On the Effectiveness of Pre-Acceptance Spam Filterning

Tatsuya Mori
Holly Esquivel
Aditya Akella
Z. Morley Mao
Yinglian Xie
Fang Yu

Technical Report #1650

February 2009

On the Effectiveness of Pre-Acceptance Spam Filtering

Tatsuya Mori[†], Holly Esquivel[*], Aditya Akella[*], Z. Morley Mao[‡], Yinglian Xie[**], Fang Yu[**]
[*]University of Wisconsin-Madison, [†]NTT, [‡]University of Michigan, [**]Microsoft Research-Silicon Valley

***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 blacklisting and whitelisting), and post-acceptance techniques which apply after a message has been accepted (e.g. content based signatures). In recent years, pre-acceptance techniques have attracted a lot of attention. In addition to cutting down spam, effective and accurate pre-acceptance filtering is crucial to reducing the load on SMTP servers.**

In this paper, we empirically study the limits of effectiveness of pre-acceptance approaches. In our study, we first classify SMTP senders into three main categories: end hosts, legitimate servers and spam gangs. We argue that both the effectiveness and the role played by pre-acceptance approaches differ significantly across spam sent by the hosts in these categories.

We find that end-hosts make up over 88% of all senders and contribute nearly 54% of all spam. Spam gangs make up less than 1.2% of all senders, but contribute more than 30% of all spam. Both these sets of spammers can be filtered using address blacklists. However, we find that the blacklists corresponding to spam gangs may have to be updated as frequently as once every few days in order to be effective. We find that legitimate servers make up less than 1% of all e-mail senders, and contribute less 0.4% of all spam. Furthermore, these servers send an overwhelming fraction of all ham. Thus, simple whitelisting can be employed to permit all e-mail from them. Whitelists of legitimate servers can be constructed relatively easily and updated infrequently.

On the whole, we find that it is possible to build effective pre-acceptance filters which can eliminate nearly 90% of all spam today.

I. Introduction

Recent reports show that unsolicited bulk e-mail, or spam, constitutes more than 90% of all messages sent or received today [3], [6]. There are many current approaches to identify and filter spam. These can be classified into two categories: those based on characterizing properties of the sending SMTP server, and those based on analyzing e-mail contents. Because these sets of approaches are applied at different stages of the receiving SMTP server accepting an e-mail, they are also called *pre-acceptance* and *post-acceptance* tests, respectively.

Pre-acceptance tests try to derive characteristics of the e-mail sender that can help in identifying the e-mail as spam. These tests apply to the initial handshake, prior to message reception. The tested characteristics include whether the host belongs in a blacklist, whether the host is "trusted", verifying the spam-sending history of the host etc. When the characterization is inconclusive, it is combined with post-acceptance tests on the message body to determine if the message is spam.

Both post and pre-acceptance approaches impose significant overhead on SMTP servers. Pre-acceptance tests often require receiving SMTP servers to query remote databases, track senders' e-mail histories, and retrieve special name server records to verify trust relationships. Additionally, post-acceptance tests also involve remote look-ups for spam signatures, as well as running expensive local tests such as 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 improve the load on SMTP servers, and enhance their ability to identify and thwart spam. It is no surprise that pre-acceptance filtering has received much recent attention.

In this paper, we attempt to quantify empirically *the limits of effectiveness of pre-acceptance approaches in filtering spam*. Specifically, we aim to answer the following questions: What is the maximal fraction of e-mails that can be filtered using pre-acceptance schemes? What guidelines should be followed in building effective pre-acceptance filters? We believe that the answers to these questions will underscore the relevance of pre-acceptance filtering in the fight against spam, and provide configuration guidelines for these approaches to achieve better spam detection accuracy.

We argue that quantifying the effectiveness of pre-acceptance filtering requires us to understand the *contribution of different categories of SMTP senders* to the overall spam and legitimate e-mail observed. This is because pre-acceptance techniques differ in how effectively they can filter spam originating from these categories. We focus on three broad categories—end hosts (or home machines), legitimate servers (e.g. SMTP servers of Enterprises or Universities) and sophisticated spammers, or "spam gangs".

Our empirical study is based on a large corpus of several million e-mails, which spans nine months of collected at UW-Madison. Our approach can be summarized as follows: For each message received at UW-Madison, we identify the category that the sending SMTP sender belongs to using simple set of heuristics. Then, we systematically compute the relative amounts of spam vs legitimate e-mails received from all hosts of a given category. Using our approaches, we are able to classify 91% of the e-mail senders. The classified spam senders contribute 86% of all spam messages.

Our investigation reveals that end-hosts make up over 88% of all e-mail senders and contribute 54% of all spam. Spam gangs make up less than 1.2% of e-mail senders, but contribute more than 32% of all spam. Our empirical analysis shows that it is possible to construct a *blacklist* encompassing spammers from these two categories. A part of the blacklist would contain offending IP address blocks, and another part would contain offending individual IP addresses of spammers. The blacklist can be constructed by analyzing naming and addressing properties of senders, supplemented in some cases by monitoring the spam sending history of individual senders or groups of senders over time. When such blacklists are employed in the pre-acceptance stage, they can ideally filter 86% of all spam.

We also identify a small collection of legitimate servers, some of which belong to popular domains. These constitute less than 1% of the e-mail senders, and contribute only 0.4% to the spam volume. However, this small collection of servers generates 70% of all ham messages. Based on these observations, we argue that a small *whitelist* of these legitimate servers should be maintained. The white list can be constructed based on monitoring sending history and on the basis of the popularity of e-mail domains. When applied in the pre-acceptance stage, we show that such a whitelist would let through a large fraction of all ham. Content-based post-acceptance filtering can then be applied to filter the small amount of spam from the legitimate servers in the whitelist.

