On the Effectiveness of IP Reputation for Spam Filtering

Holly Esquivel and Aditya Akella
University of Wisconsin-Madison
Email: esquivel@cs.wisc.edu, akella@cs.wisc.edu

Tatsuya Mori
NTT Service Integration Laboratories
Email: mori.tatsuya@lab.ntt.co.jp

***Abstract*—Modern SMTP servers apply a variety of mechanisms to stem the volume of spam delivered to users. These techniques can be broadly classified into two categories: pre-acceptance approaches, which apply prior to a message being accepted (e.g. IP reputation), and post-acceptance techniques which apply after a message has been accepted (e.g. content based signatures). We argue that the effectiveness of these measures varies based on the SMTP sender type. This paper focuses on the most light-weight pre-acceptance filtering mechanism – IP reputation. We first classify SMTP senders into three main categories: *legitimate servers*, *end-hosts*, and *spam gangs*, and empirically study the limits of effectiveness regarding IP reputation filtering for each category. Next, we develop new techniques that build custom IP reputation lists, which significantly improve the performance of existing IP reputation lists. In compiling these lists, we leverage a somewhat surprising fact that both legitimate domains and spam domains often use the DNS Sender Policy Framework (SPF) in an attempt to pass simple authentication checks. That is, good/bad IP addresses can be systematically compiled by collecting good/bad domains and looking up their SPF resource records. We also evaluate the effectiveness of these lists over time. Finally, we aim to understand the characteristics of the three categories of email senders in depth. Overall, we find that it is possible to construct IP reputation lists that can identify roughly 90% of all spam and legitimate mail, but some of the lists, i.e. the lists for spam gangs, must be updated on a constant basis to maintain this high level of accuracy.**

I. Introduction

Recent reports show that unsolicited bulk email, or spam, constitutes more than 90% of all messages sent or received today [1], [2]. Current approaches to identify and filter spam can be classified into two categories: those based on characterizing the properties of the sending SMTP server, and those based on analyzing the contents of the email. Because these two sets of approaches are applied at different stages of the receiving SMTP server accepting an email, they are also called *pre-acceptance* and *post-acceptance* tests, respectively.

Pre-acceptance tests can be further classified into two categories: those based on the reputation of IP address, i.e., IP reputation list filters, and those based on the characteristics of individual SMTP transactions, e.g., envelope from addresses, recipient addresses, and HELO/EHLO messages. Examples of the former approach are DNS Blacklists (DNSBL) [3], DNS Whitelists (DNSWL) [4], and other commercial IP reputation services such as [5] and [6]. Examples of the latter approach are greylisting [7], sender authentication [8], DNS validation [9], domain validation [10] and protocol defects [11].

978-1-4244-5489-1/10/$26.00 © 2010 IEEE

Both pre-acceptance and post-acceptance approaches impose significant overhead on SMTP servers. Pre-acceptance tests often require keeping the status of senders and retrieving special name server records in order to verify existing trust relationships before accepting an email. Additionally, post-acceptance tests involve remote look-ups for spam signatures, as well as running expensive local tests such as optical character recognition (OCR) and learning-based classifiers. Crucially, however, post-acceptance filtering is almost always preceded by pre-acceptance tests. Thus, effective and accurate pre-acceptance filtering can reduce the load on SMTP servers, and enhance their ability to identify and thwart spam. Thus, it is of no surprise that pre-acceptance filtering has recently been the focus of many spam filtering mechanisms.

In this paper, we focus our attention on the most light-weight and scalable pre-acceptance mechanism – IP reputation. The output of IP reputation can be used for effective and accurate anti-spam filtration in various ways: i.e., employing *selective* greylist filtering to IP addresses that have a *bad* reputation, bandwidth throttling of SMTP connections originating from *low* reputation scored senders, marking SMTP connections from *legitimate* senders to avoid having them falsely filtered by other anti-spam mechanisms, or employing priority queuing for SMTP connections from *legitimate* senders. DNSBLs and DNSWLs have been used to publish the IP reputation of a host. These lists have been so widely adopted that their structure, usage, and querying protocol is even being standardized in IRTF ASRG [3].

In this paper, we first classify email senders into three broad categories — *legitimate servers* (e.g. SMTP servers of Webmail providers, ISPs, enterprises, or universities), *end-hosts*, (e.g., compromised end-host machines), and *spam gangs* (e.g., sophisticated spam-sending companies). Through the extensive analysis of data sets collected at University of Wisconsin-Madison, we attempt to empirically quantify the effectiveness of current IP reputation approaches for each category. Specifically, we aim to answer the following questions: *What is the maximal fraction of emails that can be correctly classified using an IP reputation service? What guidelines should be followed in building effective custom IP reputation lists for each category?* To answer the above questions, we develop several techniques that build *custom* IP reputation lists that can account for the locality of email senders and thus complement the outcome of existing IP reputation services.

While many of the techniques used in this study are generalizations of existing techniques, we adopt them in a piecemeal fashion for each category; i.e., we combine them systematically to build accurate *custom* IP reputation lists that are suitable for an Internet edge site. Moreover, we develop a new technique that correlates domain names and IP addresses with SPF records [8]. While IP addresses are variable in nature, domain information is rather stable over time. Somewhat surprisingly our study reveals that DNS SPF records are not only utilized by legitimate (a.k.a. ham) email operators, but also some sophisticated spammers. Thus, we can systematically keep track the IP addresses of these spammers by examining the DNS SPF records for their domains.

As we shall see later, the existing and custom IP reputation lists make up roughly 90% of all spam and ham senders in our data set. We argue that quantifying the effectiveness of pre-acceptance anti-spam techniques requires us to understand the *contribution of different categories of SMTP senders* to the overall spam and ham email observed. This is because pre-acceptance filtering techniques differ in how effectively they can filter spam originating from these categories. In particular, we note the importance of understanding the characteristics of *legitimate* email senders. While many previous studies have focused on the characteristics of spam senders, to the best of our knowledge, there have been very few studies on legitimate email senders, whose properties are crucial in building effective IP whitelists. Since misclassifying legitimate messages as spam is prohibitively unacceptable, especially in enterprise email services, it is crucial that we know that legitimate senders will not be filtered in advance.

Our approach can be summarized as follows: For each message received at UW-Madison, we first test the effectiveness of off-the-shelf IP reputation services. We then present the effectiveness of our custom IP reputation mechanism. Finally, we study the properties of email senders in each category.

II. Three Categories of Email Senders

In this work, we classify email senders into three main categories, *legitimate servers*, *end-hosts*, and *spam gangs*. We note that the sender of an email in this paper means the last email relay before the message enters the receiver domain. This section describes each category, and presents examples of IP reputation services that are used to identify email senders of each category. As we shall show later, these categories make up 85-88% of all spam and ham messages seen in our data sets.

