Spamscatter: Characterizing Internet Scam Hosting Infrastructure

David S. Anderson Chris Fleizach Stefan Savage Geoffrey M. Voelker
Collaborative Center for Internet Epidemiology and Defenses
Department of Computer Science and Engineering
University of California, San Diego

Abstract

Unsolicited bulk e-mail, or SPAM, is a means to an end. For virtually all such messages, the intent is to attract the recipient into entering a commercial transaction — typically via a linked Web site. While the prodigious infrastructure used to pump out billions of such solicitations is essential, the engine driving this process is ultimately the "point-of-sale" — the various money-making "scams" that extract value from Internet users. In the hopes of better understanding the business pressures exerted on spammers, this paper focuses squarely on the Internet infrastructure used to host and support such scams. We describe an opportunistic measurement technique called *spamscatter* that mines emails in real-time, follows the embedded link structure, and automatically clusters the destination Web sites using *image shingling* to capture graphical similarity between rendered sites. We have implemented this approach on a large real-time spam feed (over 1M messages per week) and have identified and analyzed over 2,000 distinct scams on 7,000 distinct servers.

1 Introduction

Few Internet security issues have attained the universal public recognition or contempt of unsolicited bulk email — SPAM. In 2006, industry estimates suggest that such messages comprise over 80% over all Internet email with a total volume up to 85 *billion* per day [15,17]. The scale of these numbers underscores the prodigious delivery infrastructures developed by "spammers" and in turn motivates the more than $1B spent annually on anti-spam technology. However, the engine that drives this arms race is not spam itself — which is simply a means to an end — but the various money-making "scams" (legal or illegal) that extract value from Internet users.

In this paper, we focus on the Internet infrastructure used to host and support such scams. In particular, we analyze spam-advertised Web servers that offer merchandise and services (e.g., pharmaceuticals, luxury watches, mortgages) or use malicious means to defraud users (e.g., phishing, spyware, trojans). Unlike mail-relays or bots, scam infrastructure is directly implicated in the spam profit cycle and thus considerably rarer and more valuable. For example, a given spam campaign may use thousands of mail relay agents to deliver its millions of messages, but only use a single server to handle requests from recipients who respond. Consequently, the availability of scam infrastructure is critical to spam profitability — a single takedown of a scam server or a spammer redirect can curtail the earning potential of an entire spam campaign.

The goal of this paper is to characterize scam infrastructure and use this data to better understand the dynamics and business pressures exerted on spammers. To identify scam infrastructure, we employ an opportunistic technique called *spamscatter*. The underlying principle is that each scam is, by necessity, identified in the link structure of associated spams. To this end, we have built a system that mines email, identifies URLs in real time and follows such links to their eventual destination server (including any redirection mechanisms put in place). We further identify individual scams by clustering scam servers whose rendered Web pages are graphically similar using a technique called *image shingling*. Finally, we actively probe the scam servers on an ongoing basis to characterize dynamic behaviors like availability and lifetime. Using the spamscatter technique on a large real-time spam feed (roughly 150,000 per day) we have identified over 2,000 distinct scams hosted across more than 7,000 distinct servers. Further, we characterize the availability of infrastructure implicated in these scams and the relationship with business-related factors such as scam "type", location and blacklist inclusion.

The remainder of this paper is structured as follows. Section 2 reviews related measurement studies similar in topic or technique. In Section 3 we outline the struc-

ture and lifecycle of Internet scams, and describe in detail one of the more extensive scams from our trace as a concrete example. Section 4 describes our measurement methodology, including our probing system, image shingling algorithm, and spam feed. In Section 5, we analyze a wide range of characteristics of Internet scam infrastructure based upon the scams we identify in our spam feed. Finally, Section 6 summarizes our findings and concludes.

2 Related work

Spamscatter is an opportunistic network measurement technique [5], taking advantage of spurious traffic — in this case spam — to gain insight into "hidden" aspects of the Internet — in this case scam hosting infrastructure. As with other opportunistic measurement techniques, such as backscatter to measure Internet denial-of-service activity [20], network telescopes and Internet sinks [32] to measure Internet worm outbreaks [19, 21], and spam to measure spam relays [27], spamscatter provides a mechanism for studying global Internet behavior from a single or small number of vantage points.

We are certainly not the first to use spam for opportunistic measurement. Perhaps the work most closely related to ours is Ramachandran and Feamster's recent study using spam to characterize the network behavior of the spam relays that sent it [27]. Using extensive spam feeds, they categorized the network and geographic location, lifetime, platform, and network evasion techniques of spam relay infrastructure. They also evaluated the effectiveness of using network-level properties of spam relays, such as IP blacklists and suspect BGP announcements, to filter spam. When appropriate in our analyses, we compare and contrast characteristics of spam relays and scam hosts; some scam hosts also serve as spam relays, for example. In general, however, due to the different requirements of the two underground services, they exhibit different characteristics; scam hosts, for example, have longer lifetimes and are more concentrated in the U.S.

The Webb Spam Corpus effort harvests URLs from spam to create a repository of *Web spam* pages, Web pages created to influence Web search engine results or deceive users [31]. Although both their effort and our own harvest URLs from spam, the two projects differ in their use of the harvested URLs. The Webb Spam Corpus downloads and stores HTML content to create an offline data set for training classifiers of Web spam pages. Spamscatter probes sites and downloads content over time, renders browser screenshots to identify URLs referencing the same scam, and analyzes various characteristics of the infrastructure hosting scams.

Both community and commercial services consume URLs extracted from spam. Various community services mine spam to specifically identify and track phishing sites, either by examining spam from their own feeds or collecting spam email and URLs submitted by the community [1, 6, 22, 25]. Commercial Web security and filtering services, such as Websense and Brightcloud, track and analyze Web sites to categorize and filter content, and to identify phishing sites and sites hosting other potentially malicious content such as spyware and keyloggers. Sites advertised in spam provide an important data source for such services. While we use similar data in our work, our goal is infrastructure characterization rather than operational filtering.

Botnets can play a role in the scam host infrastructure, either by hosting the spam relays generating the spam we see or by hosting the scam servers. A number of recent efforts have developed techniques for measuring botnet structure, behavior, and prevalence. Cook et al. [9] tested the feasibility of using honeypots to capture bots, and proposed a combination of passive host and network monitoring to detect botnets. Bächer et al. [23] used honeynets to capture bots, infiltrate their command and control channel, and monitor botnet activity. Rajab et al. [26] combined a number of measurement techniques, including malware collection, IRC command and control tracking, and DNS cache probing. The last two approaches have provided substantial insight into botnet activity by tracking hundreds of botnets over periods of months. Ramachandran and Feamster [27] provided strong evidence that botnets are commonly used as platforms for spam relays; our results suggest botnets are not as common for scam hosting.

We developed an image shingling algorithm to determine the equivalance of screenshots of rendered Web pages. Previous efforts have developed techniques to determine the equivalence of transformed images as well. For instance, the SpoofGuard anti-phishing Web browser plugin compares images on Web pages with a database of corporate logos [7] to identify Web site spoofing. SpoofGuard compares images using robust image hashing, an approach employing signal processing techniques to create a compressed representation of an image [30]. Robust image hashing works well against a number of different image transformations, such as cropping, scaling, and filtering. However, unlike image shingling, image hashing is not intended to compare images where substantial regions have completely different content; refinements to image hashing improve robustness (e.g., [18, 28]), but do not fundamentally extend the original set of transforms.

