Taster's Choice: A Comparative Analysis of Spam Feeds

Andreas Pitsillidis[*]
apitsill@cs.ucsd.edu

Chris Kanich[†]
ckanich@cs.uic.edu

Geoffrey M. Voelker[*]
voelker@cs.ucsd.edu

Kirill Levchenko[*]
klevchen@cs.ucsd.edu

Stefan Savage[*]
savage@cs.ucsd.edu

[*]Department of Computer Science and Engineering
University of California, San Diego

[†]Department of Computer Science
University of Illinois at Chicago

ABSTRACT

E-mail spam has been the focus of a wide variety of measurement studies, at least in part due to the plethora of spam data sources available to the research community. However, there has been little attention paid to the suitability of such data sources for the kinds of analyses they are used for. In spite of the broad range of data available, most studies use a single "spam feed" and there has been little examination of how such feeds may differ in content. In this paper we provide this characterization by comparing the contents of ten distinct contemporaneous feeds of spam-advertised domain names. We document significant variations based on how such feeds are collected and show how these variations can produce differences in findings as a result.

Categories and Subject Descriptors

E.m [**Data**]: Miscellaneous; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services

General Terms

Measurement, Security

Keywords

Spam e-mail, Measurement, Domain blacklists

1. INTRODUCTION

It is rare in the measurement of Internet-scale phenomena that one is able to make comprehensive observations. Indeed, much of our community is by nature opportunistic: we try to extract the most value from the data that is available. However, implicit in such research is the assumption that the available data is *sufficient* to reach conclusions about the phenomena at scale. Unfortunately, this is not always the case and some datasets are too small or too biased to be used for all purposes. In this paper, we explore this issue in the context of a common security measurement domain: e-mail spam.

On the one hand e-mail spam is plentiful—everyone gets it—and thus is deceptively easy to gather. At the same time, the scale of the e-mail spam problem is enormous. Industry estimates (admittedly based on unknown methodology) suggest that spammers sent well over 100 billion e-mails each day in 2010 [16]. If true, then even a spam corpus consisting of 100,000 messages per day would constitute only *one ten thousandth of one percent* of what occurred globally. Thus, except for researchers at the very largest e-mail providers, we are all forced to make extrapolations by many orders of magnitude when generalizing from available spam data sources. Further, in making these extrapolations, we must assume both that our limited samples are sufficiently unbiased to capture the general behavior faithfully and that the behavior is large enough to be resolved via our measurements (concretely, that spam is dominated by small collections of large players and not vice versa). However, we are unaware of any systematic attempt to date to examine these assumptions and how they relate to commonly used data sources.

To explore these questions, we compare contemporaneous spam data from ten different data feeds, both academic and commercial, gathered using a broad range of different collection methodologies. To address differences in content, we focus on the Internet domain names advertised by spam messages in such feeds, using them as a *key* to identify like messages. Using this corpus, corresponding to over a billion messages distributed over three months, we characterize the relationships between its constituent data sources. In particular, we explore four questions about "feed quality": purity (how much of a given feed is actually spam?), coverage (what fraction of spam is captured in any particular feed?), timing (can a feed can be used to determine the start and end of a spam campaign?) and proportionality (can one use a single feed to accurately estimate the relative volume of different campaigns?).

Overall, we find that there are significant differences across distinct spam feeds and that these differences can frequently defy intuition. For example, our lowest-volume data source (comprising just over 10 million samples) captures more spam-advertised domain names than all other feeds combined (even though these other feeds contain two orders of magnitude more samples). Moreover, we find that these differences in turn translate into analysis limitations; not all feeds are good for answering all questions. In the remainder of this paper, we place this problem in context, describe our data sources

IMC'12, November 14–16, 2012, Boston, Massachusetts, USA.

and analysis, and summarize some best practices for future spam measurement studies.

2. BACKGROUND

E-mail spam is perhaps the only Internet security phenomenon that leaves no one untouched. Everybody gets spam. Both this visibility and the plentiful nature of spam have naturally conspired to support a vast range of empirical measurement studies. Some of these have focused on how to best filter spam [3, 5, 7], others on the botnets used to deliver spam [11, 30, 42], and others on the goals of spam, whether used as a vector for phishing [25], malware [12, 22] or, most commonly, advertising [17, 18].

These few examples only scratch the surface, but importantly this work is collectively not only diverse in its analyses aims, but also in the range of data sources used to drive those same conclusions. Among the spam sources included in such studies are the authors' own spam e-mail [3, 44], static spam corpora of varied provenance (e.g., Enron, TREC2005, CEAS2008) [10, 26, 34, 41, 44], open mail proxies or relays [9, 28, 29], botnet output [11, 30], abandoned e-mail domains [2, 13], collections of abandoned e-mail accounts [39], spam automatically filtered at a university mail server [4, 31, 35, 40], spam-fed URL blacklists [24], spam identified by humans in a large Web-mail system [42, 43], spam e-mail filtered by a small mail service provider [32], spam e-mail filtered by a modest ISP [6] and distributed collections of honeypot e-mail accounts [36].

These data sources can vary considerably in volume — some may collect millions of spam messages per day, while others may gather several orders of magnitude fewer. Intuitively, it seems as though a larger data feed is likely to provide better coverage of the spam ecosystem (although, as we will show, this intuition is misleading). However, an equally important concern is how differences in the *manner* by which spam is collected and reported may impact the *kind* of spam that is found.

To understand how this may be, it is worth first reflecting on the operational differences in spamming strategies. A spammer must both obtain an address list of targets and arrange for e-mail delivery. Each of these functions can be pursued in different ways, optimized for different strategies. For example, some spammers compile or obtain enormous "low-quality" address lists [15] (e.g., based on brute force address generation, harvesting of Web sites, etc.), many of which may not even be valid, while others purchase higher quality address lists that target a more precise market (e.g., customers who have purchased from an online pharmacy before). Similarly, some spam campaigns are "loud" and use large botnets to distribute billions of messages (with an understanding that the vast majority will be filtered [12]) while other campaigns are smaller and quieter, focusing on "deliverability" by evading spam filters.

These differences in spammer operations in turn can interact with differences in collection methodology. For example, spam collected via MX honeypots (accepting all SMTP connections to a quiescent domain) will likely contain broadly targeted spam campaigns and few false positives, while e-mail manually tagged by human recipients (e.g., by clicking on a "this is spam" button in the mail client) may self-select for "high quality" spam that evades existing automated filters, but also may include legal, non-bulk commercial mail that is simply unwanted by the recipient.

In addition to properties of how spam data is *collected*, how the data is *reported* can also introduce additional limitations. For example, some data feeds may include the full contents of e-mail messages, but many providers are unwilling to do so due to privacy concerns. Instead, some may redact some of the address information, while, even more commonly, others will only provide information about the spam-advertised URLs contained with a message. Even within URL-only feeds there can be considerable differences. Some data providers may include full spam-advertised URLs, while others scrub the data to only provide the fully-qualified domain name (particularly for non-honeypot data, due to concern about side-effects from crawling such data). Sometimes data is reported in raw form, with a data record for each and every spam message, but in other cases providers aggregate and summarize. For example, some providers will de-duplicate identically advertised domains within a given time window, and domain-based blacklists may only provide a single record for each such advertised domain.

Taken together, all of these differences suggest that different kinds of data feeds may be more or less useful for answering particular kinds of questions. It is the purpose of this paper to put this hypothesis on an empirical footing.

3. DATA AND METHODOLOGY

In this work we compare ten distinct sources of spam data (which we call *feeds*), ranging in their level of detail from full e-mail content to only domain names of URLs in messages. Comparisons between feeds are by necessity limited to the lowest common denominator, namely domain names. In the remainder of this paper we treat each feed as a source of spam-advertised domains, regardless of any additional information available.

By comparing feeds at the granularity of domain names, we are implicitly restricting ourselves to spam containing URLs, that is, spam that is a Web-oriented advertisement in nature, at the exclusion of some less common classes of spam (e.g., malware distribution or advance fee fraud). Fortunately, such advertising spam is the dominant class of spam today.[1]

3.1 Domains

Up to this point, we have been using the term "domain" very informally. Before going further, however, let us say more precisely what we mean: a *registered domain*—in this paper, simply a *domain*—is the part of a fully-qualified domain name that its owner registered with the registrar. For the most common top-level domains, such as COM, BIZ, and EDU, this is simply the second-level domain (e.g., "ucsd.edu"). All domain names at or below the level of registered domain (e.g., "cs.ucsd.edu") are administered by the registrant, while all domain names above (e.g., "edu") are administered by the registry. Blacklisting generally operates at the level of registered domains, because a spammer can create an arbitrary number of names under the registered domain name to frustrate fine-grained blacklisting below the level of registered domains.

[1] One recent industry report [37] places Web-oriented advertising spam for pharmaceuticals at over 93% of all total spam volume.

3.2 Types of Spam Domain Sources

The spam domain sources used in this study fall into five categories: botnet-collected, MX honeypots, seeded honey accounts, human identified, and blacklists. Each category comes with its own unique characteristics, limitations and tradeoffs that we discuss briefly here.

Botnet. Botnet datasets result from capturing instances of bot software and executing them in a monitored, controlled environment such that the e-mail they attempt to send is collected instead. Since the e-mail collected is only that sent by the botnet, such datasets are highly "pure": they have no false positives under normal circumstances.[2] Moreover, if we assume that all members of a botnet are used in a homogeneous fashion, then monitoring a single bot is sufficient for identifying the spamming behavior of the entire botnet. Botnet data is also highly accessible since a researcher can run an instance of the malware and obtain large amounts of botnet spam without requiring a relationship with any third-party security, mail or network provider [11]. Moreover, since many studies have documented that a small number of botnets are the primary source of spam e-mail messages, in principle such datasets should be ideally suited for spam studies [11, 21, 30]. Finally, these datasets have the advantage of often being high volume, since botnets are usually very aggressive in their output rate.

