Click Trajectories: End-to-End Analysis of the Spam Value Chain

Kirill Levchenko[*] Andreas Pitsillidis[*] Neha Chachra[*] Brandon Enright[*] Márk Félegyházi[‡] Chris Grier[†]
Tristan Halvorson[*] Chris Kanich[*] Christian Kreibich[†◇] He Liu[*] Damon McCoy[*]
Nicholas Weaver[†◇] Vern Paxson[†◇] Geoffrey M. Voelker[*] Stefan Savage[*]

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

[†]*Computer Science Division*
University of California, Berkeley

[◇]*International Computer Science Institute*
Berkeley, CA

[‡]*Laboratory of Cryptography and System Security (CrySyS)*
Budapest University of Technology and Economics

***Abstract*—Spam-based advertising is a business. While it has engendered both widespread antipathy and a multi-billion dollar anti-spam industry, it continues to exist because it fuels a profitable enterprise. We lack, however, a solid understanding of this enterprise's full structure, and thus most anti-spam interventions focus on only one facet of the overall spam value chain (e.g., spam filtering, URL blacklisting, site takedown). In this paper we present a holistic analysis that quantifies the full set of resources employed to monetize spam email—including naming, hosting, payment and fulfillment—using extensive measurements of three months of diverse spam data, broad crawling of naming and hosting infrastructures, and over 100 purchases from spam-advertised sites. We relate these resources to the organizations who administer them and then use this data to characterize the relative prospects for defensive interventions at each link in the spam value chain. In particular, we provide the first strong evidence of payment bottlenecks in the spam value chain; 95% of spam-advertised pharmaceutical, replica and software products are monetized using merchant services from just a handful of banks.**

I. Introduction

We may think of email spam as a scourge—jamming our collective inboxes with tens of billions of unwanted messages each day—but to its perpetrators it is a potent marketing channel that taps latent demand for a variety of products and services. While most attention focuses on the problem of spam *delivery*, the email vector itself comprises only the *visible* portion of a large, multi-faceted business enterprise. Each click on a spam-advertised link is in fact just the start of a long and complex trajectory, spanning a range of both technical and business components that together provide the necessary infrastructure needed to monetize a customer's visit. Botnet services must be secured, domains registered, name servers provisioned, and hosting or proxy services acquired. All of these, in addition to payment processing, merchant bank accounts, customer service, and fulfillment, reflect necessary elements in the spam value chain.

While elements of this chain have received study in isolation (e.g., dynamics of botnets [20], DNS fast-flux networks [17], [42], Web site hosting [1], [22]), the relationship between them is far less well understood. Yet it is these very relationships that capture the structural dependencies—and hence the potential *weaknesses*—within the spam ecosystem's business processes. Indeed, each distinct *path* through this chain—registrar, name server, hosting, affiliate program, payment processing, fulfillment—directly reflects an "entrepreneurial activity" by which the perpetrators muster capital investments and business relationships to create value. Today we lack insight into even the most basic characteristics of this activity. How many organizations are complicit in the spam ecosystem? Which points in their value chains do they share and which operate independently? How "wide" is the bottleneck at each stage of the value chain—do miscreants find alternatives plentiful and cheap, or scarce, requiring careful husbanding?

The desire to address these kinds of questions empirically—and thus guide decisions about the most effective mechanisms for addressing the spam problem—forms the core motivation of our work. In this paper we develop a methodology for characterizing the end-to-end resource dependencies ("trajectories") behind individual spam campaigns and then analyze the relationships among them. We use three months of real-time source data, including captive botnets, raw spam feeds, and feeds of spam-advertised URLs to drive active probing of spam infrastructure elements (name servers, redirectors, hosting proxies). From these, we in turn identify those sites advertising three popular classes of goods—pharmaceuticals, replica luxury goods and counterfeit software—as well as their membership in specific affiliate programs around which the overall business is structured. Finally, for a subset of these sites we perform on-line purchases, providing additional data about merchant bank affiliation, customer service, and fulfillment. Using this data we characterize the resource footprint at each step in the spam value chain, the extent of sharing between spam organizations and, most importantly, the relative prospects for interrupting spam monetization at different stages of the process.

The remainder of this paper is organized as follows. Section II provides a qualitative overview of the spam ecosystem coupled with a review of related research.

Section III describes the data sources, measurement techniques and post-processing methodology used in our study. Section IV describes our analysis of spam activities between August and October of 2010, and the implications of these findings on the likely efficacy of different anti-spam interventions, followed by our conclusions in Section V.

II. Background and Related Work

As an advertising medium, spam ultimately shares the underlying business model of all advertising. So long as the revenue driven by spam campaigns exceeds their cost, spam remains a profitable enterprise. This glib description belies the complexity of the modern spam business. While a decade ago spammers might have handled virtually all aspects of the business including email distribution, site design, hosting, payment processing, fulfillment, and customer service [33], today's spam business involves a range of players and service providers. In this section, we review the broad elements in the spam value chain, the ways in which these components have adapted to adversarial pressure from the anti-spam community, and the prior research on applied e-crime economics that informs our study.

A. How Modern Spam Works

While the user experience of spam revolves principally around the email received, these constitute just one part of a larger value chain that we classify into three distinct stages: *advertising*, *click support*, and *realization*. Our discussion here reflects the modern understanding of the degree to which specialization and affiliate programs dominate the use of spam to sell products. To this end, we draw upon and expand the narrative of the "Behind Online Pharma" project [4], which documents the experience of a group of investigative journalists in exploring the market structure for online illegal pharmaceuticals; and Samosseiko's recent overview [46] of affiliate programs, including many that we discuss in this paper.

Advertising. Advertising constitutes all activities focused on reaching potential customers and enticing them into clicking on a particular URL. In this paper we focus on the email spam vector, but the same business model occurs for a range of advertising vectors, including blog spam [39], Twitter spam [12], search engine optimization [53], and sponsored advertising [26], [27]. The delivery of email spam has evolved considerably over the years, largely in response to increasingly complex defensive countermeasures. In particular, large-scale efforts to shut down open SMTP proxies and the introduction of well-distributed IP blacklisting of spam senders have pushed spammers to using more sophisticated delivery vehicles. These include botnets [13], [20], [56], Webmail spam [9], and IP prefix hijacking [45]. Moreover, the market for spam services has stratified over time; for example, today it is common for botnet operators to rent their services to spammers on a contract basis [40].

The advertising side of the spam ecosystem has by far seen the most study, no doubt because it reflects the part of spam that users directly experience. Thus, a broad and ongoing literature examines filtering spam email based on a variety of content features (e.g., [2], [19], [43], [57]). Similarly, the network characteristics of spam senders have seen extensive study for characterizing botnet membership [58], identifying prefix hijacking [45], classifying domains and URLs [14], [32], [44], [55], [56], and evaluating blacklists [47], [48]. Finally, we note that most commercial anti-spam offerings focus exclusively on the delivery aspect of spam. In spite of this attention, spam continues to be delivered and thus our paper focuses strictly on the remaining two stages of the spam monetization pipeline.

Click support. Having delivered their advertisement, a spammer depends on some fraction of the recipients to respond, usually by clicking on an embedded URL and thus directing their browser to a Web site of interest. While this process seems simple, in practice a spammer must orchestrate a great many moving parts and maintain them against pressure from defenders.

Redirection sites. Some spammers directly advertise a URL such that, once the recipient's browser resolves the domain and fetches the content from it, these steps constitute the fullness of the promoted Web site. However, a variety of defensive measures—including URL and domain blacklisting, as well as site takedowns by ISPs and domain takedowns by registrars—have spurred more elaborate steps. Thus, many spammers advertise URLs that, when visited, redirect to additional URLs [1], [22]. Redirection strategies primarily fall into two categories: those for which a legitimate third party inadvertently controls the DNS name resource for the redirection site (e.g., free hosting, URL shorteners, or compromised Web sites), and those for which the spammers themselves, or perhaps parties working on their behalf, manage the DNS name resources (e.g., a "throwaway" domain such as `minesweet.ru` redirecting to a more persistent domain such as `greatjoywatches.com`).

Domains. At some point, a click trajectory will usually require domain name resources managed by the spammer or their accomplices. These names necessarily come via the services of a domain registrar, who arranges for the root-level registry of the associated top-level domain (TLD) to hold NS records for the associated registered domain. A spammer may purchase domains directly from a registrar, but will frequently purchase instead from a domain reseller, from a "domaineer" who purchases domains in bulk via multiple sources and sells to the underground trade, or directly from a spam "affiliate program" that makes domains available to their affiliates as part of their "startup package."

Interventions at this layer of the spam value chain depend significantly on the responsiveness of individual registrars and the pressure brought to bear [29]. For example, a recent industry study by LegitScript and KnujOn documents heavy

concentration of spam-advertised pharmacies with domains registered through a particular set of registrars who appear indifferent to complaints [28].

Name servers. Any registered domain must in turn have supporting name server infrastructure. Thus spammers must provision this infrastructure either by hosting DNS name servers themselves, or by contracting with a third party. Since such resources are vulnerable to takedown requests, a thriving market has arisen in so-called "bulletproof" hosting services that resist such requests in exchange for a payment premium [23].

Web servers. The address records provided by the spammer's name servers must in turn specify servers that host (or more commonly proxy) Web site content. As with name servers, spam-advertised Web servers can make use of bulletproof hosting to resist takedown pressure [3], [51]. Some recent interventions have focused on effectively shutting down such sites by pressuring their upstream Internet service providers to deny them transit connectivity [6].

To further complicate such takedowns and to stymie blacklisting approaches, many spammers further obfuscate the hosting relationship (both for name servers and Web servers) using fast-flux DNS [17], [41], [42]. In this approach, domain records have short-lived associations with IP addresses, and the mapping infrastructure can spread the domain's presence over a large number of machines (frequently many thousands of compromised hosts that in turn proxy requests back to the actual content server [5]). Furthermore, recently innovators have begun packaging this capability to offer it to third parties on a contract basis as a highly resilient content-hosting service [7].