A. Legitimate servers

We define a "legitimate" server as privately owned infrastructure server which has been setup with the goal of allowing legitimate users to send email. Examples of legitimate servers include the outgoing SMTP servers of large email service providers such as Hotmail, Yahoo and Google; the SMTP servers of universities, enterprises, ISPs and third party mail service providers; the mail servers of Web portals offering free email service, etc.

Some legitimate SMTP servers are exploited by spammers and are used for sending spam messages. For example, user accounts at Web-based email service providers have been abused to send spam [12]. In recent years, the top email service providers have tried to enforce stringent Acceptable Use Policies (AUP) and tight controls over the account sign-up process [13] to stem the abuse, but service operators report the continued and growing misuse of email accounts [14].

In addition, spam could be received from other legitimate SMTP servers such as the outgoing SMTP servers of ISPs and enterprises, or SMTP servers deployed by third-parties for public use (the access could be paid or free). Spam received at a network location could seem to have "originated" from such servers under two situations: (1) a user of the server sources spam (either because the user is infected by malware, or because the user himself is a spammer) or (2) spam is sent to a user who has an account on the SMTP server which is being forwarded to the receiving location.

We note that whitelisting legitimate servers is crucial in employing "selective filtration" mechanisms, i.e., senders within a whitelist bypass the pre-acceptance filtration process to avoid potential false positives. Misclassifying legitimate messages as spam messages is prohibitively unacceptable, especially in the enterprise email domain. An example of publicly available legitimate server whitelist is DNSWL [4].

B. End-hosts

It is an accepted fact that a significant fraction of the spam today originates from end-user machines which are likely infected by malware with built-in SMTP engines [12], [15]–[17]. These infected hosts either originate spam themselves, or act as SMTP proxies for the actual spammers. In either case, the end-host user is likely unaware of the spamming activity. Accurate end-host IP blacklists can help completely eliminate spam from end-hosts. Such blacklists could have two sets of entries: individual IP addresses of known compromised end-hosts and IP address blocks which are known to be used by ISPs for clients from their cable-modem, DSL, and dial-up pools. End-host IP blacklists can be considered "static" in nature, in that they observe few deletions over time. Deletions are necessary when an end-host is set up to run a legitimate SMTP service or when ISPs renumber their hosts. Both of these are unlikely events.

Examples of IP reputation services for identifying end-hosts are PBL [5] and UDMap [18]. PBL is a publicly-available DNSBL database of end-user IP addresses, which largely includes address prefixes. PBL was developed out of the Dynablock [19] blacklist, which was originally a list of dial-up IP addresses. Part of the IP addresses in PBL are maintained by network service providers participating in the PBL project. In a recent study [18], it was shown that PBL misses several prefix blocks of dynamic IP addresses. In the same study, the authors developed a new approach, called *UDMap*, specifically targeted at identifying dynamic IP address blocks automatically. UDMap uses a time-ordered log of Hotmail user activity, that gives evidence of continued spamming activity at specific IP addresses. Based on the log, UDMap computes an entropy metric which quantifies the probability that the user who used a specific IP address is also using neighboring IP addresses. Based on these probabilities, UDMap derives the dynamic IP blocks out of a set of addresses. We chose to utilize PBL because of its popularity within the spam community, and UDMap because of its ability to identify blocks of dynamic IP addresses.

C. Spam gangs

In recent years, sophisticated spammers have set-up elaborate mechanisms to provide a false sense of legitimacy to their actions, and thwart spam filters. In one highly sophisticated scheme which is becoming popular, spammers pretend to be a dummy ISP or a colocation provider [5]. These spammers purchase bandwidth from real upstream ISPs pretending that the bandwidth shall be used for Internet connectivity for their "users". In some cases, these spammers may also register several bogus domain names, and even SPF resource records (RRs). These sophisticated spammers often also buy blocks of IP addresses, such as a /24 domain.

In reality, rich infrastructures are being used to send bulk spam messages. The bogus domains and SPF RRs help the spammers thwart reverse-DNS based filtering and simple SPF checks. Multiple IP addresses are employed to load-balance spam activity and prevent receiving SMTP servers from building sufficient spam history on any single IP address from a single end-point perspective.

In less sophisticated scenarios, spammers simply register domain names and create bogus SPF RRs, in order to send spam from a small number of IP addresses. The registered domain names are used in the spam message headers to give a sense of legitimacy to the receiver. When the spamming activity is caught by upstream ISPs, the spammers simply shift their entire operation to another unsuspecting ISP, thus using different IP addresses or address blocks altogether.

Whenever an entire block of IP addresses exhibits unacceptable spam behavior (prolonged abuse for solely sending spam messages), the spam gangs behind them can be thwarted by blacklisting the entire block of IP addresses. Similarly, less sophisticated spammers can be blacklisted by studying their sending history over a period of time. In both cases, the sending history can be monitored at a single vantage point. Some well known blacklists, such as SBL [5], are constructed in to use observations from both single vantage points as well as those at multiple vantage points to improve effectiveness. However, blacklisting becomes ineffective the moment the spam gang shifts it operation to another ISP.

III. Building custom IP reputation

In this section we outline the techniques we employ to build *custom* IP reputation lists for our three categories of SMTP senders. Our techniques use a few key pieces of information available from email message headers, such as, the to and from addresses, the message time stamp, and the sender's IP address, to derive correct classifications. Many of the techniques we use are generalizations of existing techniques. However, the existing techniques are adopted in a piecemeal and systematic fashion to build accurate custom IP reputation lists that are suitable for an Internet edge site. In addition to the existing techniques, we develop a new technique that associates *good* or *bad* domains with IP addresses by looking up their DNS SPF records; which enables us to reveal both legitimate and malicious email senders. As we shall see in the next section, this technique is effective in building whitelists for legitimate senders and accurate blacklists for spammers. We also leverage techniques from recent measurement studies that indicate that spamming botnets exhibit intrinsic TCP header patterns. Precise end-host blacklists can then be constructed by leveraging these patterns.

A. Legitimate Servers

To identify if an email sender is a legitimate server, we construct a whitelist of legitimate *domains* and check if the sender belongs to this domain whitelist by looking up its DNS SPF. We use two approaches to populate the whitelist.

*1) **WL1: Legit-Popular**:* We first derive the IP addresses of servers of well-known email service providers; we call this list **WL1**. To build this list we manually compile a list of popular email provider domains that offer their services in various languages. We obtained 458 domains in all.

Next, we leverage the SPF RR in each domain to identify authorized IP addresses and address prefixes which can originate emails for that domain. We found that a large portion of today's popular email servers have deployed SPF. This observation agrees with recent measurement studies such as [20]. We include both the IP addresses and address prefixes in **WL1**. For popular domains that do not publish SPF RRs (yahoo.com is an example), we manually compile the list of authorized IP addresses based on reverse DNS look-ups of all the IP addresses we observed in our logs.