3 The life and times of an Internet scam

In this section we outline the structure and life cycle of Internet scams, and describe in detail one of the

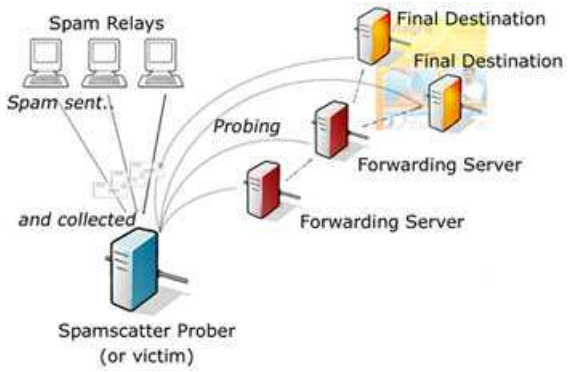

Figure 1: Components of a typical Internet scam.

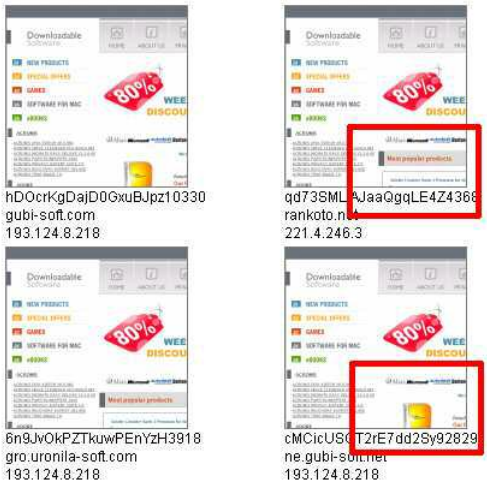

Figure 2: Screenshots, hostnames, and IP addresses of different hosts for the "Downloadable Software" scam. The highlighted regions show portions of the page that change on each access due to product rotation. Image shingling is resilient to such changes and identifies these screenshots as equivalent pages.

more extensive scams from our trace as a concrete example. This particular scam advertises "Downloadable Software," such as office productivity tools (Microsoft, Adobe, etc.) and popular games, although in general the scams we observed were diverse in what they offered (Section 5.1).

Figure 1 depicts the life of a spam-driven Internet scam. First, a spam campaign launches a vast number of unsolicited spam messages to email addresses around the world; a large spam campaign can exceed 1 billion emails [12]. In turn the content in these messages frequently advertises a *scam* — unsolicited merchandise and services available through the Web — by embedding URLs to scam Web servers in the spam; in our data, roughly 30% of spam contains such URLs (Section 5.1). An example of spam that does not contain links would be "pump-and-dump" stock spam intended to manipulate penny stock prices [3]; the recent growth of image-based stock spam has substantially reduced the fraction of spam using embedded URLs, shrinking from 85% in 2005 to 55% in 2006 [12]. These spam campaigns can be comparatively brief, with more than half lasting less than 12 hours in our data (Section 5.4). For our example software scam, over 5,000 spam emails were used to advertise it over a weeklong period.

Knowing or unsuspecting users click on URLs in spam to access content from the Web servers hosting the scams. While sometimes the embedded URL directly specifies the scam server, more commonly it indicates an intermediate Web server that subsequently redirects traffic (using HTTP or Javascript) on towards the scam server. Redirection serves multiple purposes. When spammer and scammer are distinct, it provides a simple means for tagging requests with the spammer's affiliate identifier (used by third-party merchants to compensate independent "advertisers") and laundering the spam-based origin before the request reaches the merchant (this laundering provides plausible deniability for the merchant and protects the spammer from potential conflicts over the merchant's advertising policy). If spammer and scammer are the same, a layer of redirection is still useful for avoiding URL-based blacklists and providing deployment flexibility for scam servers. In our traces, most scams use at least one level of redirection (Section 4).

On the back end, scams may use multiple servers to host scams, both in terms of multiple virtual hosts (e.g., different domain names served by the same Web server) and multiple physical hosts identified by IP address (Section 5.2). However, for the scams in our spam feed, the use of multiple virtual hosts is infrequent (16% of scams) and multiple physical hosts is rare (6%); our example software scam is one of the more extensive scams, using at least 99 virtual hosts on three physical hosts.

Finally, different Web servers (physical or virtual), and even different accesses to a scam using the same URL, can result in slightly different downloaded content for the same scam. Intentional randomness for evasion, rotating advertisements, featured product rotation, etc., add another form of aliasing. Figure 2 shows example screenshots among different hosts for the software scam. To overcome these aliasing issues, we use screenshots of Web pages as a basis for identifying all hosts participating in a given scam (Section 4.2).

A machine hosting one scam may be shared with other scams, as when scammers run multiple scams at once or the hosts are third-party infrastructure used by multiple scammers. Sharing is common, with 38% of scams be-

ing hosted on a machine with at least one other scam (Section 5.3). One of the machines hosting the software scam, for example, also hosted a pharmaceutical scam called "Toronto Pharmacy" (which happened to be hosted on a server in Guangzhou, China).

The lifetimes of scams are much longer than spam campaigns, with 50% of scams active for at least a week (Section 5.4). Furthermore, scam hosts have high availability during their lifetime (most above 99%) and appear to have good network connectivity (Section 5.5); the lifetime of our software scam ran for the entire measurement period and was available 97% of the time. Finally, scam hosts tend to be geographically concentrated in the United States; over 57% of scam hosts from our data mapped to the U.S. (Section 5.6.2). Such geographic concentration contrasts sharply with the location of spam relay hosts; for comparison, only 14% of spam relays used to send the spam to our feed are located in the U.S. Figure 3 shows the geographic locations of the spam relays and scam hosts for the software scam. The three scam hosts were located in China and Russia, whereas the 85 spam relays were located around the world in 30 countries.

The lifetimes, high availability, and good network connectivity, as well as the geographic diversity of spam relays compared with scam hosts, all reflect the fundamentally different requirements and circumstances between the two underground services. Spam relays require no interaction with users, need only be available to send mail, but must be great enough in number to mitigate the effects of per-host blacklists. Consequently, spam relays are well suited to "commodity" botnet infrastructure [27]; one recent industry estimate suggests that over 80% of spam is in fact relayed by bots [13]. By contrast, scam hosts are naturally more centralized (due to hosting a payment infrastructure), require interactive response time to their target customers, and may — in fact — be hosting legal commerce. Thus, scam hosts are much more likely to have high-quality hosting infrastructure that is stable over long periods.

4 Methodology

This section describes our measurement methodology. We first explain our data collection framework for probing scam hosts and spam relays, and then detail our image shingling algorithm for identifying equivalent scams. Finally, we describe the spam feed we use as our data source and discuss the inherent limitations of using a single viewpoint.

4.1 Data collection framework

We built a data collection tool, called the *spamscatter prober*, that takes as input a feed of spam emails, extracts the sender and URLs from the spam messages, and probes those hosts to collect various kinds of information (Figure 1). For spam senders, the prober performs a ping, traceroute, and DNS-based blacklist lookup (DNSBL) once upon receipt of each spam email. The prober performs more extensive operations for the scam hosts. As with spam senders, it first performs a ping, traceroute, and DNSBL lookup on scam hosts. In addition, it downloads and stores the full HTML source of the Web page specified by valid URLs extracted from the spam (we do not attempt to de-obfuscate URLs). It also renders an image of the downloaded page in a canonical browser configuration using the KHTML layout engine [14], and stores a screenshot of the browser window. For scam hosts, the prober repeats these operations periodically for a fixed length of time. For the trace in this paper, we probed each host and captured a screenshot while visiting each URL every three hours. Starting from when the first spam email introduces a new URL into the data set, we probe the scam host serving that URL for a week independently of whether the probes fail or succeed.