We also investigate the challenges in building and updating pre-acceptance filters, specifically, the black- and whitelists mentioned above. Our investigation shows that whitelists based on sending history can be constructed by monitoring server e-mail patterns over relatively short durations (say, one week). Also, the whitelists can be updated relatively infrequently. However, in order to construct some blacklists, especially those covering spam gang operations, the spamming activity of hosts may have to be monitored over multiple weeks. Also, the blacklist must be updated as frequently as once every few days in order to be effective.

Thus, we find that it is possible to construct pre-acceptance filters that can eliminate nearly 90% all spam.

In Section II, we describe various pre-acceptance filtering techniques, as well as the three categories of spam senders we study. In Section III, we describe the techniques we employed to characterize the senders in our data sets into these categories. In Section IV, we present the results from applying our analysis to the UW-Madison logs. We discuss related work in Section V and conclude in Section VI.

II. Spam Techniques and Pre-Acceptance Filtering

This section first describes the most common pre-acceptance filtering techniques. Following this, the three categories of SMTP e-mail spam senders are described. Finally, we outline our empirical approach to studying the effectiveness of pre-acceptance techniques.

A. Pre-acceptance Techniques

As mentioned earlier, pre-acceptance techniques attempt to derive properties of the e-mail sender, which can potentially aid in filtering spam from the sender. There are four common techniques in use today.

(1) Blacklists (BLs). Each list contains IP addresses of known spam senders. SMTP servers can query these lists to determine if an e-mail is likely spam based on the IP of the sender. Since most lists are queried using DNS, these blacklists are also referred to as DNSBLs.

There are several different categories of blacklists. The first is a list of IP address blocks of dynamic and dial-up IPs. These lists are constructed based on knowledge of how ISPs allocate addresses to end-users. In some cases, the lists are maintained by the owner of block or ISP. The second is a list of compromised hosts which are abused as mail relays by spammers. These are collected by monitoring activity of compromised hosts at spam blackholes and honeypots, as well as based on feedback from victims of spam. The third is a list of spam gangs (more below) or sophisticated spammers. This list is constantly updated as spammer activity changes.

(2) Sender Authorization or Whitelisting (WL). These techniques attempt to filter spam by verifying if the sender e-mail address is authentic. The verification is done by querying DNS Resource Record (RR), e.g. Sender Policy Framework resource records (SPF RRs), for the domain in the sender address. The SPF RRs contain a list of IP addresses which are authorized to send e-mail for the domain. This approach is similar to maintaining a distributed IP-address whitelist. As our evaluation shows, sender authorization testing is inconclusive when checking for spam because many spammers have started to register bogus SPF RRs.

(3) Sender Analysis. This includes approaches which track properties of the sender over time, such as the spamming history and number of e-mails received per day. As a standalone mechanism, sender analysis is effective only if the sender has a history of sending spam messages predominantly (say $> 90\%$ of the messages). Sender analysis is usually employed to aid in the construction of spammer blacklists.

Another common approach is to check if the sender has a valid RDNS name. If no RDNS mapping exists, the sender IP is identified as most likely an end-host. This test is often employed to associate a spam weight with the e-mail received, but never used by itself to filter spam. The weight is used to aid post-acceptance filtering.

(4) Greylisting. In this popular approach, the recipient drops the first SMTP connection from a sender and stores the IP address of the sender in a greylist. An SMTP connection request is allowed only if the IP address requesting the connection is in the greylist, because this indicates a retry from the IP address. A significant fraction of spammers do not retry and thus get filtered. In most typical scenarios, greylisting precedes most other tests such as checking in a blacklist, whitelisting, and sender analysis.

Although greylisting is very effective today, spammers are becoming increasingly sophisticated and can easily emulate full SMTP functionality. Also, greylisting could leave SMTP servers susceptible to resource exhaustion attacks and hence it is not employed by a significant fraction of SMTP servers. Thus, we believe that the other three approaches will grow in importance for pre-acceptance filtering.

Our study focuses largely on the effectiveness of the first three passive pre-acceptance approaches, all of which are widely employed as anti-spam defenses.

B. Three Categories of Spam Senders

End-User machines. 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 [11], [16], [17], [23]. These infected hosts either originate spam themselves, or act as SMTP proxies for the actual spammers. In either case, the end-host user is unaware of the spamming activity. Accurate end-host IP blacklists can help completely eliminate spam from end-hosts. 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.

Spam Gangs. In recent years, sophisticated spammers have set-up elaborate mechanisms to provide false sense of legitimacy to their actions, and thwart spam filters.

In one highly sophisticated scheme which is becoming very popular [7], spammers pretend to be a dummy ISP or a colocation provider. These spammers purchase bandwidth from real upstream ISPs pretending that the bandwidth shall be used to enable Internet connectivity for their "users". In some cases, the spammers may also register several bogus domain names, and even SPF RRs. The sophisticated spammers often also buy a block of IP addresses, typically a /24.

In reality, the bandwidth is used to send spam. The bogus domains and SPF RRs help the spammers thwart reverse-DNS based filtering and SPF checks, respectively. Multiple IP addresses are employed to load-balance spam activity and prevent receiving SMTP servers from building sufficient history on any single IP address.

In less sophisticated scenarios, spammers simply register domain names and create bogus SPF RRs, and send spam from a small number of IP addresses. The registered domain names are used in the spam message headers.