*2) **WL2: SPF-good**:* Our second whitelist consists of email senders who have an impeccable email sending history. As shown earlier, we use a commercial spam detection software, which assigns each message a spam score. To compile this list, we first enumerate all the domains that appeared in the from-address field of the emails in our logs. From this list of domains, we pick out the subset of domains that have SPF RR. For each IP address that is associated with the SPF of a domain, we check its email activity over a one month period as observed in our data set. If, for each IP belonging to a domain, the number of messages sent out by the IP is larger than 10 and the fraction of spam messages is less than 0.1, then the list of entire IPs of the domain is added to **WL2**. We note that these thresholds are tuned conservatively to decrease false positives. We also note that the good domains we observed remain quite stable over time.

B. Compromised end-hosts

In order to identify if a sender is a compromised end-host, we first employ a blacklist that reorganizes some commonly-employed naming heuristics used by spam detection software. Our second blacklist leverages a technique from a recent measurement study on the characteristics of spamming botnets.

*1) **BL1: Hostname**:* The first heuristic flags a sender as a potential end-host based on the naming conventions of ISPs; that is, we employ a sequential test on successive IP addresses. Most IPs belonging to DSL, cable or dial-up pools are named by the ISPs using *sequential or similar names*. For example, sequential hosts using the ISP bnsi.net are named 12-5-51-80.static.bnsi.net, 12-5-51-81.static.bnsi.net, 12-5-51-82.static.bnsi.net, and so forth. This is likely done for easy administrative purposes and inventory tracking (see [21] for common best practices). To leverage these naming conventions in identifying end-hosts, we perform reverse look-ups for each sender IP (say 12.5.51.81), and for the IP addresses immediately preceding (12.5.51.80) and following the IP (12.5.51.82). We check the similarity of these names by computing the *Levenshtein Distance* (LD) [22] between the names for IP, IP-1, and IP+1. This metric measures the distance between two strings by counting the minimum number of operations needed to transform one string into the other. An operation may be insertion, deletion, or substitution of a character.

We consider an end-host IP to have passed the Neighbor Naming Test if $LD(IP, IP-1) < \theta_{LD}$ and $LD(IP, IP+1) < \theta_{LD}$ for some small threshold θ_{LD}. Setting $\theta_{LD} = 6$ covers most of the the naming conventions identified in [21]. Because the above step relies purely on similarity of names, it could suffer from false positives. Thus, each sender IP that passes the above heuristic is also subjected to two additional tests to identify if it was falsely identified.

First, we look for the reverse DNS (RDNS) name of the sender-IP to carry specific keywords which indicate that the IP belongs to the cable, DSL or dial-up provider networks (e.g. dsl, cable, telecom, telekom, ppp, dhcp, catv, wireless, broadband, 56k etc.). IPs which do not have these keywords in their RDNS names fail the Keyword-based test. We also use the numbers in IP addresses as keywords. If RDNS includes at least two numbers (or their hex expression) which are associated with the last two bytes of IP address, we pass the IP address on to the next test.

Second, we further check if the RDNS name includes keywordswhich indicate that the IP is likely to belong to a legitimate infrastructure server (such as mail, smtp, mx (but not those matching the pattern /.mx$/), web, www, dns, name, etc.). If the keywords are found, then the sender IP fails the Keyword-based test. Although some spammers intentionally use these keywords in the RDNS of their servers for evasion purpose, we exclude the senders with these keywords to avoid potential false positives. The excluded hosts employed by bad ISPs will be covered by our blacklists for spam gangs. If an IP passes the Neighbor Naming Test and the two Keyword-based tests then it is added to **BL1**.

*2) **BL2: Srizbi**:* Our second BL leverages a recent measurement finding that states that a particular type of spamming botnet, known as *full-kernel malware*, exhibits intrinsic TCP fingerprint signatures [23]–[25]. We monitor the TCP headers of incoming SMTP sessions, and add the sender IP address to **BL2** if the sender exhibits a TCP fingerprint that is associated with Srizbi botnet, which is known as the worst spamming botnet during the measurement period [24], [25].

*3) Verification of **BL1**:* Because we use naming characteristics, our above categorization could have both false positives and false negatives. In general, it is very difficult to completely check the validity of the end-hosts we identified using the naming tests, as many ISPs consider this sensitive information. Nevertheless, we check the accuracy of our approach by applying it to *known* dynamic IP addresses listed in PBL and UDMap, and quantify the fraction of dynamic IPs that are also identified by our simple tests. To do this, we first collected a list of sender-IPs found in email logs collected at UW-Madison for March 2008 that also appeared in PBL and UDMap. From these, we picked a random subset of 1M IPs which had RDNS names. We applied the neighbor naming and keyword tests to these IPs. A total of 980K IPs (98%) passed the neighbor naming test with $\theta_{LD} = 6$. Of these, 930K IPs (93%) passed the keyword test. The high overlap between the IPs identified by our approach and the UDMap and PBL lists indicates that our heuristic is fairly accurate.

We also tried to check if the IPs that were identified by the above two heuristics, but *not found* in UDMap or PBL were likely to be end-hosts. In particular, we used passive OS fingerprinting logs compiled during March 2008 to identify the OSes of the hosts in this category. We found that over 98% of the senders in this category used variants of Windows operating systems (including Srizbi-infected hosts), excluding server editions. Furthermore, we examined the emailing patterns of senders observed in our March 2008 data which used the variants of Windows or Srizbi signatures, and found that most of the emails received from these senders were spam. While not conclusive, these sets of checks provide an indication that our tests are capturing end-hosts IP addresses with high probability.

C. Spam gangs

We propose two blacklists which attempt to detect key operational modes of spam gangs. Our first blacklist identifies hyper-active spamming servers which consist of large IP blocks. Our second heuristic identifies spam gangs that misuse DNS SPF RRs to evade simple SPF checks.

*1) **BL3: Bad Blocks**:* As mentioned in previous section, professional spam gangs employ blocks of IP addresses to send spam. We look for this property in the email logs we collected. In particular, we employ the following steps: (1) First, sender IP addresses are mapped to BGP prefixes using global BGP tables [26]. (2) Prefixes are then picked which have at least k active IP addresses in them. Similar to Xie et al [18], we adopt $k = 8$ which is often the minimum unit size for IP address assignment. (3) Next, let n be the total number of addresses, active or inactive, in an address block. Let a_1 be the first active IP address in integer format, and a_n be the highest active IP address. Select blocks such that $|a_n - a_1 + 1| \leq n(1+\epsilon)$, which implies that the successive active IPs in an address block are nearly consecutive in the IP space. We set $\epsilon = 0.05$. We found that a larger threshold, such as, $\epsilon = 0.1$ can cover more bad blocks with good accuracy. However, we adopt the smaller threshold to reduce the chance of falsely identified blocks. A much smaller threshold, say $\epsilon = 0.02$, did not affect the results adversely. Among the collected blocks, we select a block if the block as a whole sent out more than 100 messages in a month, with a collective spam ratio exceeding 90% to add to**BL3**.

