Understanding the Domain Registration Behavior of Spammers

Shuang Hao
Georgia Tech
shao@cc.gatech.edu

Matthew Thomas
Verisign, Inc.
mthomas@verisign.com

Vern Paxson
ICSI & UC Berkeley
vern@cs.berkeley.edu

Nick Feamster
Georgia Tech
feamster@cc.gatech.edu

Christian Kreibich
ICSI & Lastline, Inc.
christian@icir.org

Chris Grier
ICSI & UC Berkeley
grier@cs.berkeley.edu

Scott Hollenbeck
Verisign, Inc.
shollenbeck@verisign.com

ABSTRACT
Spammers register a tremendous number of domains to evade blacklisting and takedown efforts. Current techniques to detect such domains rely on crawling spam URLs or monitoring lookup traffic. Such detection techniques are only effective after the spammers have already launched their campaigns, and thus these countermeasures may only come into play after the spammer has already reaped significant benefits from the dissemination of large volumes of spam. In this paper we examine the registration process of such domains, with a particular eye towards features that might indicate that a given domain likely has a malicious purpose at registration time, before it is ever used for an attack. Our assessment includes exploring the characteristics of registrars, domain life cycles, registration bursts, and naming patterns. By investigating zone changes from the .com TLD over a 5-month period, we discover that spammers employ bulk registration, that they often re-use domains previously registered by others, and that they tend to register and host their domains over a small set of registrars. Our findings suggest steps that registries or registrars could use to frustrate the efforts of miscreants to acquire domains in bulk, ultimately reducing their agility for mounting large-scale attacks.

Categories and Subject Descriptors
C.2.3 [Computer-Communication Networks]: Network Operations—Network monitoring; K.6.5 [Security and Protection]; K.4.1 [Computers and Society]: Public Policy Issues—Abuse and crime involving computers

General Terms
Measurement, Security

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IMC’13, October 23–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-1953-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2504730.2504753.

Keywords
DNS; Domain Registration; Spam; Blacklist

1. INTRODUCTION
Spammers and other miscreants often make heavy use of domains registered in the DNS to direct victims to Web sites that host scams, malware, and other malicious content. To mitigate these threats, network operators employ blacklisting of domains, but the rate at which new domains appear makes developing decisions sufficiently quickly to blacklist domains particularly challenging. Ideally, such decisions could be made at registration time rather than at usage time, enabling “proactive blocking”. However, developing registration-time decisions about which new domains will likely see subsequent malicious employment appears quite daunting given the very large rate at which new domains appear (tens of thousands per day for .com). Instead, existing DNS reputation systems use either evidence of malicious use (e.g., appearance of names in a spam trap) or the characteristics of DNS lookup traffic [1, 4]. Such systems generally must observe a significant volume of DNS lookups before determining the reputation to associate with a domain.

In this work we seek to understand the nature of spammer domain registrations with an ultimate goal of hampering the ease with which attackers currently acquire large volumes of registered domains. We do so by analyzing a range of registration-time features as manifest in changes seen every 5 minutes to .com over a 5-month period. For partial ground truth in assessing which of these registrations reflected spammer activity, we draw upon domains identified by several large blacklist feeds associated with email spam campaigns.

Our study develops the following findings:

• We confirm the earlier finding that only a handful of registrars account for the majority of spammer domains [18]. 70% of spammer domains came from 10 registrars, though these registrars accounted for only about 20% of all new domains added to the zone. Thus, miscreants appear to prefer those specific registrars, and positive actions from these registrars could have significant impact in impeding the use of large volumes of newly registered domains for spam activity.

• Groups of domains registered by a given registrar at a single time exhibit two statistically distinct patterns. We show
that groups of registrations very often follow a distribution well-described by a compound Poisson process, but many registrars also exhibit registration “spikes” that this process would produce only with exceedingly low probability.

- **Spammer domains occur in such “spikes” with much more prevalence than in general (non-spike) registration activity.** This finding suggests that spammers find economic and/or management benefit to registering domains in large batches, and thus detection procedures that leverage the presence of such spikes could force spammers to adopt less efficient approaches for their registrations.

- **Spammers frequently re-register expired domains that originally had a clean history.** Presumably using such names alleviates spammers of the burden of generating plausible-looking names (though we also observe algorithmically generated names), providing textual diversity as well as a benign past reputation that may aid in initially avoiding detection.

We hope these findings will ultimately lead to the development of a detection procedure that can accurately identify names intended for malicious use at time-of-registration rather than only later at time-of-use.

The remainder of this paper is organized as follows. §2 introduces the taxonomy and §3 surveys the related work. §4 describes the datasets that we collected and used in our analysis. In §5 use the datasets to illuminate the benefits of identifying spammer domains at registration time. We then proceed with investigating the prospects for doing so, starting in §6, which analyzes the distribution of spammer and non-spammer domains across registrars and DNS servers. §7 presents our findings regarding bulk registration and our approach to identify registration spikes. §8 associates the domain life cycle with registration to dissect spammers’ strategies to acquire domains.

## 2. BACKGROUND: DNS REGISTRATION PROCESS AND LIFE CYCLE

To set the context for our work, here we sketch the process by which malicious parties (and others) register domains and the subsequent life cycle regarding use of the domains.

Figure 1 shows the domain registration process. There are three roles in the figure: registrants (domain registration applicants), registrars (e.g., GoDaddy), registries (e.g., Verisign). Registries are responsible for managing the registration of domain names within the top-level domains (TLDs) and generating the zone files that list domain names and their authoritative nameservers. For example, Verisign serves as the registrar for .com, CNNIC for .cn, and DENIC for .de [32]. In this work we focus on .com, the largest TLD [30], which has long reflected a major target abused by miscreants for spamming activities [27].