As we mentioned earlier, many spam URLs simply point to sites that forward the request onto another server. There are many possible reasons for the forwarding behavior, such as tracking users, redirecting users through third-party affiliates or tracking systems, or consolidating the many URLs used in spam (ostensibly to avoid spam filters) to just one. Occasionally, we also noticed forwarding that does not end, either indicating a misconfiguration, programming error, or a deliberate attempt to avoid spidering.

The prober accommodates a variety of link forwarding practices. While some links direct the client immediately to the appropriate Web server, others execute a series of forwarding requests, including HTTP 302 server redirects and JavaScript-based redirects. To follow these, the prober processes received page content to extract simple META refresh tags and JavaScript redirect statements. It then tracks every intermediate page between the initial link and the final content page, and marks whether a page is the end of the line for each link. Properly handling forwarding is necessary for accurate scam monitoring. Over 68% of scams used some kind of forwarding, with an average of 1.2 forwards per URL.

4.2 Image shingling

Many of our analyses compare content downloaded from scam servers to determine if the scams are equivalent. For example, scam hosts may serve multiple indepen-

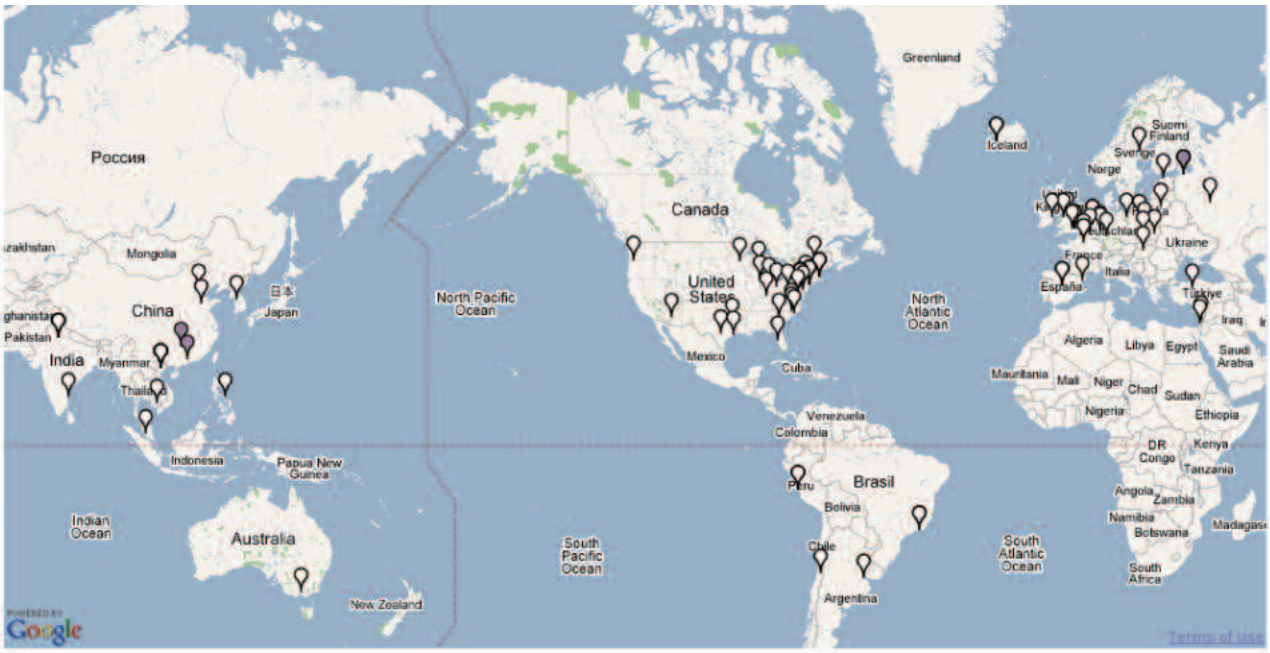

Figure 3: Geographic locations of the spam relays and scam server hosts for the "Downloadable Software" scam. The three scam servers are located in China and Russia and shown with dark grey points. The 85 spam relays are located around the world in more than 30 different countries, and are shown in white.

dent scams simultaneously, and we cannot assume that URLs that lead to the same host are part of the same scam. Similarly, scams are hosted on multiple virtual servers as well as distributed across multiple machines. As a result, we need to be able to compare content from scam servers on different hosts to determine whether they are part of the same scam. Finally, even for content downloaded from the same URL over time, we need to determine whether the content fundamentally changes (e.g., the server has stopped hosting the scam but returns valid HTTP responses to requests, or it has transitioned to hosting a different scam altogether).

Various kinds of aliasing make determining scam equivalence across multiple hosts, as well as over time, a challenging problem. One possibility is to compare spam messages within a window of time to identify emails advertising the same scam. However, the randomness and churn that spammers introduce to defeat spam filters makes it extremely difficult to use textual information in the spam message to identify spam messages for the same scam (e.g., spam filters continue to struggle with spam message equivalence). Another possibility is to compare the URLs themselves. Unfortunately, scammers have many incentives not to use the same URL across spams, and as a result each spam message for a scam might use a distinct URL for accessing a scam server. For instance, scammers may embed unique tracking identifiers in the query part of URLs, use URLs that contain domain names to different virtual servers, or simply randomize URLs to defeat URL blacklisting.

A third option is to compare the HTML content downloaded from the URLs in the spam for equivalence. The problem of comparing Web pages is a fundamental operation for any effort that identifies similar content across sites, and comparing textual Web content has been studied extensively already. For instance, text shingling techniques were developed to efficiently measure the similarity of Web pages, and to scale page comparison to the entire Web [4, 29]. In principle, a similar method could be used to compare the HTML text between scam sites, but in practice the downloaded HTML frequently provides insufficient textual information to reliably identify a scam. Indeed, many scams contained little textual content at all, and instead used images entirely to display content on the Web page. Also, many scams used frames, iframes, and JavaScript to display content, making it difficult to capture the full page context using a text-based Web crawler.

Finally, a fourth option is to render screenshots of the content downloaded from scam sites, and to compare the screenshots for equivalence. Screenshots are an attractive basis for comparison because they sidestep the aforementioned problems with comparing HTML source. However, comparing screenshots is not without

its own difficulties. Even for the same scam accessed by the same URL over time — much less across different scam servers — scam sites may intentionally introduce random perturbations of the page to prevent simple image comparison, display rotating advertisements in various parts of a page, or rotate images of featured products across accesses. Figure 2 presents an example of screenshots from different sites for the same scam that show variation between images due to product rotation.

Considering the options, we selected screenshots as the basis for determining spam equivalence. To overcome the problems described earlier, we developed an image-clustering algorithm, called image shingling, based on the notion of shingling from the text similarity literature. Text shingling decomposes a document into many segments, usually consisting of a small number of characters. Various techniques have been developed to increase the efficiency and reduce the space complexity of this process [11]. Next, these hashed "shingles" are sorted so that hashes for documents containing similar shingles are close together. The ordering allows all the documents that share an identical shingle to be found quickly. Finally, documents are clustered according to the percentage of shared shingles between them. The power of the algorithm is that it essentially performs $O(N^2)$ comparisons in $O(N\ lg\ N)$ time.