Stores and Affiliate Programs. Today, spammers operate primarily as advertisers, rarely handling the back end of the value chain. Such spammers often work as affiliates of an online store, earning a commission (typically 30–50%) on the sales they bring in [46]. The affiliate program typically provides the storefront templates, shopping cart management, analytics support, and even advertising materials. In addition, the program provides a centralized Web service interface for affiliates to track visitor conversions and to register for payouts (via online financial instruments such as WebMoney). Finally, affiliate programs take responsibility for contracting for payment and fulfillment services with outside parties. Affiliate programs have proven difficult to combat directly—although, when armed with sufficient legal jurisdiction, law enforcement has successfully shut down some programs [8].

Realization. Finally, having brought the customer to an advertised site and convinced them to purchase some product, the seller realizes the latent value by acquiring the customer's payment through conventional payment networks, and in turn fulfilling their product request.

Payment services. To extract value from the broadest possible customer base, stores try to support standard credit card payments. A credit card transaction involves several parties in addition to the customer and merchant: money is transferred from the *issuing bank* (the customer's bank) to the *acquiring bank* (the bank of the merchant) via a *card association network* (i.e., Visa or MasterCard). In addition to the acquiring bank, issuing bank, and card association, the merchant frequently employs the services of a *payment processor* to facilitate this process and act as the technical interface between the merchant and the payment system.

Card associations impose contractual restrictions on their member banks and processors, including the threat of fines and de-association; but to our knowledge little public documentation exists about the extent to which the associations apply this pressure in practice nor the extent to which it plays an important role in moderating the spam business. Evidence from this study suggests that any such pressure is currently insufficient to stop this activity.

Fulfillment. Finally, a store arranges to fulfill an order[1] in return for the customer's payment. For physical goods such as pharmaceuticals and replica products, this involves acquiring the items and shipping them to the customer. Global business-to-business Web sites such as Alibaba, ECPlaza, and ECTrade offer connections with a broad variety of vendors selling a range of such goods, including prepackaged drugs—both brand (e.g., Viagra) and off-brand (e.g., sildenafil citrate capsules)—and replica luxury goods (e.g., Rolex watches or Gucci handbags). Generally, suppliers will offer direct shipping service ("drop shipping"), so affiliate programs can structure themselves around "just in time" fulfillment and avoid the overhead and risk of warehousing and shipping the product themselves.[2] Fulfillment for virtual goods such as software, music, and videos can proceed directly via Internet download.

B. Pharmacy Express: An Example

Figure 1 illustrates the spam value chain via a concrete example from the empirical data used in this study.

On October 27th, the Grum botnet delivered an email titled *VIAGRA® Official Site* (❶). The body of the message includes an image of male enhancement pharmaceutical tablets and their associated prices (shown). The image provides a URL tag and thus when clicked (❷) directs the user's browser to resolve the associated domain name, `medicshopnerx.ru`. This domain was registered by REGRU-REG-RIPN (a.k.a. `reg.ru`) on October 18th (❸)—it is still active as of this writing. The machine providing name service resides in China, while hosting resolves to a

[1] In principle, a store could fail to fulfill a customer's order upon receiving their payment, but this would both curtail any repeat orders and would lead to chargebacks through the payment card network, jeopardizing their relationship with payment service providers.

[2] Individual suppliers can differ in product availability, product quality, the ability to manage the customs process, and deliver goods on a timely basis. Consequently, affiliate programs may use different suppliers for different products and destinations.

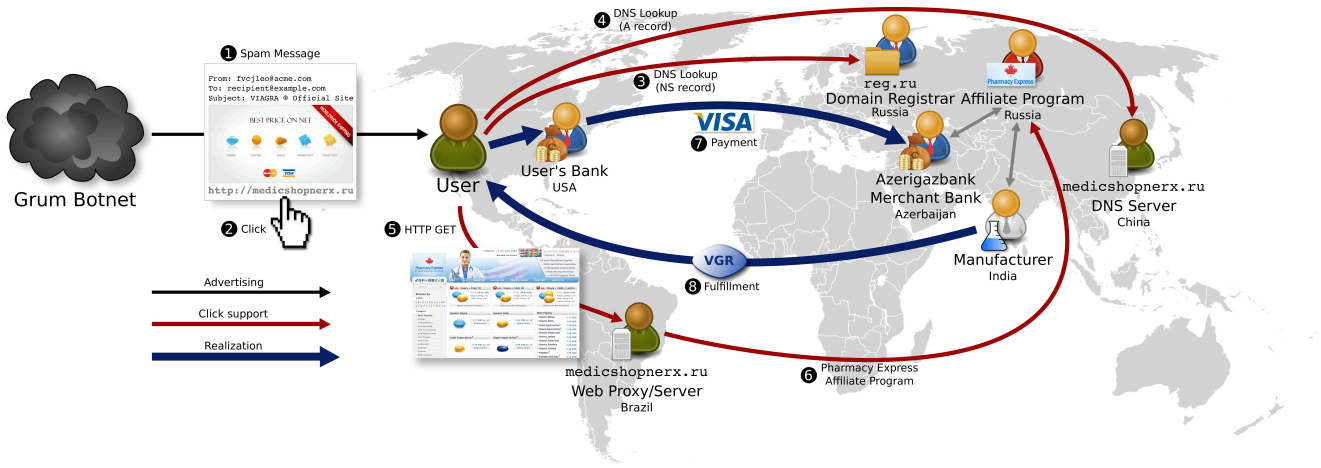

Figure 1: Infrastructure involved in a single URL's value chain, including advertisement, click support and realization steps.

machine in Brazil (❹). The user's browser initiates an HTTP request to the machine (❺), and receives content that renders the storefront for "Pharmacy Express," a brand associated with the Mailien pharmaceutical affiliate program based in Russia (❻).

After selecting an item to purchase and clicking on "Checkout", the storefront redirects the user to a payment portal served from `payquickonline.com` (this time serving content via an IP address in Turkey), which accepts the user's shipping, email contact, and payment information, and provides an order confirmation number. Subsequent email confirms the order, provides an *EMS* tracking number, and includes a contact email for customer questions. The bank that issued the user's credit card transfers money to the acquiring bank, in this case the Azerigazbank Joint-Stock Investment Bank in Baku, Azerbaijan (BIN 404610, ❼). Ten days later the product arrives, blister-packaged, in a cushioned white envelope with postal markings indicating a supplier named PPW based in Chennai, India as its originator (❽).

C. Cybercrime economics

Alongside the myriad studies of the various components employed in spam (e.g., botnets, fast flux, etc.), a literature has recently emerged that focuses on using economic tools for understanding cybercrime (including spam) in a more systematic fashion, with an aim towards enabling better reasoning about effective interventions. Here we highlight elements of this work that have influenced our study.

Some of the earliest such work has aimed to understand the scope of underground markets based on the value of found goods (typically stolen financial credentials), either as seen on IRC chatrooms [10], forums [59], malware "drop-zones" [16], or directly by intercepting communications to botnet C&C servers [50]. Herley and Florêncio critique this line of work as not distinguishing between claimed and true losses, and speculate that such environments inherently reflect "lemon markets" in which few participants are likely to acquire significant profits (particularly spammers) [15]. While this hypothesis remains untested, its outcome is orthogonal to our focus of understanding the structure of the value chain itself.

Our own previous work on spam conversion also used empirical means to infer parts of the return-on-investment picture in the spam business model [21]. By contrast, this study aims to be considerably more comprehensive in breadth (covering what we believe reflect most large spam campaigns) and depth (covering the fullness of the value chain), but offering less precision regarding specific costs.

Finally, another line of work has examined interventions from an economic basis, considering the efficacy of site and domain takedown in creating an economic impediment for cybercrime enterprises (notably phishing) [6], [35], [36]. Molnar *et al.* further develop this approach via comparisons with research on the illicit drug ecosystem [34]. Our work builds on this, but focuses deeply on the spam problem in particular.

III. Data Collection Methodology

In this section we describe our datasets and the methodology by which we collected, processed, and validated them. Figure 2 concisely summarizes our data sources and methods. We start with a variety of full-message spam feeds, URL feeds, and our own botnet-harvested spam (❶). Feed parsers extract embedded URLs from the raw feed data for further processing (❷). A DNS crawler enumerates various resource record sets of the URL's domain, while a farm of Web crawlers visits the URLs and records HTTP-level interactions and landing pages (❸). A clustering tool clusters pages by content similarity (❹). A content tagger labels the content clusters according to the category of goods sold, and the associated affiliate programs (❺). We then make targeted purchases from each affiliate program (❻), and store the feed data and distilled and derived metadata in a database

