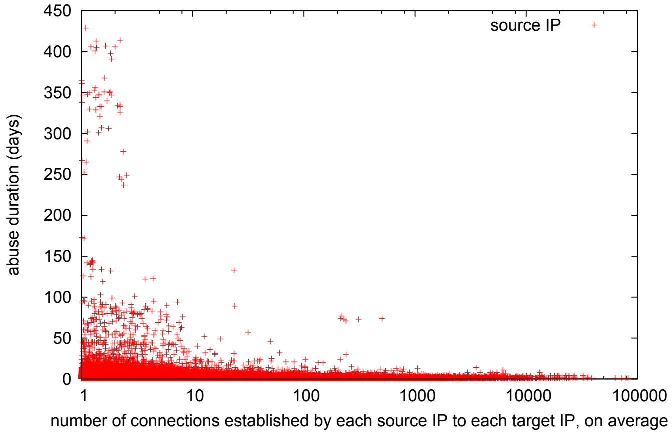

Figure 11: number of connections established by each source IP to each target IP x duration of abuse (days)

In our observations, when a honeypot's open relay is abused, the most common case is to find such abuses only during a fraction of the duration of the campaign. Figure 12 shows open proxy and open relay abuses during a typical campaign. In that case the honeypots were abused as open relays only on the intermediary period of the campaign lifetime. Figure 13 shows the cumulative distribution function of the fraction of the duration of a campaign in which abuses to the honeypots as open relays were observed. We can see that, in 50% of the campaigns, our honeypots were abused as open relays during less than 50% of the days; during the rest of the time, only open proxy abuses were observed. The targets of chains change over time in any given campaign, what agrees with the behavior previously detected: spammers avoid overloading the hosts which contact directly victim's Mail Transfer Agents.

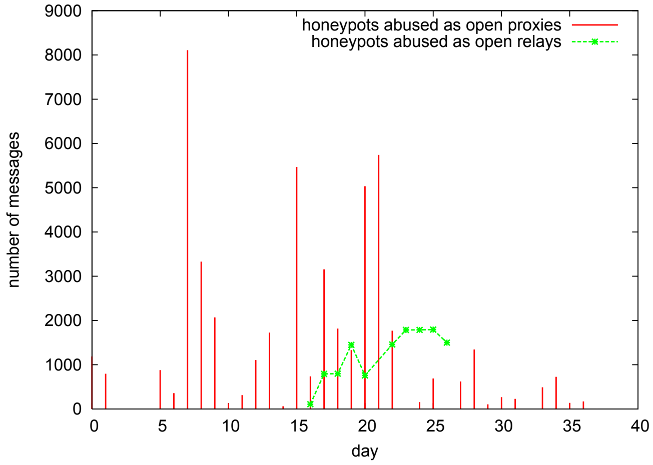

Figure 12: Number of messages abusing honeypot's proxy and relay for a given campaign

Actually, it has been observed on our dataset that, overall, the number of messages sent to open proxies are far higher than the volume sent to open relays [4]. As open relays/compromised machines are on the last hop of the chain and they contact directly victim's mail servers, it is impor-

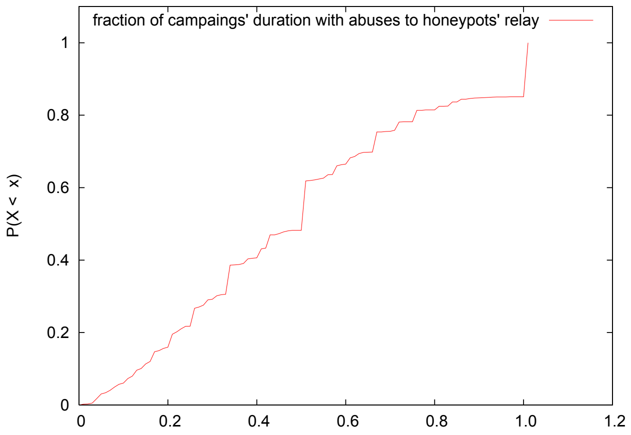

Figure 13: Fraction of the number of days in each spam campaign in which abuses to the honeypots as open relays were observed, in % (CDF)

tant for spammers to spread the abuses among those machines as much as they can, and limit their use to a shorter period, since after being identified (and added to blacklists) they become useless to the spammer's objectives. As open proxies are exploited on a previous link of the chain, they can be abused more intensively and for longer, since they are not noticed by the destination servers. After reaching open proxies, each campaign distribute their connections through a number of different open relays/compromised machines. This result complements what has been observed in [14]. The authors observed that their open relay sinkholes usually receive few connections from each source IP, probably originated from botnets or open proxies. Although we were not able to confirm the chains linking botnets to open relays, we observed chains between open proxies and open relays, and a single open proxy abuses many open relays and compromised machines, for the reasons discussed above.