ICANN accredits registrars, which contract with TLD registries to provide registration service to the public. Presently around 900 registrars exist across all TLDs, the bulk of which serve .com (and often other TLDs) [25]. A registrant selects a designated registrar to register a domain. The designated registrar in turn connects to the registry’s SRS (Shared Registration System) via EPP (Extensible Provisioning Protocol, RFC5730 [14]) or RRP (Registry Registrar Protocol, RFC3632 [13]) to manage the zones. The registry updates the corresponding DNS zone information in the database and uses RZU (Rapid Zone Update) to add the DNS information in the top-level domain nameservers. Domain registration operates in a real-time fashion, resulting in only a short interval between registration requests and domains becoming active in the zone.

Figure 2 indicates the life cycle of a second-level domain in the .com zone. We show only a simplified cycle; see ICANN’s registry agreements [15] for a full description of the possible states of a domain. In order to obtain a domain, registrants need to select available domain names that are not registered and in use. The registration term usually ranges from 1 year to 10 years. If the registrant chooses to renew a domain, the expiration date will be extended and the domain remains in the zone files. The renewal could occur at any time during the registration period, the Auto-Renew Grace Period (for domains the registrar has already marked for renewal) and the Redemption Grace Period (for domains marked for deletion). If the registrant chooses not to renew, the domain expires, gets removed from the zone and becomes available for others to register. Two special periods mark the beginning and end of a domain’s life cycle. The 5-day Add Grace Period begins immediately after domain registration, allowing the registrant to change their mind, undo the registration, and receive full credit for the registration fee [16]. To limit domain tasting abuse, i.e., taking advantage of no-cost trial periods for domains, registrars limit the number of registrations a registrant may revert per month. The domain enters a 5-day Pending Delete Period that prevents further alterations to the domain’s status before it gets unregistered and becomes available for re-registration [29, 22]. We explore how this life cycle relates to the registration of spammer domains in §8.

## 3. RELATED WORK

**DNS Resolution.** Most previous DNS-based detection studies have focused on analyzing lookup traffic. These studies date back many years to work performed by Danzig et al. [6] and Jung et al. [17], both of which examined lookup behavior from the vantage point of lookups produced by individual sites. Notos [1] and EXPOSURE [4] leverage DNS lookup behavior within a local network to formulate reputations associated with domains. Kopis monitors the traffic to authoritative nameservers and top-level domain servers to achieve global visibility to detect abnormal activities [2]. Our previous work actively probed the DNS records and studied the lookup traffic of second-level domains after their registration to characterize those associated with Internet attacks [12]. In contrast, our
work here analyzes the registration patterns of spammer domains to illuminate potential opportunities for early detection of spammer domains.

**Registrars and Registration.** Several lines of recent work have examined the roles of registrars and registries. Coull et al. studied the registration abuse phenomenon, speculating, tasting, and front-running (where registrars abuse insider information to obtain a domain, thus locking out other registrars) [5]. Liu et al. investigated the effect of registry policy changes and registrar actions taken to terminate domains used by illicit on-line pharmacies [19], finding domain pricing and domain takedowns to have at least temporary deterring effect. Feleghyhazi et al. inferred groups of malicious domains based on the DNS servers and daily registration concurrency associated with known-bad seed domains [10]. While closest to the theme of our work, their approach’s need to identify such seeds complicates its application at registration time. In addition, we provide a significantly more granular analysis (5-minute zone changes, modeling the registration process of individual registrars, statistical identification of “spikes”), and data broader in scope (non-spam registrations vs. spam registrations, domain life cycle, naming patterns).

**Domain Generation Algorithms.** In an effort to resist takedown of centralized command and control (C&C) servers, recent botnets such as Conficker, Kraken and Bobax have used domain generation algorithms (DGAs) to query a set of domains. The botmaster needs to only register one domain of the set to enable bots to access the C&C server. Yadav et al. studied different metrics to identify botnets by finding randomly generated domain names [31]. Antonakakis et al. developed clustering techniques to discover new DGA variants and compromised hosts [3]. Unlike these works, we examine the domain names as actually registered in the .com TLD in an attempt to identify lexical patterns distinguishing spammer domains.

### 4. DATA COLLECTION

In this section we describe the datasets used in our analysis, which we summarize in Table 1. Our primary dataset consists of changes made to the .com zone every five minutes for a 5-month period, March–July 2012. In addition, we interpret the significance (i.e., spammer or otherwise) of new registrations in .com based on any subsequent appearance of a given domain in either a “spam trap” we operate or a well-known blacklist. We term a newly registered domain that appears in either the spam trap or on one of the blacklists as a spammer domain, and a domain that does not as a non-spammer domain.

**Domain Registration.** Verisign operates the .com zone under contract to ICANN. Changes to the zone appear in Domain Name Zone Alert (DNZA) files, which indicate (1) the addition of new domains, (2) the removal of existing domains, and (3) changes to existing domains in terms of revisions to their associated nameservers. Our data includes captures of the DNZA files as recorded every five minutes, time periods we refer to as epochs.

**Registrars and History.** Domain registrations must be executed by an ICANN-accredited registrar chosen by the user registering the domain (the “registrant”). The registrant pays the registrar a fee for this service. In general, we have no visibility into the registrants associated with particular domains (sometimes WHOIS information provides their identities, but numerous registrars provide a “private registration” service that masks this information). One can however obtain information about a given domain’s registrar based on WHOIS information, or using third-party services such as Domain-Tools [8]. Thus, we can only attempt to tease out the registration behavior of individual users as inferable from the registration activities of individual registrars.

A given domain added to the zone might reflect either a first-time registration or a re-registration of a previously registered domain. We can distinguish these two based on historical WHOIS information; for re-registered domains, we can obtain when the domain was previously deleted from the zone [28].

**Identifying Spammer Domains.** In general we would like to associate with domains a label indicating whether an attacker registered the domain for spamming purposes. Since we lack comprehensive ground truth regarding the ultimate use of domains, to this end we use two proxies: subsequent appearance of a newly registered domain in: (1) an email spam campaign, or (2) a domain blacklist.

For the first of these, we operated a spam trap, i.e., our own domain with an associated mail server that has no legitimate email addresses. We can confidently consider all emails sent to the spam trap as spam. Although the spam contains non-spam related domains (e.g., youtube.com), by restricting our focus to domains recently registered (March–July 2012) we can filter down the domains appearing in the spam trap to those very likely used for spamming.