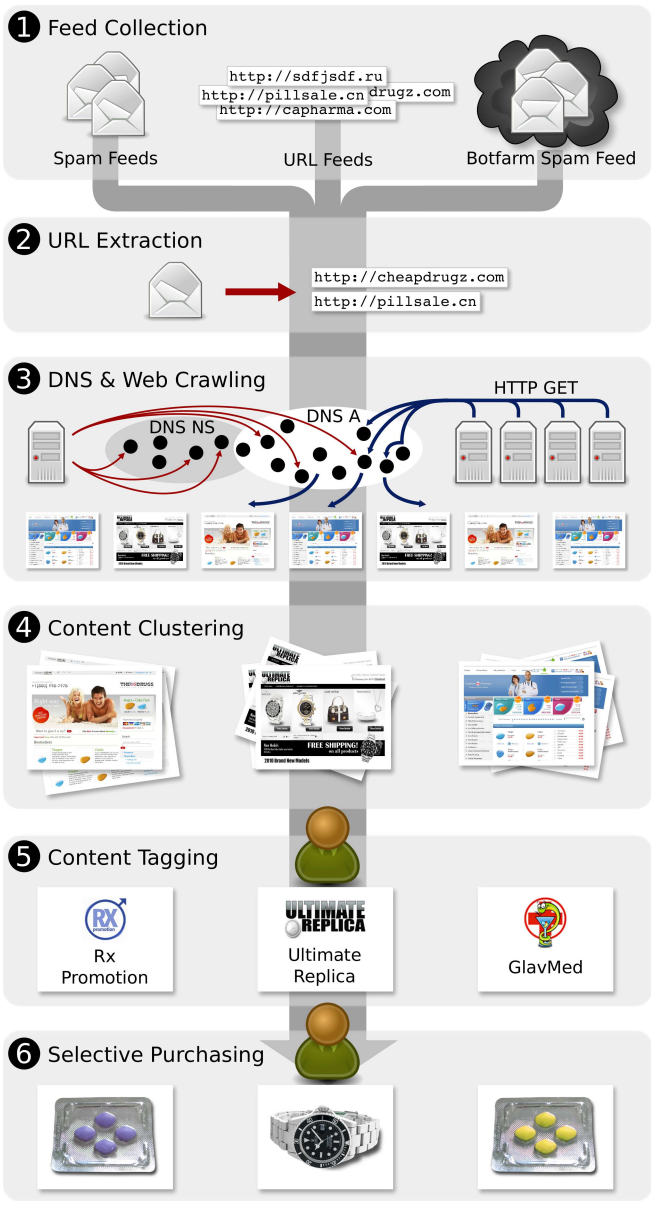

Figure 2: Our data collection and processing workflow.

for subsequent analysis in Section IV. (Steps ❺ and ❻ are partially manual operations, the others are fully automated.)

The rest of this section describes these steps in detail.

A. Collecting Spam-Advertised URLs

Our study is driven by a broad range of data sources of varying types, some of which are provided by third parties, while others we collect ourselves. Since the goal of this study is to decompose the spam ecosystem, it is natural that our seed data arises from spam email itself. More specifically, we focus on the URLs embedded within such email, since these are the vectors used to drive recipient traffic to particular Web sites. To support this goal, we obtained seven distinct URL feeds from third-party partners (including multiple commercial anti-spam providers), and harvested URLs from our own botfarm environment.

Feed Name	*Feed Description*	*Received URLs*	*Distinct Domains*
Feed A	MX honeypot	32,548,304	100,631
Feed B	Seeded honey accounts	73,614,895	35,506
Feed C	MX honeypot	451,603,575	1,315,292
Feed D	Seeded honey accounts	30,991,248	79,040
Feed X	MX honeypot	198,871,030	2,127,164
Feed Y	Human identified	10,733,231	1,051,211
Feed Z	MX honeypot	12,517,244	67,856
Cutwail	Bot	3,267,575	65
Grum	Bot	11,920,449	348
MegaD	Bot	1,221,253	4
Rustock	Bot	141,621,731	13,612,815
Other bots	Bot	7,768	4
Total		968,918,303	17,813,952

Table I: Feeds of spam-advertised URLs used in this study. We collected feed data from August 1, 2010 through October 31, 2010.

For this study, we used the data from these feeds from August 1, 2010 through October 31, 2010, which together comprised nearly 1 billion URLs. Table I summarizes our feed sources along with the "type" of each feed, the number of URLs received in the feed during this time period, and the number of distinct registered domains in those URLs. Note that the "bot" feeds tend to be focused spam sources, while the other feeds are spam sinks comprised of a blend of spam from a variety of sources. Further, individual feeds, particularly those gathered directly from botnets, can be heavily skewed in their makeup. For example, we received over 11M URLs from the Grum bot, but these only contained 348 distinct registered domains. Conversely, the 13M distinct domains produced by the Rustock bot are artifacts of a "blacklist-poisoning" campaign undertaken by the bot operators that comprised millions of "garbage" domains [54]. Thus, one must be mindful of these issues when analyzing such feed data in aggregate.

From these feeds we extract and normalize embedded URLs and insert them into a large multi-terabyte Postgres database. The resulting "feed tables" drive virtually all subsequent data gathering.

B. Crawler data

The URL feed data subsequently drives active crawling measurements that collect information about both the DNS infrastructure used to name the site being advertised and the Web hosting infrastructure that serves site content to visitors. We use distinct crawlers for each set of measurements.

DNS Crawler: We developed a DNS crawler to identify the name server infrastructure used to support spam-advertised domains, and the address records they specify for hosting those names. Under normal use of DNS this process would be straightforward, but in practice it is significantly

complicated by *fast flux* techniques employed to minimize central points of weakness. Similar to the work of [18], we query servers repeatedly to enumerate the set of domains collectively used for click support (Section II-A).

From each URL, we extract both the fully qualified domain name and the registered domain suffix (for example, if we see a domain `foo.bar.co.uk` we will extract both `foo.bar.co.uk` as well as `bar.co.uk`). We ignore URLs with IPv4 addresses (just 0.36% of URLs) or invalidly formatted domain names, as well as duplicate domains already queried within the last day.

The crawler then performs recursive queries on these domains. It identifies the domains that resolve successfully and their authoritative domains, and filters out unregistered domains and domains with unreachable name servers. To prevent fruitless domain enumeration, it also detects wildcard domains (`abc.example.com`, `def.example.com`, etc.) where all child domains resolve to the same IP address. In each case, the crawler exhaustively enumerates all A, NS, SOA, CNAME, MX, and TXT records linked to a particular domain.

The crawler periodically queries new records until it converges on a set of distinct results. It heuristically determines convergence using standard maximum likelihood methods to estimate when the probability of observing a new unique record has become small. For added assurance, after convergence the crawler continues to query domains daily looking for new records (ultimately timing out after a week if it discovers none).

Web Crawler: The Web crawler replicates the experience of a user clicking on the URLs derived from the spam feeds. It captures any application-level redirects (HTML, JavaScript, Flash), the DNS names and HTTP headers of any intermediate servers and the final server, and the page that is ultimately displayed—represented both by its DOM tree and as a screenshot from a browser. Although straightforward in theory, crawling spam URLs presents a number of practical challenges in terms of scale, robustness, and adversarial conditions.

For this study we crawled nearly 15 million URLs, of which we successfully visited and downloaded correct Web content for over 6 million (unreachable domains, blacklisting, etc., prevent successful crawling of many pages).[3] To manage this load, we replicate the crawler across a cluster of machines. Each crawler replica consists of a controller managing over 100 instances of Firefox 3.6.10 running in parallel. The controller connects to a custom Firefox extension to manage each browser instance, which incorporates the Screengrab! extension [38] to capture screen shots (used for manual investigations). The controller retrieves batches of URLs from the database, and assigns URLs to Firefox instances in a round-robin fashion across a diverse set of IP address ranges.[4]

[3]By comparison, the spam hosting studies of Anderson *et al.* and Konte *et al.* analyzed 150,000 messages per day and 115,000 messages per month respectively [1], [22].

Stage	*Count*	
Received URLs	968,918,303	
Distinct URLs	93,185,779	(9.6%)
Distinct domains	17,813,952	
Distinct domains crawled	3,495,627	
URLs covered	950,716,776	(98.1%)

Table II: Summary results of URL crawling. We crawl the registered domains used by over 98% of the URLs received.

Table II summarizes our crawling efforts. Since there is substantial redundancy in the feeds (e.g., fewer than 10% of the URLs are even unique), crawling every URL is unnecessary and resource inefficient. Instead, we focus on crawling URLs that cover the set of registered domains used by all URLs in the feed. Except in rare instances, all URLs to a registered domain are for the same affiliate program. Thus, the crawler prioritizes URLs with previously unseen registered domains, ignores any URLs crawled previously, and rate limits crawling URLs containing the same registered domain—both to deal with feed skew as well as to prevent the crawler from being blacklisted. For timeliness, the crawler visits URLs within 30 minutes of appearing in the feeds.

We achieve nearly complete coverage: Over 98% of the URLs received in the raw feeds use registered domains that we crawl. Note that we obtain this coverage even though we crawled URLs that account for only 20% of the nearly 18 million distinct registered domains in the feeds. This outcome reflects the inherent skew in the feed makeup. The vast majority of the remaining 80% of domains we did not crawl, and the corresponding 2% URLs that use those domains, are from the domain-poisoning spam sent by the Rustock bot and do not reflect real sites (Section III-A).

C. Content Clustering and Tagging

The crawlers provide low-level information about URLs and domains. In the next stage of our methodology, we process the crawler output to associate this information with higher-level spam business activities.

Note that in this study we exclusively focus on businesses selling three categories of spam-advertised products: pharmaceuticals, replicas, and software. We chose these categories because they are reportedly among the most popular goods advertised in spam [31]—an observation borne out in our data as well.[5]

[4]Among the complexities, scammers are aware that security companies crawl them and blacklist IP addresses they suspect are crawlers. We mitigate this effect by tunneling requests through proxies running in multiple disparate IP address ranges.

[5]We did not consider two other popular categories (pornography and gambling) for institutional and procedural reasons.

Stage	*Pharmacy*	*Software*	*Replicas*	*Total*
URLs	346,993,046	3,071,828	15,330,404	365,395,278
Domains	54,220	7,252	7,530	69,002
Web clusters	968	51	20	1,039
Programs	30	5	10	45

Table III: Breakdown of clustering and tagging results.

To classify each Web site, we use *content clustering* to match sites with lexically similar content structure, *category tagging* to label clustered sites with the category of goods they sell, and *program tagging* to label clusters with their specific affiliate program and/or storefront brand. We use a combination of automated and manual analysis techniques to make clustering and tagging feasible for our large datasets, while still being able to manageably validate our results.