MX honeypot. MX honeypot spam is the result of configuring the MX record for a domain to point to an SMTP server that accepts all inbound messages. Depending on how these domains are obtained and advertised, they may select for different kinds of spam. For example, a newly registered domain will only capture spam using address lists created via brute force (i.e., sending mail to popular user names at every domain with a valid MX). By contrast, MX honeypots built using abandoned domains or domains that have become quiescent over time may attract a broader set of e-mail, but also may inadvertently collect legitimate correspondence arising from the domain's prior use. In general MX honeypots have low levels of false positives, but since their accounts are not in active use they will only tend to capture spam campaigns that are very broadly targeted and hence have high volume. Since high-volume campaigns are easier to detect, these same campaigns are more likely to be rejected by anti-spam filters. Thus, some of the most prevalent spam in MX-based feeds may not appear frequently in Web mail or enterprise e-mail inboxes.

Seeded honey accounts. Like MX honeypots, seeded honey accounts capture unsolicited e-mail to accounts whose sole purpose is to receive spam (hence minimizing false positives). However, unlike MX honeypots, honey accounts are created across a range of e-mail providers, and are not limited to addresses affiliated with a small number of domains. However, since these e-mail addresses must also be seeded—distributed across a range of vectors that spammers may use to harvest e-mail address lists (e.g., such as forums, Web sites and mailing lists)—the "quality" of a honey account feed is related both to the number of accounts and how well the accounts are seeded. The greater operational cost of creating and seeding these accounts means that researchers generally obtain honey account spam feeds from third parties (frequently commercial anti-spam providers).

Honey accounts also have many of the same limitations as MX-based feeds. Since the accounts are not active, such feeds are unlikely to capture spam campaigns targeted using social network information (i.e., by friends lists of real e-mail users) or by mailing lists obtained from compromised machines [14]. Thus, such feeds mainly include low-quality campaigns that focus on volume and consequently are more likely to be captured by anti-spam filters.

Human identified. These feed are those in which humans actively nominate e-mail messages as being examples of spam, typically through a built-in mail client interface (i.e., a "this is spam" button). Moreover, since it is primarily large Web mail services that provide such user interfaces, these datasets also typically represent the application of human-based classification at very large scale (in our case hundreds of millions of e-mail accounts). For the same reason, human identified spam feeds are not broadly available and their use is frequently limited to large Web mail providers or their close external collaborators.

Human identified spam feeds are able to capture "high quality" spam since, by definition, messages that users were able to manually classify must also have evaded any automated spam filters. As mentioned before, however, such feeds may underrepresent the high-volume campaigns since they will be pre-filtered before any human encounters them. Another limitation is that individuals do not have a uniform definition of what "spam" means and thus human identified spam can include legitimate commercial e-mail as well (i.e., relating to an existing commercial relationship with the recipient). Finally, temporal signals in human-identified spamfeeds are distorted because identification occurs at human time scales.

Domain blacklists. Domain blacklists are the last category of spam-derived data we consider and are the most opaque since the method by which they are gathered is generally not documented publicly.[3] In a sense, blacklists are meta-feeds that can be driven by different combinations of spam source data based on the organization that maintains them. Among the advantages of such feeds, they tend to be broadly available (generally for a nominal fee) and, because they are used for operational purposes, they are professionally maintained. Unlike the other feeds we have considered, blacklists represent domains in a binary fashion—either a domain is on the blacklist at time t or it is not. Consequently, while they are useful for identifying a range of spam-advertised domains, they are a poor source for investigating questions such as spam volume.

While these are not the only kinds of spam feeds in use by researchers (notably omitting automatically filtered spam taken from mail servers, which we did not pursue in our work due to privacy concerns), they capture some of the most popular spam sources as well as a range of collection mechanisms.

3.3 False Positives

No spam source is pure and all feeds contain false positives. In addition to feed-specific biases (discussed above), there is a range of other reasons why a domain name appearing in a spam feed may have little to do with spam.

[2] However, see Section 4.1 for an example of domain poisoning carried out by the Rustock botnet.

[3] However, they are necessarily based on some kind of real-time spam data since their purpose is to identify spam-advertised domains that can then be used as a dynamic feature in e-mail filtering algorithms.

Feed	Type	Domains	Unique
Hu	Human identified	10,733,231	1,051,211
URIBL	Blacklist	n/a	144,758
DBL	Blacklist	n/a	413,392
MX$_1$	MX honeypot	32,548,304	100,631
MX$_2$	MX honeypot	198,871,030	2,127,164
MX$_3$	MX honeypot	12,517,244	67,856
Ac$_1$	Seeded honey accounts	30,991,248	79,040
Ac$_2$	Seeded honey accounts	73,614,895	35,506
Bot	Botnet	158,038,776	13,588,727
Hyb	Hybrid	451,603,575	1,315,292

Table 1: Summary of spam domain sources (feeds) used in this paper. The first column gives the feed mnemonic used throughout.

First, false positives occur when legitimate messages are inadvertently mixed into the data stream. This mixing can happen for a variety of reasons. For example, MX domains that are lexically similar to other domains may inadvertently receive mail due to sender typos (see Gee and Kim for one analysis of this behavior [8]). The same thing can occur with honeypot accounts (but this time due to username typos). We have also experienced MX honeypots receiving legitimate messages due to a user specifying the domain in a dummy e-mail address created to satisfy a sign-up requirement for an online service (we have found this to be particularly an issue with simple domain names such as “test.com”).

The other major source of feed pollution is chaff domains: legitimate domains that are not themselves being advertised but co-occur in spam messages. In some cases these are purposely inserted to undermine spam filters (a practice well documented by Xie *et al.* [42]), in other cases they are simply used to support the message itself (e.g., image hosting) or are non-referenced organic parts of the message formatting (e.g., DTD reference domains such as w3.org or microsoft.com). Finally, to bypass domain-based blacklists some spam messages will advertise “landing” domains that in turn redirect to the Web site truly being promoted. These landing domains are typically either compromised legitimate Web sites, free hosting Web services (e.g., Google’s Blogspot, Windows Live domains or Yahoo’s groups) or Web services that provide some intrinsic redirection capability (e.g., bit.ly) [18]. We discuss in more detail how these issues impact our feeds in Section 4.1.

3.4 Meet the Feeds

Table 1 lists the feeds used in this study. We assign each feed a concise label (e.g., Ac$_2$) identifying its type, as described earlier. Of these ten feeds, we collected MX$_1$ and Bot directly. We receive both blacklist feeds (DBL and URIBL) by subscription. Provider agreements preclude us from naming the remaining feeds (Ac$_1$, MX$_2$, Ac$_2$, MX$_3$, Hyb, Hu). One feed, Hyb, we identify as a “hybrid.” We do not know the exact collection methodology it uses, but we believe it is a hybrid of multiple methods and we label it as such.

Referring to Table 1, the *Domains* column shows the total number of samples we received from the feed during the three-month period under consideration. Thus, the Hu feed included only a bit over ten million samples, while the Bot feed produced over ten times that number. The *Unique* column gives the number of unique registered domain names in the feed.

With the exception of the two blacklists, we collected the feeds used in this paper in the context of the Click Trajectories project [18] between August 1st, 2010 and October 31st, 2010. The Click Trajectories project measured the full set of resources involved in monetizing spam—what we call the spam value chain. One of the resources in the value chain is Web hosting. To identify the Web hosting infrastructure, we visited the spam-advertised sites using a full-fidelity Web crawler (a specially instrumented version of Firefox), following all redirections to the final storefront page. We then identified each storefront using a set of hand-generated content signatures, thus allowing us to link each spam URL to the goods it was advertising.

We use the results of this Web crawl to determine whether a spam domain, at the time it is advertised, led to a live Web site.[4] Furthermore, we determined if this site was the storefront of either a known online pharmacy selling generic versions of popular medications, a known replica luxury goods store, or a known “OEM” software store selling unlicensed versions of popular software. These three categories — pharmaceuticals, replica goods, software — are among the most popular classes of goods advertised via spam [18, 20]. Based on this information, we refer to domains as *live* if at least one URL containing the domain led to a live Web site, and *tagged* if the site was a known storefront.

Finally, because we obtained the blacklist feeds after the completion of the Click Trajectories work, we could not systematically crawl all of their domains. Thus the entries listed in the table only include the subset that *also* occurred in one of the eight base feeds. While this bias leads us to undercount the domains in each feed (thus underrepresenting their diversity), this effect is likely to be small. The DBL feed contained 13,175 additional domains that did not occur in any of our other base feeds (roughly 3% of its feed volume) while the URIBL feed contained only 3,752 such domains (2.5% of its feed volume).

4. ANALYSIS

We set out to better understand the differences among sources of spam domains available to the researcher or practitioner. The value of a spam domain feed, whether used in a production system for filtering mail or in a measurement study, ultimately lies in how well it captures the characteristics of spam. In this paper we consider four qualities: *purity*, *coverage*, *proportionality*, and *timing*.

Purity is a measure of how much of a feed is actually spam domains, rather than benign or non-existent domains.

Coverage measures how much spam is captured by a feed. That is, if one were to use the feed as an oracle for classifying spam, coverage would measure how much spam is correctly classified by the oracle.

Proportionality is how accurately a feed captures not only the domains appearing in spam, but also their relative frequency. If one were tasked with identifying the top 10 most spammed domains, for example, proportionality would be the metric of interest.

Timing is a measure of how accurately the feed represents the period during which a domain appears in spam.