When the spamming activity of sophisticated spammers is caught by upstream ISPs, the spammers shift their entire operation to another unsuspecting ISP, thus using different IP addresses or address blocks altogether. Even the less sophisticated gangs are known to move their operations around.

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. However, monitoring is more effective when multiple recipients collaborate and combine their observations.

Some well known blacklists, such as SBL, are constructed in this fashion using observations both at single vantage points as well as those at multiple vantage points. However, blacklisting becomes ineffective the moment the spam gang shifts it operation to another ISP. Thus, in contrast to end-host blacklists, spam gang lists must be refreshed constantly.

“Legitimate” SMTP servers. User accounts at Web-based e-mail service providers could be abused to send spam. In recent years, the top e-mail service providers have tried to enforce stringent AUPs and tight controls over the account sign-up process [21] to stem the abuse, but service operators report continued and growing misuse of e-mail accounts [2].

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) and (2) spam to a user who has an account on the SMTP server is being forwarded to the receiving location.

No pre-acceptance approach is effective against spam originated from legitimate servers because such servers contribute a significant amount of legitimate e-mail themselves, they implement the SMTP protocol accurately, and have the necessary authorization records installed. To filter such spam, post-acceptance filters must be used.

Our Approach. The above discussion indicates that in order to quantify the effectiveness of pre-acceptance filtering, one must understand the contribution of the above categories of hosts to the overall spam and legitimate e-mail observed. If end-hosts contribute significantly to the overall spam, and a very small amount of legitimate e-mail originates from them, 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 legitimate e-mail, then pre-acceptance is only of limited help and effective content based filters must be developed. If spam gangs contribute most spam, then blacklists may be very effective at filtering the spam, but the blacklists may have 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 e-mail message into one of these three categories using simple tests on the IP address, RDNS mappings, SPF RRs and overall spamming activity. Many of the techniques we employ are generalizations of existing techniques. However, the existing techniques are adopted in a piecemeal fashion, with different SMTP servers employing different subsets of techniques. We combine them in a systematic fashion to gain an accurate understanding of the contributions of different SMTP sender types to the overall spam.

III. Techniques for Characterizing Spam Senders

We outline the techniques we employ to classify spam senders into one of the three categories defined above. Our techniques use a few key pieces of information available in e-mail message headers such as the from and to addresses and the message timestamp, along with the sender’s IP address, to derive the classification.

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 e-mail. Examples of legitimate servers include the outgoing SMTP servers of large e-mail service providers such as Hotmail, Yahoo and Google, the mail servers of Web portals offering free e-mail service, the servers of universities and enterprises, third party mail server providers, banks sending e-mails to their clients, servers of ISPs, etc.

To identify if an e-mail sender is a legitimate server, we first construct a whitelist of legitimate servers and check if the sender belongs in this whitelist. We use two approaches to populate the whitelist.

The Legit-Popular Whitelist. We first derive the IP addresses of the servers of well-known e-mail service providers; we call this list Legit-Popular. To build this list, we compile a list of popular e-mail provider domains by searching the Web for the term “free email” in various languages and retrieving the domain names for the top 1000 results of each search. Some domains were clearly false positives (e.g. www.emailaddresses.com which is an e-mail address directory) and were manually pruned. 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 e-mails that use the domain in the from address. We found that many of today’s popular e-mail servers, including the large e-mail service providers, enterprise, and academic have adopted the SPF framework. We include both the IP addresses and the address prefixes in the Legit-Popular whitelist. 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 lookups of all the IP addresses we observed in our logs.

Overall, our Legit-Popular list consists of 7108 single addresses and 171 prefixes; of the latter, 90 were /24 network prefixes and 5 were /16 network prefixes.

The SPF-good Whitelist. Our second whitelist consists of e-mail senders who have an impeccable e-mail sending history.

To compile this list, we first enumerate all the domains that appeared in the from-address field of the e-mails in our logs. From this list of domains, we pick out the subset of domains that have SPF RRs. For each IP address that is associated with the SPF RR of a domain, we check its e-mail activity over a one month period, as observed at one of our vantage points. 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 entire list of IPs is added to SPF-good. As shown later, we use a commercial spam detection software, which assigns each message a score that indicates the probability of being spam.

B. End-Hosts

In order to identify if a sender is an end-host, we use two checks in a sequential fashion: first, we check the sender's IP address against blacklists of address prefixes which are purportedly assigned by ISPs to their end-users. Second, we recreate some commonly-employed heuristics used by spam detection software, to identify if the sender is an end-host. We apply the second test to senders which fail the first.

1) End-Host IP Blacklists: We use two sets of *address blacklists* for testing if the IP address on an e-mail belongs to an end-host: PBL [8] and UDmap [22].

PBL is a publicly-available DNSBL database of end-user IP addresses, which largely includes address prefixes. PBL was developed out of the Dynablock [5] blacklist, which was originally developed as 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 [22], 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. We obtained the UDMap list corresponding to the time period over which our analysis was conducted.[1]

2) More End-hosts: As Xie et. al point out in [22], the UDMap tool may not identify several active blocks of dynamic IP addresses. This is because the tool is based on user login activity tracked over a limited window of time at a single network vantage point. The PBL blacklist is itself known to be incomplete in various ways [22].