*2) **BL4: SPF-bad:*** As mentioned earlier, modern spammers have been publishing their own SPF-compliant domains, and sending spam messages from IP addresses associated with these domains. Thus, they are increasing the chances of evading SPF pre-acceptance filters. We leverage this fact to construct a blacklist of spam gang members. In particular, for all the domains appearing in the "from address" of the collected emails over a period of a month we resolve the SPF RRs and compile the list of valid IP addresses which are associated with the domains. We note here that unsophisticated spammers often create fake domain names, and insert both the domain names and the corresponding SPF records into the DNS system. They seldom spoof the sending domain (e.g. use yahoo.com in the from address), because it is easy to catch such spoofing. We also check the email history of each IP. If the number of messages sent out by an IP is larger than 10 and the fraction of spam messages is larger than 0.75[1], then the IP address is inserted into **BL4**.

IV. Data description

This section describes the data set we use for our analysis. We collected email logs at the University of Wisconsin-Madison's Department of Information Technology (DoIT) mail servers over a period of nine months between July 1, 2007 and March 31, 2008. According to University network administrators, these mail servers receive 80% of all external emails, i.e. emails originating outside the university.

For each email, we log the metadata, such as the from address, to address and message time stamp, along with the size of the email and the IP addresses of the sending mail relay. Each message is assigned a spam score between 0 and 1 using sophisticated checks for each email; the score indicates the probability of the message being spam. Like all spam detection software, it may suffer both from false positives and false negatives, although we expect these proportions to be small. DoIT's mail servers also receive several emails forwarded from other internal university servers, which we ignore because of difficultly in inferring the true source of the spam. We also note that greylist filtering is applied before a message is accepted for delivery. Roughly 80% of spam senders are filtered at this stage; we do not analyze these filtered hosts in this work.

We use a threshold score of 0.75 to identify a message as spam. Emails with a spam score below 0.25 are considered ham. The default setting for identifying spam for all user accounts on the University's mail servers is 0.5. In contrast, our threshold choices are much more conservative to ensure that our empirical study is not affected by misclassified emails.

To indicate the suitability of the thresholds we chose, in Figure 1 we show a CDF of scores assigned to the emails received during September 2007. As can be seen, our choice of thresholds clearly segregates email into spam and ham. Over the nine month period, an average of 40 million emails were received per month at UW-Madison, of which 27 million were classified as spam (68%) and 12 million were classified as ham (29%), and only 3% were unclassified (i.e., with a spam score between 0.25 and 0.75) using these thresholds.

The overall volumes of email, spam and ham, and the number of sender IPs observed remained roughly stable over nine month duration.

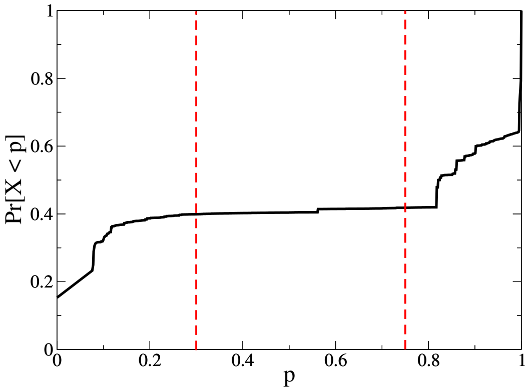


Fig. 1. CDF of email scores in the data set for September 2007. Dashed lines indicate the thresholds of $p = 0.25$ and $p = 0.75$.

V. Effectiveness of IP reputation

This section studies the effectiveness of IP reputation services from the viewpoint of an Internet edge site. In the following, we present the analysis of a one month log (March 2008) and omit the remaining analysis for brevity (the results for these months were similar). Each IP reputation list (PBL, SBL,UDMap, DNSWL) was gathered during the same period as when the SMTP logs were collected. We first study the performance of existing IP reputation services, namely, we use DNSWL [4], spamhaus BLs [5], and UDMap [18]. We then present how our techniques complement these existing solutions. Finally, we study the effectiveness of the history-based IP reputation lists over time.

A. Performance of IP reputation lists

1) Legit SMTP servers: We first study the effectiveness of whitelists, namely DNSWL [4], **WL1: Legit-popular**, and **WL2: SPF-good**. For DNSWL, we use all the trust levels. These trust levels indicate how often spam is sent from these servers where high indicates that the server never sends spam and none indicates that it may sometimes send spam. Table I shows the results of applying the whitelists. In the table, we study the number of IP addresses, spam, ham and unclassified messages which are covered by each of IP reputation lists. First, we find that small number of senders originate the ham

[1] We evaluated both less and more conservative thresholds, i.e., 0.5 and 0.9, found the choice of values was not sensitive to the overall results.

messages seen in the data set. Second, DNSWL identifies 56% of all ham messages, while **WL1** and **WL2** covers roughly 80% of all ham messages. That is, the custom whitelists cover more ham messages and also have a lower number of false positives compared to DNSWL. In particular, **SPF-good** has the lowest (potential) false-positives (0.2%=72K/31M). Finally, **Union** is the aggregation of the three whitelists. By leveraging DNSWL and the custom whitelists, we can cover more than 88% of total ham messages.

TABLE I
Effectiveness of whitelists (March 2008).

List	#IPs	#Spam	#Ham	#Unclassified
Total	5,160,210	31,831,274	11,834,098	826,862
DNSWL	23,762	484,855	6,648,228	231,581
Legit-popular	34,227	131,376	9,578,685	332,570
SPF-good	30,060	72,498	9,455,952	320,333
Union	49,612	546,141	10,400,068	387,810

2) End-hosts: Next, we study the effectiveness of the end-host blacklists: PBL, UDMap, **BL1: Hostname**, and **BL2: Srizbi**. Table II shows the results of these lists. First, the existing BLs, PBL and UDMap, are certainly effective in stopping spam from IPs belonging to dynamic IP addresses. These addresses contribute a huge volume of spam overall (42%). Second, we see that the amount of spam messages covered by the custom BLs is roughly 18% and 13% of the total spam messages. Finally, we note that the end-hosts covered by existing BLs and the custom BLs are not identical. Thus, the fraction of the spam messages covered by the **Union** increases the coverage up to 55% of all spam messages. This simple observation indicates that *blacklists of dynamic-IP address blocks can filter 55% of all spam.* This observation is similar to that made in past work on the effectiveness of such blacklists [18], [27]. We note that accuracy of the **Srizbi** list is especially high; we note that only 0.25% of total messages caught by the list are classified as (potential) ham[2] messages.