[4] With feeds containing URLs, we visited the received URL. Otherwise, for feeds containing domains only, we prepended `http://` to the domain to form a URL.

Feed	DNS	HTTP	Tagged	ODP	Alexa
Hu	88%	55%	6%	1%	1%
DBL	100%	72%	11%	<1%	<1%
URIBL	100%	85%	22%	2%	2%
MX$_1$	96%	83%	20%	9%	8%
MX$_2$	6%	5%	<1%	<1%	<1%
MX$_3$	97%	83%	16%	9%	7%
Ac$_1$	95%	82%	20%	8%	5%
Ac$_2$	96%	88%	33%	10%	11%
Bot	<1%	<1%	<1%	<1%	<1%
Hyb	64%	51%	2%	12%	10%

Table 2: Positive and negative indicators of feed purity. See Section 4.1 for discussion.

Most often with timing we care about how quickly a domain appears in the feed after it appears in spam in the wild.

Unfortunately, all of the measures above presuppose the existence of an ultimate "ground truth," a platonic absolute against which all feeds can be compared. Sadly, we have no such feed: barring the challenges of even defining what such a feed would contain, the practical difficulty of capturing all spam (however defined) is immense. We can still gain useful insight, however, by comparing feeds to each other. In particular, for coverage and timing, we combine all of our feeds into one aggregate super-feed, taking it as our ideal. For proportionality, we measure the relative frequency of spam domains in incoming mail seen by a large Web mail provider, allowing us to compare the relative frequencies of domains in a spam feed to their frequencies in a representative e-mail feed.

In the remainder of this section, we evaluate the spam domain feeds available to us (summarized in Table 1) with respect to the qualities described above.

4.1 Purity

The purity of a feed is a measure of how much of the feed is actually spam, rather than benign or non-existent domains. Very simply, purity is the fraction of the feed that are spam domains. We refer to these spam domains appearing in the feed as *true positives*, and non-spam domains appearing in the feed as *false positives*. Purity is thus akin to *precision* in Information Retrieval or *positive predictive value* in Statistics.

The importance of purity varies from application to application. If the feed is used to directly filter spam (by marking any message containing a domain appearing in the feed as spam), then purity is of paramount importance. On the other hand, for a measurement study, where spam domains are visited and further analyzed, low purity may tax the measurement system, but generally has little impact on the results once filtered.

Operationally, the nature of the false positives matters as well. While non-existent domains appearing in the feed are merely a nuisance, benign domains can lead to highly undesirable false positives in the filtering context.

Table 2 shows several purity indicators for each feed. The first three (*DNS*, *HTTP*, and *Tagged*) are positive indicators—larger numbers mean higher purity. The last two (Alexa and ODP) are negative indicators, with larger numbers meaning lower purity.

4.1.1 Non-existent Domains

The DNS column shows the proportion of domains in the feed that were registered, based on several major TLDs. Specifically, we checked the DNS zone files for the COM, NET, ORG, BIZ, US, AERO, and INFO top-level domains between April 2009 and March 2012, which bracket the measurement period by 16 months before and 16 months after. Together these TLDs covered between 63% and 100% of each feed. We report the number of domains in these TLDs that appeared in the zone file.

Blacklists, seeded honey accounts, and two of the three MX honeypot feeds consisted largely of real domains (over 95%). Human-identified spam and the hybrid feed were lower, at 88% and 64%, levels at which non-registered domains pose little harm operationally or experimentally.

Two feeds—Bot and MX$_2$—exhibit unusually low registration levels, however. Most of these relate to a single phenomenon, a period of several weeks during which the Rustock botnet was sending randomly-generated domains [19, 38]. Such bogus domains cost spammers nearly nothing to generate, while costing spam filter maintainers and spam researchers considerably more in dealing with them.

The HTTP column shows the fraction of domains in the feed that responded to an HTTP request (with a code 200 reply) made to any of the URLs received from the feed during the measurement period. Like the DNS registration measure, HTTP responses indicate that a feed contains live URLs (whether spam or not). Some amount of HTTP failures are inevitable, and we see success rates in the 51% to 88% range for most feeds, with the exception of the same two feeds—Bot and MX$_2$—discussed above.

4.1.2 Known Spam

An HTTP response still does not mean that a domain is not a benign domain accidentally included in the list. To get at the true spam domains, we turn to the Web content tagging carried out in the context of the Click Trajectories project [18]. Recall from Section 3.4 that these are domains that lead to storefronts associated with known online pharmacies, replica stores, or software stores.

Such domains constituted 11–33% of domains in high-purity feeds. Note that while these domains are less than a third of all domains in a feed, they cover the bulk of the spam by volume [18].

4.1.3 Benign Domains

Finally, the ODP and Alexa columns indicate the number of domains in the feed that appeared in Open Directory Project [27] listings and the Alexa top 1 million Web sites [1] list. We expect that nearly all of these domains are benign, and their appearance in a feed is erroneous.[5]

There are at least three reasons why a benign domain might appear in spam. Benign domains may be included in a message by the spammer. A phishing e-mail, for example, may contain some legitimate links to the service being phished. In some cases, a legitimate e-mail may be inadvertently sent to a honeypot or honey account. For example, if an MX honey-

[5] While nothing prohibits a spam domain from appearing on the Alexa list or in the Open Directory listings, these domains are usually short-lived because their utility, and therefore use, is reduced with domain blacklisting. We expect both lists to be overwhelmingly composed of domains incompatible with spam advertising.

pot uses an abandoned, previously-used domain, it may still receive legitimate traffic from its former life. A third cause of benign domains appearing in spam are legitimate services being used by the spammer as a redirection mechanism. By using a URL shortening service, for example, the spammer can evade domain blacklists by hiding behind an established domain.

Using spam domain feeds to drive a production spam filtering system thus runs the risk of false positives. Because blacklists are intended specifically for this task, they have the fewest false positives: only 2% of domains in the URIBL feed, and less than 1% of DBL domains, intersected the ODP and Alexa lists.

4.1.4 Removing Impurities

In the analysis ahead of us, these impurities skew the results and thus obscure the picture. To better understand the useful contributions of each feed, we remove all non-responsive domains and all domains we believe are likely benign. Specifically, for each feed, we take only the set of domains for which we receive at least one successful HTTP response and from this set remove all domains appearing on the Open Directory and Alexa list. (These are the domains listed in the *HTTP* column of Table 2 less those counted in the *ODP* and *Alexa* columns.) We call these ***live*** domains.

In several instances, data collected by the Click Trajectories project [18] allows us to see deeper into the nature of domain, namely into the affiliate programs and affiliates behind each domain. For this analysis, however, we are limited to the set of ***tagged*** domains. We remove Alexa and ODP domains from this set as well. Table 3 shows the number of distinct domains of each type in the feed. In the remainder of the paper, we state explicitly whether a measurement uses live or tagged domains.

We have chosen to explicitly remove Alexa-listed and ODP-listed domains from the set of live and tagged domains used in the remainder of the paper. As discussed in Section 4.1.3, live and even tagged domains may contain domains in the Alexa and ODP listings. An otherwise benign domain may be tagged if it is abused by a spammer as a redirection mechanism, as noted above. Unfortunately, the stakes are high when it comes to such false positives. These same Alexa and ODP domains—comprising less than 2% of the domains in a blacklist—are disproportionately more popular than spam domains. Figure 3 shows the fraction of spam messages containing such domains. In many feeds, these handful of benign domains comprise as much as 90% of live domain volume. Working at the granularity of registered domains, even a single URL redirecting to a spam site can affect the standing of an entire domain.

Practitioners must take great care in choosing which domains to blacklist and whether to blacklist each instance at the registered name or finer granularity. It is not the purpose of this paper, however, to design the perfect blacklist or blacklisting mechanism, and so we leave the question of how best to deal with potential false positives without a full and satisfactory resolution. For our analysis, we take the conservative approach and remove such suspect domains.

4.2 Coverage

Roughly speaking, coverage is a measure of how many spam domains a feed contains. In an operational context, greater coverage—more spam domains—means more spam filtered. For a measurement study or system evaluation, more spam domains means more comprehensive results. In this section, we consider how coverage varies across our ten spam domain feeds. But domains do not exist in a vacuum: they are a projection of external entities into the domain name system, and it is often these entities that are the object of our interest. In the world of spam, these take the form of affiliate programs and affiliates. In Section 4.2.3 we compare feeds on the visibility they provide into that world.

4.2.1 Domains

Table 3 shows the number of live and tagged domains in each feed in the *Total* column. Recall that *live* domains are those that resulted in at least one successful Web visit to a URL containing the domain, while *tagged* domains are those for which the final Web site is a known storefront (Section 3.4).

In absolute terms, whether one considers live domains or tagged domains, the largest contributor of unique instances is the human-identified spam domain feed Hu, despite also being the smallest feed in terms of absolute volume (see Table 1). The reason for this coverage is undoubtedly that this particular provider has hundreds of millions of accounts and thus their customers are likely to be targets of virtually any spam campaign. In turn, we believe that the reason this feed has low volume is that as users identify e-mails as "spammy" the included domains are used to filter subsequent inbound messages. Thus, high-volume campaigns are unlikely to have high representation in such a feed.

Clearly, if one had to choose only one feed to provide maximum coverage, it would be that feed. Unfortunately, outside large mail providers, such data is not widely available to the research community. Instead, the readily-available blacklists—DBL and URIBL—are an excellent alternative, providing more tagged domains than any other feed besides Hu.