Table III summarizes the results of this process. It lists the number of received URLs with registered domains used by the affiliate programs we study, the number of registered domains in those URLs, the number of clusters formed based on the contents of storefront Web pages, and the number of affiliate programs that we identify from the clusters. As expected, pharmaceutical affiliate programs dominate the data set, followed by replicas and then software. We identify a total of 45 affiliate programs for the three categories combined, that are advertised via 69,002 distinct registered domains (contained within 38% of all URLs received in our feeds). We next describe the clustering and tagging process in more detail.

Content clustering: The first step in our process uses a clustering tool to group together Web pages that have very similar content. The tool uses the HTML text of the crawled Web pages as the basis for clustering. For each crawled Web page, it uses a q-gram similarity approach to generate a fingerprint consisting of a set of multiple independent hash values over all 4-byte tokens of the HTML text. After the crawler visits a page, the clustering tool computes the fingerprint of the page and compares it with the fingerprints representing existing clusters. If the page fingerprint exceeds a similarity threshold with a cluster fingerprint (equivalent to a Jaccard index of 0.75), it places the page in the cluster with the greatest similarity. Otherwise, it instantiates a new cluster with the page as its representative.

Category tagging: The clusters group together URLs and domains that map to the same page content. The next step of category tagging broadly separates these clusters into those selling goods that we are interested in, and those clusters that do not (e.g., domain parking, gambling, etc). We are intentionally conservative in this step, potentially including clusters that turn out to be false positives to ensure that we include all clusters that fall into one of our categories (thereby avoiding false negatives).

We identify interesting clusters using generic keywords found in the page content, and we label those clusters with *category tags*—"pharma", "replica", "software"—that correspond to the goods they are selling. The keywords consist of large sets of major brand names (Viagra, Rolex, Microsoft, etc.) as well as domain-specific terms (herbal, pharmacy, watches, software, etc.) that appear in the storefront page. These terms are tied to the content being sold by the storefront site, and are also used for search engine optimization (SEO). Any page containing a threshold of these terms is tagged with the corresponding keyword. The remaining URLs do not advertise products that we study and they are left untagged.

Even with our conservative approach, a concern is that our keyword matching heuristics might have missed a site of interest. Thus, for the remaining untagged clusters, we manually checked for such false negatives, i.e., whether there were clusters of storefront pages selling one of the three goods that should have a category tag, but did not. We examined the pages in the largest 675 untagged clusters (in terms of number of pages) as well as 1,000 randomly selected untagged clusters, which together correspond to 39% of the URLs we crawled. We did not find any clusters with storefronts that we missed.[6]

Program tagging: At this point, we focus entirely on clusters tagged with one of our three categories, and identify sets of distinct clusters that belong to the same affiliate program. In particular, we label clusters with specific *program tags* to associate them either with a certain affiliate program (e.g., EvaPharmacy—which in turn has many distinct storefront brands) or, when we cannot mechanically categorize the underlying program structure, with an individual storefront "brand" (e.g., Prestige Replicas). From insight gained by browsing underground forum discussions, examining the raw HTML for common implementation artifacts, and making product purchases, we found that some sets of the these brands are actually operated by the same affiliate program.

In total, we assigned program tags to 30 pharmaceutical, 5 software, and 10 replica programs that dominated the URLs in our feeds. Table IV enumerates these affiliate programs and brands, showing the number of distinct registered domains used by those programs, and the number of URLs that use those domains. We also show two aggregate programs, Mailien and ZedCash, whose storefront brands we associated manually based on evidence gathered on underground Web forums (later validated via the purchasing process).[7] The "feed volume" shows the distribution of the affiliate programs as observed in each of the spam "sink" feeds (the feeds *not* from bots), roughly approximating the

[6]The lack of false negatives is not too surprising. Missing storefronts would have no textual terms in their page content that relate to what they are selling (incidentally also preventing the use of SEO); this situation could occur if the storefront page were composed entirely of images, but such sites are rare.

[7]Note, ZedCash is unique among programs as it has storefront brands for each of the herbal, pharmaceutical and replica product categories.

Affiliate Program		*Distinct Domains*	*Received URLs*	*Feed Volume*
RxPrm	RX–Promotion	10,585	160,521,810	24.92%
Mailn	Mailien	14,444	69,961,207	23.49%
PhEx	Pharmacy Express	14,381	69,959,629	23.48%
EDEx	ED Express	63	1,578	0.01%
ZCashPh	ZedCash (Pharma)	6,976	42,282,943	14.54%
DrMax	Dr. Maxman	5,641	32,184,860	10.95%
Grow	Viagrow	382	5,210,668	1.68%
USHC	US HealthCare	167	3,196,538	1.31%
MaxGm	MaxGentleman	672	1,144,703	0.41%
VgREX	VigREX	39	426,873	0.14%
Stud	Stud Extreme	42	68,907	0.03%
ManXt	ManXtenz	33	50,394	0.02%
GlvMd	GlavMed	2,933	28,313,136	10.32%
OLPh	Online Pharmacy	2,894	17,226,271	5.16%
Eva	EvaPharmacy	11,281	12,795,646	8.7%
WldPh	World Pharmacy	691	10,412,850	3.55%
PHOL	PH Online	101	2,971,368	0.96%
Aptke	Swiss Apotheke	117	1,586,456	0.55%
HrbGr	HerbalGrowth	17	265,131	0.09%
RxPnr	RX Partners	449	229,257	0.21%
Stmul	Stimul-cash	50	157,537	0.07%
Maxx	MAXX Extend	23	104,201	0.04%
DrgRev	DrugRevenue	122	51,637	0.04%
UltPh	Ultimate Pharmacy	12	44,126	0.02%
Green	Greenline	1,766	25,021	0.36%
Vrlty	Virility	9	23,528	0.01%
RxRev	RX Rev Share	299	9,696	0.04%
Medi	MediTrust	24	6,156	0.01%
ClFr	Club-first	1,270	3,310	0.07%
CanPh	Canadian Pharmacy	133	1,392	0.03%
RxCsh	RXCash	22	287	<0.01%
Staln	Stallion	2	80	<0.01%
	Total	54,220	346,993,046	93.18%
Royal	Royal Software	572	2,291,571	0.79%
EuSft	EuroSoft	1,161	694,810	0.48%
ASR	Auth. Soft. Resellers	4,117	65,918	0.61%
OEM	OEM Soft Store	1,367	19,436	0.24%
SftSl	Soft Sales	35	93	<0.01%
	Total	7,252	3,071,828	2.12%
ZCashR	ZedCash (Replica)	6,984	13,243,513	4.56%
UltRp	Ultimate Replica	5,017	10,451,198	3.55%
Dstn	Distinction Replica	127	1,249,886	0.37%
Exqst	Exquisite Replicas	128	620,642	0.22%
DmdRp	Diamond Replicas	1,307	506,486	0.27%
Prge	Prestige Replicas	101	382,964	0.1%
OneRp	One Replica	77	20,313	0.02%
Luxry	Luxury Replica	25	8,279	0.01%
AffAc	Aff. Accessories	187	3,669	0.02%
SwsRp	Swiss Rep. & Co.	15	76	<0.01%
WchSh	WatchShop	546	2,086,891	0.17%
	Total	7,530	15,330,404	4.73%
	Grand Total	69,002	365,395,278	100%

Table IV: Breakdown of the pharmaceutical, software, and replica affiliate programs advertising in our URL feeds.

distribution that might be observed by users receiving spam.[8]

To assign these affiliate program tags to clusters, we manually crafted sets of regular expressions that match the page contents of program storefronts. For some programs, we defined expressions that capture the structural nature of the software engine used by all storefronts for a program (e.g., almost all EvaPharmacy sites contained unique hosting conventions). For other programs, we defined expressions that capture the operational modes used by programs that used multiple storefront templates (e.g., GlavMed).[9] For others, we created expressions for individual storefront brands (e.g., one for Diamond Replicas, another for Prestige Replicas, etc.), focusing on the top remaining clusters in terms of number of pages. Altogether, we assigned program tags to clusters comprising 86% of the pages that had category tags.

We manually validated the results of assigning these specific program tags as well. For every cluster with a program tag, we inspected the ten most and least common page DOMs contained in that cluster, and validated that our expressions had assigned them their correct program tags. Although not exhaustive, examining the most and least common pages validates the pages comprising both the "mass" and "tail" of the page distribution in the cluster.

Not all clusters with a category tag ("pharma") had a specific program tag ("EvaPharmacy"). Some clusters with category tags were false positives (they happened to have category keywords in the page, but were not storefronts selling category goods), or they were small clusters corresponding to storefronts with tiny spam footprints. We inspected the largest 675 of these clusters and verified that none of them contained pages that should have been tagged as a particular program in our study.

D. Purchasing

Finally, for a subset of the sites with program tags, we also purchased goods being offered for sale. We attempted to place multiple purchases from each major affiliate program or store "brand" in our study and, where possible, we ordered the same "types" of product from different sites to identify differences or similarities in suppliers based on contents (e.g., lot numbers) and packaging (nominal sender, packaging type, etc.). We attempted 120 purchases, of which 76 authorized and 56 settled.[10]

Of those that settled, all but seven products were delivered. We confirmed via tracking information that two undelivered packages were sent several weeks after our mailbox lease had ended, two additional transactions received no follow-up email, another two sent a follow-up email stating that the order was re-sent after the mailbox lease had ended,

[8]We remove botnet feeds from such volume calculations because their skewed domain mix would bias the results unfairly towards the programs they advertise.

[9]We obtained the full source code for all GlavMed and RX–Promotion sites, which aided creating and validating expressions to match their templates.

[10]Almost 50% of these failed orders were from ZedCash, where we suspect that our large order volume raised fraud concerns. In general, any such biases in the order completion rate do not impact upon our analysis, since our goal in purchasing is simply to establish the binding between individual programs and realization infrastructure; we obtained data from multiple transactions for each major program under study.

and one sent a follow-up email stating that our money had been refunded (this refund, however, had not been processed three months after the fact).