TABLE II
Effectiveness of End-host blacklists (March 2008).

List	#IPs	#Spam	#Ham	#Unclassified
Total	5,160,210	31,831,274	11,834,098	826,862
PBL+UDMap	4,014,156	13,619,609	146,334	140,134
Hostname	978,400	5,878,251	76,018	71,676
Srizbi	1,105,008	4,051,060	10,418	51,722
Union	4,388,812	17,530,909	224,903	199,842

3) Spam gangs: Next, we study the effectiveness of the spam gang blacklists SBL [5], **BL3: Bad blocks**, and **BL4: SPF-bad**. The coverage of these lists is shown in Table III. Among the three blacklists, **SPF-bad** is most effective in catching spam messages, and the addresses from this list contributed more than 35% of the total number of spam messages. In contrast, SBL was not effective for our data set. It is somewhat surprising that many spam senders are intentionally (mis)using the SPF authentication mechanism.

4) Effectiveness of Blacklists: Finally, we aggregate the BLs to analyze their effectiveness. Table IV shows the coverage of these blacklists when grouped together. While the existing BLs cover more email senders (possibly spammers), more spam messages are covered by the custom BLs. In total, the aggregated WLs and BLs IPs contribute more than 88% of email senders, more than 85% of total spam messages, and more than 88% of ham messages. Using these classified results, we study the sources of email in section VI.

TABLE III
Effectiveness of Spam Gang blacklists (March 2008).

List	#IPs	#Spam	#Ham	#Unclassified
Total	5,160,210	31,831,274	11,834,098	826,862
SBL	7,297	342,989	1,402	62
Bad blocks	33,573	3,150,770	19,275	10,835
SPF-bad	111,682	11,436,122	71,802	34,980
Union	132,760	11,931,074	84,250	39,720

TABLE IV
Effectiveness of aggregated blacklists (March 2008).

List	#IPs	#Spam	#Ham	#Unclassified
Total	5,160,210	31,831,274	11,834,098	826,862
PBL+UDMap+SBL	4,020,214	13,957,464	147,736	140,196
BL1+BL2+BL3+BL4	1,981,858	18,805,907	164,144	144,141
Union	4,472,718	26,818,235	305,704	231,935

B. The effectiveness of IP reputation lists over time

Among the custom IP reputation lists we build, three lists are compiled based on the history of hosts, i.e., **WL2: SPF-good**, **BL3: Bad Blocks**, and **BL4: SPF-Bad**. As we have shown, these lists play a crucial role in catching ham or spam messages. Our evaluation of these lists considered the ideal situation where an oracle computes the lists based on sending patterns over a large time interval, and the list is applied to all email received in the same interval. (Note that the senders in other categories such as popular servers and end-hosts did not require us to track history – these senders can be classified on the basis of static host properties like host names and long-term popularity.) Our evaluation thus shows the ideal extent to which pre-acceptance filtering could be useful.

In this section, we examine the practical challenges in building and maintaining the reputation lists. In particular, we examine if history-based lists collected on the basis of an email observed during one time interval will be effective for future time intervals, and we study how to choose the appropriate size of the interval and the update frequencies for the various lists. Since history-based lists are extracted from the observation of addresses in a specific time interval, the lists may not be able to cover the newly observed addresses in the next time interval. It is also useful to address the update frequency required to achieve good performance for each of the lists.

To answer these questions, we compiled the lists based on the email logs collected over the three distinct time intervals, namely one day (31 Sept., 2007), one week (25–31 Sept., 2007), and one month (Sept., 2007). We applied the whitelist and the blacklists to the email logs for the next month (Oct., 2007) and computed the number of senders identified using the lists during each day in the next month. We also computed the number of ham emails (for the SPF-good whitelist) and spam emails (for the two blacklists) contributed by the identified senders.

[2] While not conclusive, we note that these ham messages could be actually spam messages that are not caught by the spam detection software we used.

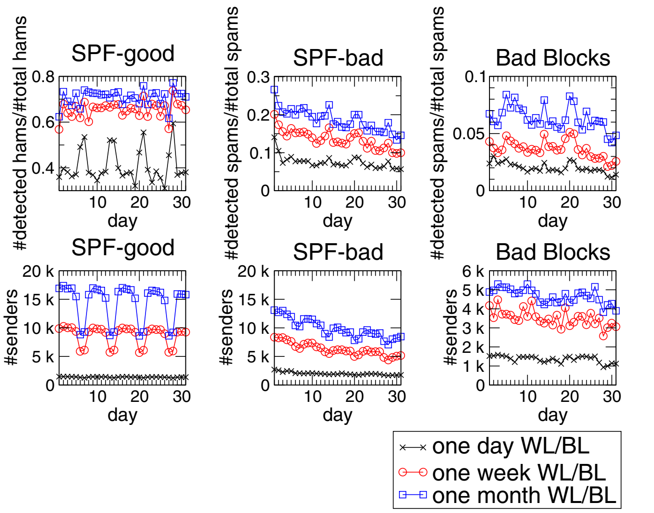


Fig. 2. Effectiveness of the history-based custom whitelist and blacklists. The coverage ratio of ham and spam messages for each list (top) and the number of senders for each list (bottom) is presented. Note the weekly patterns in the graph plotting the number of legitimate senders identified over time – fewer legitimate senders appear on weekends.

Figure 2 shows the results of these custom lists. First, we note that the number of legitimate senders identified each day increases as longer intervals are used to compute the SPF-good list. The list computed using one month's worth of data performs the best. The one computed using one week of data also performs reasonably well, especially in terms of the volume of ham originating from the identified senders. *This indicates that legitimate senders which contribute the most to the overall volume of ham can be whitelisted using a short time window.*

The differences between the one week and the one month lists are much more pronounced for the SPF-bad and the Bad Blocks blacklists. In these two cases, the one-month lists performs significantly better. This indicates that *in order to identify spam gangs and add them to these blacklists their sending behaviors need to be monitored over a fairly long time interval spanning a few weeks to a month.*

While the one month list performs the best in the case of SPF-bad and Bad Blocks, the number of senders identified each day using the one-month list and the volume of spam email filtered each day by applying this list falls significantly over time. In particular, the performance of the SPF-bad list drops as soon as the list is stale by 3 or more days. For the Bad Blocks list, the performance suffers once it is stale by 2 weeks or so. This is in contrast to the SPF-good list where both the number of legitimate senders and the volume of ham email they sent were both fairly stable over the entire month. *This indicates that the SPF-bad and the Bad Blocks blacklists which correspond to spam gangs must be updated periodically. Updating these lists once every few days for the former, and every 2 weeks or so for the latter would ensure that the blacklists are effective. In contrast, SPF-good can be updated much less frequently.*