An interesting observation is that some works recognize botnets in spam campaigns that are originated from several sources at the same time [23] may be, actually, be identifying also chains between open proxy and compromised machines like the ones shown on this paper, which generate the same impression to the recipient: widespread sources disseminating a spam campaign. Spammers that spread their abuses to several open relays and compromised machines will pose limitations on the effectiveness of DNS blacklists similar to the ones brought by botnets, discussed in [17].

4.5 Are chains becoming more frequent?

We used the fact that we have an older dataset (collected between June/2006 and July/2007) and a newer one (colleted between October/2008 and April/2009) to compare general characteristics of the abuses and check if the behaviors observed in 2006 and 2007 still persist. Table 5 shows some relevant numbers related to both datasets.

The first comparison that arises is that the number of messages each source IP tries to deliver through the honeypots is kept at the same level, indicating that, since 2006, abuses to open proxies and open relays have not reduced. The average number of new source IPs abusing the honeypots also remain stable. Moreover, 21% of the source IPs that abused the honeypots in 2006 or 2007 *still* attempt connections in 2008 and 2009; those spammers are successfully hiding behind chains of open proxies and compromised machines.

Table 5: Overview of the data for 2006/2007 and 2008/2009 dataset

dataset	2006–07	2008–09
messages per day	111,111	175,121
new source IPs per day	65.5	84.8
avg. num. of target IPs contacted	20.8	87.5
new target IPs per day	441.6	1007.4
num. of connections to each target IP	195.6	43.6

When we verified the intensity of chains have varied among the two datasets, however, a signficant difference showed up: the number of different hosts abused by each spammer as the target of their HTTP connections raised by a ratio of 4.8. The number of hosts abused daily by spammers also increased (from 441.6 to 1007.4), and each one of these hosts are targeted by fewer connections (43.6).

A work from 2004 [7] showed that from 2000 to 2004 spammers increased the distribution of their spam workload across mail relays, establishing fewer connections to each mail relay across the time. A similar behavior is noticed on our dataset, from 2006/2007 to 2008/2009. This may be an indication of a *spam arms race* between spammers and blacklists. As the number of blacklists increased, spammers needed to respond adequately. Another explanation for this increase, in our datasets, is that the raise of botnets forced spammers that rely on open proxies, open relays and user compromised machines to reach a similar level of dissemination of spams through many hosts at the same time.

5. CONCLUSIONS AND FUTURE WORK

In this work, we studied the chains of machines built by spammers to deliver their messages anonymously. The main contribution of this paper is to show how investigation of chains can unveil spammer behaviors, such as the establishments of chains of open proxies with open relays, compromised machines and other open proxies. Our analysis was based on the examination of Proxy HTTP and SMTP connections established by spammers to a set of low-interaction honeypots emulating open proxies and open relays. We showed that spammers that chain open proxies with open relays and compromised machines along the network generate a traffic pattern similar to the observed for botnets.

We draw attention to the fact that our observations are just samplings of the dissemination of each spam campaign; our view of the data is, by nature, incomplete, limited to the small set of honeypots we have deployed.

Future work include extending the characterization of spamming chains in two main directions. First, we will investigate the relations between botnets and open proxies/open relays, by comparing the campaigns and traffic characteristcs observed on both types of spam dissemination strategies. We will also deploy honeypots in other countries' networks, to analyze chains from a global vantage point. Spamming strategies characterizations are usually done from a single, local viewpoint, and considering the chains involving various types of abuses (botnets, open proxies, open relays, malware-infected machines) established with machines spreaded in different countries may help us to define a single, global view of spammer behavior.

We are in the process of deploying honeypots in other countries, rather than Brazil, and then analyze spam dissemination and chaining from a global vantage point. Researchers interested in having access to our dataset and participate on the international phase of our project should feel free to contact one of the authors.

6. ACKNOWLEDGMENTS

This work was partially supported by NIC.br, UOL, CNPq, CAPES, FAPEMIG and FINEP.