Operational protocol: We placed our purchases via VPN connections to IP addresses located in the geographic vicinity to the mailing addresses used. This constraint is necessary to avoid failing common fraud checks that evaluate consistency between IP-based geolocation, mailing address and the Address Verification Service (AVS) information provided through the payment card association. During each purchase, we logged the full contents of any checkout pages as well as their domain names and IP addresses (frequently different from the sites themselves). We provided contact email addresses hosted on domain names purchased expressly for this project, as several merchants did not allow popular Web-based email accounts during the purchase process. We recorded all email sent to these accounts, as well as the domain names and IP addresses of any customer service sites provided. We also periodically logged into such sites to record the current status of our purchases. For physical goods, we always selected the quickest form of delivery, while software was provided via the Internet (here too we recorded the full information about the sites used for software fulfillment).

All of our purchases were conducted using prepaid Visa payment cards contracted through a specialty issuer. As part of our relationship with the issuer, we maintained the ability to create new cards on demand and to obtain the authorization and settlement records for each transaction. We used a unique card for each transaction.

We had goods shipped to a combination of individual residences and a suite address provided by a local commercial mailbox provider. We regularly picked up, tagged, and photographed shipments and then stored them in a centralized secure facility on our premises. We stored software purchases on a secure hard drive, checked for viruses using Microsoft Security Essentials and Kaspersky Free Trial, and compared against other copies of the same software (including a reference version that we owned).

Legal and ethical concerns: This purchasing portion of our study involved the most careful consideration of legal and ethical concerns, particularly because this level of active involvement has not been common in the academic community to date. We worked with both our own project legal advisors and with general counsel to design a protocol for purchasing, handling, analyzing and disposing of these products within a legal framework that minimizes any risk of harm to others. While the full accounting of the legal considerations are outside the scope of this paper, most of our effort revolved around item selection and controls. For example, we restricted our pharmaceutical purchasing to non-prescription goods such as herbal and over-the-counter products, and we restricted our software purchases to items for which we already possessed a site license (also communicating our intent with the publisher). We did not use any received products (physical or electronic) and, aside from a few demonstration lots, they are scheduled to be destroyed upon the completion of our analyses.

Finally, while these controls are designed to prevent any explicit harm from resulting through the study, a remaining issue concerns the ethics of any implicit harm caused by supporting merchants (through our purchasing) who are themselves potentially criminal or unethical. Since our study does not deal with human subjects our institutional review board did not deem it appropriate for their review. Thus, our decision to move forward is based on our own subjective evaluation (along with the implicit oversight we received from university counsel and administration). In this, we believe that, since any such implicit support of these merchants is small (no individual affiliate program received more than $277 dollars from us), the potential value from better understanding their ecosystem vastly outweighs the potential harm.[11]

IV. Analysis

A major goal of our work is to identify any "bottlenecks" in the spam value chain: opportunities for disrupting monetization at a stage where the fewest alternatives are available to spammers (and ideally for which switching cost is high as well). Thus, in this section we focus directly on analyzing the degree to which affiliate programs *share* infrastructure, considering both the *click support* (i.e., domain registration, name service and Web hosting service) and *realization* (i.e., payment and fulfillment) phases of the spam value chain. We explore each of these in turn and then return to consider the potential effectiveness of interventions at each stage.

A. Click Support

As described in Section III we crawl a broad range of domains—covering the domains found in over 98% of our spam feed URLs—and use clustering and tagging to associate the resulting Web sites with particular affiliate programs. This data, in combination with our DNS crawler and domain WHOIS data, allows us to associate each such domain with an affiliate program and its various click support resources (registrar, set of name server IP addresses and set of Web hosting IP addresses). However, before we proceed with our analysis, we first highlight the subtleties that result from the use of Web site *redirection*.

Redirection: As we mentioned, some Web sites will redirect the visitor from the initial domain found in a spam message to one or more additional sites, ultimately resolving the final Web page (we call the domain for this page the "final domain"). Thus, for such cases one could choose to measure the infrastructure around the "initial domains" or the "final domains".

[11] This is similar to the analysis made in our previous study of the CAPTCHA-solving ecosystem [37].

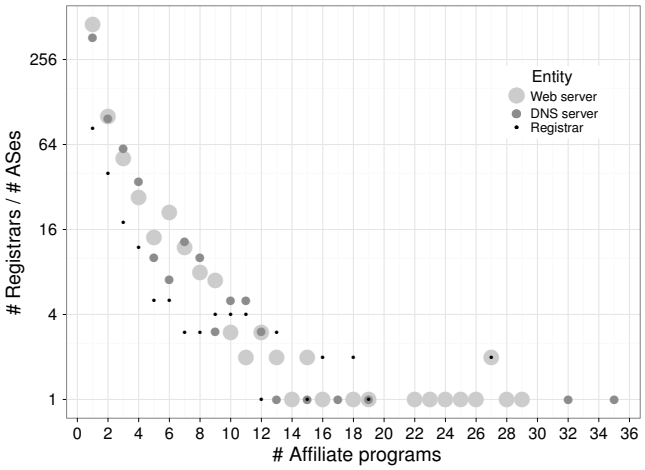

Figure 3: Sharing of network infrastructure among affiliate programs. Only a small number of registrars host domains for many affiliate programs, and similarly only a small number of ASes host name and Web servers for many programs. (Note y-axis is log scale.)

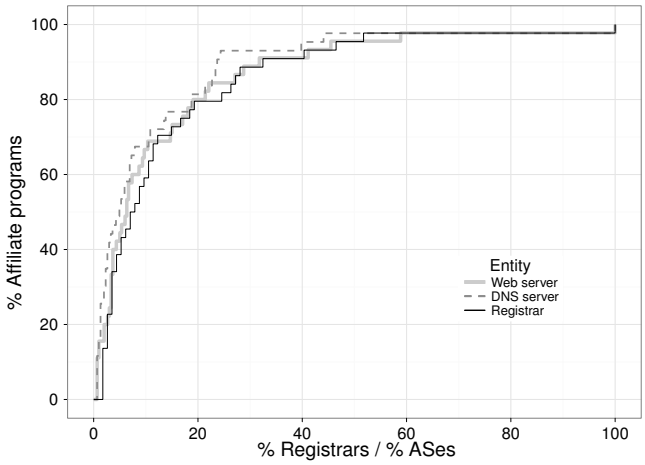

Figure 4: Distribution of infrastructure among affiliate programs. Only a small percentage of programs distribute their registered domain, name server, and Web server infrastructure among many registrars and ASes, respectively.

To explain further, 32% of crawled URLs in our data redirected at least once and of such URLs, roughly 6% did so through public URL shorteners (e.g., `bit.ly`), 9% through well-known "free hosting" services (e.g., `angelfire.com`), and 40% were to a URL ending in `.html` (typically indicating a redirect page installed on a compromised Web server).[12] Of the remainder, the other common pattern is the use of low-quality "throw away" domains, the idea being to advertise a new set of domains, typically registered using random letters or combinations of words, whenever the previous set's traffic-drawing potential is reduced due to blacklisting [24].

Given this, we choose to focus entirely on the final domains precisely because these represent the more valuable infrastructure most clearly operated by an affiliate.

Returning to our key question, we next examine the set of resources used by sites for each affiliate program. In particular, we consider this data in terms of the service organization who is responsible for the resource and how many affiliate programs make use of their service.

Network infrastructure sharing: A spam-advertised site typically has a domain name that must be resolved to access the site.[13] This name must in turn be allocated via a registrar, who has the authority to shutdown or even take back a domain in the event of abuse [30]. In addition, to resolve and access each site, spammers must also provision servers to provide DNS and Web services. These servers receive network access from individual ISPs who have the authority to disconnect clients who violate terms of service policies or in response to complaints.

[12]In our data, we identified over 130 shortener services in use, over 160 free hosting services and over 8,000 likely-compromised Web servers.

[13]Fewer than half a percent use raw IP addresses in our study.

Figure 3 shows that network infrastructure sharing among affiliate programs—when it occurs—is concentrated in a small number of registrars and Autonomous Systems (ASes).[14] Many registrars and ASes host infrastructure for just one or two affiliate programs, only a small number host infrastructure for many affiliate programs, and no single registrar or AS hosts infrastructure for a substantial fraction of the programs overall. (As we will see in Section IV-C however, this situation can change drastically when we weight by the volume of spam advertising each domain.) Specifically, Figure 3 shows the number of registrars (y-axis) that serve registered domains for a given number of affiliate programs (x-axis). Over 80 registrars, for instance, serve domains for a single affiliate program, while just two registrars (NauNet and China Springboard) serve domains for over 20 programs. For name servers and Web servers, it shows the number of ASes hosting servers for a given number of affiliate programs. Over 350 and 450 ASes host DNS and Web servers, respectively, for a single affiliate program; yet, just two and nine ASes host DNS and Web servers, respectively, for over 20 programs (including Hanaro Telecom, China Communication, and ChinaNet).

Although most registrars and ASes host infrastructure for just one affiliate program, each program could still engage many such registrars to serve their domains and many such ASes to host their DNS and Web servers. Figure 4 shows, though, that programs do not in general distribute their infrastructure across a large set of registrars or ASes: for most programs, each of them uses only a small fraction of registrars and ASes found in our data set. Specifically, Figure 4 shows the cumulative distribution of the fraction of registrars and ASes in our data set used by affiliate

[14]We use the AS number as a proxy for ISP.