In order to identify other end-host senders which evade the first check against UDMap and PBL, we use a collection of popular heuristics that some advanced spam filtering software (such as Sophos Pure Message) employ to identify if the sender IP address belongs to an end-host. These heuristics leverage common naming conventions employed by ISPs to label hosts in dial-up, dsl and cable Internet pools [9]. Obviously, they only apply to spam senders whose IPs have a valid reverse name. The heuristics apply on a per-IP basis.

We apply the following two heuristics:

(1) Neighbor Naming Test: The first heuristic flags a sender as a potential end-host based on naming conventions of ISPs. Most IPs belonging to DSL, Cable or Dial-up pools are named by the ISPs using *sequential or similar names*, e.g., hosts using 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, etc. This is done mostly for administrative purposes and inventory tracking (see [9] for common best practices). To leverage this naming convention in identifying end-hosts, we perform reverse lookups for each sender IP (say 12.5.51.81), and for the IP addresses immediately preceding (12.5.51.80) and immediately following the IP (12.5.51.82). We then check the similarity of these names by computing the *Levenshtein Distance* (LD) [4] between the names for IP, IP-1, and of IP, IP+1. This metric, commonly used in information theory, measures the edit 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 a spammer 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 [9].

(2) Keyword-based test: Because the above step relies purely on similarity of names, it could suffer from false positives. In particular, e-mail service providers and end-networks may name their mail servers using sequential names. Thus, each sender IP that passes the above heuristic is also subjected to two additional tests to identify if it was a false positive. First, we look for the RDNS name of the sender-IP to carry specific keywords which indicate that they belong in 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.

If the IP passes the test, then we further check if the RDNS name include keywords in the most specific portion of the RDNS name which indicate that the IP is likely to belong to an infrastructure server (such as mail, smtp, mx (but not /ṁx$/), web, www, dns, name, etc.). If the keywords are found, then the sender IP is considered to fail the keyword based test.

To summarize, to identify if a sender is an end-host, we first check if it belongs in UDMap or PBL. If it does, then we conclude that it is an end-host. If it doesn't, then we check whether the IP passes both the neighbor test and keyword-based test. If it does, then the IP belongs to an end-host; otherwise, it doesn't.

Verifying the tests. Because we use naming characteristics, our above categorization could have both false positives and false negatives. Next, we present a small set of checks which suggest that our categorization of end-hosts using the above tests is fairly accurate. In general, it is very difficult to completely check the validity of the end-hosts we identified using the naming tests. There is very limited knowledge of how ISPs manage the naming for such hosts, 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 or UDMap and quantify what fraction of the dynamic IPs are also identified by our simple tests. To do this, we first collected a list of sender-IPs found in e-mail logs collected at UW-Madison for Mar 2008 that also appeared in PBL and UDMap end-host blacklists. From these, we picked a random subset of 1M IPs which had reverse DNS names. We applied neighborhood naming and key-word tests to these IPs. A total of 980K IPs (98%) passed the neighborhood naming

[1] UDmap uses a time-ordered log of Hotmail user activity, that gives evidence of continued related activity at specific IP addresses. Based on the log, UDMap computes an entropy metric to quantify the probability of user who appeared to use IP address also using neighboring IPs. Based on this probability, UDMap derives the dynamic IP block.

test with $\theta_{LD} = 6$. Of these, 930K IPs (93%) overall 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 checked 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 Mar 2008 to identify the OSes of the hosts in this category. We found that over 85% of the senders in this category used variants of the Windows operating system. Furthermore, we examined the e-mailing patterns of senders observed in our Mar 2008 data which used the same variants of Windows, and found that in a large fraction of cases, the e-mails received from these senders were overwhelmingly spam.

While not conclusive, these sets of checks provide an indication that our tests are capturing end-hosts IP addresses with very high probability.

C. Spam Gangs

In order to verify a sender-IP is part of a spam gang operation, we use two approaches. The first approach is based on checking membership in existing blacklists. The second approach employs heuristics which attempt to identify key operational mechanisms of spam gangs such as employing a large block of addresses simultaneously, and/or registering fake sender authorization records.

1) Spam Gang Blacklist: SBL [7] is a publicly available DNSBL that collects verified spam gangs and spam support services. As we noted earlier, spam gangs constantly change their ISPs and operations to evade detection. As a consequence, the SBL ends up being incomplete.

2) Identifying Spam Gang Characteristics: In addition to the above, we use two heuristics which attempt to detect key operational modes of spam gangs, such as the ones described in Section II-B. Using each heuristic, we construct our own blacklists of spam gang operations. Our first heuristic identifies sophisticated spam gangs which abuse large IP blocks. Our second heuristic identifies the less sophisticated ones.

We construct and use the blacklist in a two-pass operation. In the first pass, we apply the heuristics to our entire e-mail logs to construct the respective blacklists. Then, in a second pass, we check the sender-IPs found in our logs against the blacklist to verify it it belongs to a spam gang operation.

Blacklist 1: Blocks of hyper-active spammers or "Bad Blocks". As mentioned in Section II-B, sophisticated spammers employ blocks of IP addresses to send spam. We look for this property in the e-mail logs we collected. In particular, we employ the following steps:

(1) Map sender IP addresses into BGP prefixes using global BGP tables [1]. (2) Pick prefixes which have at least k active IP addresses. Similar to Xie et al, we adopt $k = 8$ which is often the minimum unit for IP address assignment. (3) 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 $n \geq (1-\epsilon) * |a_n - a_1 + 1|$, implying that the successive active IPs in an address block are (almost) footnote consecutive in the IP space. We set $\epsilon = 0.05$ [2].

Among the collected blocks, we pick out the ones which are heavy spam senders. In particular, 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%.