Our image shingling algorithm applies a similar process to the image domain. The algorithm first divides each image into fixed size chunks in memory; in our experiments, we found that an image chunk size of 40x40 pixels was an effective tradeoff between granularity and shingling performance. We then hash each chunk to create an image shingle, and store the shingle on a global list together with a link to the image (we use the MD4 hash to create shingles due to its relative speed compared with other hashing algorithms). After sorting the list of shingles, we create a hash table, indexed by shingle, to track the number of times two images shared a similar shingle. Scanning through the table, we create clusters of images by finding image pairs that share at least a threshold of similar images.

To determine an appropriate threshold value, we took one day's worth of screenshots and ran the image shingling algorithm for all values of thresholds in increments of 1%. Figure 4 shows the number of clusters created per threshold value. The plateau in the figure starting at 70% corresponds to a fair balance between being too strict, which would reduce the possibility of clustering nearly similar pages, and being too lenient, which would cluster distinct scams together. Manually inspecting the clusters generated at this threshold plateau and the cluster membership changes that occur at neighboring threshold values, we found that a threshold of 70% minimized false negatives and false positives for determining scam page equivalence.

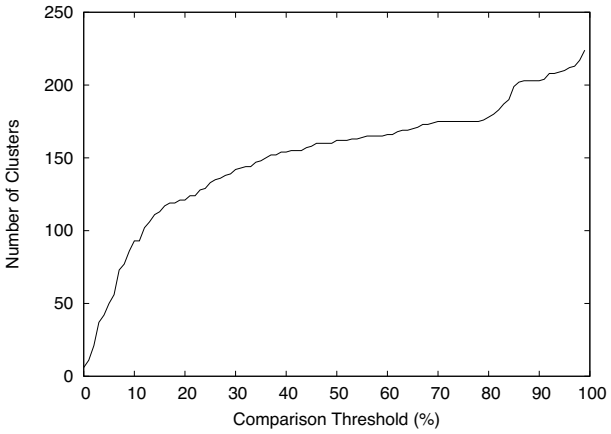

Figure 4: The choice of a threshold value for image shingling determines the number of clusters.

We have developed a highly optimized version of this basic algorithm that, in practice, completes an all-pairs comparison in roughly linear time. In practice, image shingling is highly effective at clustering similar scam pages, while neatly side-stepping the adversarial obfuscations in spam messages, URLs, and page contents. Clearly, a determined scammer could introduce steps to reduce the effectiveness of image shingling as described (e.g., by slightly changing the colors of the background or embedded images on each access, changing the compression ratio of embedded images, etc.). However, we have not witnessed this behavior in our trace. If scammers do take such steps, this methodology will likely need to be refined.

4.3 Spam feed and limitations

The source of spam determines the scams we can measure using this methodology. For this study, we have been able to take advantage of a substantial spam feed: all messages sent to any email address at a well-known four-letter top-level domain. This domain receives over 150,000 spam messages every day. We can assume that any email sent to addresses in this domain is spam because no active users use addresses on the mail server for the domain. Examining the "From" and "To" addresses of spam from this feed, we found that spammers generated "To" email addresses using a variety of methods, including harvested addresses found in text on Web pages, universal typical addresses at sites, as well as name-based dictionary address lists. Over 93% of "From" addresses were used only once, suggesting the use of random source addresses to defeat address-based spam blacklists.

Characteristic	*Summary Result*
Trace period	11/28/06 – 12/11/06
Spam messages	1,087,711
Spam w/ URLs	319,700 (30% of all spam)
Unique URLs	36,390 (11% of all URLs)
Unique IP addresses	7,029 (19% of unique URLs)
Unique scams	2,334 (6% of unique URLs)

Table 1: Summary of spamscatter trace.

We analyze Internet scam hosting infrastructure using spam from only a single, albeit highly active, spam feed. As with other techniques that use a single network viewpoint to study global Internet behavior, undoubtedly this single viewpoint introduces bias [2,8]. For example, the domain that provides our spam feed has no actual users who read the email. Any email address harvesting process that evaluates the quality of email addresses, such as correlating spam email targets with accesses on scam sites, would be able to determine that sending spam to these addresses yields no returns (that is, until we began probing).

While measuring the true bias of our data is impossible, we can anecdotally gauge the coverage of scams from our spam feed by comparing them with scams identified from an entirely different spam source. As a comparison source, we used the spam posted to the Usenet group news.admin.net-abuse.sightings, a forum for administrators to contribute spam [22]. Over a single 3-day period, January 26–28th, 2007, we collected spam from both sources. We captured 6,977 spam emails from the newsgroup and 113,216 spam emails from our feed. The newsgroup relies on user contributions and is moderated, and hence is a reliable source of spam. However, it is also a much smaller source of spam than our feed.

Next we used image shingling to distill the spam from both sources into distinct scams, 205 from the newsgroup and 1,687 from our feed. Comparing the scams, we found 25 that were in both sets, i.e., 12% of the newsgroup scams were captured in our feed as well. Of the 30 most-prominent scams identified from both feeds (in terms of the number of virtual hosts and IP addresses), ten come from the newsgroup feed. These same ten, furthermore, were also in our feed. Our goal was not to achieve global coverage of all Internet scams, and, as expected, we have not. The key question is how representative our sample is; without knowing the full set of scams (a very challenging measurement task), we cannot gauge the representativeness of the scams we find. Characterizing a large sample, however, still provides substantial insight into the infrastructure used to host scams. And it is further encouraging that many of the most extensive scams in the newsgroup feed are also found in ours. Moving forward, we plan to incorporate other sources of spam to expand our feed and further improve representativeness.

Scam category	*% of scams*
Uncategorized	29.57%
Information Technology	16.67%
Dynamic Content	11.52%
Business and Economy	6.23%
Shopping	4.30%
Financial Data and Services	3.61%
Illegal or Questionable	2.15%
Adult	1.80%
Message Boards and Clubs	1.80%
Web Hosting	1.63%

Table 2: Top ten scam categories.

5 Analysis

We analyze Internet scam infrastructure using scams identified from a large one-week trace of spam messages. We start by summarizing the characteristics of our trace and the scams we identify. We then evaluate to what extent scams use multiple hosts as distributed infrastructure; using multiple hosts can help scams be more resilient to defenses. Next we examine how hosts are shared across scams as an indication of infrastructure reuse. We then characterize the lifetime and availability of scams. Scammers have an incentive to use host infrastructure that provides longer lifetimes and higher availability; at the same time, network and system administrators may actively filter or take down scams, particularly malicious ones. Lastly, we examine the network and geographic locations of scams; again, scammers can benefit from using stable hosts that provide high availability and good network connectivity.

Furthermore, since spam relay hosts are an integral aspect of Internet scams, where appropriate in our analyses we compare and contrast characteristics of spam relays and scam hosts.

5.1 Summary results

We collected the spam from our feed for a one-week period from November 28, 2006 to December 4, 2006. For every URL extracted from spam messages, we probed the host specified by the URL for a full week (independent of whether the host responded or not) starting from the moment we received the spam. As a result, the prober monitored some hosts for a week beyond the receipt of the last spam email, up until December 11. Table 1 summarizes the resulting spamscatter trace. Starting with over 1 million spam messages, we extracted

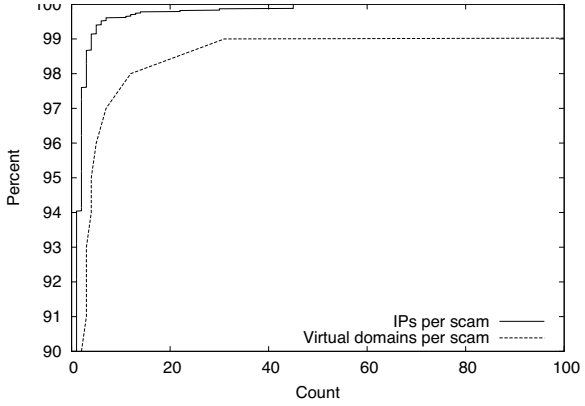

Figure 5: Number of IP address and virtual domains per scam.