For the second, we subscribed to three major DNS blacklists, URIBL, SURBL, and Spamhaus DBL. During our subsequent analysis we found strong indications that the Spamhaus DBL very
likely uses registration-time features to establish the reputation of a domain (see the discussion in §5.3). Given that, then since part of our focus is to assess to what degree registration-time features correlate with a domain’s subsequent employment in an abusive context, for our purposes we cannot soundly use a domain’s presence on the Spamhaus DBL as such an indicator. Consequently, we omit this source from our analysis other than to demonstrate the indicators that it uses such features for blacklisting.

**Summary of Data.** Table 2 shows the number of second-level .com domains registered in each month, and the subset of those registrations that later appeared in either our spam trap or on one of the two blacklists. Of the 12,824,401 second-level domains registered in the .com zone over five months, 134,455 reflect spammer domains.

### 5. LONGEVITY OF SPAMMER DOMAINS

In this section we look at several facets of the time periods over which spammers employ their domains: the age of domains (time since registration) when they appear on blacklists or in spam campaigns; the amount of time during which domains continue to see use once spammers begin to employ them; and the amount of time between a recent registration of a spammer domain and its subsequent appearance. Our examination shows that detecting spammer domains at the time of their registration can offer significant advantages.

#### 5.1 Age of Domains Used for Spamming

We first consider the degree to which spammers employ relatively “fresh” (recently registered) domains in their spam campaigns. If spammers primarily rely upon long-lived domains, then we cannot hope to gain much benefit from disrupting the registration of new spammer domains.

Figure 3 shows the distribution of the amount of time elapsed between the registration of a given .com spammer domain and its appearance in either the URIBL blacklist or our spam trap. (We omit results for SURBL because we lack sufficiently long data from it to compute a comparable distribution.) In particular, for all .com domains that appeared in either URIBL or our spam trap from March–July 2012, we determine the domain’s date of registration, and plot the difference between that time and the first such appearance. Overall, 35–40% of the domains were registered within the past 30 days, and 40–50% within 60 days. (In addition, since listings in URIBL will likely lag behind actual use, and the spam trap will include some long-lived benign domains such as google.com, the age of actual spammer domains at time of first use will skew somewhat lower than these figures.)

Because domain registrations represent a direct cost for spammers, the fact that spammers frequently employ domains registered quite recently (within a few months) indicates that they have an ongoing need to acquire new domains. Thus, if we impair their registration activities, we can add friction to their general enterprise. In addition, given the quantity of domains that spammers use, we would expect that their need to continually acquire new domains will incline them towards registering new domains in batches, a feature that we analyze in §7.1.

#### 5.2 Duration-of-Use in Spam Campaigns

Another facet of spammer domain longevity concerns for how long a spammer uses a given domain. If domains see only brief periods of use, then activity-based blacklisting will fail to effectively block the spammer’s fruitful employment of a domain unless the blacklisting occurs very soon after the onset of use. If so, then the benefits of identifying spammer domains prior to use, such as at-time-of-registration, rise.

We can assess duration-of-use from our spam trap data, and indeed we find that more than 60% of the .com domains observed in the spam trap appear during only a single day (75% for ≤ 10 days, and only 5% for ≥ 60 days). This “single-shot” nature of most of the spammer domains complicates blacklisting efforts—though it also may reflect the efficacy of such efforts at narrowing the window during which spammers can profitably use their domains—and highlights the benefit of at-time-of-registration detection.

#### 5.3 Lifetime of Recently Registered Domains

Finally, we examine the use by spammers of newly registered domains given that we flagged it as a spammer domain (and thus it necessarily appeared in our spam trap, or in one of our black-

### Table 2: Monthly data statistics.

<table>
<thead>
<tr>
<th>Days between registration and appearance</th>
<th>New domains</th>
<th>Subset of new domains appearing in spam trap</th>
<th>Subset of new domains appearing in URIBL</th>
<th>Subset of new domains appearing in SURBL</th>
<th>Subset of new domains appearing on blacklists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days between registration and appearance</td>
<td>March 2012</td>
<td>2,832,867</td>
<td>6,072</td>
<td>12,372</td>
<td>18,875</td>
</tr>
<tr>
<td>Days between registration and appearance</td>
<td>April 2012</td>
<td>2,596,192</td>
<td>3,970</td>
<td>12,111</td>
<td>21,824</td>
</tr>
<tr>
<td>Days between registration and appearance</td>
<td>May 2012</td>
<td>2,641,466</td>
<td>4,091</td>
<td>10,726</td>
<td>21,616</td>
</tr>
<tr>
<td>Days between registration and appearance</td>
<td>June 2012</td>
<td>2,383,010</td>
<td>2,861</td>
<td>8,651</td>
<td>21,872</td>
</tr>
<tr>
<td>Days between registration and appearance</td>
<td>July 2012</td>
<td>2,389,636</td>
<td>2,958</td>
<td>8,875</td>
<td>29,525</td>
</tr>
</tbody>
</table>

*Note: Registrations over 5 months.*

---

**Figure 3:** Distribution of days between domain registration and appearance of a .com domain in either our spam trap (red) or URIBL (blue).
lists, during the coming months). Figure 4 shows the distribution of the time between the registration of such spammer domains and their appearance in our spam trap or in a blacklist.\(^1\) We see that a number of days may pass after registration prior to the domain’s appearance. This delay again indicates we might gain significant benefit from identifying spammer domains at the time of registration, before any activity takes place.

The figure also shows the delay between registration and blacklist for the Spamhaus DBL. We observed that Spamhaus blacklists domains much earlier than URIBL—or even than the appearance of the domain in our spam trap. We confirmed with Spamhaus that they base their DBL entries in part on information gathered at registration time to facilitate more proactive blacklisting. Because Spamhaus uses registration-time features to construct their blacklist, the presence of a Spamhaus domain in a blacklist does not provide independent evidence that features we consider for registration-time detection indeed will have power for identifying spammer domains. Given this lack of independence, we refrain from further analysis treating the appearance of a domain on the Spamhaus blacklist as separate confirmation of the domain as one employed by spammers.