Bank Name	*BIN*	*Country*	*Affiliate Programs*
Azerigazbank	404610	Azerbaijan	GlvMd, RxPrm, PhEx, Stmul, RxPnr, WldPh
B&N	425175	Russia	ASR
B&S Card Service	490763	Germany	MaxGm
Borgun Hf	423262	Iceland	Trust
Canadian Imperial Bank of Commerce	452551	Canada	WldPh
Cartu Bank	478765	Georgia	DrgRev
DnB Nord (Pirma)	492175	Latvia	Eva, OLPh, USHC
Latvia Savings	490849	Latvia	EuSft, OEM, WchSh, Royal, SftSl
Latvijas Pasta Banka	489431	Latvia	SftSl
St. Kitts & Nevis Anguilla National Bank	427852	St. Kitts & Nevis	DmdRp, VgREX, Dstn, Luxry, SwsRp, OneRp
State Bank of Mauritius	474140	Mauritius	DrgRev
Visa Iceland	450744	Iceland	Staln
Wells Fargo	449215	USA	Green
Wirecard AG	424500	Germany	ClFr

Table V: Merchant banks authorizing or settling transactions for spam-advertised purchases, their Visa-assigned Bank Identification Number (BIN), their location, and the abbreviation used in Table IV for affiliate program and/or store brand.

programs. For 50% of the affiliate programs, their domains, name servers, and Web servers are distributed over just 8% or fewer of the registrars and ASes, respectively; and 80% of the affiliate programs have their infrastructure distributed over 20% or fewer of the registrars and ASes. Only a handful of programs, such as EvaPharmacy, Pharmacy Express, and RX Partners, have infrastructure distributed over a large percentage (50% or more) of registrars and ASes.

To summarize, there are a broad range of registrars and ISPs who are used to support spam-advertised sites, but there is only limited amounts of organized sharing and different programs appear to use different subsets of available resource providers.[15]

B. Realization

Next, we consider several aspects of the realization pipeline, including post-order communication, authorization and settlement of credit card transactions, and order fulfillment.

We first examined the hypothesis that realization infrastructure is the province of *affiliate programs* and not individual affiliates. Thus, we expect to see consistency in payment processing and fulfillment between different instances of the same affiliate program or store brand. Indeed, we found only two exceptions to this pattern and purchases from different sites appearing to represent the same affiliate program indeed make use of the same merchant bank and same pharmaceutical drop shipper.[16] Moreover, key customer support features including the email templates and order number formats are consistent across brands belonging to the same program. This allowed us to further confirm our understanding that a range of otherwise distinct brands all belong to the same underlying affiliate program, including most of the replica brands: Ultimate Replica, Diamond Replicas, Distinction Replica, Luxury Replica, One Replica, Exquisite Replicas, Prestige Replicas, Aff. Accessories; most of the herbal brands: MaxGentleman, ManXtenz, Viagrow, Dr. Maxman, Stud Extreme, VigREX; and the pharmacy: US HealthCare.[17]

Having found strong evidence supporting the dominance of affiliate programs over free actors, we now turn to the question how much realization infrastructure is being shared *across* programs.

Payment: The sharing of payment infrastructure is substantial. Table V documents that, of the 76 purchases for which we received transaction information, there were only 13 distinct banks acting as Visa acquirers. Moreover, there is a significant concentration even among this small set of banks. In particular, most herbal and replica purchases cleared through the same bank in St. Kitts (a by-product of ZedCash's dominance of this market, as per the previous discussion), while most pharmaceutical affiliate programs used two banks (in Azerbaijan and Latvia), and software was handled entirely by two banks (in Latvia and Russia).

Each payment transaction also includes a standardized "Merchant Category Code" (MCC) indicating the type of goods or services being offered [52]. Interestingly, most affiliate program transactions appear to be coded correctly.

[15]We did find *some* evidence of clear inter-program sharing in the form of several large groups of DNS servers willing to authoritatively resolve collections of EvaPharmacy, Mailien and OEM Soft Store domains for which they were outside the DNS hierarchy (i.e., the name servers were never referred by the TLD). This overlap could reflect a particular affiliate advertising for multiple distinct programs and sharing resources internally or it could represent a shared service provider used by distinct affiliates.

[16]In each of the exceptions, at least one order cleared through a different bank—perhaps because the affiliate program is interleaving payments across different banks, or (less likely) because the store "brand" has been stolen, although we are aware of such instances.

[17]This program, currently called ZedCash, is only open by invitation and we had little visibility into its internal workings for this paper.

Supplier	*Item*	*Origin*	*Affiliate Programs*
Aracoma Drug	Orange bottle of tablets (pharma)	WV, USA	ClFr
Combitic Global Caplet Pvt. Ltd.	Blister-packed tablets (pharma)	Delhi, India	GlvMd
M.K. Choudhary	Blister-packed tablets (pharma)	Thane, India	OLPh
PPW	Blister-packed tablets (pharma)	Chennai, India	PhEx, Stmul, Trust, ClFr
K. Sekar	Blister-packed tablets (pharma)	Villupuram, India	WldPh
Rhine Inc.	Blister-packed tablets (pharma)	Thane, India	RxPrm, DrgRev
Supreme Suppliers	Blister-packed tablets (pharma)	Mumbai, India	Eva
Chen Hua	Small white plastic bottles (herbal)	Jiangmen, China	Stud
Etech Media Ltd	Novelty-sized supplement (herbal)	Christchurch, NZ	Staln
Herbal Health Fulfillment Warehouse	White plastic bottle (herbal)	MA, USA	Eva
MK Sales	White plastic bottle (herbal)	WA, USA	GlvMd
Riverton, Utah shipper	White plastic bottle (herbal)	UT, USA	DrMax, Grow
Guo Zhonglei	Foam-wrapped replica watch	Baoding, China	Dstn, UltRp

Table VI: List of product suppliers and associated affiliate programs and/or store brands.

For example, all of our software purchases (across all programs) were coded as 5734 (*Computer Software Stores*) and 85% of all pharmacy purchases (again across programs) were coded as 5912 (*Drug Stores and Pharmacies*). ZedCash transactions (replica and herbal) are an exception, being somewhat deceptive, and each was coded as 5969 (*Direct Marketing—Other*). The few other exceptions are either minor transpositions (e.g., 5921 instead of 5912), singleton instances in which a minor program uses a generic code (e.g., 5999, 8999) with a bank that we only observed in one transaction, and finally Greenline which is the sole pharmaceutical affiliate program that cleared transactions through a US Bank during our study (completely miscoded as 5732, *Electronic Sales*, across multiple purchases). The latter two cases suggest that some minor programs with less reliable payment relationships do try to hide the nature of their transactions, but generally speaking, category coding is correct. A key reason for this may be the substantial fines imposed by Visa on acquirers when miscoded merchant accounts are discovered "laundering" high-risk goods.

Finally, for two of the largest pharmacy programs, GlavMed and RX–Promotion, we also purchased from "canonical" instances of their sites advertised on their online support forums. We verified that they use the same bank, order number format, and email template as the spam-advertised instances. This evidence undermines the claim, made by some programs, that spammers have stolen their templates and they do not allow spam-based advertising.

Fulfillment: Fulfillment for physical goods was sourced from 13 different suppliers (as determined by declared shipper and packaging), of which eight were again seen more than once (see Table VI). All pharmaceutical tablets shipped from India, except for one shipped from within the United States (from a minor program), while replicas shipped universally from China. While we received herbal supplement products from China and New Zealand, most (by volume) shipped from within the United States. This result is consistent with our expectation since, unlike the other goods, herbal products have weaker regulatory oversight and are less likely to counterfeit existing brands and trademarks. For pharmaceuticals, the style of blister packs, pill shapes, and lot numbers were all exclusive to an individual nominal sender and all lot numbers from each nominal sender were identical. Overall, we find that only modest levels of supplier sharing between pharmaceutical programs (e.g., Pharmacy Express, Stimul-cash, and Club-first all sourced a particular product from PPW in Chennai, while RX–Promotion and DrugRevenue both sourced the same drug from Rhine Inc. in Thane). This analysis is limited since we only ordered a small number of distinct products and we know (anecdotally) that pharmaceutical programs use a network of suppliers to cover different portions of their formulary.

We did not receive enough replicas to make a convincing analysis, but all ZedCash-originated replicas were low-quality and appear to be of identical origin. Finally, purchased software instances were bit-for-bit identical between sites of the same store brand and distinct across different affiliate programs (we found no malware in any of these images). In general, we did not identify any particularly clear bottleneck in fulfillment and we surmise that suppliers are likely to be plentiful.

C. Intervention analysis

Finally, we now reconsider these different resources in the spam monetization pipeline, but this time explicitly from the standpoint of the defender. In particular, for any given registered domain used in spam, the defender may choose to intervene by either blocking its advertising (e.g., filtering spam), disrupting its click support (e.g., takedowns for name servers of hosting sites), or interfering with the realization step (e.g., shutting down merchant accounts).[18] But which of these interventions will have the most *impact*?

[18]In each case, it is typically possible to employ either a "takedown" approach (removing the resource comprehensively) or cheaper "blacklisting" approach at more limited scope (disallowing access to the resource for a subset of users), but for simplicity we model the interventions in the takedown style.

Ideally, we believe that such anti-spam interventions need to be evaluated in terms of two factors: their overhead to implement and their business impact on the spam value chain. In turn, this business impact is the sum of both the replacement cost (to acquire new resources equivalent to the ones disrupted) and the opportunity cost (revenue forgone while the resource is being replaced). While, at this point in time, we are unable to precisely quantify all of these values, we believe our data illustrates gross differences in scale that are likely to dominate any remaining factors.

To reason about the effects of these interventions, we consider the registered domains for the affiliate programs and storefront brands in our study and calculate their relative volume in our spam feeds (we particularly subtract the botnet feeds when doing this calculation as their inherent bias would skew the calculation in favor of certain programs). We then calculate the fraction of these domain trajectories that could be completely blocked (if only temporarily) through a given level of intervention at several resource tiers:

Registrar. Here we examine the effect if individual registrars were to suspend their domains which are known to be used in advertising or hosting the sites in our study.