TABLE V
CONTRIBUTION OF EACH CATEGORY

List	#IPs	#Spam	#Ham
Total	100 %	100 %	100 %
Legit Servers	1.0 %	1.7 %	87.9 %
End-hosts	85.0 %	55.0 %	0.5 %
Spam gangs	1.6 %	28.6 %	0.6 %
Hijacked prefixes	0.4 %	0.4 %	0.2 %
Open Relays/Proxies	0.9 %	2.6 %	0.1 %
Unclassified	11.1 %	11.7 %	10.7 %

VI. UNDERSTANDING THE SOURCES OF SPAM

As we discussed in section I, in order to quantify the effectiveness of pre-acceptance filtering mechanisms, one must understand the contribution of the above categories of hosts to the overall spam and legitimate email observed. If end-hosts contribute significantly to the overall spam while only originating a small amount of legitimate email then using simple static IP address blacklists will be very effective at limiting the overall spam volume. At the other extreme, if legitimate servers contribute a significant fraction of both spam and ham email, then pre-acceptance filtering is of limited help and effective content based filters must be developed. If spam gangs contribute the most spam, then blacklists may be very effective at filtering the spam, but the lists need to be refreshed very often. Also, multiple recipients may need to cooperate in order to quickly blacklist a spam gang operation.

In our study, we attempt to characterize the sender of each email message into one of the categories using techniques outlined in the previous section.

A. High-level differences among sender categories and their implications on pre-acceptance filtering

Our high level classification of sender IPs for our data set is outlined in the previous section is shown in Table V. This table also shows the relative volume of emails, spam and ham messages, originated by senders in each category. Since a sender could be caught by multiple lists, we apply the lists in the order of legit servers, end-hosts, spam gangs, hijacked prefixes, and open relays/proxies. The last two categories are other types of spam senders we investigate in section VI-C.

First, we focus on legitimate servers. From Table V, we note that legitimate servers constitute a small fraction of sender IPs (1%). However, they contribute a significant fraction of all ham messages (87.9%). The good news is that an *overwhelming volume of all ham* email originates from these servers. More importantly, the legitimate servers contribute a small volume of spam (1.7%). These observations indicate that it is highly beneficial to construct a whitelist of these servers and accept all email originating from these senders as the small amount of spam can be filtered using post-acceptance tests.

We focus next on end-hosts. From Table V, we note that *the majority of all spam,* — roughly 55% — seems to originate from end-hosts. End-hosts also make up an *overwhelming fraction of email sender IPs (85%)* and the proportion of legitimate messages they send is small (0.5%). These observations indicate that constructing a blacklist for end-hosts is quite beneficial. A good property of our custom BLs is that it uses static characteristics of senders, i.e., the RDNS of the IP address and the TCP header combination; thus, it does not require frequent classification updates. Our observations regarding end-hosts are in agreement with prior work [15], [18], which also has shown that a large fraction of spam originates from end-hosts. However, unlike our work, these studies do not examine where the rest of the spam originates from, and the implications these spam messages pose on pre-acceptance filtering.

We turn our attention to spam gangs. To the best of our knowledge, prior work has not examined the role of spam gangs relative to the email spam problem. From Table V, we note that spam gangs make up a very small fraction of senders (1.6%). On the other hand, they contribute 28.6% of all email spam. The average spam ratio for spam gangs is very high – nearly 96%. Thus, spam gangs are serious offenders in terms of the overall spam volume, and attention must be paid to developing mechanisms to thwart their activity.

We note that several spammers in the spam gang category rely on registering fake SPF records in an attempt to bypass filtering (which are caught by **SPF-bad**). These spammers make up majority of the IPs in the spam gang category in our data set. These senders look "legitimate" in many ways - they have SPF records and legitimate looking domain and host names which appear very different than the ones typically applied to end-hosts. In order to identify and blacklist these spammers, an SMTP server has to track their spam sending history, similar to the approach we employed in building the SPF-bad blacklist. As soon as high spamming history is identified, we mark the domain and extract its associated IP addresses from SPF record. As we studied in section V-B, in order to obtain enough history for compiling an effective list we need to monitor the logs over a fairly long time interval spanning a few weeks to a month. We argue that collecting the history information at multiple vantage points will help in accelerating this process.

Finally, we note that a small number of ham emails appear to originate from end-hosts and spam gangs put together (the former contributes 0.5% and the latter contributes 0.6%). Although this fraction is quite small relative to the volume of spam from these two sets of the senders, the fraction should ideally be *zero*, because spam gangs are dedicated to sending unsolicited emails, and end-hosts are rarely configured as SMTP servers [18]. Upon examining these messages, we found that the set of sender IPs responsible for sending these legitimate messages had an average spam ratio of 96.5%. This indicating that these messages come from heavy spammers. Thus, it is unlikely that we classified legitimate servers into end-hosts or spam gangs. In fact, the above observation indicates that some spammers are able to get a significant amount of emails through the spam detection software.

B. Spamming Characteristics

In this section, we analyze the spam-sending characteristics of the hosts in each of the three categories as observed at UW-Madison. In particular, we look at the activity profiles of spam senders. We analyze the number of messages received from each address, the number of days each sender was observed, and their respective spam history. For history, we monitor the relative amounts of spam and ham received from each sender IP. Our observations highlight the role of sender-analysis mechanisms, in particular those based on sending history, in identifying specific categories of spam senders. They also highlight the need for cooperation among spam recipients to construct effective blacklists.