Finally, we examine the extent to which any particular blacklist covers the full set of spammer domains, and the extent to which these blacklists overlap with one another. Figure 5 shows the intersection of spammer domains registered from March–July 2012, based on the information from our spam trap, and from URIBL and SURBL. We observe that each data source identified many spammer domains that did not appear in the other information sources. This lack of overlap presumably indicates that different blacklist sources use different criteria to determine whether to include a domain in its blacklist.

\(^1\)It is important to keep in mind the distinction between Figure 3 and Figure 4. The former is conditioned on any domain appearing in either the spam trap or the blacklist during the five-month period; the latter is conditioned on the appearance of any domain registered during the five-month period.

6. SPAM DOMAIN INFRASTRUCTURE

In this section we briefly look at the infrastructure supporting individual spammer domain registrations: the registrars used to register these domains, the DNS servers initially selected to resolve the domains, and how this infrastructure compares with that used for non-spammer domains. Our analysis supports the following:

- Nearly 70% of spammer domains originated from 10 registrars, while those registrars accounted for only 20% of all newly registered domains over the 5 months of our data.
- Spammer domains primarily use the regular authoritative DNS servers operated by the registrar, at least initially. This finding suggests that efforts to proactively blacklist spammer domains should focus on registrar-level analysis rather than DNS-server analysis.

We now develop these findings in more detail.

6.1 Registrars Used for Spammer Domains

We first examine the proportion of registrations at each registrar that correspond to spammer domains and how this proportion varies by registrar. Table 3 shows the registrars, ranked by the number of spammer domains that they registered over the five-month period of our study (shown in the second column); the third column shows the percentage of known spammer domains registered by that registrar. The fourth column indicates the cumulative percentage of spammer domains for the top registrars. Interestingly, 46% of the spammer domains correspond to just two registrars. This statistic implies that the positive actions from a small set of registrars might significantly frustrate the use of newly registered spammer domains. We treated GoDaddy separately because it manages significantly more domains than other registrars; hence, even though it registers a significant number of spammer domains, the number of spammer domains that it registers remains a small fraction of the total number of domains that it registers.

Our findings agree with a similar study by Levchenko et al. [18], and also with an independent study that ranks the registrars serving rogue Internet pharmacies [24]. Indeed, three registrars—Moniker, Tucows, and Bizcn.com, Inc.—appeared in both studies as top-ten registrars for spammer domains.

Next, we explore how the registrars compare, in terms of the number of spammer and non-spammer domains that they register. Figure 6 shows the number of spammer and non-spammer domains that each registrar registered over the course of our study; each dot represents a registrar. Dots above the diagonal show registrars that registered more non-spammer domains than spammer
GoDaddy sellers, provides free private spamming operation known as “Quick Cart Pro.”

A range of fees. For example, eNom offers cheaper prices for bulk registration, and

The percentage of domains that a nameserver hosts that are spammer domains. This metric represents the extent to which an authoritative nameserver sees use mainly in support of spamming activity. A toxicity of 100% indicates that the nameservers appear to operate solely under miscreant control; the presence of such nameservers in a new domain registration could effectively identify new spammer domains.

Owners of a given domain typically use multiple DNS servers to host the same domain to achieve redundancy in case of failure [20]. Intuitively, a group of domains hosted by the same set of authoritative nameservers likely have some relationship. We compute the Jaccard index to measure the similarity of the authoritative nameservers in terms of the domains they host. Suppose there exist N name-servers, each of which hosts a set of domains $D_i$. We compute $| \cap_{i=1}^{N} D_i | / | \cup_{i=1}^{N} D_i |$. A higher Jaccard index for a pair of nameservers indicates a high overlap in terms of the set of domains that those nameservers resolve. This association may ultimately help with identifying groups of nameservers commonly used to host spammer domains.

The percentage of domains hosted on a nameserver that belong to a particular registrar. We define the registrar with the highest association score for a nameserver as the primary registrar for that nameserver.

Figure 7 shows the cumulative distribution of spammer domains over the DNS servers. The X-axis shows the indexes of DNS servers ordered by their toxicity: 8,543 of 242,790 authoritative DNS servers assigned to those domains. We thought that we might find that spammer domains are disproportionately hosted on certain sets of authoritative DNS servers, which we assessed using three metrics:

- Toxicity:
- Duplication:
- Association:

 APIs that allow users to manipulate the domain zone entries, and Moniker allows up to 500 domain registrations at a time.

6.2 Authoritative Nameservers

The zone updates we use for our study include NS records associated with each new domain. During March–July 2012, we observed 12,824,401 newly registered .com domains, but only 242,790 authoritative DNS servers assigned to those domains. We thought that we might find that spammer domains are disproportionately hosted on certain sets of authoritative DNS servers, which we assessed using three metrics:

- Toxicity: The percentage of domains that a nameserver hosts that are spammer domains. This metric represents the extent to which an authoritative nameserver sees use mainly in support of spamming activity. A toxicity of 100% indicates that the nameservers appear to operate solely under miscreant control; the presence of such nameservers in a new domain registration could effectively identify new spammer domains.

- Duplication: Owners of a given domain typically use multiple DNS servers to host the same domain to achieve redundancy in case of failure [20]. Intuitively, a group of domains hosted by the same set of authoritative nameservers likely have some relationship. We compute the Jaccard index to measure the similarity of the authoritative nameservers in terms of the domains they host. Suppose there exist N nameservers, each of which hosts a set of domains $D_i$. We compute $| \cap_{i=1}^{N} D_i | / | \cup_{i=1}^{N} D_i |$. A higher Jaccard index for a pair of nameservers indicates a high overlap in terms of the set of domains that those nameservers resolve. This association may ultimately help with identifying groups of nameservers commonly used to host spammer domains.