Exclusive domains. So far, we have been comparing feeds in terms of their absolute coverage: the total number of spam domains contributed. Given a choice of one feed, one may well pick the largest one by this measure. A feed's value, however, may be in its differential contribution, that is, in the domains it provides that are in no other feed. We term domains that occur in exactly one feed *exclusive* domains. Across our feeds, 60% of all live domains and 19% of all tagged domains were exclusive to a single feed.

Table 3 shows number of exclusive domains provided by each feed in the *Excl.* column. The relationship between the numbers of distinct domains in a feed and the number of exclusive domains is also shown graphically in Figure 1; the left plot shows this relationship for live domains, and the right plot shows it for tagged domains. In both plots, the x axis denotes to the number of distinct domains on a logarithmic scale, while the y axis denotes number of exclusive domains in each feed on a logarithmic scale. Dotted lines denote fixed exclusive domain proportions. For example, the Hyb feed lies just under the 100% line, indicating that most of its live domains—just over 65%—are exclusive.

Figure 1 makes apparent that the Hu and Hyb feeds make the greatest contribution in terms of the distinct number of domains they contribute as well as the number of domains exclusive to each feed. The number of tagged domains is about an order of magnitude less in each feed than the number of live domains, suggesting that spam belonging to

	All Domains		Live Domains		Tagged Domains	
Feed	Total	Excl.	Total	Excl.	Total	Excl.
Hu	1,051,211	534,060	564,946	191,997	64,116	11,356
DBL	413,355	0	298,685	0	46,058	0
URIBL	144,721	0	119,417	0	30,891	0
MX$_1$	100,631	4,523	72,634	1,434	19,482	29
MX$_2$	2,127,164	1,975,081	93,638	6,511	18,055	4
MX$_3$	67,856	6,870	49,659	2,747	10,349	2
Ac$_1$	79,040	3,106	58,002	798	15,242	2
Ac$_2$	35,506	3,049	26,567	972	11,244	31
Bot	13,588,727	13,540,855	21,265	3,728	2,448	0
Hyb	1,315,292	1,069,074	496,893	322,215	25,993	1,285

Table 3: Feed domain coverage showing total number of distinct domains (*Total* column) and number of domains exclusive to a feed (*Excl.* column).

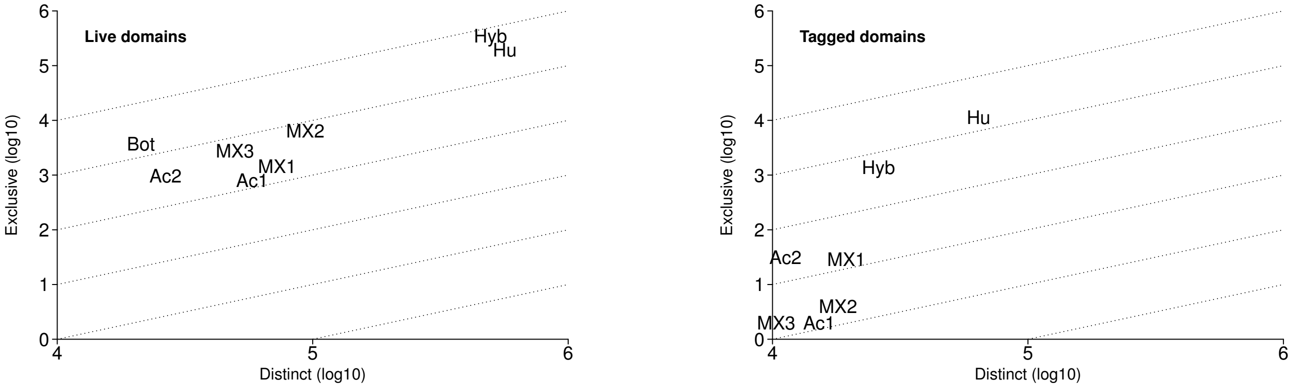


Figure 1: Relationship between the total number of domains contributed by each feed and the number of domains exclusive to each.

the categories represented by the tagged domains—online pharmacies, replica shops, and counterfeit software stores—is a small fraction of all spam. This is not so, however. As we will see in Section 4.3, these domains dominate the feeds in volume.

Figure 1 and Table 3 put the Bot feed in perspective. Although extremely valuable in identifying which domains are being spammed by which botnet, its contribution to the big picture is more limited. None of its tagged domains were exclusive, not a surprising fact given that bots are renowned for indiscriminate high-volume spamming. The roughly 3,700 exclusive live domains in the Bot feed are likely the result of the domain poisoning described earlier (Section 4.1), as fewer than 1% of all domains were legitimate (Table 2).

Pairwise comparison. In the preceding discussion of exclusive contribution, we were implicitly asking which feed, if it were excluded, would be missed the most. Next we consider the question of each feed's differential contribution with respect to another feed. Equivalently, we are asking how many domains from one feed are also in another. (Removing non-responsive and benign domains is particularly important for a meaningful comparison here.)

Figure 2 shows pairwise domain overlap as a matrix, with live domains plotted on the left and tagged domains on the right. For two feeds A and B, the cell in row A column B shows how many domains (percent) from feed B are in feed A, as well as the absolute number of such domains. Formally, the top and bottom numbers show

$$|A \cap B|/|B| \quad \text{and} \quad |A \cap B|.$$

For example, in the left-hand matrix, the cell in row Ac$_1$ column MX$_1$ indicates that Ac$_1$ and MX$_1$ have approximately 47,000 live domains in common, and that this number is 65% of the MX$_1$ feed. Note that these same 47,000 live domains constitute 81% of the Ac$_1$ feed (row MX$_1$ column Ac$_1$). In addition, the right-most column, labeled *All* contains the union of all domains across all feeds. The numbers in the *All* column thus indicate what proportion of all spam domains (the union of all feeds) is covered by a given feed.

Figure 2 once again highlights the coverage of the Hu and Hyb feeds. The Hyb feed covers 51% of all live domains (the union of all non-blacklist feeds), while the Hu feed covers 58%; the two feeds together covering 98% (not shown in matrix) of all live domains. When restricted to tagged domains only (Figure 2 right), the coverage of the Hu feed is an astounding 96%, while the contribution of Hyb drops to 39%. In fact, restricting the domains to tagged only (right-hand matrix) excludes many of the benign domains appearing in Hyb from the *All* column, *improving* the coverage of most feeds with respect to *All*.

Figure 2 also reveals that most feeds—especially Ac$_1$, MX$_1$, MX$_2$, and MX$_3$—are quite effective at capturing bot-generated spam domains. These feeds range from 12% to 21% bot-generated (tagged domains), although the true number is

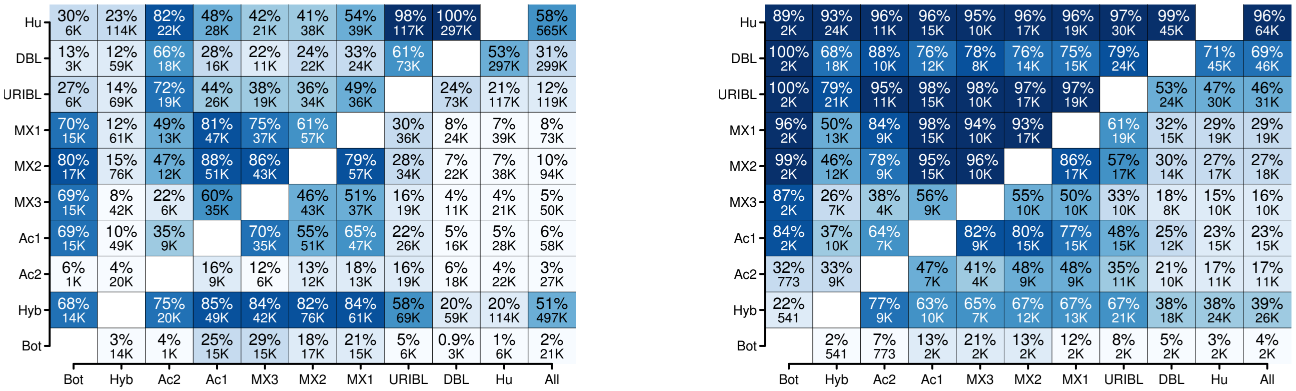


Figure 2: Pairwise feed domain intersection, shown for live (left) and tagged domains (right).

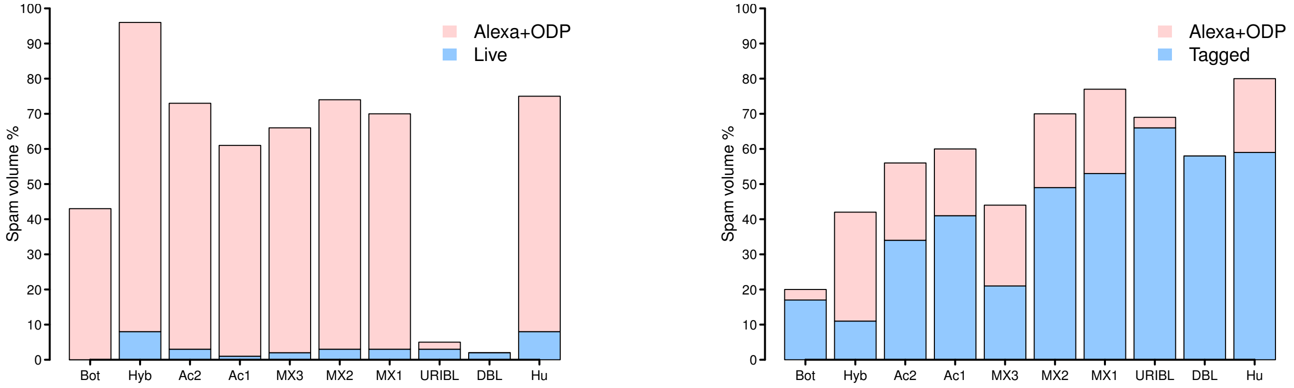


Figure 3: Feed volume coverage shown for live (left) and tagged domains (right).

likely higher given the limited set of bots included in the Bot feed. In turn, URIBL is quite effective at capturing these honeypot feeds (MX$_1$, MX$_2$, MX$_3$, Ac$_1$, and Ac$_2$), and both blacklists considerably overlap each other. Moreover, blacklists have a non-trivial overlap with the Hu feed. Despite these higher numbers, though, a gap still exists, as blacklists cannot replace the human identified dataset. Overall, this is a strong indicator of the strength of human-identified feeds, while also stressing the significance of blacklists.