Scam category	*# of domains*	*# of IPs*
Watches	3029	3
Pharmacy	695	4
Watches	110	3
Pharmacy	106	1
Software	99	3
Male Enhancement	94	2
Phishing	91	14
Viagra	90	1
Watches	81	1
Software	80	45

Figure 6: The ten largest virtual-hosted scams and the number of IP addresses hosting the scams.

36,390 unique URLs. Using image shingling, we identified 2,334 scams hosted on 7,029 machines. Spam is very redundant in advertising scams: on average, 100 spam messages with embedded URLs lead to only seven unique scams.

What kinds of scams do we observe in our trace? We use a commercial Web content filtering product to determine the prevalence of different kinds of scams. For every URL in our trace, we use the Web content filter to categorize the page downloaded from the URL. We then assign that category to the scams referenced by the URL.

Table 2 shows the ten most-prevalent scam categories. Note that we were not able to categorize all of the scams. We did not obtain access to the Web content filter until a few weeks after taking our traces, and 30% of the scams had URLs that timed out in DNS by that time ("Uncategorized" in the table). Further, 12% of the scams did not categorize due to the presence of dynamic content. The remaining 58% of scams fell into over 60 categories. Of these the most prevalent scam category was "Information Technology", which, when examining the screenshots of the scam sites, include click affiliates, survey and free merchandise offers and some merchandise for sale (e.g., hair loss, software). Just over 2% of the scams were labeled as malicious sites (e.g., containing malware).

5.2 Distributed infrastructure

We start by evaluating to what extent scams use multiple hosts as distributed infrastructure. Scams might use multiple hosts for fault-tolerance, for resilience in anticipation of administrative takedown or blacklisting, for geographic distribution, or even for load balancing. Also, reports of large-scale botnets are increasingly common, and botnets could provide a large-scale infrastructure for hosting scams; do we see evidence of botnets being used as a scalable platform for scam hosting?

We count multiple scam hosting from two perspectives, the number of virtual hosts used by a scam and the number of unique IP addresses used by those virtual hosts. Overall, the scams from our trace are typically hosted on a single IP address with one domain name. Of the 2,334 scams, 2,195 (94%) were hosted on a single IP address and 1,960 (84%) were hosted on a single domain name. Only a small fraction of scams use multiple hosting. Figure 5 shows the tails of the distributions of the number of virtual hosts and IP addresses used by the scams in our trace, and Table 6 lists the top ten scams with the largest number of domains and IP addresses. Roughly 10% of the scams use three or more virtual domains, and 1% use 15 or more. The top scams use hundreds of virtual domains, with one scam using over 3,000. Of the 6% of scams hosted on multiple IP addresses, only a few used more than ten, with one scam using 45. The relatively prevalent use of virtual hosts suggests that scammers are likely concerned about URL blacklisting and use distinct virtual hosts in URLs sent in different spam messages to defeat such blacklists.

The scams in our trace do not use hosting infrastructure distributed across the network extensively. Most scams are hosted on a single IP address, providing a potentially convenient single point for network-based interdiction either via IP blacklisting or network filtering. Assuming that scammers adapt to defenses to remain effective, such filtering does not appear to be applied extensively. Scam serving workloads are apparently low enough that a single host can satisfy offered load sufficiently to reap the benefits of the scam. Finally, if scams do use botnets as hosting infrastructure, then they are not used to scale a single scam. A scammer could potentially use a botnet to host multiple different scams, hosting each scam on a separate distinct bot, but our methodology would not identify this case.

Those few scams hosted on multiple IP addresses, however, are highly distributed. Scams with multiple IP addresses were most commonly distributed outside of

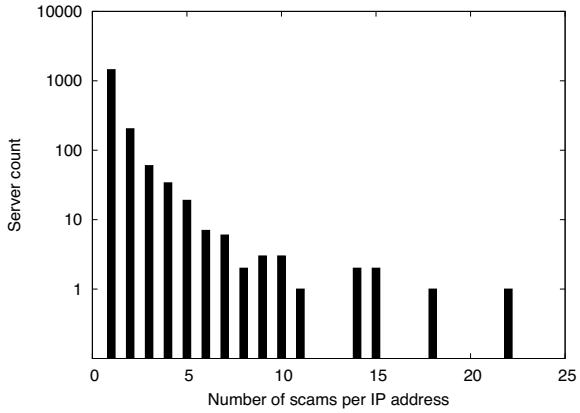

Figure 7: The number of scams found on a server IP address.

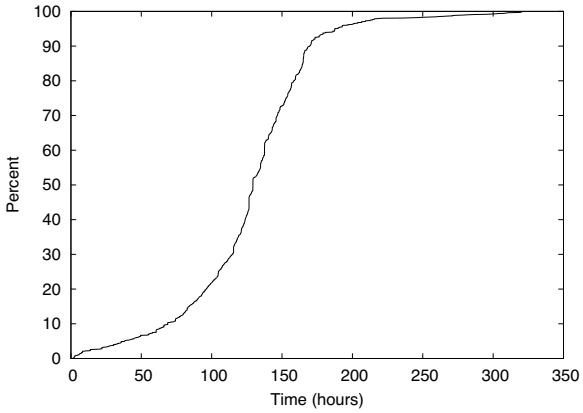

Figure 8: Overlap time for scam pairs on a server.

the same /24 prefix. Of the 139 distributed scams, all the hosts in 86% of the scams were located entirely on distinct /24 networks. Moreover, 64% of the distributed scams had host IP addresses that were all in entirely different ASes. As an example, one distributed scam was a phishing attack targeting a bank. The phishing Web pages were identical across 14 hosts, all in different /24 networks. The attack employed 91 distinct domain names. The domain names followed the same naming convention using a handful of common keywords followed by a set of numbers, suggesting the hosts were all involved in the distributed attack. The fully distributed nature of these scams suggests that scammers were concerned about resilience to defenses such as blacklisting.

5.3 Shared infrastructure

While we found that most scams are hosted on a single machine, a related question is whether these individual machines in turn host multiple scams, thereby sharing infrastructure across them. For each hosting IP address in our trace, we counted the number of unique scams hosted on that IP address at any time in the trace. Figure 7 shows these results as a logscale histogram. Shared infrastructure is rather prevalent: although 1,450 scams (62%) were hosted on their own machines, the remaining 38% of scams were hosted on machines hosting at least one other scam. Ten servers hosted ten or more scams, and the top three machines hosted 22, 18, and 15 different scams. This sharing of infrastructure suggests that scammers frequently either run multiple different scams on hosts that they control, or that hosts are made available (sold, rented, bartered) to multiple scammers.

Host type	*Classification*	*% of hosts recognized*
Spam relay	Open proxy	72.3%
	Spam host	5.86%
Scam host	Open proxy	2.06%
	Spam host	14.9%

Table 3: Blacklist classification of spam relays and scam hosts.