- Association: The percentage of domains hosted on a nameserver that belong to a particular registrar. We define the registrar with the highest association score for a nameserver as the primary registrar for that nameserver.

Figure 7 shows the cumulative distribution of spammer domains over the DNS servers. The X-axis shows the indexes of DNS servers ordered by their toxicity: 8,543 of 242,790 DNS servers hosted spammer domains. The figure has two Y axes. The blue dashed curve shows the toxicity of each DNS server and corresponds to the Y-axis on the left side. The red solid curve shows the cumulative percentage of spammer domains for the set of nameservers ranked by their toxicity, and maps to the Y-axis values on the right side of the plot. The nameservers hosting only spammer ones, and dots below the diagonal line reflect a ratio towards higher spammer domain registrations than non-spammer. The figure labels the top registrars for spammer domains, and shows that 19 of the 919 registrars in our study registered more than 1,000 spammer domains. We see that spammers often use popular registrars to register spammer domains, perhaps because doing so may make it more difficult to identify spammer domains solely based on the registrar. On the other hand, for some registrars that do not register many domains, their fraction of spammer domains can be strikingly high (ABSystems, in particular 3).

We speculate that the decisions by spammers regarding domain registrations are driven by both economic concerns (price of registration) as well as the ease of managing multiple domain registrations. Regarding the first of these, we note that registrars charge a range of fees. For example, eNom sets special prices for resellers. GoDaddy offers cheaper prices for bulk registration, and INTERNET. bs provides free private WHOIS protection. The management features that each registrar provides determine how easily a customer can manage a domain; for example, eNom provides

3The badness of ABSystems comes as no surprise. This registrar effectively acts as the DNS infrastructure division of a large spamming operation known as “Quick Cart Pro.”

<table>
<thead>
<tr>
<th>Registrar</th>
<th>Number</th>
<th>Percentage</th>
<th>Cumulative pct.</th>
<th>Percentage</th>
<th>Cumulative pct.</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNom, Inc.</td>
<td>36,245</td>
<td>27.03</td>
<td>27.03</td>
<td>7.62</td>
<td>7.62</td>
</tr>
<tr>
<td>Moniker Online Services, Inc.</td>
<td>25,488</td>
<td>19.01</td>
<td>46.05</td>
<td>0.67</td>
<td>8.30</td>
</tr>
<tr>
<td>Tucows.com Co.</td>
<td>5,996</td>
<td>4.47</td>
<td>50.52</td>
<td>6.28</td>
<td>14.57</td>
</tr>
<tr>
<td>INTERNET. bs Corp.</td>
<td>5,786</td>
<td>4.32</td>
<td>54.83</td>
<td>0.70</td>
<td>15.27</td>
</tr>
<tr>
<td>Bizcn.com, Inc.</td>
<td>5,638</td>
<td>4.21</td>
<td>59.04</td>
<td>0.70</td>
<td>15.97</td>
</tr>
<tr>
<td>Trunkoz Technologies Pvt Ltd. d/b/a OwnRegistrar.com</td>
<td>4,577</td>
<td>3.41</td>
<td>62.45</td>
<td>0.05</td>
<td>16.02</td>
</tr>
<tr>
<td>PDR Ltd. d/b/a PublicDomainRegistry.com</td>
<td>3,595</td>
<td>2.68</td>
<td>65.13</td>
<td>3.08</td>
<td>19.10</td>
</tr>
<tr>
<td>OnlineNIC, Inc.</td>
<td>2,857</td>
<td>2.13</td>
<td>67.26</td>
<td>0.70</td>
<td>19.80</td>
</tr>
<tr>
<td>Center of Ukrainian Internet Names</td>
<td>2,781</td>
<td>2.07</td>
<td>69.33</td>
<td>0.06</td>
<td>19.86</td>
</tr>
<tr>
<td>Register.com, Inc.</td>
<td>2,540</td>
<td>1.89</td>
<td>71.22</td>
<td>2.18</td>
<td>22.04</td>
</tr>
<tr>
<td>GoDaddy.com, LLC</td>
<td>5,532</td>
<td>4.13</td>
<td>75.35</td>
<td>30.75</td>
<td>53.79</td>
</tr>
</tbody>
</table>

Table 3: The 10 registrars that registered the greatest number of spammer domains.
domains (i.e., with toxicity 100%) only account for about 10% of all spammer domains.

Table 4 lists the top nameservers associated with spammer domains, in terms of the total number of spammer domains that they host. We group servers together if their Jaccard index exceeds 0.95, to ensure grouping similarity; we use a regular expression to represent groups of similar domains. For nameservers in common groups, we calculate the metrics based on common domains. The second column shows the number of common spammer domains for each DNS server group; we rank the groups in descending order of the number of common spammer domains for that group. The third and fourth columns indicate the server toxicity and the Jaccard index, respectively. When spammer domains are sheltered in large registrars (like Moniker or eNom), these registrars provide and operate their authoritative DNS servers, also hosting a large number of legitimate domains.

It becomes clear from these results that although spammers prefer certain nameservers in some cases, no clear-cut separation exists between nameservers used for spamming and those used for benign purposes. Hence, our earlier hope fails to pan out: we do not see how to fruitfully leverage the nameservers associated with domain registrations to identify spammer domains. In the next section, we turn to exploring to what degree the patterns that spammer registrations exhibit can help distinguish spammer domain registrations from benign ones.

7. DETECTING REGISTRATION SPIKES

In this section we examine the extent to which spammers register domains in abnormally large batches (“spikes”). We first present evidence that suggests that domains associated with spamming are registered in groups. We then show that the number of domains that a given registrar registers in a given five-minute interval usually follows a distribution well-modeled by a compound Poisson process—but that many registrars also exhibit registration spikes that this process would produce only with exceedingly low probability. From this we conclude that these spikes represent a different underlying process than that corresponding to routine activity.

After deriving a model to explain both normal and anomalous registration activity, we show that spammers tend to register batches of domains in such spikes more often than do non-spammers. Our results suggest that methods to reliably identify registration spikes can serve as a useful feature for proactively detecting names that are subsequently used in spam campaigns.