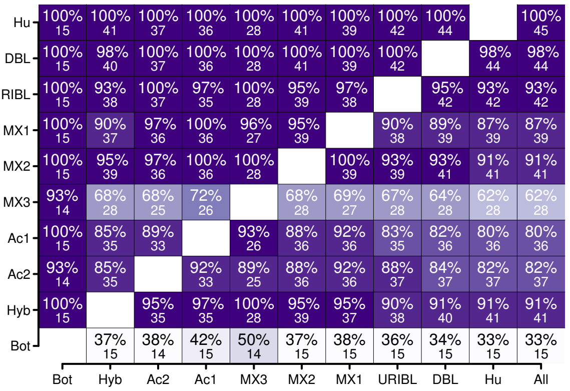
4.2.2 *Volume*

While there are millions of URLs and thousands of domains spammed daily, the number of messages in which each appears can vary dramatically. We call the number of messages advertising a domain the *volume* of that domain. Here we consider the coverage of each feed with respect to the relative volume of spam it covers. To estimate this quantity, we solicited the help of a large Web mail provider to measure the volume of spam domains at their incoming mail servers.

The incoming mail oracle. We refer to this data source as our *incoming mail oracle*. For this measurement, we collected all live domains seen across all feeds, and submitted them to the cooperating mail provider. The provider reported back to us the number of messages (normalized) containing each spam domain, as seen by their incoming mail servers over five days during the measurement period. This provider handles mail for hundreds of millions of users. Although the measurement collected is not a perfectly uniform sample of all spam globally, we believe it to be a reasonable representative. Given the limited duration of the measurement—five days versus three months of feed data—these results should be interpreted with caution.

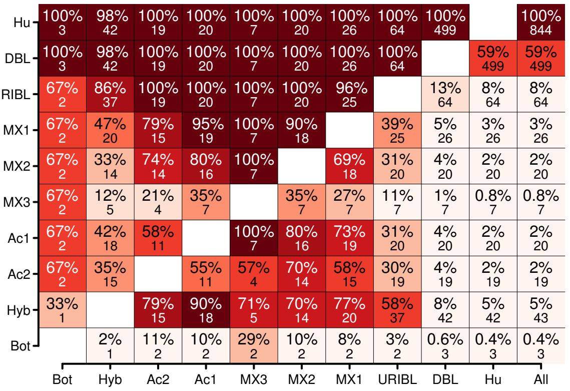
Figure 3 shows the volume of spam covered by the live and tagged domains in each feed. Recall that both live and tagged domains specifically exclude domains listed in the Alexa 1 million and domains appearing in the Open Directory Project listings (Section 4.1.4). In the figure, we've included the volume due to these Alexa and ODP domains occurring in each feed, shown stacked on top of the live and tagged volume bars. Before removing Alexa and ODP domains, the volume of live domains is dominated by these potential false positives. Among tagged domains, the volume attributed to Alexa and ODP domains (before exclusion) is much lower. These are domains which may have been used by the spammer as a redirection mechanism, either by abusing a legitimate service or via compromise. Of the feeds, the blacklists show the highest purity, as noted in Section 4.1.

With the Alexa and ODP domains excluded from the set of tagged domains, the URIBL blacklist provides the greatest coverage, followed by the Hu feed and DBL blacklist. At the opposite end, the Hyb feed provides only about a sixth of the coverage (by tagged domain volume) compared to URIBL, DBL, and Hu. Although it has nearly an order of magnitude more domains, its spam volume coverage is less than the Bot feed. One possibility is that this feed contains spam domains not derived from e-mail spam.



	Bot	Hyb	Ac2	Ac1	MX3	MX2	MX1	URIBL	DBL	Hu	All
Hu	100% 15	100% 41	100% 37	100% 36	100% 28	100% 41	100% 39	100% 42	100% 44		100% 45
DBL	100% 15	98% 40	100% 37	100% 36	100% 28	100% 41	100% 39	100% 42		98% 44	98% 44
URIBL	100% 15	93% 38	100% 37	97% 35	100% 28	95% 39	97% 38		95% 42	93% 42	93% 42
MX1	100% 15	90% 37	97% 36	100% 36	96% 27	95% 39		90% 38	89% 39	87% 39	87% 39
MX2	100% 15	95% 39	97% 36	100% 36	100% 28		100% 39	93% 39	93% 41	91% 41	91% 41
MX3	93% 14	68% 28	68% 25	72% 26		68% 28	69% 27	67% 28	64% 28	62% 28	62% 28
Ac1	100% 15	85% 35	89% 33		93% 26	88% 36	92% 36	83% 35	82% 36	80% 36	80% 36
Ac2	93% 14	85% 35		92% 33	89% 25	88% 36	92% 36	88% 37	84% 37	82% 37	82% 37
Hyb	100% 15		95% 35	97% 35	100% 28	95% 39	95% 37	90% 38	91% 40	91% 41	91% 41
Bot		37% 15	38% 14	42% 15	50% 14	37% 15	38% 15	36% 15	34% 15	33% 15	33% 15

Figure 4: Pairwise feed similarity with respect to covered affiliate programs.



	Bot	Hyb	Ac2	Ac1	MX3	MX2	MX1	URIBL	DBL	Hu	All
Hu	100% 3	98% 42	100% 19	100% 20	100% 7	100% 20	100% 26	100% 64	100% 499		100% 844
DBL	100% 3	98% 42	100% 19	100% 20	100% 7	100% 20	100% 26	100% 64		59% 499	59% 499
URIBL	67% 2	86% 37	100% 19	100% 20	100% 7	100% 20	96% 25		13% 64	8% 64	8% 64
MX1	67% 2	47% 20	79% 15	95% 19	100% 7	90% 18		39% 25	5% 26	3% 26	3% 26
MX2	67% 2	33% 14	74% 14	80% 16	100% 7		69% 18	31% 20	4% 20	2% 20	2% 20
MX3	67% 2	12% 5	21% 4	35% 7		35% 7	27% 7	11% 7	1% 7	0.8% 7	0.8% 7
Ac1	67% 2	42% 18	58% 11		100% 7	80% 16	73% 19	31% 20	4% 20	2% 20	2% 20
Ac2	67% 2	35% 15		55% 11	57% 4	70% 14	58% 15	30% 19	4% 19	2% 19	2% 19
Hyb	33% 1		79% 15	90% 18	71% 5	70% 14	77% 20	58% 37	8% 42	5% 42	5% 43
Bot		2% 1	11% 2	10% 2	29% 2	10% 2	8% 2	3% 2	0.6% 3	0.4% 3	0.4% 3

Figure 5: Pairwise feed similarity with respect to covered RX-Promotion affiliate identifiers.

4.2.3 *Affiliate Programs*

Up to this point, our focus has been the domains occurring in feeds, with the implicit understanding that domains represent a spam campaign. The relationship between a campaign and the domains it uses can be complex: a domain may be used in multiple campaigns, and a campaign may continuously cycle through several domains.

In fact, there is another level of structure beyond domains and campaigns: affiliate programs. Today, spammers operate primarily as advertisers, working with an affiliate program and earning a commission (typically 30–50%). The affiliate program handles Web site design, payment processing, customer service and fulfillment [33].

The prior Click Trajectories measurement effort [18] identified 45 leading affiliate programs specializing in pharmaceutical sales, replica luxury goods, and "OEM" software (this classification includes all the major players in each category that advertise via spam). We use the classification results of this project to define the tagged domains (Section 4.1). Here, we explore the tags themselves, that is, the affiliate programs associated with domains. Specifically, we consider the coverage of each feed with respect to *affiliate programs*.

Figure 4 shows the proportion of programs covered by each feed, relative to other feeds and all feeds combined. The representation is the same as Figure 2: each cell indicates the number of programs represented by the two feeds given by the row and column labels, and the top number of each cell expresses this intersection relative to the second program (identified by the column). For example, the cell in row MX$_1$ column Hyb indicates that 37 affiliate programs are seen in both feeds, and that these 37 programs represent approximately 90% of the 41 programs appearing in the Hyb feed.

Generally, most feeds do a good job covering all programs. The MX$_3$ feed has the second worst coverage, covering only 62% of all programs. Not surprisingly, the Bot feed has the worst coverage: only 15 programs. This poor coverage is partly because botnet operators frequently act as affiliates themselves and thus only advertise for a modest number of programs where they have spent the time to establish themselves. Even a botnet for rent will have a modest number of users and thus any given botnet will typically spam for a small number of programs in a given time.

4.2.4 *RX-Promotion*

In the affiliate marketing model for spam-advertised goods, the store site is usually hosted by the affiliate program itself, and not the spammer. For the affiliate program to determine which affiliate should receive credit for a sale, the URL itself must uniquely identify the affiliate. The most common mechanism is to assign each (major) affiliate a handful of dedicated domains.[6] Any sales generated via those domains are automatically credited to the appropriate affiliate.

One program, RX-Promotion, embeds an affiliate identifier in the page source of the storefront itself. This embedding allowed us to extract affiliate identifiers and map them to domains. In total, we were able to identify 846 distinct affiliate identifiers.