5.3.1 Sharing over time

We further examined these shared servers to determine if they host different scams sequentially or if, in fact, servers are used concurrently for different scams. For each pair of scams hosted on the same IP address, we compared their active times and durations with each other. When they overlapped, we calculated the duration of overlap. We found that scams sharing hosts shared them at the same time: 96% of all pairs of scams overlapped with each other when they remained active. Figure 8 shows the distribution of time for which scams overlapped. Over 50% of pairs of scams overlapped for at least 125 hours. Further calculating the ratio of time that scams sharing hosts were active, we found that overlapped scams did not necessarily start and end at the same time: only 10% of scam pairs fully overlapped each other.

5.3.2 Sharing between scam hosts and spam relays

More broadly, how often do the same machines serve as both spam relays as well as scam hosting? Hosts used for both spam and scams suggest, for instance, that either the spammer and the scammer are the same party, or that a third party controls the infrastructure and makes it available for use by different clients. We can only estimate the extent to which hosts play both roles, but we

estimate it in two ways. First, we determine the IP addresses of all of the hosts that send spam into our feed. We then compare those addresses with the IP addresses of the scam hosts. Based upon this comparison, we find only a small amount of overlap (9.7%) between the scam hosts and spam relays in our trace.

Scam hosts could, of course, serve as spam relays that do not happen to send spam to our feed. For a more global perspective, we identify whether the spam and scam hosts we observe in our trace are blacklisted on well-known Internet blacklists. When the prober sees an IP address for the first time (either from a host sending spam or from a scam host), it performs a blacklist query on that IP address using the DNSBLLookup Perl module [16].

Table 3 shows the percentage of blacklisted spam relays and scam hosts. This perspective identifies a larger percentage (17%) of scam hosts as also sending spam than we found by comparing scam hosts and open relays within our spam feed, but the percentage is still small overall. The blacklists are quite effective, though, at classifying the hosts that send spam to our feed: 78% of those hosts are blacklisted. The query identifies most of the spam hosts as open spam relays — servers that forward mail and mask the identity of the true sender — whereas most blacklisted scam hosts are identified as just sending spam directly. These results suggest that when scam hosts are also used to send spam, they are rarely used as an open spam service.

5.4 Lifetime

Next we examine how long scams remain active and, in the next section, how stable they are while active. The lifetime of a scam is a balance of competing factors. Scammers have an incentive to use hosting infrastructure that provides longer lifetimes and higher availability to increase their rate of return. On the other hand, for example, numerous community and commercial services provide feeds and products to help network administrators identify, filter or take down some scam sites, particularly phishing scams [1,6,22,25].

We define the lifetime of a scam as the time between the first and last successful timestamp for a probe operation during the two-week measurement period, independent of whether any probes failed in between (we look at the effect of failed probe attempts on availability below). We use two types of probes to examine scam host lifetime from different perspectives (Section 4). Periodic ping probes measure host network lifetime, and periodic HTTP requests measure scam server lifetime. Recall that we probe all hosts for a week after they appear in our spam feed — and no longer — to remove any bias towards hosts that appear early in the measurement study.

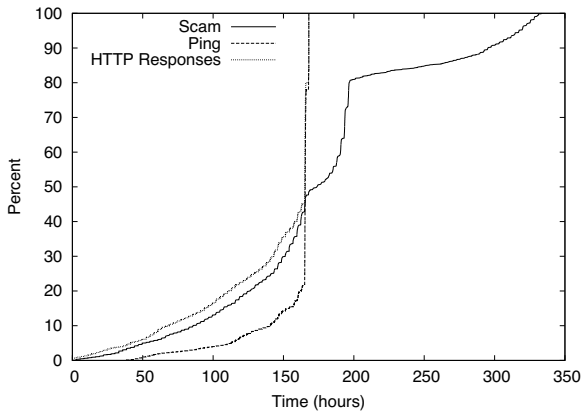

Figure 9: Lifetimes of individual scam hosts and Web servers, as well overall lifetimes of scams across multiple hosts.

For comparison, we also calculate the lifetimes of entire scams. For scams that use multiple hosts, their lifetimes start when the first host appears in our trace and end with the lifetime of the last host to respond. As a result, scam lifetimes can exceed a week.

How long are scams active? Figure 9 shows the distributions of scam lifetime based upon these probes for the scams in our trace. For ping probes, we show the distribution of just those scam hosts that responded to pings (67% of all scam hosts). Scam hosts had long network lifetimes. Over 50% of hosts responded to pings for nearly the entire week that we probed them, and fewer than 10% of hosts responded to pings for less than 80 hours. Given how close the distributions are, scam Web servers had only slightly shorter lifetimes overall. These results suggest that scam hosts are taken down soon after scam servers.

Comparing the distribution of scam lifetimes to the others, we see that scams benefit from using multiple hosts. The 50% of scams whose lifetimes exceed a week indicate that the lifetimes of the individual scam hosts do not entirely overlap each other. Indeed, individual hosts for some scams appeared throughout the week of our measurement study, and the overall scam lifetime approached the two weeks.

5.4.1 Lifetime by category

A substantial amount of community and commercial effort goes into identifying malicious sites, such as phishing scams, and placing those sites on URL or DNS/IP blacklists. Thus, we would expect that the hosting infrastructure for clearly malicious scams would be more transient than for other scams. To test this hypothesis, we used the categorization of scams to create a group of malicious scams that include the "Illegal or Question-

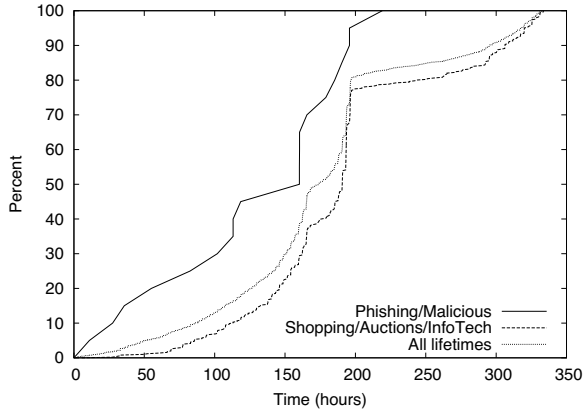

Figure 10: Scam lifetime distributions for malicious and shopping scams.

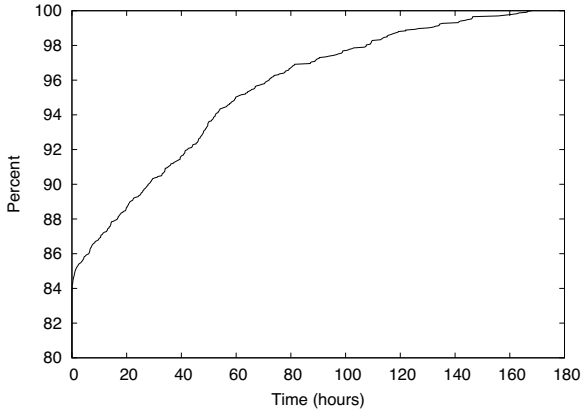

Figure 11: The duration of a spam campaign.

able" and "Phishing" categories labeled by the Web content filter (32 scams). For comparison, we also broke out another group of more innocuous shopping scams that include the "Shopping", "Information Technology", and "Auction" categories (701 scams).

We examined the lifetimes and prevalence on blacklists of these scams. Figure 10 shows the lifetime distributions of the malicious and shopping groups of scams, and includes the distribution of all scams from Figure 9 for reference. The malicious scams have a noticeably shorter lifetime than the entire population, and the shopping scams have a slightly longer lifetime. Over 40% of the malicious scams persist for less than 120 hours, whereas the lifetime for the same percentage of shopping scams was 180 hours and the median for all scams was 155 hours. These results are consistent with malicious scam sites being identified and taken down faster than other scam sites, although we cannot verify the causality.