7.1 Bulk Registrations by Spammers

Spammers acquire large volumes of domains to remain agile when conducting their operations [18]. We frequently observe spammer domains registered in spikes as large as hundreds of domains within a single five-minute .com update. We speculate that such registration behavior occurs due to: (1) convenience; (2) bulk pricing from registrars [11, 21]; or (3) the use of stolen credit cards to purchase a large number of domains in a short period of time, since the fraudulent purchases will trigger detection and result in voiding of the stolen credit card.

We use the term bulk registration to refer to the behavior of registering a batch of domains during a period of a few minutes. Since the registrants’ information and behavior are not directly observable in the zone updates, we can only infer that a group of domains may represent a bulk registration by observing updates with multiple domains within the same (registrar, 5-minute-epoch) tuple. The granularity of data that we have provides only an approximation of the behavior of each registrant because different registrants may register simultaneously from the same registrar, and the same registrant could spread their registrations across multiple registrars. Still, we observe the general tendency for spammers to perform registrations in batches, as we develop below.

Figure 8 shows the distribution of the number of spammer domains registered in the same epoch. The X-axis shows the number of spammer domains observed for a given registrar within a single epoch, and the Y-axis shows the cumulative percentage of spammer domains within such epochs. The inlay shows that 50% of the spammer domains were registered in groups of ten or more. We find that only 20% of the spammer domains got registered in
7.2 Detecting Abnormal Registration Batches

The evidence from the previous section suggests that spammers often register multiple domains at the same time. This phenomenon motivates us to identify a way to determine whether a registrar’s set of registrations during a given epoch is “abnormally large”. If so, then we posit that those registrations are more likely to reflect domain registration activity by spammers. We aim to identify registration activity behavior that qualitatively differs from routine (and thus, we presume, likely benign) activity. As noted above, we refer to such a set of registrations as a “spike”.

Developing a model for registration batch size. Our challenge is to determine that a given set of registrations crosses the line into “abnormally large”, thus constituting a spike. The difficulty we face is that simple approaches for spike detection can lack soundness. For example, simply setting a single threshold (registrations per epoch) may cause us to miss numerous spikes that appear in the registrations for some of the smaller registrars, since for those registrars, an abnormally large set of registrations might not be at all that large in terms of absolute volume. If, on the other hand, we instead pick a fixed quantile, such as “treat as spikes all registrations larger than the 99th percentile of a given registrar’s registration sizes”, then we will necessarily define a subset of each registrar’s activity as “abnormally large”—failing to capture the notion of a qualitative difference.

Instead, we strive to develop a principled approach to identify qualitatively different (abnormally large) registration epochs, as follows. We hypothesize that a single model can capture the bulk of a registrar’s registration activity (i.e., the distribution of how many names the registrar registers during each of its epochs). We then look for epochs during which, according to that model, the volume of names registered was exceedingly unlikely (in a probabilistic sense). We deem such epochs as qualitatively different, and classify the corresponding set of registrations as a spike.

The first question we face concerns what sort of model to use to capture regular registrar activity. If registrars receive a steady stream of customers who act independently of one another, and each registers a single name, then a Poisson process should capture the corresponding activity well: during each epoch, the registrar registers a number of names corresponding to the number of customers who arrived since the last epoch. For this model, all we need to identify is the rate at which the customers arrive at the registrar with their requests. However, we would expect that diurnal patterns would cause the arrival rate at each registrar to vary over the course of each day, and indeed from inspection we find that this is the case. We adjust for this consideration by separately computing for each registrar the mean number of names they registered for each hour of the day, analogous to the nonhomogeneous Poisson processes used previously in characterizing network traffic [23]. For example, we determine a registrar’s registration rate per epoch as the average over all epochs between 10 a.m. and 11 a.m., and then use that rate to parameterize a Poisson process to capture the number of registrations that we expect to occur in each such epoch.

We explored this simple Poisson model and found that while it works well for some registrars, for many registrars it often fails to produce convincing fits to the body of the distribution of names registered per epoch. This provides evidence that customers do not arrive at a registrar independently from one another, the rates at which they do vary significantly more rapidly than on a per-hour basis, and/or customers sometimes register more than a single name at a time in normal activity.

The first two of these possibilities appear somewhat at odds with how we expect users to function under normal circumstances, which leads us to consider instead the third option. By employing a compound Poisson process, we can capture customers who arrive independently at a fixed rate, but each of whom makes a number of registrations drawn from a given distribution. The general compound Poisson formalism does not require a particular family for this second distribution. However, we achieved quite good results by using a second Poisson distribution; we later discovered that previous work has also modeled consumer purchase behaviors using a compound Poisson process that employs a second Poisson distribution [26].

In summary, for normal registrar activity we capture the number of domain registrations per epoch as $Y = \sum_{i=1}^{N} X_i$, where $N$ represents the number of registrants during the epoch, and follows one Poisson distribution, and $X_i$ (1 ≤ i ≤ N) are i.i.d. Poisson distributions capturing the number of domains each registrant registers. For this model, we have:

$$E(Y) = E(N)E(X_i)$$
$$Var(Y) = E(N)[Var(X_i) + E(X_i)^2]$$
Figure 9: Compound Poisson processes fitted to the count of registrations per epoch for 4 registrars (hourly window, 10AM–11AM ET).

Fitting the distribution. Given this model, we now turn to how to fit it to a given registrar’s activity. (A reminder, we do this for each hour of the day separately, to accommodate diurnal patterns.) We need to estimate $N$’s parameter, $\lambda_N$, and that for the $X_i$, $\lambda_X$. Given they are Poisson distributions, we have: $E(N) = \lambda_N$, $E(X_i) = \lambda_X$, and $\text{Var}(X_i) = \lambda_X$, and therefore:

$$\lambda_X = \frac{\text{Var}(Y)}{E(Y)} - 1, \quad \lambda_N = \frac{E(Y)}{\lambda_X}.$$  

Of course, we cannot simply compute these estimates from each registrar’s registration process because our goal is precisely to try to identify registration events that do not conform to the registrar’s usual activity. For any given registrar, we do not know whether any of these qualitatively different events even exist, but we have strong confidence that they do exist for at least some registrars.