7. REFERENCES

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

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

[3] Boneh, D. The difficulties of tracing spam email. `http://www.ftc.gov/reports/rewardsys/expertrpt_boneh.pdf`, September 2004.

[4] Calais, P. H., Pires, D., Guedes, D., Wagner Meira, J., Hoepers, C., and Steding-Jessen, K. A campaign-based characterization of spamming strategies. In *Proceedings of Fifth the Conference on e-mail and anti-spam (CEAS)* (2008).

[5] Gansterer, W. N., and Ilger, M. Analyzing uce/ube traffic. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce* (New York, NY, USA, 2007), ACM, pp. 195–204.

[6] Gomes, L. H., Cazita, C., Almeida, J. M., Almeida, V., and Wagner Meira, J. Characterizing a spam traffic. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conf. on Internet measurement* (New York, NY, USA, 2004), ACM, pp. 356–369.

[7] Jung, J., and Sit, E. An empirical study of spam traffic and the use of dns black lists. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2004), ACM, pp. 370–375.

[8] Kolcz, A., and Chowdhury, A. Hardening fingerprinting by context. In *Proceedings of the 4th Conference on Email and Anti-Spam (CEAS). Mountain View, CA.* (2007).

[9] Kreibich, C., Kanich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., and Savage, S. On the spam campaign trail. In *LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats* (Berkeley, CA, USA, 2008), USENIX Association, pp. 1–9.

[10] Lee, W., Wang, C., and Dagon, D. Honeynet-based botnet scan traffic analysis. In *Botnet Detection*, W. Lee, C. Wang, and D. Dagon, Eds., vol. Volume 36. Springer Berlin / Heidelberg, 2007, pp. 25–44.

[11] Li, F., and Hsieh, M.-H. An empirical study of clustering behavior of spammers and group-based anti-spam strategies. *Proceedings of the Third Conference on Email and Anti-Spam (CEAS). Mountain View, CA.* (July 2006).

[12] Naraine, R. Is the botnet battle already lost? `http://www.eweek.com/print_article2/0,1217,a=191391,00.asp`, February 2007.

[13] Oudot, L. Fighting spammers with honeypots. `http://www.securityfocus.com/infocus/1747`, November 2003.

[14] Pathak, A., Hu, Y. C., and Mao, Z. M. Peeking into spammer behavior from a unique vantage point. In *LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats* (Berkeley, CA, USA, 2008), USENIX Association, pp. 1–9.

[15] Provos, N. A virtual honeypot framework. In *13th USENIX Security Symposium* (2004).

[16] Provos, N., and Holz, T. *Virtual Honeypots: From Botnet Tracking to Intrusion Detection*, 1st ed. Addison-Wesley Professional, July 2007. ISBN-13: 978-0321336323.

[17] Ramachandran, A., Dagon, D., and Feamster, N. Can dns-based blacklists keep up with bots? In *Proceedings of 3rd the Conference on e-mail and anti-spam (CEAS)* (2006).

[18] Ramachandran, A., and Feamster, N. Understanding the network-level behavior of spammers. In *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications* (New York, NY, USA, 2006), ACM, pp. 291–302.

[19] SpamCop. Botnets. `http://forum.spamcop.net/scwik/BotNet`, 2007.

[20] Steding-Jessen, K., Vijaykumar, N. L., and Montes, A. Using low-interaction honeypots to study the abuse of open proxies to send spam. *INFOCOMP Journal of Computer Science* (2008).

[21] Tan, P., Steinbach, M., and Kumar, V. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., 2005.

[22] Weaver, N., Paxson, V., Staniford, S., and Cunningham, R. A taxonomy of computer worms. In *WORM '03: Proceedings of the 2003 ACM workshop on Rapid malcode* (New York, NY, USA, 2003), ACM, pp. 11–18.

[23] Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., and Osipkov, I. Spamming botnets: signatures and characteristics. In *SIGCOMM '08: Proceedings of the ACM SIGCOMM 2008 conference on Data communication* (New York, NY, USA, 2008), ACM, pp. 171–182.

[24] Yeh, C.-C., and Lin, C.-H. Near-duplicate mail detection based on url information for spam filtering. In *Information Networking. Advances in Data Communications and Wireless Networks* (November 2006), Springer Berlin / Heidelberg, pp. 842–851.