As further evidence, we also examined the prevalence of malicious scams on the DNS blacklists we use in Section 5.3.2, and compare it to the blacklisting prevalence of all scams and the shopping scams. Over 28% of the malicious scams were blacklisted, roughly twice as often as the shopping scams (12% blacklisted) and all scams (15%). Again, these results are consistent with the lifetimes of malicious scams — being blacklisted twice as frequently could directly result in shorter scam lifetimes.

5.4.2 Spam campaign lifetime

A related aspect to scam lifetime are the "spam campaigns" used to advertise scams and attract clients. We captured 319,700 spam emails with links in our trace, resulting in 2,334 scams; on average, then, each scam was advertised by 137 spam emails. We use these repeated spam emails to determine the lifetime of spam campaigns for a scam by measuring the time between the first and last spam email messages advertising that scam. Figure 11 shows the distribution of the spam campaign lifetimes. Compared to the lifetime of scam sites, most spam campaigns are relatively short. Over 50% of the campaigns last less than 12 hours, over 90% last less than 48 hours, and 99% last less than three days. Roughly speaking, the lifecycle of a typical scam starts with a short spam campaign lasting half of a day while the scam site remains up for at least a week.

The relative lifetimes of spam campaigns and scam hosts again reflect the different needs of the two services. Compared with scam hosts, spam relays need to be active for much shorter periods of time to accomplish their goals. Spammers need only a window of time to distribute spam globally; once sent, spam relays are no longer needed for that particular scam. Scam hosts, in contrast, need to be responsive and available for longer periods of time to net potential clients. Put another way, spam is blanket advertising that requires no interaction with users to deliver, whereas scam hosting is a service that fundamentally depends upon user interaction to be successful. In contrast, scam hosts benefit more from stable infrastructure that remains useful and available for much longer periods of time.

5.5 Stability

A profitable scam requires stable infrastructure to serve potential customers at any time, and for as long as the scam is active. To gauge the stability of scam hosting infrastructure, we probed each scam host periodically for a week to measure its availability. When downloading pages from the hosts, we also used p0f to fingerprint host operating systems and link connectivity.

We computed scam availability as the number of successful Web page downloads divided by the total number

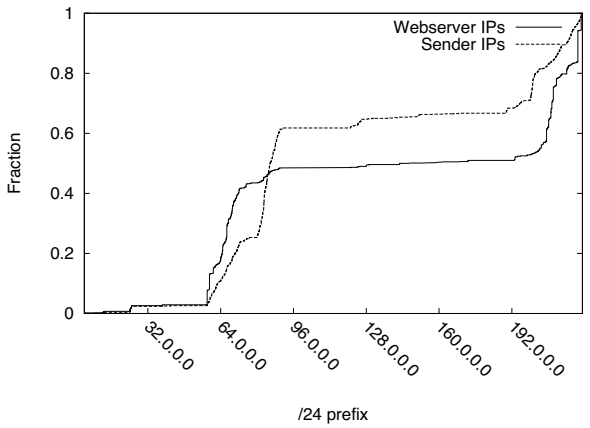

Figure 12: IP addresses, binned by /24 prefix, for spam sending relays and scam host servers.

of download attempts within the overall lifetime of the scam; if a scam lasted for only three days, we computed availability only during those days. Scams had excellent availability: over 90% of scams had an availability of 99% or higher. Of the remaining, most had availabilities of 98% or higher. As fingerprinted by p0f, more scams ran on Unix or server appliances (43%) than Windows systems (30%), and all of them had reported good link connectivity. These results indicate that scam hosting is quite reliable within the lifetime of a scam.

5.6 Scam location

We next examine both the network and geographic locations of scam hosts. For comparison, we also examine the locations of the spam relays that sent the spam in our trace. Comparing them highlights the extent to which the different requirements of the two services reflect where around the world and in the network they are found.

5.6.1 Network location

The network locations of spam relays and scam hosts are more consistent. Figure 12 shows the cumulative distribution of IP addresses for spam relays and scam hosts in our trace. Consistent with a similar analysis of spam relays in [27], the distributions are highly non-uniform. The IP addresses of most spam relays and scam hosts fall into the two same ranges, 58.* to 91.* and 200.* to 222.*. However, within those two address ranges hosts for the two services have different concentrations. The majority of spam relays (over 60%) fall into the first address range and are distributed somewhat evenly except for a gap between 70.* and 80.*. Roughly half of the scam hosts also fall into the first address range, but most of those fall into the 64.* to 72.* subrange and relatively few in the second half of the range. Similarly, scams are more uniformly distributed within the second address range as well.

Scam host country	*% of all servers*
United States	57.40%
China	7.23%
Canada	3.70%
Great Britain	3.07%
France	3.06%
Germany	2.52%
Russia	1.80%
South Korea	1.77%
Japan	1.60%
Taiwan	1.53%
Other	16.32%

Table 4: Countries of scam hosts.

Spam relay country	*% of all relays*
United States	14.50%
France	7.06%
Spain	6.75%
China	6.65%
Poland	5.68%
India	5.42%
Germany	5.00%
South Korea	4.67%
Italy	4.44%
Brazil	3.86%
Other	30.97%

Table 5: Countries of spam relays.

5.6.2 Geographic location

How do these variations in network address concentrations map into geographic locations? The effectiveness of scams could relate to (at least perceived) geographic location. As one anecdote, online pharmaceutical vendors utilized hosting servers inside the United States to imply to their customers that they were providing a lawful service [24].

Using Digital Element's NetAcuity tool [10], we mapped the IP addresses of scam hosts to latitude and longitude coordinates. Using these coordinates, we then identified the country in which the host was geographically located. Table 4 shows the top ten countries containing scam hosts in our trace. Interestingly, the NetAcuity service reported that nearly 60% of the scam hosts are located in the United States. Overall, 14% were located in Western Europe and 13% in Asia. For compar-

ison, Table 5 shows the top ten countries containing spam relays. The geographic distributions for spam relays are quite different than scam hosts. Only 14% of spam relays are located in the United States, whereas 28% are located in Western Europe and 16% in Asia. We also found the top ASes for scam hosts and senders, but found no discernible pattern and omit the results for brevity.

The strong bias of locating scam hosts in the United States suggests that geographic location is more important to scammers than spammers. There are a number of possible reasons for this bias. One is the issue of perceived enhanced credibility by scammers mentioned above. Another relates to the difference in requirements for the two types of services. As discussed in Section 5.4.2, spam relays can take advantage of hosts with much shorter lifetimes than scam hosts. As a result, spam relays are perhaps more naturally suited to being hosted on compromised machines such as botnets; the compromised machine need only be under control of the spammer long enough to launch the spam campaign. Scam hosts benefit more from stability, and hosts and networks within the United States can provide this stability.

6 Conclusion

This paper does not study spam itself, nor the infrastructure used to deliver spam, but rather focuses on the scam infrastructure that is nourished by spam. We demonstrate the *spamscatter* technique for identifying scam infrastructure and how to use approximate image comparison to cluster servers according to individual scams — sidestepping the extensive content and networking camouflaging used by spammers.

From a week-long trace of a large real-time spam feed (roughly 150,000 per day), we used the spamscatter technique to identify and analyze over 2,000 distinct scams hosted across more than 7,000 distinct servers. We found that, although large numbers of hosts are used to advertise Internet scams using spam campaigns, individual scams themselves are typically hosted on only one machine. Further, individual machines are commonly used to host multiple scams, and occasionally serve as spam relays as well. This practice provides a potentially convenient single point for network-based interdiction either via IP blacklisting or network filtering.