We thus refine the process of fitting the distribution based on the following intuition. Because the events we seek to detect reflect abnormally large registration batches, they will occur in the upper tail of the distribution of all of a registrar’s registrations. Therefore, for each registrar we progressively apply different truncation thresholds (proportion of the upper tail to discard) to see whether omitting extreme tail values provides us with a better fit of the remaining data to the compound Poisson process. Note that if all of the registration events indeed conform to the same compound Poisson process, then we would expect to do no better—and possibly a bit worse—as we discard upper tail events, since these in fact simply reflect the natural extremes of the process.

We use KL divergence to assess how well the model fits a truncated portion of a registrar’s activity: given two probability distributions $P$ and $Q$, the KL divergence of $Q$ from $P$ is $D_{KL}(P||Q) = \sum_i \log(\frac{P(i)}{Q(i)}) P(i)$, which captures the information lost when we use $Q$ to approximate $P$. The smaller the KL divergence, the better a model fits the data. For each registrar, we compute the KL divergence for all of its data versus that for a compound Poisson processes fitted to that data; the same but discarding the upper 99.5% tail; again, but discarding the upper 99.0% tail; etc., through the 90% tail (i.e., we discard the top 10% of largest registration events). We then take as the best fit the tail truncation (if any) that provides the lowest KL divergence.

To demonstrate the fitting results, Figure 9 shows the models and epoch distributions of the first 4 registrars listed in Table 3 regarding to the hourly window between 10 AM and 11 AM US Eastern Time.
 abilities, we propose defining a spike as a registration size with a probability threshold to define spikes. For example, if when truncating the upper 1% tail, we find that truncated model predicts probabilities for the points in the tail as 0.5%, then in fact those points are not so extreme, given the model, and we should not consider them as reflecting qualitatively different behavior. On the other hand, if those points have predicted probabilities of 0.005%, then they are quite unlikely, bolstering evidence that they represent a fundamentally different process.

In the next section, we refine the model by assessing each of these probabilities and explain how we make a final determination of whether a given registration epoch reflects a truly abnormal "spike".

7.3 Refining Threshold Probabilities

The compound Poisson model that we have derived enables us to assess the probability of observing a given number of domains registered by a registrar in a five-minute epoch. A low probability indicates a rare event; if the probability is sufficiently low, we can then conclude the presence of an “abnormal” spike. Figure 10 shows the number of domains that were registered in spikes depending on how low we set this probability; the X-axis is log-scaled. The blue dashed curve shows the proportion for all newly registered .com domains, and the red solid curve shows the same statistic for spammer domains. Spammer domains appear in spikes with a much higher likelihood.

We also observe that the slope of the curves increases significantly at a probability of $10^{-3}$, suggesting a modal change at that point. For this range of probabilities, the model incorporates spikes that arise simply due to stochastic fluctuations of the normal model, rather than reflecting qualitatively different registration behavior. To avoid mis-classifying registration events for this range of probabilities, we propose defining a spike as a registration size with probability $\leq 10^{-4}$. With that definition, we find about 15% of all domains were registered in spikes; in contrast, 42% of spammer domains were registered in such spikes.

8. Domain Registration Patterns

Spammers can potentially use different strategies to decide on which names to register for use in their campaigns. In this section we analyze the history of domains registered by spammers to assess the different approaches they use. We first define different types of registrations in terms of the domain life cycle discussed in §2. We then show that some types are significantly more likely than others to correlate with spammer domains. We finish with a look at the nature of the names spammers choose when creating new domains.

8.1 Domain Categories

The most basic property of a domain registration concerns whether the domain is brand-new, i.e., has never appeared in the zone before, and thus now gets registered for the first time. Such domains have no registration history. On the other hand, a re-registration reflects a name that previously appeared in the zone that the registrant now registers once more after its expiry from the previous owners. For re-registration domains we possess registration history, such as previous registrar(s), registration time(s) and deletion time(s).

We further characterize re-registered domains as either drop-catch or retread. The former refers to a domain re-registered immediately after its expiry, a phenomenon that occurs frequently [7]. Conversely, if some time elapses between a domain’s prior deletion and its re-registration, then we term it as a “retread”. Thus, the “drop-catch” and “retread” categories are mutually exclusive, and together comprise all members of the “re-registration” category.

8.2 Prevalence of Registration Patterns

How common is each registration pattern? We define a domain registration as drop-catch if the domain was deleted and re-registered in the same 5-minute epoch. If more time elapses for a re-registered domain, then we consider it a retread. Among the spammer .com domains that were registered over the 5 months, 68% were brand-new, 30% were retread, and 2% were drop-catch.

Which registrations are more likely to reflect spammer domains? To better understand the role of each type of registration in spamming activity, we investigate the conditional probability that a registration reflects a spammer domain, given a specific category of registration. For example, to calculate the conditional probability of observing a spammer domain given that the registration is a retread, we divide the count of domains that are both retread and spammer by the count of retread domains.

This procedure then gives us the conditional probabilities of being a spammer domain given that the registration is retread, drop-catch, or brand-new as 1.34%, 0.33%, and 1.01%, respectively. Retread and brand-new have higher conditional probabilities of reflecting spammer domains compared to drop-catch registrations. One possible explanation for this could be that spammers simply use drop-catch domains less often in their spam campaigns. Usually registrars charge higher prices to purchase drop-catch domains; for example, three major drop catching services—Namejet.com, Pool.com, and Snapnames.com—charge $59, $60, and $69, respectively, for drop-catch registrations, significantly higher than typical domain registration rates of around $8–12 per registration [9].
Thus, drop-catch registrations appear significantly less economical for spammers.