Hosting. We use the same analysis, but instead look at the number of distinct ASs that would need to be contacted (who would then need to agree to shut down all associated hosts in their address space) in order to interrupt a given volume of spam domain trajectories. We consider both name server and Web hosting, but in each case there may be multiple IP addresses recorded providing service for the domain. We adopt a "worst case" model that *all* such resources must be eliminated (i.e., every IP seen hosting a particular domain) for that domain's trajectory to be disrupted.

Payments. Here we use the same approach but focused on the role played by the acquiring banks for each program. We have not placed purchases via each domain, so we make the simplifying assumption that bank use will be consistent across domains belonging to the same brand or affiliate program. Indeed this is strongly borne out in our measurements. For the two small exceptions identified earlier, we assign banks proportionally to our measurements.

Figure 5 plots this data as CDFs of the spam volume in our feeds that would be disrupted using these approaches. For both registrars and hosters there are significant concentrations among the top few providers and thus takedowns would seem to be an effective strategy. For example, almost 40% of spam-advertised domains in our feeds were registered by NauNet, while a single Romanian provider, Evolva Telecom, hosts almost 9% of name servers for spam-advertised domains and over 10% of the Web servers hosting their content; in turn, over 60% of these had payments handled via a single acquirer, Azerigazbank.

However, these numbers do not tell the entire story. Another key issue is the availability of alternatives and their switching cost.

For example, while only a small number of individual IP addresses were *used* to support spam-advertised sites, the supply of hosting resources is vast, with thousands of hosting providers and millions of compromised hosts.[19] The switching cost is also low and new hosts can be provisioned on demand and for low cost.[20]

By contrast, the situation with registrars appears more promising. The supply of registrars is fewer (roughly 900 gTLD registrars are accredited by ICANN as of this writing) and there is evidence that not all registrars are equally permissive of spam-based advertising [28]. Moreover, there have also been individual successful efforts to address malicious use of domain names, both by registries (e.g., CNNIC) and when working with individual registrars (e.g., eNom [25]). Unfortunately, these efforts have been slow, ongoing, and fraught with politics since they require global cooperation to be effective (only individual registrars or registries can take these actions). Indeed, in recent work we have empirically evaluated the efficacy of *past* registrar-level interventions and found that spammers show great agility in working around such actions [29]. Ultimately, the low cost of a domain name (many can be had for under $1 in bulk) and ease of switching registrars makes such interventions difficult.

Finally, it is the banking component of the spam value chain that is both the least studied and, we believe, the most critical. Without an effective mechanism to transfer consumer payments, it would be difficult to finance the rest of the spam ecosystem. Moreover, there are only two networks—Visa and Mastercard—that have the consumer footprint in Western countries to reach spam's principal customers. While there are thousands of banks, the number who are willing to knowingly process what the industry calls "high-risk" transactions is far smaller. This situation is dramatically reflected in Figure 5, which shows that just three banks provide the payment servicing for over 95% of the spam-advertised goods in our study.

More importantly, the replacement cost for new banks is high, both in setup fees and more importantly in time and overhead. Acquiring a legitimate merchant account directly with a bank requires coordination with the bank, with the card association, with a payment processor and typically involves a great deal of due diligence and delay (several days

[19]Note, spam hosting statistics can be heavily impacted by the differences in spam volume produced by different affiliates/spammers. For example, while we find that over 80% of all spam received in this study leads to sites hosted by just 100 distinct IP addresses, there are another 2336 addresses used to host the remaining 20% of spam-advertised sites, many belonging to the same affiliate programs but advertising with lower volumes of spam email.

[20]The cost of compromised proxies is driven by the market price for compromised hosts via Pay-Per-Install enterprises, which today are roughly $200/1000 for Western hosts and $5–10/1000 for Asian hosts [49]. Dedicated bulletproof hosting is more expensive, but we have seen prices as low as $30/month for virtual hosting (up to several hundred dollars for dedicated hosting).

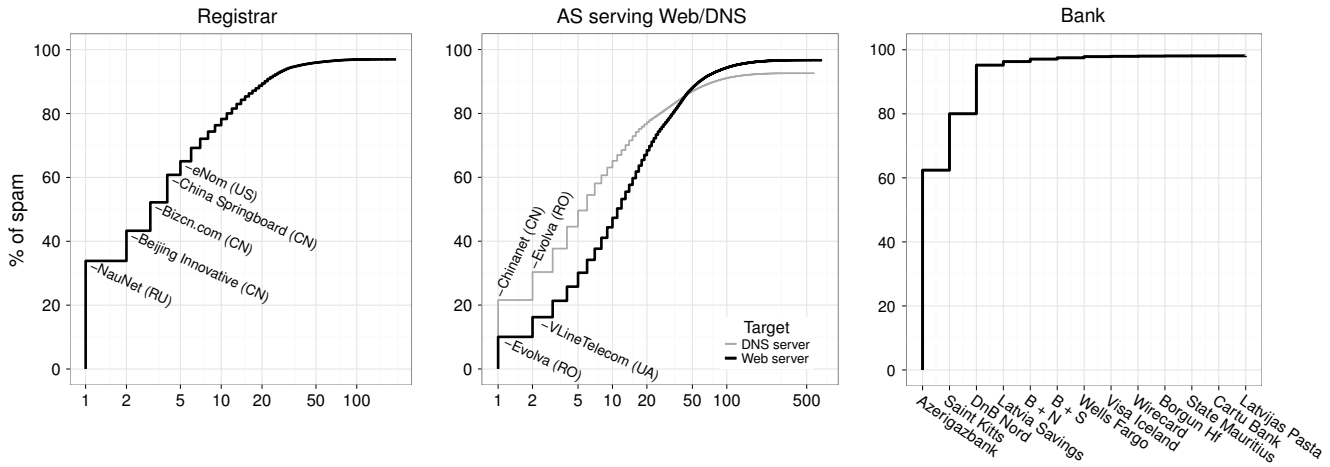

Figure 5: Takedown effectiveness when considering domain registrars (left), DNS and Web hosters (center) and acquiring banks (right).

or weeks). Even for so-called third-party accounts (whereby a payment processor acts as middleman and "fronts" for the merchant with both the bank and Visa/Mastercard) we have been unable to locate providers willing to provide operating accounts in less than five days, and such providers have significant account "holdbacks" that they reclaim when there are problems.[21] Thus, unlike the other resources in the spam value chain, we believe payment infrastructure has far fewer alternatives and far higher switching cost.

Indeed, our subsequent measurements bear this out. For four months after our study we continued to place orders through the major affiliate programs. Many continued to use the same banks four months later (e.g., all replica and herbal products sold through ZedCash, all pharmaceuticals from Online Pharmacy and all software from Auth. Soft. Resellers). Moreover, while many programs did change (typically in January or February 2011), they still stayed within same set of banks we identified earlier. For example, transactions with EvaPharmacy, Greenline, and OEM Soft Store have started clearing through B&N Bank in Russia, while Royal Software, EuroSoft and Soft Sales, have rotated through two different Latvian Banks and B & S Card Service of Germany. Indeed, the *only* new bank appearing in our follow-on purchases is Bank Standard (a private commercial bank in Azerbaijan, BIN 412939); RX–Promotion, GlavMed, and Mailien (a.k.a. Pharmacy Express) all appear to have moved to this bank (from Azerigazbank) on or around January 25th. Finally, one order placed with DrugRevenue failed due to insufficient funds, and was promptly retried through two different banks (but again, from the same set). This suggests that while cooperating third-party payment processors may be able to route transactions through merchant accounts at difference banks, the set of banks currently available for such activities is quite modest.

[21]To get a sense of the kinds of institutions we examined, consider this advertisement of one typical provider: "We have ready-made shell companies already incorporated, immediately available."

D. Policy options

There are two potential approaches for intervening at the payment tier of the value chain. One is to directly engage the merchant banks and pressure them to stop doing business with such merchants (similar to Legitscript's role with registrars [25], [28]). However, this approach is likely to be slow—very likely slower than the time to acquire new banking facilities. Moreover, due to incongruities in intellectual property protection, it is not even clear that the sale of such goods is illegal in the countries in which such banks are located. Indeed, a sentiment often expressed in the spammer community, which resonates in many such countries, is that the goods they advertise address a real need in the West, and efforts to criminalize their actions are motivated primarily by Western market protectionism.

However, since spam is ultimately supported by Western money, it is perhaps more feasible to address this problem in the West as well. To wit, if U.S. issuing banks (i.e., banks that provide credit cards to U.S. consumers) were to refuse to *settle* certain transactions (e.g., card-not-present transactions for a subset of Merchant Category Codes) with the banks identified as supporting spam-advertised goods, then the underlying enterprise would be dramatically demonetized. Furthermore, it appears plausible that such a "financial blacklist" could be updated very quickly (driven by modest numbers of undercover buys, as in our study) and far more rapidly than the turn-around time to acquire new banking resources—a rare asymmetry favoring the anti-spam community. Furthermore, for a subset of spam-advertised goods (regulated pharmaceuticals, brand replica products, and pirated software) there is a legal basis for enforcing such a policy.[22] While we suspect that the political challenges for

[22]Herbal products, being largely unregulated, are a more complex issue.

such an intervention would be significant—and indeed merit thoughtful consideration—we note that a quite similar action has already occurred in restricting U.S. issuers from settling certain kinds of online gambling transactions [11].

V. Conclusion

In this paper we have described a large-scale empirical study to measure the spam value chain in an end-to-end fashion. We have described a framework for conceptualizing resource requirements for spam monetization and, using this model, we have characterized the use of key infrastructure—registrars, hosting and payment—for a wide array of spam-advertised business interests. Finally, we have used this data to provide a normative analysis of spam intervention approaches and to offer evidence that the payment tier is by far the most concentrated and valuable asset in the spam ecosystem, and one for which there may be a truly effective intervention through public policy action in Western countries.