Affiliate coverage. Figure 5 shows the pairwise feed comparison matrix for RX-Promotion affiliate identifier coverage. Similar to affiliate program coverage, the human-identified feed Hu contributes the largest number of distinct affiliates. In this case, however, the difference between Hu and other feeds is more pronounced, with more than 40% of the affiliates found exclusively in Hu. The remaining feeds follow the same pattern as before. The MX honeypots (especially MX$_1$ and MX$_2$) continue to offer slightly greater coverage than seeded honey accounts (Ac$_1$ and Ac$_2$). Most striking is the paucity of affiliate IDs in the botnet feeds, confirming our earlier suspicion that botnet-originated affiliate program spam is associated with a single individual (the botnet operator).

Revenue coverage. The ultimate measure of an affiliate's success is the revenue he generates for the program. By this measure, a feed's value lies not in how many affiliates it covers, but in how much revenue it covers.

In a recent dispute between rival pharmacy affiliate programs, a number of RX-Promotion documents were leaked to the public [23]. One such document available to us, lists annual revenue generated by each RX-Promotion affiliate in 2010. Using these revenue statistics, we calculate affiliate coverage *weighted by affiliate revenue*. Figure 6 shows the revenue-weighted affiliate coverage of each feed.

The domains advertised in the smaller campaigns only found in Hu and DBL generate an order of magnitude more revenue than the sites advertised by bots and typically five times more than those sites seen in MX and honey account feeds. Overall, the results generally follow the affiliate cover-

[6] New domains must be constantly registered and assigned, as domains quickly become ineffective because of blacklisting.

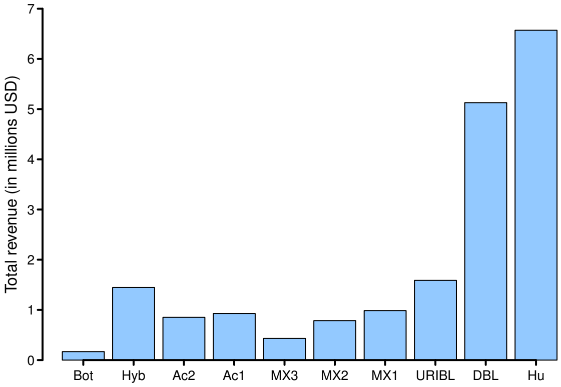
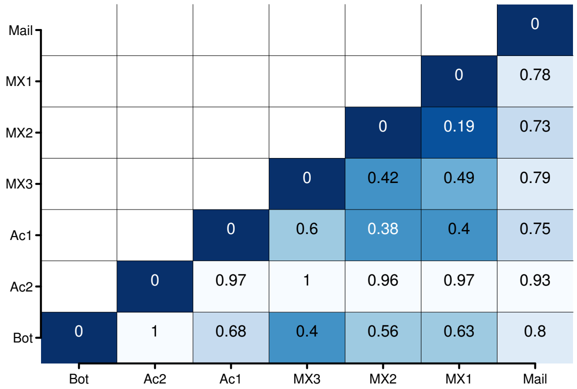


Figure 6: RX-Promotion affiliate coverage of each feed weighted by each affiliate's 2010 revenue.



	Bot	Ac2	Ac1	MX3	MX2	MX1	Mail
Mail							0
MX1						0	0.78
MX2					0	0.19	0.73
MX3				0	0.42	0.49	0.79
Ac1			0	0.6	0.38	0.4	0.75
Ac2		0	0.97	1	0.96	0.97	0.93
Bot	0	1	0.68	0.4	0.56	0.63	0.8

Figure 7: Pairwise variational distance of tagged domains frequency across all feeds. Shading is inverted (larger values are darker).

age shown in Figure 5, although the revenue-weighted results indicate a bias toward higher-revenue affiliates. While DBL covers only 59% of Hu affiliates, these affiliates represent over 78% of revenue covered by Hu.

4.3 Proportionality

An anti-spam system seeks to identify as many spam messages as possible, and in this context volume is a natural measure of a domain's importance. A blacklist that identifies the top 100 spammed domains by volume will identify more spam than a list of the same size consisting of infrequent domains. Similarly, domain take-downs are best prioritized to target high-volume domains first. To make these judgments, a spam domain feed must contain not only the domains themselves, but also their observed volume.

It happens that some of our feeds do provide volume information: each domain is listed with the number of times a domain was seen in spam, allowing relative domain volume and rank to be estimated. This section considers only feeds with volume information; the Hyb feed, Hu feed and both blacklist feeds (DBL and URIBL) have no associated volume information and are thus excluded from this analysis.

Empirical domain distribution and rank. The volumes associated with each domain define an empirical distribution on domains. That is, if a domain i has reported volume c_i in a feed, then the empirical domain probability distribution is c_i/m, where m is the total volume of the feed (i.e., $m = \sum_i c_i$).

Variation distance. Variation distance (also called "statistical difference" in some areas of Computer Science) is a straightforward metric frequently used to compare distributions. Formally, given two probability distributions (feeds) P and Q, let p_i be the empirical probability of domain i in P, and q_i the probability of the same domain in Q. (If a domain does not occur in a feed, its empirical probability is 0.) The variation distance is given by:

$$\delta = \frac{1}{2}\sum_i |p_i - q_i|.$$

Variation distance takes on values between 0 and 1, where $\delta = 0$ if and only if $P = Q$ (domains have the same probability in both), and $\delta = 1$ if P and Q are disjoint (no domains in common). Figure 7 shows pairwise measures of variation distance of tagged domains. (Because we round values to two decimal places, a variational distance of 1 in the figure may still allow for some domain overlap.)

Kendall rank correlation coefficient. Variation distance places more weight on more frequently occurring domains. In some cases, only the relative ranks of domains are of interest, and not the magnitudes of the empirical probabilities. The Kendall rank correlation coefficient (also called Kendall's tau-b) allows us to compare the relative ranking of domains between two distributions. In the case where all probabilities are distinct,

$$\tau = \frac{1}{n(n-1)}\sum_{i \neq j} \operatorname{sgn}\big[(p_i - p_j)(q_i - q_j)\big].$$

where $\operatorname{sgn}(x)$ is the familiar signum function. The sum is over all domains common to both feeds being compared, and n is the number of such domains. The Kendall rank correlation coefficient takes on values between -1 and 1, with 0 indicating no correlation, 1 indicating perfect positive correlation, and -1 indicating perfect negative correlation. If there are ties, i.e., $p_i = p_j$ or $q_i = q_j$ for some $i \neq j$, the denominator $n(n-1)$ must be adjusted to keep the full range between -1 to 1; we refer the reader to an appropriate Statistics textbook for details.

Figure 8 shows the pairwise tagged domain Kendall rank correlation coefficient between all feed pairs.

Pairwise comparison. Figures 7 and 8 show how well each pair of feeds agree in domain volume and rank. (The *Mail* column will be described shortly.) Qualitatively, both variation distance and Kendall rank correlation coefficient show similar results. The MX honeypot feeds and the Ac$_1$ honey account feeds exhibit similar domain distributions; these four also have many domains in common as seen in Figure 2.

The Bot feed brings a small number of domains, many of which also occur in the MX honeypot feeds and the Ac$_1$ feed (Figure 2). The volume of these domains, however, is significant; so much so, that in terms of domain proportions, the MX$_3$ feed is more like the Bot feed than any other feed, including the remaining MX honeypots.

The similarity in coverage and empirical domain probability distributions indicates that, roughly speaking, one MX honeypot feed is as good as another. By their nature, MX honeypots are targets of high-volume spammers who spam

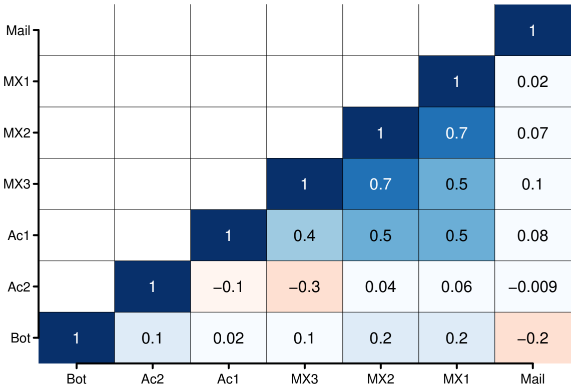


Figure 8: Pairwise Kendall rank correlation coefficient of tagged domain frequency across all feed pairs.

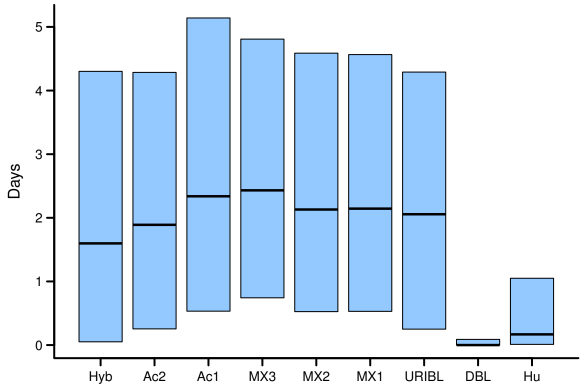


Figure 9: Relative first appearance time of domains in each feed. Campaign start time calculated from all feeds except Bot. Solid lines are medians; boxes range from the 25th to the 75th percentile.

randomly-generated names at all registered domains. By this process, it is just as easy to stumble upon one MX honeypot as another.

Comparison to real mail. In Section 4.2.2 we reported on the fraction of incoming spam—as seen by a major Web mail provider—covered by each feed. Here we use the same incoming mail oracle to determine the real-world relative volumes of spam domains, and compare those numbers to the relative domain volumes reported by each feed. We use only tagged domains appearing in at least one spam feed in the comparison: in the calculation of δ and τ, we set $p_i = 0$ for any domain i not appearing in the union of all spam feeds.