Blacklist 2: SPF-bad. As mentioned earlier, modern sophisticated spammers have been publishing their own SPF-conformant domains and sending spam messages from the block of IP addresses associated with these domains; thus they are increasing chances of evading pre-acceptance filters. We leverage this fact to construct a blacklist of spam gang members. In particular, from all the domains appearing in the "from addresses" of e-mails collected over period of a month at one of our vantage points, we resolve the SPF RRs and compile the list of valid IP addresses which are authorized to be 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 RRs 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 e-mail 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[3], then the IP address is inserted into our SPF-bad blacklist.

To summarize, in order to identify if a sender IP belongs to a spam gang operation, we first check its membership in SBL. If found, the IP belongs to spam gangs. If not, we check in the two blacklists we constructed above. If the IP is found in either blacklist, we conclude that it belongs to a spam gang.

D. Applying the Heuristics

To minimize false positives and misclassifications, we employ the techniques defined above in a specific order for each sender IP. In particular, we always apply the legitimate server tests first. Based on the way we constructed the two whitelists that are employed in this approach, we feel that there is very little scope of false positives (i.e. IPs being wrongly classified as legitimate). The remaining the two sets of tests can be applied in any order. We chose to first apply the end-host tests, followed by tests for spam gangs.

Each IP is classified into a category based on the first set of tests it passes. For each classified IP, we also tally the volume of spam and ham it contributes. Finally, we aggregate the volumes of ham and spam for all IPs in each category. IPs which fail all test are considered unclassified.

IV. Analysis of E-mail logs

Data description. We collected e-mail logs at the University of Wisconsin-Madison's Department of Information Technology 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 e-mails i.e. e-mails originating outside the university.

For each e-mail, we log the metadata, such as the from address, the to address and the message timestamp, along with

[2] We found that larger threshold, such as, $\epsilon = 0.1$ can cover more bad blocks with good accuracy. However, we adopt the smaller theshold to make the results of the heuristics as strict as possible and to reduce the chance of blocks being falsely identified. A much smaller threshold, say $\epsilon = 0.02$, did not affect the results adversely.

[3] 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.

the size of the e-mail and the IP addresses of the sending mail relay. The likelihood of an e-mail being spam is tracked using Sophos Pure Message (SPM) – a commercial spam detection technique. SPM assigns each message a spam score between 0 and 1 using sophisticated checks for each e-mail; the score indicates the probability of the message being spam. Like all spam detection software, SPM may suffer both from false positives and false negatives, although we expect that these proportions to be small. The Department of IT's mail servers also receive several e-mails forwarded from other university-internal servers. It is difficult to infer the true source of spam from the meta-data for the forwarded e-mails and hence we ignore them in our analysis.

We use a threshold of 0.75 on the spam score to identify spam. E-mails 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 choices of the thresholds are much more conservative. Our conservative choice ensures that our empirical study is is not affected by misclassification of e-mails.

To indicate the suitability of the thresholds we chose, in Figure 1 we show the CDF of scores assigned to e-mails received during September 2007. As can be seen, our choice of thresholds clearly segregates e-mail into spam and ham. Over the nine month period, an average of 40 million e-mails 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 classified as gray (i.e., with a spam score between 0.25 and 0.75) using these thresholds.

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

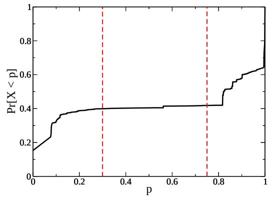

Fig. 1. CDF of E-mail score in the dataset for September 2007. Dashed lines indicate the thresholds of $p = 0.25$ and $p = 0.75$.

A. Sender Characterization and Pre-acceptance filtering

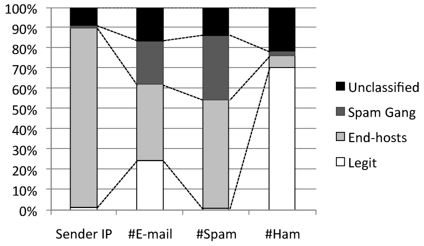

Fig. 2. Classification of Spam e-mail in the dataset for September 2007.

In this section, we present the results from applying the techniques outlined in the previous section to the data collected in September 2007.

Our high level classification of sender IPs observed in this data as outlined in the previous section is shown pictorially in Figure 2. This figure also shows the relative volume of e-mails, spam, and legitimate (ham) messages originated by senders in each category.

First, we examine the overall success of our heuristics in classifying the sender IPs into the three categories, namely, legitimate servers, spam gangs and end-hosts. The overall success rate of our heuristic determines whether or not our analysis of pre-acceptance filtering is complete and free from unknown biases. Fortunately, our heuristics achieve very good coverage: In particular, we were able to classify 91% of sender IPs into one of the three categories. These sender IPs contributed 83% of all e-mail, 86% of all spam e-mail, and 78% of legitimate e-mail.

We present the detailed breakdown of the sender IPs and the e-mails they originate in Table I. We refer to it in our analysis below. We first present a high-level overview of the contribution of the three categories to the overall e-mail, spam and ham volume, followed by an in-depth analysis of each category and a discussion of implications for pre-acceptance filtering.