Acknowledgments

This is, again, the most ambitious measurement effort our team has attempted and even with 15 authors it would have been impossible without help from many other individuals and organizations. First and foremost, we are indebted to our spam data providers: Jose Nazario, Chris Morrow, Barracuda Networks, Abusix and a range of other partners who wish to remain anonymous. Similarly, we received operational help, wisdom and guidance from Joe Stewart, Kevin Fall, Steve Wernikoff, Doug McKenney, Jeff Williams, Eliot Gillum, Hersh Dangayach, Jef Pozkanzer, Gabe Lawrence, Neils Provos, Kevin Fu and Ben Ransford among a long list of others. On the technical side of the study, we thank Jon Whiteaker for an early implementation of the DNS crawler and Brian Kantor for supporting our ever expanding needs for cycles, storage and bandwidth. On the purchasing side of the study, we are deeply indebted to the strong support of our card issuer and their staff. On the oversight side, we are grateful to Erin Kenneally and Aaron Burstein for their legal guidance and ethical oversight, to our Chief Counsel at UCSD, Daniel Park, and UC's Systemwide Research Compliance Director, Patrick Schlesinger, for their open-mindedness and creativity, and finally to Marianne Generales and Art Ellis representing UCSD's Office of Research Affairs for helping to connect all the dots.

This work was supported in part by National Science Foundation grants NSF-0433668, NSF-0433702, NSF-0831138 and CNS-0905631, by the Office of Naval Research MURI grant N000140911081, and by generous research, operational and/or in-kind support from Google, Microsoft, Yahoo, Cisco, HP and the UCSD Center for Networked Systems (CNS). Félegyházi contributed while working as a researcher at ICSI. McCoy was supported by a CCC-CRA-NSF Computing Innovation Fellowship.

References

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

[2] 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.

[3] J. Armin, J. McQuaid, and M. Jonkman. Atrivo — Cyber Crime USA. http://fserror.com/pdf/Atrivo.pdf, 2008.

[4] Behind Online Pharma. From Mumbai to Riga to New York: Our Investigative Class Follows the Trail of Illegal Pharma. http://behindonlinepharma.com, 2009.

[5] C. Castelluccia, M. A. Kaafar, P. Manils, and D. Perito. Geolocalization of Proxied Services and its Application to Fast-Flux Hidden Servers. In *Proc. of 9th IMC*, 2009.

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

[7] Dancho Danchev's Blog — Mind Streams of Information Security Knowledge. The Avalanche Botnet and the TROYAK-AS Connection. http://ddanchev.blogspot.com/2010/05/avalanche-botnet-and-troyak-as.html, 2010.

[8] Federal Trade Commission. FTC Shuts Down, Freezes Assets of Vast International Spam E-Mail Network. http://ftc.gov/opa/2008/10/herbalkings.shtm, 2008.

[9] W. Feng and E. Kaiser. kaPoW Webmail: Effective Disincentives Against Spam. In *Proc. of 7th CEAS*, 2010.

[10] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants. In *Proc. of 14th ACM CCS*, 2007.

[11] Gamblingplanet.org. Visa blocks gaming transactions for US players. http://www.gamblingplanet.org/news/Visa-blocks-gaming-transactions-for-US-players/022310, 2010.

[12] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The Underground on 140 Characters or Less. In *Proc. of 17th ACM CCS*, 2010.

[13] G. Gu, J. Zhang, and W. Lee. BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic. In *Proc. of 15th NDSS*, 2008.

[14] S. Hao, N. Feamster, A. Gray, N. Syed, and S. Krasser. Detecting Spammers with SNARE: Spatio-Temporal Network-Level Automated Reputation Engine. In *Proc. of 18th USENIX Security*, 2009.

[15] C. Herley and D. Florencio. Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy. In *Proc. of 8th WEIS*, 2009.

[16] T. Holz, M. Engelberth, and F. Freiling. Learning More About the Underground Economy: A Case-Study of Keyloggers and Dropzones. In *Proc. of 15th ESORICS*, 2009.

[17] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling. Measuring and Detecting Fast-Flux Service Networks. In *Proc. of 15th NDSS*, 2008.

[18] X. Hu, M. Knysz, and K. G. Shin. RB-Seeker: Auto-detection of Redirection Botnets. In *Proc. of 16th NDSS*, 2009.

[19] D. Irani, S. Webb, J. Giffin, and C. Pu. Evolutionary Study of Phishing. In *eCrime Researchers Summit*, pages 1–10, 2008.

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

[21] 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.

[22] M. Konte, N. Feamster, and J. Jung. Dynamics of Online Scam Hosting Infrastructure. In *Proc. of 10th PAM*, 2009.

[23] Krebs on Security. Body Armor for Bad Web Sites. http://krebsonsecurity.com/2010/11/body-armor-for-bad-web-sites/, 2010.

[24] 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.

[25] LegitScript and eNom. LegitScript Welcomes Agreement with eNom (DemandMedia). http://www.legitscript.com/blog/142, 2010.

[26] LegitScript and KnujOn. No Prescription Required: Bing.com Prescription Drug Ads. http://www.legitscript.com/download/BingRxReport.pdf, 2009.

[27] LegitScript and KnujOn. Yahoo! Internet Pharmacy Advertisements. http://www.legitscript.com/download/YahooRxAnalysis.pdf, 2009.

[28] LegitScript and KnujOn. Rogues and Registrars: Are some Domain Name Registrars safe havens for Internet drug rings? http://www.legitscript.com/download/Rogues-and-Registrars-Report.pdf, 2010.

[29] 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.

[30] B. Livingston. Web registrars may take back your domain name. http://news.cnet.com/2010-1071-281311.html, 2000.

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

[32] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Identifying Suspicious URLs: An Application of Large-Scale Online Learning. In *Proc. of 26th ICML*, 2009.

[33] B. S. McWilliams. *Spam Kings: The Real Story Behind the High-Rolling Hucksters Pushing Porn, Pills and @*#?% Enlargements*. O'Reilly Media, Sept. 2004.

[34] D. Molnar, S. Egelman, and N. Christin. This Is Your Data on Drugs: Lessons Computer Security Can Learn From The Drug War. In *Proc. of 13th NSPW*, 2010.

[35] T. Moore and R. Clayton. The Impact of Incentives on Notice and Take-down. In *Proc. of 7th WEIS*, 2008.

[36] T. Moore, R. Clayton, and H. Stern. Temporal Correlations between Spam and Phishing Websites. In *Proc. of 2nd USENIX LEET*, 2009.

[37] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G. M. Voelker, and S. Savage. Re: CAPTCHAs — Understanding CAPTCHA Solving from an Economic Context. In *Proc. of 19th USENIX Security*, 2010.

[38] A. Mutton. Screengrab! http://www.screengrab.org/, 2010.

[39] Y. Niu, Y.-M. Wang, H. Chen, M. Ma, and F. Hsu. A Quantitative Study of Forum Spamming Using Context-based Analysis. In *Proc. of 14th NDSS*, 2007.

[40] C. Nunnery, G. Sinclair, and B. B. Kang. Tumbling Down the Rabbit Hole: Exploring the Idiosyncrasies of Botmaster Systems in a Multi-Tier Botnet Infrastructure. In *Proc. of 3rd USENIX LEET*, 2010.

[41] E. Passerini, R. Paleari, L. Martignoni, and D. Bruschi. FluXOR: Detecting and Monitoring Fast-Flux Service Networks. In *Proc. of 5th DIMVA*, 2008.

[42] R. Perdisci, I. Corona, D. Dagon, and W. Lee. Detecting Malicious Flux Service Networks through Passive Analysis of Recursive DNS Traces. In *Proc. of 25th ACSAC*, 2009.

[43] 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.

[44] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu. On Network-level Clusters for Spam Detection. In *Proc. of 17th NDSS*, 2010.

[45] A. Ramachandran and N. Feamster. Understanding the Network-Level Behavior of Spammers. In *Proc. of ACM SIGCOMM*, 2006.

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

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

[48] S. Sinha, M. Bailey, and F. Jahanian. Improving SPAM Blacklisting through Dynamic Thresholding and Speculative Aggregation. In *Proc. of 17th NDSS*, 2010.

[49] K. Stevens. The Underground Economy of the Pay-Per-Install (PPI) Business. http://www.secureworks.com/research/threats/ppi, 2009.

[50] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna. Your Botnet Is My Botnet: Analysis of a Botnet Takeover. In *Proc. of 16th ACM CCS*, 2009.

[51] B. Stone-Gross, C. Kruegel, K. Almeroth, A. Moser, and E. Kirda. FIRE: FInding Rogue nEtworks. In *Proc. of 25th ACSAC*, 2009.

[52] Visa Commercial Solutions. Merchant Category Codes for IRS Form 1099-MISC Reporting. http://usa.visa.com/download/corporate/resources/mcc_booklet.pdf.

[53] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen. Spam Double-Funnel: Connecting Web Spammers with Advertisers. In *Proc. of 16th WWW*, 2007.

[54] 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.

[55] C. Whittaker, B. Ryner, and M. Nazif. Large-Scale Automatic Classification of Phishing Pages. In *Proc. of 17th NDSS*, 2010.

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

[57] L. Zhang, J. Zhu, and T. Yao. An Evaluation of Statistical Spam Filtering Techniques. *ACM Trans. on ALIP*, 3(4), 2004.

[58] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum. BotGraph: Large-Scale Spamming Botnet Detection. In *Proc. of 6th NSDI*, 2009.

[59] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying Malicious Websites and the Underground Economy on the Chinese Web. In *Proc. of 7th WEIS*, 2008.