The lifecycle of a typical scam starts with a short spam campaign lasting half of a day while the scam site remains up for at least a week. The relative lifetimes of spam campaigns and scam hosts reflect the different requirements of the two underground services. Spam is blanket advertising that requires no interaction with users to deliver, whereas scam hosting is a service that fundamentally depends upon user interaction to be successful. Finally, mapping the geographic locations of scam hosts, we found that they have a strong bias to being located in the United States. The strong bias suggests that geographic location is more important to scammers than spammers, perhaps due to the stability of hosts and networks within the U.S.

Acknowledgments

We would like to thank a number of people who made contributions to this project. We are particularly grateful to Weidong Cui and Christian Kreibich, who maintained the spam feed we used for our analyses, the anonymous party who gave us access to the spam feed itself, and Vern Paxson for discussions and feedback. Kirill Levchenko suggested image-based comparison of Web pages as an equivalence test, and Colleen Shannon assisted us with Digital Element's NetAcuity tool. Finally, we would like to also thank the the anonymous reviewers for their comments, the CCIED group for useful feedback on the project, and Chris X. Edwards for system support. Support for this work was provided in part by NSF under CyberTrust Grant No. CNS-0433668 and AFOSR MURI Contract F49620-02-1-0233.

References

[1] ANTI-PHISHING WORKING GROUP. Report Phishing. `http://www.antiphishing.org/`.

[2] BARFORD, P., BESTAVROS, A., BYERS, J., AND CROVELLA, M. On the marginal utility of network topology measurements. In *In Proceedings of ACM SIGCOMM Internet Measurement Workshop* (Oct. 2001).

[3] BÖHME, R., AND HOLZ, T. The effect of stock spam on financial markets. In *Proceeedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006)* (June 2006).

[4] BRODER, A. Z. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences (SEQUENCES'97)* (June 1997), pp. 21–29.

[5] CASADO, M., GARFINKEL, T., CUI, W., PAXSON, V., AND SAVAGE, S. Opportunistic measurement: Extracting insight from spurious traffic. In *Proceedings of the 4th ACM Workshop on Hot Topics in Networks (HotNets-IV)* (College Park, MD, Nov. 2005).

[6] CASTLECOPS. Fried Phish: Phishing Incident Reporting and Termination (PIRT). `http://www.castlecops.com/pirt`.

[7] CHOU, N., LEDESMA, R., TERAGUCHI, Y., BONEH, D., AND MITCHELL, J. C. Client-side defense against Web-based identity theft. In *Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS '04)* (Feb. 2004).

[8] COOKE, E., BAILEY, M., MAO, Z. M., WATSON, D., JAHANIAN, F., AND MCPHERSON, D. Toward understanding distributed blackhole placement. In *Workshop on Rapid Malcode (WORM'04)* (Oct. 2004).

[9] COOKE, E., JAHANIAN, F., AND MCPHERSON, D. The zombie roundup: Understanding, detecting, and disrupting botnets. In *Proceedings of the First Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI'05)* (July 2005).

[10] DIGITAL ELEMENT. NetAcuity IP Intelligence. http://www.digital-element.net/ip_intelligence/ip_intelligence.html/.

[11] FETTERLY, D., MANASSE, M., AND NAJORK, M. On the evolution of clusters of near-duplicate Web pages. In *Proceedings of the First Latin American Web Congress* (Nov. 2003), pp. 37–45.

[12] GILLIS, T. Internet Security Trends for 2007. Ironport Whitepaper, 2007.

[13] IRONPORT INC. Spammers continue innovation. IronPort press release, June 28, 2006. http://www.ironport.com/company/ironport_pr_2006-06-28.html.

[14] KDE. Khtml layout engine. http://www.kde.org/.

[15] KEIZER, G. Spam volume jumps 35% in November, Dec. 2006. http://informationweek.com/news/showArticle.jhtml;?articleID=196701527.

[16] MATHER, T. Net::DNSBLLookup Perl module. http://search.cpan.org/~tjmather/Net-DNSBLLookup-0.03/.

[17] MESSAGELABS. 2006: The year spam raised its game and threats got personal, Dec. 2006. http://www.messagelabs.com/publishedcontent/publish/about_us_dotcom_en/%news___events/press_releases/DA_174397.html.

[18] MONGA, V., AND EVANS, B. L. Robust perceptual image hashing using feature points. In *Procedings of the IEEE International Conference on Image Processing (ICIP'04)* (Oct. 2004), pp. 677–680.

[19] MOORE, D., PAXSON, V., SAVAGE, S., SHANNON, C., STANIFORD, S., AND WEAVER, N. Inside the Slammer worm. *IEEE Security and Privacy 1*, 4 (July 2003), 33–39.

[20] MOORE, D., SHANNON, C., BROWN, D., VOELKER, G. M., AND SAVAGE, S. Inferring Internet denial-of-service activity. *ACM Transactions on Computer Systems 24*, 2 (May 2006), 115–139.

[21] MOORE, D., SHANNON, C., AND BROWN, J. Code-Red: a case study on the spread and victims of an Internet worm. In *Proceedings of the ACM/USENIX Internet Measurement Workshop (IMW)* (Marseille, France, Nov. 2002).

[22] NEWS.ADMIN.NET-ABUSE.SIGHTINGS. USENET newsgroup for discussion of spam. http://www.nanae.org/.

[23] PAUL BÄHER, THORSTEN HOLZ, MARKUS KÖTTER AND GEORG WICHERSKI. Know your enemy: Tracking botnets. In *The Honeynet Project & Research Alliance* (Mar. 2005).

[24] PHILADELPHIA INQUIRER. Special reports: Drugnet. http://www.philly.com/mld/inquirer/news/special_packages/pill/.

[25] PHISHTANK. Join the fight against phishing. http://www.phishtank.com/.

[26] RAJAB, M. A., ZARFOSS, J., MONROSE, F., AND TERZIS, A. A multifaceted approach to understanding the botnet phenomenon. In *Proceedings of the ACM Internet Measurement Conference* (Rio de Janeiro, Brazil, Oct. 2006).

[27] RAMACHANDRAN, A., AND FEAMSTER, N. Understanding the network-level behavior of spammers. In *Proceedings of the ACM SIGCOMM Conference* (Pisa, Italy, Sept. 2006).

[28] ROOVER, C. D., VLEESCHOUWER, C. D., LEFEBVRE, F., AND MACQ, B. Robust image hashing based on radial variance of pixels. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'05)* (Sept. 2005), pp. 77–80.

[29] SHIVAKUMAR, N., AND GARCIA-MOLINA, H. Finding near-replicas of documents and servers on the Web. In *Proceedings of the First International Workshop on the Web and Databases (WebDB'98)* (Mar. 1998).

[30] VENKATESAN, R., KOON, S. M., JAKUBOWSKI, M. H., AND MOULIN, P. Robust image hashing. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'00)* (Sept. 2000).

[31] WEBB, S., CAVERLEE, J., AND PU, C. Introducing the Webb spam corpus: Using email spam to identify Web spam automatically. In *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS)* (Mountain View, 2006).

[32] YEGNESWARAN, V., BARFORD, P., AND PLONKA, D. On the design and use of Internet sinks for network abuse monitoring. In *Proceedings of Recent Advances on Intrusion Detection* (Sept. 2004).