Sender type	#IPs	#E-mails	#Spam	#Ham	% Spam
Total	***3.6M***	***37.7M***	***24.6M***	***12.3M***	***65.09%***
Popular	0.22 %	2.84 %	0.25 %	7.95 %	5.73%
SPF-good	0.75 %	21.25 %	0.12 %	62.35 %	0.38%
Legit: sub total	***0.98 %***	***24.08 %***	***0.37 %***	***70.30 %***	***1.01%***
PBL or UDMap	84.00 %	26.36 %	38.02 %	2.99 %	93.89%
RDNS heuristics	4.86 %	11.48 %	15.98 %	2.77 %	90.60%
End-hosts: sub total	***88.86 %***	***37.84 %***	***54.00 %***	***5.76 %***	***92.88%***
SBL	0.34 %	2.28 %	3.31 %	0.38 %	94.47%
Bad blocks	0.26 %	3.49 %	5.17 %	0.34 %	96.47%
SPF-Bad	0.64 %	15.71 %	23.33 %	1.43 %	96.66%
Spam Gang: sub total	***1.24 %***	***21.47 %***	***31.80 %***	***2.15 %***	***96.39 %***
Unclassified	***8.92 %***	***16.60 %***	***13.82 %***	***21.79 %***	***54.21 %***

TABLE I
CLASSIFICATION OF SPAM E-MAIL IN THE DATASET FOR SEPTEMBER 2007.

1) High-level differences among sender categories: We focus first on end-hosts. From Figure 2 and Table I, we note that *a majority of all spam,* — roughly 54% — seems to originate from end-hosts. End-hosts also make up an *overwhelming fraction of e-mail sender IPs (89%)* but the proportion of e-mail they send is not as high (38%). Our observations regarding end-hosts are in agreement with prior work (See [17], [22], for example) which has also 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 for pre-acceptance filtering.

We focus next on legitimate servers. From Figure 2 and Table I, we note that legitimate servers constitute a small fraction of the sender IPs (1%). However, they contribute a significant fraction of all e-mails (24%). The good news is that an *overwhelming volume of all ham* or legitimate e-mail originates from these servers (70%). More importantly, the legitimate servers contribute a very low volume spam overall (under 0.4%). The average spam ratio - which is the volume of spam over the total volume of e-mail - is shown in the last column and is just 1% for the legitimate servers.

We discuss the implications of these observations on pre-acceptance filtering in the next section.

We turn our attention to spam gangs. To the best of our knowledge, prior work has not examined the role of spam gangs in the e-mail spam problem. From Figure 2 and Table I, we note that spam gangs make up only a very small fraction of senders (1.24%). Quite susprisingly, they contribute 21% of all e-mail, *almost all of which is spam*. Crucially, 32% of *all e-mail spam* originates from this small collection of senders. The average spam ratio for the 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 develop mechanisms to thwart their activity. In the next section, we show how pre-acceptance filtering can play a crucial role.

Finally, we note that roughly 8% of ham or legitimate e-mail appears to originate from end-hosts and spam gangs put together (the former contributes 6% and the latter contributes 2%). Although this fraction is small relative to the volume of spam from these two sets of the senders, the fraction should ideally be *zero* (this is because spam gangs are dedicated to sending unsolicited e-mails, and end-hosts are rarely configured as SMTP servers [22]). 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% and a standard deviation of 8.38%, 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 it is more likely that some spammers are able to get a significant amount of e-mail through without being detected by the Sophos spam detection software employed at UW-Madison. Such false negatives could have an impact on the ability of SMTP servers in constructing effective pre-acceptance filters (e.g. sending history based blacklists) which we discuss in the next section.

For completeness, we studied if these high level observations held true for other time periods when data was collected. In what follows, we compare the observations derived for data collected in Sep 2007 against those derived for Mar 2008. While we omit the detailed results for brevity, we note that our techniques were able to classify senders quite successfully. Overall, 84% of senders were classified and 81% of all e-mail originated from them. In terms of the overall contributions from the different categories, that the proportions remain qualitatively the same across Sept 2007 and Mar 2008.

2) In-depth analysis and the role of pre-acceptance: Next, we delve deeper into the three categories of e-mail senders and examine their contribution to the overall spam and ham volumes. We also discuss the implications of our observations on pre-acceptance filtering and discuss how appropriate pre-acceptance filters can be built.

First, we note from Table I that 84% of the sender IPs belong to the dynamic IP address prefix blacklists, i.e. PBL and UDMap. These contribute a huge volume of the spam overall (38%). This simple observation indicates that *blacklists of dynamic-IP address blocks can filter 38% of all spam.* This observation is similar to that made by past work on the effectiveness of such blacklists [18], [22].

The remaining end-hosts were identified using our naming-based heuristics (the heuristics are applied after first checking against PBL and UDMap). This category of end-hosts contributes 16% of all spam. While blacklisting is certainly effective in stopping spam from these end-hosts, the senders may have to be *individually* blacklisted, as opposed to blacklisting entire ranges of IP addresses. In order to create a blacklist of these senders, an SMTP server could use the same heuristics as we do, namely, heuristics based on neighbor names and key-word based tests. Note that these are "static" heuristics and can be applied even if a host sent just a few e-mails (i.e., it is not necessary to track spamming history over many messages). In fact, many SMTP servers already maintain and use such blacklists. Thus, *up-to-date blacklists of individual end-hosts could further filter 16% of all spam.*

We now turn to legitimate server. Since legitimate servers contribute a majority of ham and almost no spam, it may be highly beneficial to construct a whitelist of these servers and accept all e-mail originating from senders in the list. The small amount of spam originating from the legitimate servers can be filtered using post-acceptance tests; the low volume of e-mails involved in post-acceptance tests implies that the overhead on the receiving SMTP servers will be minimal if whitelisting is used for the legitimate servers.