The *Mail* column in Figures 7 and 8 shows these results. The MX$_2$ feed comes closest to approximating the domain volume distribution of live mail, with Ac$_1$ coming close behind. As with coverage, the Ac$_2$ feed stands out as being most unlike the rest.

4.4 Timing

For both sides of the spam conflict, time is of the essence. For a spammer, the clock starts ticking as soon as a domain is advertised. It is only a matter of time before the domain is blacklisted, drastically reducing the deliverability of spam. While a domain is still clean, the spammer must maximize the number of messages delivered to potential customers. On the other side, blacklist maintainers strive to identify and blacklist spam domains as quickly as possible to maximize the volume of spam captured.

In this section we consider how well each spam feed captures the timing of spam campaigns. Specifically, we identify how quickly each feed lists spam domains, and, for feeds driven by live mail, how accurately they identify the end of a spam campaign. Unless noted otherwise, we restrict our analysis to tagged domains because we have the most confidence in their provenance.

Ideally, we would like to compare the time a domain first appears in spam with the time it first appears in a spam feed. Lacking such perfect knowledge about the start of each spam campaign, we instead take the *earliest* appearance time of a domain across all feeds as the *campaign start time*, and the *last* appearance time of a domain in live mail-based feeds as the *campaign end time*. For this analysis, we exclude the Bot feed because its domains have little overlap with the other feeds. As a result, including them greatly diminishes the number of domains that appear in the intersection of the feeds, and hence the number of domains that we can consider.

Taking the campaign start time and end time as described above, we define the *relative first appearance time* for a domain in a particular feed to the time between campaign start and its first appearance in the feed. In other words, we take campaign start time as “time zero” and calculate the relative first appearance time relative to this time. Put another way, the relative first appearance time is thus the *latency* of a feed with respect to a domain.

4.4.1 First Appearance Time

Figure 9 shows the distribution of relative first appearance times of domains in each feed. The bottom of the box corresponds to the 25th percentile, the top denotes the 75th percentile, and the solid bar inside the box denotes the median.

Both Hu and DBL are excellent early warnings of spam campaigns since they see most domains soon after they are used. The Hu feed sees over 75% of the domains within a day after they appear in any feed, and 95% within three days; DBL is delayed even less, with over 95% appearing on the blacklist within a day. Once again, the nature of these feeds lends themselves to observing wide-spread spam activity quickly: Hu has an enormous net for capturing spam, while DBL combines domain information from many sources. In contrast, the other feeds have much later first appearance times: they do not see roughly half of the domains until two days have passed, 75% until after four days, and 95% after ten. Operationally, by the time many of the domains appear in these feeds, spammers have already had multiple days to monetize their campaigns.

Of course, these results depend on both the set of domains that we consider and the sets of feeds we use to define campaign start times. When performing the same analysis on the larger set of live domains that appear in the same set of feeds, the first appearance times remain very similar to Figure 9: even for the broader set of domains, Hu and DBL see the domains multiple days earlier than the other feeds.

However, changing the set of *feeds* we consider does change relative first appearance times. Figure 10 shows similar

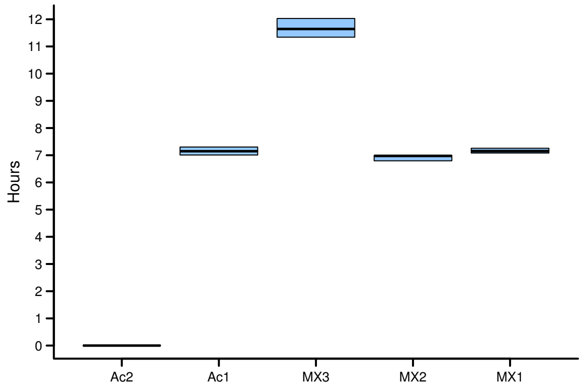


Figure 10: Relative first appearance time of domains in each feed. Campaign start time calculated from MX honeypot and honey account feeds only. Solid lines are medians; boxes range from the 25th to the 75th percentile.

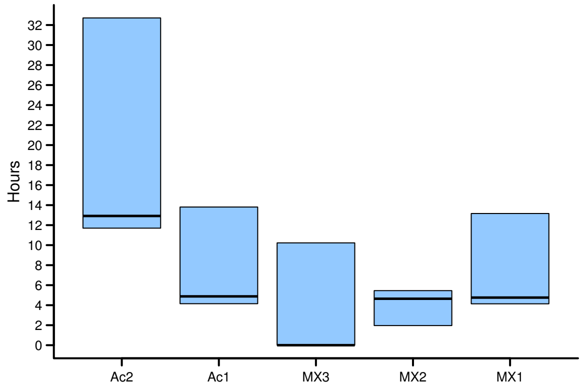


Figure 11: Distribution of differences between the last appearance of a domain in a particular and the domain campaign end calculated from an aggregate of the same five feeds. Solid lines are medians; boxes range from the 25th to the 75th percentile.

results as Figure 9, but with the Hu, Hyb, and blacklist feeds removed. (We chose these feeds because, as discussed further below, they all contain domains reported by users, which affects the last appearance times of domains.) Restricting the feeds we use to determine campaign start times reduces the total set of domains, but also increases the likelihood that a domain appears in all traces. When we focus on just the MX honeypot and account traces in this way, we see that relative to just each other they continue to have consistent first appearance times with each other, but the relative first appearance times are now very short (roughly less than a day). As with other metrics, these results show that timing estimates are quite relative and fundamentally depend on the feeds being considered.

4.4.2 Last Appearance Time

Last appearance times are often used to estimate when spam campaigns end. Figure 11 shows the time between the last appearance of a domain in a feed and the domain's campaign end time. As with Figure 10 we focus on only a subset of the feeds where the last appearance of a domain likely corresponds to when a spammer stopped sending messages using the domain: the MX honeypots and honeypot account feeds. Feeds like Hu, Hyb, and the blacklists all have domains reported by users. Since user reports fundamentally depend on when users read their mail and report spam, they may report spam long after a spammer has sent it.

Consistent with the first appearance times for the honeypot feeds, the feeds are similar to each other for last appearance times as well. The difference with the baseline are relatively short (a day or less), but have longer tails (the 95th percentiles for most are over a week).

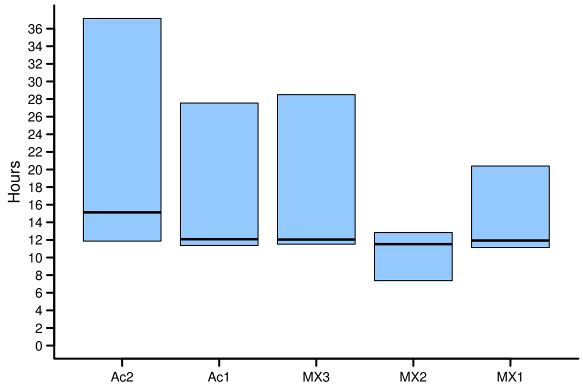


Figure 12: Distribution of differences between domain lifetime estimated using each feed and the domain campaign duration computed from an aggregate of those same five feeds. Solid lines are medians; boxes range from the 25th to the 75th percentile.

4.4.3 Duration

Another common metric for spam campaigns is their duration: how long spammers advertise domains to attract customers. Figure 12 shows the differences in time durations of domains advertised in spam as observed by each feed relative to estimated campaign duration (campaign end time minus campaign start time). For each feed we calculate the *lifetime* of a domain in the feed using the first and last appearance of a domain just in that feed. Then we compute the difference between the domain lifetime in a feed and the estimated campaign duration. (Campaign duration is computed from the same five feeds and is always at least as long as the domain lifetime in any feed.) The box plots in the graph summarize the distributions of these differences across all domains in each feed.

The differences in duration estimates for the honeypot feeds are also consistent with their first and last appearance time estimates. The duration estimates across feeds are similar to each other, the duration estimates differ from the baseline by less than a day for half of the domains and roughly a day for 75% of the domains. The distribution tails are longer, though, with outliers underestimating durations by multiple weeks.

5. CONCLUSION

Most measurement studies focus on using data to infer new facts about the world. This goal is why we measure things—to put truth on an empirical footing. However, occasionally it is necessary to perform introspective studies such as this one to understand the limits of what we can conclude from available data.

While our analysis is not comprehensive, we have found significant variation among the ten feeds we did study. Based on these findings we recommend that researchers consider four different challenges whenever using spam data:

- Limited purity. Even the best spam feeds include benign domains and these domains should be anticipated in analyses. We should identify the "kinds" of benign domains that appear in a dataset and determine if their existence will bias results—in particular when spam feed data will be correlated with other data sources.

- Coverage limitations. MX and honey account spam sources are inherently biased towards loud broad campaigns. If we desire a broader view of what is advertised via spam and are unable to strike an arrangement with a large e-mail provider, operational domain blacklists are the next best source of such information.

- Temporal uncertainty. Studies of spam campaign timing should recognize how timing error can be introduced via different feeds. Botnet-based feeds are among the best for timing information, but naturally coverage is limited. Other feeds provide highly accurate "onset" information (e.g., blacklists and human-identified feeds) but may not provide a correspondingly accurate ending timestamp. This area is one where combining the features of different feeds may be appropriate.

- Lack of proportionality. It is tempting to measure the prevalence of one kind of spam in a feed and extrapolate to the entire world—"25% of all spam advertises eBooks!" or "My spam filter can block 99.99% of all spam". However, the significant differences in the makeup of the feeds we have studied suggests that any such conclusion is risky. For example, spam filter results trained on botnet output may have little relevance to a large Web mail provider. In general, we advise making such claims based on knowledge of the source data set. For example, MX-based honeypots may be appropriate for characterizing relative prevalence among distinct high volume spam campaigns.