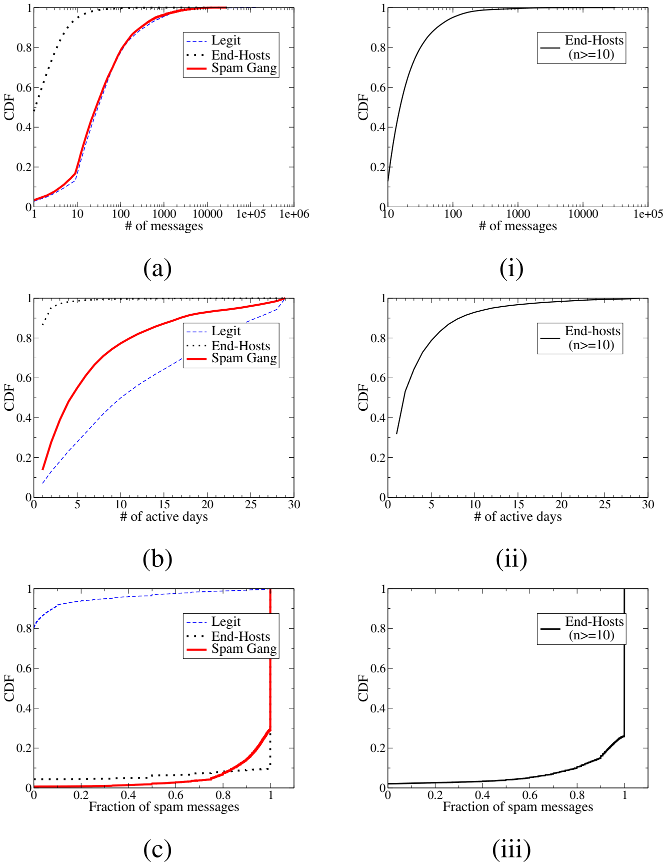
The results are shown in Figure 3. First, we focus on end-hosts. From Figure 3(a), we note that a majority of end-hosts send very few messages (90% send 7 or fewer messages in a month), and are active for a small number days (90% are observed on a just one day). Thus, history-based analysis is unlikely to be useful in capturing the spamming activity of end-hosts. As we mentioned earlier, blacklists for end-hosts can be constructed by having ISPs provide up-to-date lists of IPs assigned to end-users, as well as through the use of honeypots. In addition, collaboration among victims of spam can also help quickly build blacklists of end-hosts.

Since the distribution of email messages for end-hosts is very skewed (the majority of end-hosts sent out only one message in the month), we also plot the distribution of end-hosts that sent out $n \geq 10$ messages to the right (Figure 3 (i)–(iii)). We also note that majority of end-hosts are short-lived (Figure 3(b)) and are the sources of spam messages (Figure 3(c)). Figure 3 (i)–(iii) show that these characteristics also apply to end-hosts that send more than 10 messages – (1) more than 80% of the end-hosts are active less than a week (Figure 3(ii)) and (2) more than 70% of end-hosts send purely spam (Figure 3(iii)).

Spam gangs and legitimate servers send many more messages per host, and show very similar profiles to each other in terms of overall email volume. In both cases, 45% of the senders send 30 or more messages in a month (Figure 3(a)). However, these two categories differ significantly in terms of *when* the messages are sent (Figure 3(b)); the activity of legitimate servers is more spread out on average with 50% of the servers appearing 10 or more days each. In contrast, 50% of the senders in the spam gang category have appeared fewer than 4 days in the month. We do note that a significant fraction of IPs in the spam gang category (15%) are active for more than 20 days.

From Figure 3(c), we see that legitimate servers and spam gang senders also differ significantly in terms of the relative amounts of spam and ham sent. Almost all legitimate servers send a very small amount of spam relative to the amount of overall email; at most 20% of the email sent is spam. In contrast, 70% of senders in the spam gangs category send purely spam.

Taken together, these observations suggest that: (1) The spam ratio is useful in identifying spam gang operations. (2) The activity of the majority of senders in the spam gang category is concentrated over a few days. Thus, history-based approaches based at a single vantage point must rely on relatively short observation periods of under a day to flag



(a) (i)

(b) (ii)

(c) (iii)

Fig. 3. Characteristics of senders in the UW-Madison data set for the month of September 2007. Presented is the # of messages per address(top), # of active days per address (middle), and fraction of spam messages (bottom) for Legit servers, End-hosts, and Spam gangs (left) and End-hosts that sent out $n \geq 10$ messages in the month(right).

the activity of these senders, or recipients must collaborate to blacklist these senders.

C. Other spam sources

The previous section showed that there are roughly 15% of unclassified spam sources. One might assume that these senders are unsuitable for IP reputation filtering. We question this assumption by studying unclassified senders in detail. In particular, we check whether some unclassified senders share certain properties that can still make them amenable to IP reputation filtering.

1) Hijacked Prefixes: Route hijack-based spamming is another important category of spam senders. Past studies have shown that hijacking events last for a short duration, and hosts are abused for malicious activities such as spamming [15], [28], [29]. In recent years, real time hijack detection techniques have been developed that could be used to pass this information to the SMTP server for pre-acceptance spam filtering. We collect BGP announcement data from the RouteViews Project [26] for the same period during which the email data was collected. Then we apply algorithms developed by Hu et al. [29] to identify potential hijacking prefixes and the hijacking duration as observed by a server at University of Michigan. For each hijack instance, we collect the IP addresses inside the hijacked prefix that have sent emails within the hijacking duration.

Using this technique, we identified 19,121 senders who sent 129,798 spam messages, 25,731 legitimate messages, and 2,318 unclassified messages. Thus, these senders made very little contribution overall (see Table V). Interestingly, the fraction of ham messages over total messages sent from this category is fairy high (roughly 15%). The main reason behind that is that the prefix hijacking can only effect a certain region of the network. More specifically, the hijacked prefix must have a shorter or preferred route from our mail server, than the original prefix location. Hence, the hijacking event detected at the University of Michigan may not apply directly to University of Wisconsin data. Thus, we may conclude that senders of this category need to be filtered using post-acceptance tests.

2) Open Proxies / Open Relays: Open proxies and open mail relay servers have been abused by spammers with the rise of commercial Internet services in 1990's. Many DNSBL projects such as [5], [19], [30] have made great effort in compiling lists of these servers in the wild. As Xie et al. indicated, open proxies has been commonly used to *launder* the sources of email spam [31].

For our analysis, we used the lists of open proxy servers from several popular DNSBLs including XBL [5], DSBL [30], and NJABL [19]. We tested the unclassified senders using the union of these three lists. In total, they captured 48,057 senders, 812,089 of spam messages, 8,398 of legitimate messages, and 3,782 of unclassified messages. Again, the contribution of this category was not high (see Table V), but these senders are suitable for IP reputation filtering.

VII. Related work

There have been several studies related to spamming activity on the Internet, and various spam filtering solutions have been proposed. We provide an overview of these studies followed by a discussion of how our study complements them.

Several studies have attempted to mitigate spam using non-content based filtering methods. Sender characteristics are used in a similar approach proposed by Ramachandran et al. They employ clustering algorithms over the sending patterns of multiple senders in a given time window as an indicator of whether a sender is a spammer [32]. Venkataraman et al. showed that network-aware clustering of IP address spaces coupled with the spam history ratio of individual IP senders is effective in classifying email senders as spammers or not based on their IP [33]. In another non-content based approach, Beverly and Solins showed that transport-layer characteristics of email senders, e.g., number of retransmissions, minimum window advertised, and initial round trip time estimate, are effective in identifying spammers [34].