There could be multiple ways of constructing such whitelist and our in-depth analysis of the legitimate senders indicates how the whitelist may be constructed. For instance, from Table I we note that servers in SPF-good category form a majority of the legitimate senders (0.75% out of 1.0%). Senders in SPF-good originate a significant fraction of ham (62%) and almost no spam (0.12%). Recall that we constructed the SPF-good list by monitoring the spam sending activity of senders with SPF RRs over a one month period. An SMTP server can employ a similar sending-history based technique to populate a whitelist of mail senders. In the next section, we show that the SPF-good whitelist can in fact be constructed by monitoring sending behavior over roughly a one week period, and that the computed list could be used fairly effectively for a month or more without requiring any updates.

Similarly, we note that senders in the popular sender category contribute a further 8% of all ham and just 0.25% of all spam. They make up 0.22% of all senders. Their average spam ratio is also low (5.7%). A whitelist of popular servers can be constructed by an SMTP using techniques similar to the ones we used.

Overall, *if an SMTP server maintained an accurate whitelist of legitimate servers (including popular servers and servers with SPF RRs and good sending history), then the list would suffice to let through nearly 70% of all legitimate e-mail and just 0.37% of spam. Furthermore the list would only contain a small number of IP addresses*.

Our conclusion regarding legitimate servers is also supported by Venkataraman et al. in [20], who show that IP addresses which have a long-lived sending history contribute the majority of legitimate e-mail.

Next, we discuss the role of spam gangs and the effectiveness of pre-acceptance filters in thwarting spam originating from these senders. From Table I, we first note that nearly 20% of the sender IPs in this category (0.26% out of 1.24%) belong to sophisticated spammers who use blocks of IP addresses (labeled "bad blocks"). These spammers originate 5% of all spam e-mail. The average spam ratio of these senders is 97%.

Recall that we constructed the "bad blocks" list by monitoring spamming activity over an IP range for a certain duration of time. An SMTP server can similarly track historical spamming activity from IP address blocks and construct a blacklist of such sophisticated spam gangs. Assuming such a blacklist can be constructed in an accurate fashion, it can be combined with the address blocks corresponding to dynamic IP addresses to create a “master blacklist” of address blocks from which only spam and almost no ham can originate. *An aggregate prefix based blacklist of this form can filter 43% of all spam (38% coming from end-hosts with dynamic IPs, and the rest from bad blocks or sophisticated spam gangs)*. In practice, however, maintaining an up-to-date blacklist of the sophisticated spammers is challenging because the list must be updated very frequently. But it is possible to build and maintain such a list, as we show in the next section.

We note that several spammers in the spam gang category rely on registering fake SPF RRs in an attempt to bypass filtering (labeled “SPF-bad”). These less sophisticated spammers make up over half of the IPs in the spam gang category (0.64% out of 1.24%). Surprisingly, these spammers contribute 23% of all spam. The average spam ratio of these spammers is 0.97. These spammers are quite difficult to identify. These senders look “legitimate” in many ways - they have SPF RRs and legitimate looking domain names which appear very different than the ones typically applied to end-hosts. In order to identify and blacklist these spammers, an SMTP server will have to track their spam sending history, similar to the approach we employed in building the SPF-bad blacklist. As soon as poor spamming history is identified (i.e. the host’s spam ratio becomes worse than, say, 0.75, and a sufficient number of e-mails have been received), each sender must be blacklisted individually. (Note that prior to blacklisting, detecting spam from a sender would require post-acceptance content-based tests. Once blacklisted, the sender can be filtered in the pre-acceptance stage.) In the next section, we examine the time-scales at which the sender history must be tracked and the effectiveness of the blacklists constructed in this fashion.

Finally, we note that SBL, the public list of spam gang IPs, covers a very small fraction of all e-mails (just 2.3%), indicating that the blacklist is grossly incomplete with regards to spam gang membership.

We repeated the in-depth analysis using data from the March 2008 data set and found that the observations for each category were similar to those obtained using the September 2007 data. We omit the results for brevity.

B. The effectiveness of history-based White- and Black-lists

In the previous section, we referred to three history based lists (white or black) – SPF-Good, Bab Blocks and SPF-Bad, and pointed out the crucial role they play in pre-acceptance filtering. Our evaluation of these lists considered an ideal situation where an oracle computes the lists based on sending patterns over a large time interval and the list is applied to all e-mail 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 lists. In particular, we examine if history-based lists collected on the basis of e-mail 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.

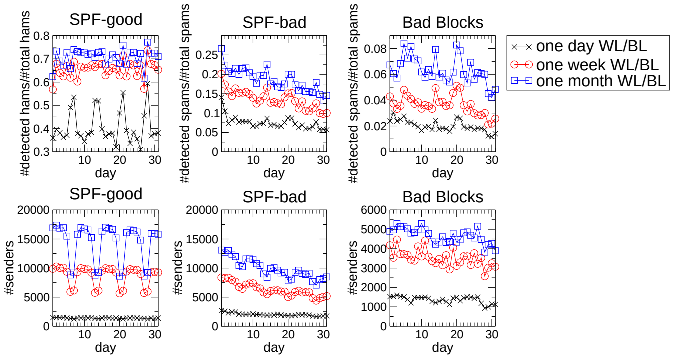

Fig. 3. Effectiveness of history-based whitelist and blacklists: the coverages of hams and spams for each list (top) and senders for each list (bottom). Note the weekly patterns in the graph plotting the number of legitimate senders identified over time – fewer legitimate senders appear on weekends.