While it is important to be aware of the limitations and challenges of spam feeds, an even more interesting question is what feeds one should use for related studies. The clear answer, as shown by our results, is that there is no perfect feed. Instead, the choice should be closely related to the questions we are trying to answer. It is still possible, though, to provide some general guidelines that would apply for most cases:

- Human identified feeds, which are provided by large mail providers, will usually be the best choice for most studies. They provide a clear advantage when it comes to coverage, due to their wide exposure, and allow for visibility inside low-volume campaigns. They do so with reasonable purity, but due to the presence of the human factor, filtering is required. On the other hand, we should avoid human identified feeds when we are interested in timing, and especially last appearance information.

- If it is not possible to get access to human identified feeds, due to their limited availability, high-quality blacklist feeds offer very good coverage and first appearance information. They also offer the best purity since they are usually commercially maintained, and have low false positives as their primary goal. Similar to human identified feeds, they are less useful for studies that rely on last appearance or duration information.

- When working with multiple feeds, the priority should be to obtain a set that is as diverse as possible. Additional feeds of the same type offer reduced added value, and this situation is especially true in the case of MX honeypot feeds.

- It is very challenging to obtain accurate information regarding volume and provide conclusions that apply to the entirety of the spam problem. Given our limited view into the global spam output, all results are inherently tied to their respective input datasets.

In a sense, the spam research community is blessed by having so many different kinds of data sources available to it. In many other measurement regimes the problem of bias is just as great, but the number of data sources on hand is far fewer. However, with great data diversity comes great responsibility. It is no longer reasonable to take a single spam feed and extrapolate blindly without validation. Our paper provides a basic understanding of the limitations of existing feeds and provides a blueprint for refining this understanding further.

Acknowledgments

We would like to thank the named and anonymous providers of our feeds, whose willingness to share data with us made a paper such as this possible. We are also grateful to Brian Kantor and Cindy Moore who have managed our systems and storage needs.

This work was supported by National Science Foundation grants NSF-0433668, NSF-0433702, NSF-0831138, by Office of Naval Research MURI grant N000140911081, and by generous research, operational and in-kind support from the UCSD Center for Networked Systems (CNS).

6. REFERENCES

[1] Alexa. Alexa top 500 global sites. `http://www.alexa.com/topsites`, June 2011.

[2] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. Spamscatter: Characterizing Internet Scam Hosting Infrastructure. In *Proc. of 16th USENIX Security*, 2007.

[3] I. Androutsopoulos, J. Koutsias, K. Chandrinos, G. Paliouras, and C. D. Spyropoulos. An Evaluation of Naive Bayesian Anti-Spam Filtering. In *Proc. of 1st MLNIA*, 2000.

[4] R. Beverly and K. Sollins. Exploiting Transport-Level Characteristics of Spam. In *Proc. of 5th CEAS*, 2008.

[5] X. Carreras and L. Màrquez. Boosting Trees for Anti-Spam Email Filtering. In *Proceedings of RANLP-2001*, 2001.

[6] R. Clayton. How much did shutting down McColo help? In *Proc. of 6th CEAS*, 2009.

[7] H. Drucker, D. Wu, and V. N. Vapnik. Support vector machines for spam categorization. In *Proc. of IEEE Transactions on Neural Networks*, 1999.

[8] G. Gee and P. Kim. Doppleganger Domains. `http://www.wired.com/images_blogs/threatlevel/2011/09/Doppelganger.Domains.pdf`, 2011.

[9] P. H. C. Guerra, D. Guedes, W. M. Jr., C. Hoepers, M. H. P. C. Chaves, and K. Steding-Jessen. Spamming Chains: A New Way of Understanding Spammer Behavior. In *Proc. of 6th CEAS*, 2009.

[10] P. H. C. Guerra, D. Guedes, W. M. Jr., C. Hoepers, M. H. P. C. Chaves, and K. Steding-Jessen. Exploring the Spam Arms Race to Characterize Spam Evolution. In *Proc. of 7th CEAS*, 2010.

[11] J. P. John, A. Moshchuk, S. D. Gribble, and A. Krishnamurthy. Studying Spamming Botnets Using Botlab. In *Proc. of 6th NSDI*, 2009.

[12] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: An Empirical Analysis of Spam Marketing Conversion. In *Proc. of 15th ACM CCS*, 2008.

[13] M. Konte, N. Feamster, and J. Jung. Dynamics of Online Scam Hosting Infrastructure. In *PAM*, 2009.

[14] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the Spam Campaign Trail. In *Proc. 1st USENIX LEET*, 2008.

[15] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamcraft: An Inside Look at Spam Campaign Orchestration. In *Proc. of 2nd USENIX LEET*, 2009.

[16] M. Lee. Why My Email Went. `http://www.symantec.com/connect/blogs/why-my-email-went`, 2011.

[17] N. Leontiadis, T. Moore, and N. Christin. Measuring and Analyzing Search-Redirection Attacks in the Illicit Online Prescription Drug Trade. In *Proc. of USENIX Security*, 2011.

[18] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proc. of IEEE Symposium on Security and Privacy*, 2011.

[19] H. Liu, K. Levchenko, M. Félegyházi, C. Kreibich, G. Maier, G. M. Voelker, and S. Savage. On the Effects of Registrar-level Intervention. In *Proc. of 4th USENIX LEET*, 2011.

[20] M86 Security Labs. Top Spam Affiliate Programs. `http://www.m86security.com/labs/traceitem.asp?article=1070`, 2009.

[21] Marshal8e6 TRACELabs. Marshal8e6 Security Threats: Email and Web Threats. `http://www.marshal.com/newsimages/trace/Marshal8e6_TRACE_Report_Jan2009.pdf`, 2009.

[22] M. M. Masud, L. Khan, and B. Thuraisingham. Feature Based Techniques for Auto-Detection of Novel Email Worms. In *Proc. of 11th PACKDDD*, 2007.

[23] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko. PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs. In *Proc. of the USENIX Security Symposium*, 2012.

[24] D. K. McGrath and M. Gupta. Behind Phishing: An Examination of Phisher Modi Operandi. In *Proc. of 1st USENIX LEET*, 2008.

[25] T. Moore and R. Clayton. Examining the Impact of Website Take-down on Phishing. In *Proceedings of the Anti-Phishing Working Group's 2nd annual eCrime Researchers Summit.* ACM, 2007.

[26] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia. Exploiting Machine Learning to Subvert Your Spam Filter. In *Proc. of 1st USENIX LEET*, 2008.

[27] ODP – Open Directory Project. `http://www.dmoz.org`, September 2011.

[28] A. Pathak, Y. C. Hu, , and Z. M. Mao. Peeking into Spammer Behavior from a Unique Vantage Point. In *Proc. of 1st USENIX LEET*, 2008.

[29] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan. Botnet Spam Campaigns Can Be Long Lasting: Evidence, Implications, and Analysis. In *Proc. of 9th ACM SIGMETRICS*, 2009.

[30] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. Voelkera, V. Paxson, N. Weaver, and S. Savage. Botnet Judo: Fighting Spam with Itself. In *Proc. of 17th NDSS*, 2010.

[31] Z. Qian, Z. Mao, Y. Xie, and F. Yu. On network-level clusters for spam detection. In *Proc. of 17th NDSS*, 2010.

[32] A. Ramachandran, N. Feamster, and S. Vempala. Filtering Spam with Behavioral Blacklisting. In *Proc. of 14th ACM CCS*, 2007.

[33] D. Samosseiko. The Partnerka — What is it, and why should you care? In *Proc. of Virus Bulletin Conference*, 2009.

[34] F. Sanchez, Z. Duan, and Y. Dong. Understanding Forgery Properties of Spam Delivery Paths. In *Proc. of 7th CEAS*, 2010.

[35] S. Sinha, M. Bailey, and F. Jahanian. Shades of Grey: On the effectiveness of reputation-based blacklists. In *Proc. of 3rd MALWARE*, 2008.

[36] O. Thonnard and M. Dacier. A Strategic Analysis of Spam Botnets Operations. In *Proc. of 8th CEAS*, 2011.

[37] Trustwave. Spam Statistics – Week ending Sep 2, 2012. `https://www.trustwave.com/support/labs/spam_statistics.asp`, September 2012.

[38] G. Warner. Random Pseudo-URLs Try to Confuse Anti-Spam Solutions. `http://garwarner.blogspot.com/2010/09/random-pseudo-urls-try-to-confuse-anti.html`, Sept. 2010.

[39] C. Wei, A. Sprague, G. Warner, and A. Skjellum. Identifying New Spam Domains by Hosting IPs: Improving Domain Blacklisting. In *Proc. of 7th CEAS*, 2010.

[40] A. G. West, A. J. Aviv, J. Chang, and I. Lee. Spam Mitigation Using Spatio-temporal Reputations From Blacklist History. In *Proc of 26th. ACSAC*, 2010.

[41] J. Whissell and C. Clarke. Clustering for Semi-Supervised Spam Filtering. In *Proc. of 8th CEAS*, 2011.

[42] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming Botnets: Signatures and Characteristics. In *Proceedings of ACM SIGCOMM*, 2008.

[43] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, I. Osipkov, G. Hulten, and J. Tygar. Characterizing Botnets from Email Spam Records. In *Proc. of 1st USENIX LEET*, 2008.

[44] J. Zittrain and L. Frieder. Spam Works: Evidence from Stock Touts and Corresponding Market Activity. Social Science Research Network, March 2007.