Complementing these works, and adding to the body of literature on spam filtering techniques, our paper shows the effectiveness of simple techniques based on blacklisting and whitelisting. We also provide a discussion of the techniques required for, and challenges involved in creating and updating these lists, and we discuss how these techniques (in particular, those pertaining to updates) must be tuned depending on the specific category of hosts that are generating the spam. We note here that Jung et al. in [35] made similar observations regarding keeping a small collection of popular DNSBLs (DNS blacklist of spam domains) up-to-date, but they do not examine the update time-scales relative to the specific category of spam senders in question. Also, they do not provide a discussion of the overall effectiveness of blacklist-based approaches in mitigating overall spam totals. A final key difference between our study and the prior works is that prior studies do not quantify how much spam originates from legitimate servers, and they do not study the role of whitelisting.

In recent years, botnets have emerged as a major tool for sending spam from end-hosts. Ways to identify the spamming bots have been explored in [12], [15]–[17], [23]–[25]. Note that, in contrast with these studies, our characterization of end-hosts is very broad and could involve hosts from botnets of various sizes, as well as other infected individual home computers. Thus, the botnet studies complement our work by allowing for a deeper analysis of the role played by infected end-hosts in generating spam.

VIII. Conclusion

Effective IP reputation filtering can significantly lower the load on email delivery systems today. Unlike previous studies, we classified SMTP senders into three main categories: *legitimate servers*, *end-hosts*, and *spam gangs*, and empirically studied the limits of effectiveness of IP reputation mechanisms for each category from an Internet edge view. We next developed new techniques that build custom IP reputation lists for each category, which significantly improves the performance of existing IP reputation lists. We also examined the implications of our analysis to pre-acceptance filtering mechanisms and the time-scales at which the activity of sending SMTP servers must be monitored in order to construct effective blacklists and whitelists. Overall, we find that it is possible to construct IP reputation lists that can cover 90% of all spam and ham, but IP lists for spam gangs must be updated on a constant basis for accuracy.

Acknowledgment

We thank Mike Blodgett, Jesse Thompson, and Jeffrey Savoy for their assistance in collecting the campus SMTP traffic. We thank Yinglian Xie and Fang Yu for providing a part of UDMap data set and Z. Morley Mao for providing hijacked prefix data set. This work was supported in part by an NSF CAREER Award (CNS-0746531), an NSF NeTS FIND Award (CNS-0626889) and a NDSEG Fellowship.

References

[1] “Its not about the spam,” http://googleblog.blogspot.com/2007/10/its-not-about-spam.html.

[2] “Spam Reaches All-Time High of 95% of All Email,” http://www.commtouch.com/Site/News_Events/pr_content.asp?news_id=942&cat_id=1.

[3] J. Levine, *DNS Blacklists and Whitelists*, IRTF Anti-Spam Rsearch Group, Nov 2008, internet Draft draft-irtf-asrg-dnsbl-08.txt.

[4] “DNS Whitelist - Protect against false positives,” http://www.dnswl.org/.

[5] “The Spamhaus Project,” http://www.spamhaus.org/.

[6] “SenderBase,” http://www.senderbase.org/.

[7] “Greylisting,” http://www.greylisting.org/.

[8] “Sender Policy Framework,” http://www.openspf.org/.

[9] “IRTF The Anti Spam Research Group wiki, DNS validation,” http://wiki.asrg.sp.am/wiki/DNS_validation.

[10] “Spam domain blacklist,” http://www.joewein.de/sw/blacklist.htm.

[11] R. Clayton, “Stopping outgoing spam by examining incoming server logs,” 2005.

[12] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, “Spamming botnets: signatures and characteristics,” in *SIGCOMM*, 2008.

[13] L. von Ahn, M. Blum, and J. Langford, “Telling humans and computers apart automatically,” in *Commun. ACM, 47(2):56–60, 2004.*, 2004.

[14] “Hotmail Operators: Private Communication.”

[15] A. Ramachandran and N. Feamster, “Understanding the network-level behavior of spammers,” in *SIGCOMM*, 2006.

[16] F. Li and M.-H. Hsieh, “An empirical study of clustering behavior of spammers and group-based anti-spam strategies,” in *CEAS 2006: Third Conference on Email and Anti-Spam*, 2006.

[17] K. Chiang and L. Lloyd, “A case study of the rustock rootkit and spam bot,” in *The First Workshop in Understanding Botnets*, 2007.

[18] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber, “How dynamic are ip addresses?” in *SIGCOMM*, 2007.

[19] “Not Just Another Bogus List,” http://www.njabl.org/.

[20] L. Eggert, “SPF Deployment Trends,” https://fit.nokia.com/lars/meter/spf.html.

[21] “Suggested generic DNS naming schemes for large networks and unassigned hosts,” http://tools.ietf.org/wg/dnsop/draft-msullivan-dnsop-generic-naming-schemes-00.txt, April 2006.

[22] “Levenshtein Distance,” http://en.wikipedia.org/wiki/Levenshtein_distance/.

[23] H. Esquivel, T. Mori, and A. Akella, “Router-level spam filtering using tcp fingerprints: Architecture and measurement-based evaluation,” in *CEAS*, 2009.

[24] T. Mori, H. Esquivel, A. Akella, A. Shimoda, and S. Goto, “Understanding the world.s worst spamming botnet,” *University of Wisconsin Madison Tech Report TR1660*, June 2009.

[25] H. Stern, “The rise and fall of reactor mailer,” in *Proc. MIT Spam Conference 2009*, Mar 2009.

[26] “BGP Tables from the University of Oregon RouteViews Project,” http://moat.nlanr.net/AS/data.

[27] A. Ramachandran, N. Feamster, and D. Dagon, “Revealing botnet membership using DNSBL counter-intelligence,” in *SRUTI*, 2006.

[28] X. Hu and Z. M. Mao, “Accurate real-time identification of ip prefix hijacking,” in *IEEE Symposium on Security and Privacy*, 2007.

[29] C. Zheng, L. Ji, D. Pei, J. Wang, and P. Francis, “A light-weight distributed scheme for detecting ip prefix hijacks in real-time,” *SIGCOMM*, 2007.

[30] “Distributed Sender Black List,” http://dsbl.org/usage.

[31] M. Xie, H. Yin, and H. Wang, “Thwarting Email Spam Laundering,” *ACM Transactions on Information and System Security*, 2008.

[32] A. Ramachandran, N. Feamster, and S. Vempala, “Filtering spam with behavioral blacklisting,” in *CCS*, 2007.

[33] S. Venkataraman, S. Sen, O. Spatscheck, P. Haffner, and D. Song, “Exploiting network structure for proactive spam mitigation,” in *USENIX Security*, 2007.

[34] R. Beverly and K. Sollins, “Exploiting transport-level characteristics of spam,” in *CEAS*, Aug. 2008.

[35] J. Jung and E. Sit, “An empirical study of spam traffic and the use of DNS black lists,” in *IMC*, 2004.