To answer these questions, we compiled the lists, i.e., the SPF-good whitelist and the SPF-Bad and Bad Blocks blacklists, based on the e-mail logs collected over the three distinct time intervals, namely one day (31 Sep., 2007), one week (25–31 Sep., 2007), and one month (Sep., 2007). We applied the whitelist and the blacklists to the e-mail 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 e-mails (for the SPF-good whitelist) and the number of spam e-mails (for the two blacklists) contributed by the identified senders.

Figure 3 shows the results. 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 e-mail 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 e-mail they sent were both fairly stable over the entire month. *This indicates that the SPF-Bad and the Bad blocks black lists which correspond to spam gangs must be updated periodically, once every few days for the former, and every 2 weeks or so for the latter in order to ensure that the blacklists are effective. In contrast, SPF-Good can be updated much less frequently.*

C. Measurement Summary

Our measurements indicate the following:

(1) A large fraction of all spam originates from end-hosts, who also make up nearly 90% of the senders. Spam gangs make up a small fraction of the senders but contribute a disproportionately large amount of spam. Legitimate servers send most of the legitimate e-mail and very little spam.

(2) Up-to-date blacklists of *dynamic address prefixes* can filter at least 38% of all spam. Up-to-date blacklists of sophisticated spam gangs which use blocks of IPs can filter at least 5% of spam. Thus, accurate IP prefix blacklists can filter 43% of spam. Up-to-date blacklists of *individual* end-host IPs can filter up to 16% of all spam. Up-to-date blacklists of individual spam gang IPs can filter up to 23% of spam. Thus, blacklists of individual IPs can filter an additional 39% of spam.

In all, blacklists can potentially filter 86% of all spam.

(3) Whitelists covering a small fraction of legitimate senders (1% of senders) are sufficient to let an overwhelming fraction of all ham through. These servers contribute very little spam.

(4) The blacklists and the white lists differ significantly in terms of the time intervals over which e-mail activity must be observed, and the rate at which the lists must be updated. A fairly good whitelist of legitimate servers can be constructed by monitoring sending history over a one week period and compiling a list of popular servers. In contrast, to construct an effective blacklist of spam gangs an SMTP server must monitor sending activity over much longer time periods. The white list can be updated infrequently, while the two black lists corresponding to spam gangs must be updated once every few days or couple of weeks depending on the list.

V. Related work

There have been several studies relating 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. Clayton demonstrated that analyzing e-mail sender characteristics, such as, the number of delivery failures, null return path SMTP envelope, and variations of HELO messages, using simple heuristics can be effective in detecting spammers [12], [13]. Sender characteristics are used in a similar approach proposed Ramachandran et al., who used clustering algorithms based on the sending patterns of multiple senders over a given time window as an indicator of whether a sender is a spammer [19]. Venkataraman et al. showed that network-aware clustering of IP address space coupled with the spam ratio history of individual IP senders is effective in classifying e-mail senders as spammers or not based on their IP [20]. In another non-content based approach, Beverly and Solins showed that transport-layer characteristics of e-mail senders, e.g., number of retransmissions, minimum window advertised, and initial round trip time estimate, are effective in identifying spammers [10].

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 the 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 [15] make similar observations regarding keeping a small collection of popular DNSBLs (DNS Blacklist of spam domains) up-to-date, but they do not examine how the time-scales for update depend on 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 the overall spam.

A final key difference between our study and the prior works is that prior studies don't quantify how much spam originates from legitimate servers and don't 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 end-host bots have been explored in [11], [16], [17], [23]. 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.

Similar to identifying botnet characterization, researchers have aimed to identify IP hijacking events in [14], [17], [24] and the e-mail sending activity of spammers who use hijacked prefixes. We note that our characterization of spam senders does not include senders who may have used hijacked prefixes.

VI. Conclusion

Effective pre-acceptance filtering can significantly lower the load on SMTP servers today. In this paper, we analyzed the limits of effectiveness of pre-acceptance spam filtering. We studied a dataset consisting of nine months of e-mail activity to first discover the spam sending activity of three categories of hosts — end-hosts, hosts belonging to spam gangs and legitimate servers. We leveraged the observations regarding the sending properties of these hosts to evaluate the overall effectiveness of pre-acceptance filters. We also examined the time-scales at which the activity of sending SMTP servers must be monitored in order to construct effective black and white lists. Overall, we find that it is possible to construct pre-acceptance filters that can eliminate 90% of all spam, but some of the filters must be updates on a very constant basis for accuracy.

References

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

[2] Hotmail Operators: Private Communication.

[3] Its not about the spam. http://googleblog.blogspot.com/2007/10/its-not-about-spam.html.

[4] Levenshtein Distance. http://en.wikipedia.org/wiki/Levenshtein_distance/.

[5] Not Just Another Bogus List. http://www.njabl.org/.

[6] 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.

[7] Spamhaus Block List. http://www.spamhaus.org/sbl/index.lasso.

[8] The Spamhaus Project. http://www.spamhaus.org/.

[9] 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.

[10] R. Beverly and K. Sollins. Exploiting transport-level characteristics of spam. In *CEAS*, Aug. 2008.

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

[12] R. Clayton. Stopping spam by extrusion detection. In *CEAS 2004: First Conference on Email and Anti-Spam*, 2004.

[13] R. Clayton. Stopping outgoing spam by examining incoming server logs, 2005.

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

[15] J. Jung and E. Sit. An empirical study of spam traffic and the use of DNS black lists. In *IMC*, 2004.

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

[18] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *SRUTI*, 2006.

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

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

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

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

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

[24] 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.