Interestingly, if we also condition on the registration event occurring in a spike, certain types of registrations become much more likely to reflect spammer domains. Figure 11 shows the conditional probability of observing a spammer domain given a specific category and registration occurring in a spike (as defined in §7). The X-axis indicates the probability thresholds in log scale. Each curve shows the conditional probability of a spammer domain for different spike detection threshold values. The red solid curve shows that the conditional probability of a spammer domain given that the registration is a drop-catch and appears in a spike reaches as high as 6%, significantly higher than the conditional probability of a given spammer domain being a drop-catch alone (1.34%). This observation indicates spammers are adept at finding previously used but expired domains and re-registering them in bulk. The black dashed curve shows the same statistic for brand-new domains; for this category, the conditional probability of a spammer domain roughly doubles when observing a spike. The highest conditional probability for spammer domains occurs around a threshold of 10^{-4}, which indicates that when spammers register new domains in bulk, the spikes may not be as large as they are for registration spikes for other categories of domains. This difference may arise from the difficulty of finding large numbers of brand-new domain names that are suitable for use in spam campaigns. Finally, the conditional probabilities for spammer domains occurring in drop-catch registrations are small and do not vary significantly depending on the detection threshold.

8.3 Retread Registration Patterns

We have seen that spammers commonly re-register expired domains, especially when performing bulk registrations. Information about domain expiration is publicly released via various channels [22, 29]; spammers, of course, have access to this information and appear to exploit it when selecting the domains to register for subsequent spam campaigns. The majority of the retread registrations that reflect spammer domains were deleted from the zone within 90 days, which indicates that spammers tend to select domains that have expired recently (although not so recently as to qualify as drop-catch domains). We now examine the registration patterns of retread domain registrations in more detail.

Do spammers perform reconnaissance to determine whether a re-registered domain has been previously blacklisted? We study whether the re-registration that reflect spammer domains have typically appeared in spammer activity in the time period before the spammer decided to re-register the domain. This analysis allows us to better understand whether spammers specifically aim to re-register expired domains with clean histories. We use both the blacklist reports and spam trap observations from the preceding five months—from October 2011 to February 2012—as our source of historical information about spammer domains. (The SURBL blacklist data was only available for October and November 2011, but we also use it for historical information about spammer domains.) Only 6.8% of the retread registrations during March–July 2012 that reflected spammer domains had ever previously appeared in a blacklist, which suggests that spammers indeed deliberately re-register expired domains with clean histories.

How long are retread domains dormant between periods of registration? Next, we investigate the amount of time that typically elapses between domain expiration and a re-registration. The distribution of the dormancy periods for re-registration that do not reflect spammer domains is much more uniform than the distribution for spammer domains, which tend to be reused more quickly after they expire. Figure 12 shows a cumulative distribution of the dormancy period for re-registration; more than 65% of spammer domains were dormant for less than 90 days. If we condition on re-registration that reflect spammer domains for less than three months, and registered in moderately sized spikes (according to a threshold probability of 10^{-4}), the conditional probability of a re-registered domain being a spammer domain is 7.7%, again significantly higher than the conditional probability of being a spammer domain based on a re-registration alone (1.34%).

8.4 Naming Patterns for Brand-New Domains

We now study the naming conventions that spammers use when registering brand-new domains, focusing in particular on such do-
mains registered in spikes (once again using $10^{-4}$ as the threshold probability for defining a spike). We first compare the proportion of spammer and non-spammer domains that are registered in spikes that contain English words. To do so, we compare the domain names against a dictionary, looking for matches against words of at least four letters. We find that about 84% of the brand-new spammer domain registrations occurring in spikes contain an English word, versus about 82% of such non-spammer registrations. Thus, it appears that spammers create names that for the most part will appear plausible, as opposed to employing simple algorithms to crank out gibberish. Perhaps spammers seek to avoid detection by domain reputation algorithms that use entropy as a feature (e.g., [3]), or aim to diminish user suspicions and increase the likelihood that users will visit the corresponding website.

We hypothesized that when spammers register new domains in bulk, that they may register domains that represent various combinations of English words that relate to the campaign itself, perhaps with slight variations (e.g., one might expect a spam campaign involving watches to involve the registration of many domains containing the word “watch”). To test this hypothesis, we counted the number of brand-new domains in the same registration spike that share a common word, again considering only English words that are at least four characters long. Figure 13 shows the results of this analysis; many domains in the same spike share no common English words, yet the spammer domains show a slightly higher tendency to have common words in spikes. For example, about 40% of brand-new domains that appear in the same spike contain a common subword overall, yet slightly more than 50% of brand-new spammer domains contain a common subword when they appear in the same registration spike.

9. SUMMARY

In this work we have analyzed the domain registration behavior of spammers, including both the infrastructure that they use to register their domains and the patterns that they exhibit when registering them. Our motivation in exploring these behaviors is ultimately to facilitate _time-of-registration_ detection of such domains, enabling proactive blocking. We found that nearly half of spammer domains are less than 3 months old; spammer domains are often only used for short periods of time; and current blacklists (with the exception of Spamhaus DBL) identify spammer domains at _time-of-use_ rather than _time-of-registration_.

We based our study on a large, fine-grained dataset that reflected all changes to the .com zone over a five-month period, as seen during five-minute intervals. We then analyzed this data in conjunction with several spam trap and blacklist feeds as _post facto_ indicators of spammer domains. After confirming the previous finding that just a handful of registrars account for the bulk of spammer domain registrations, we examined the _registration process_ of each registrar, finding two distinct types of registration activity. In the first, predominant mode, the number of domains registrars register is well-described by a compound Poisson process. By fitting such a process to the bulk of a registrar’s registration epochs, we can associate probabilities with “outlier” epochs that register large numbers of domains, allowing us to identify registration spikes that _qualitatively differ_ from the registrar’s usual registration practices. We then showed that spammers often register their domains in such spikes, whereas non-spammers do so much less frequently.

Spammers also often prefer to _re-register_ domains that previously existed in the zone but subsequently expired. While spammers do not engage in “drop-catching” (immediately re-registering domains that have just expired), they prefer domains that have recently expired (within the past few months) and that in their previous life did not appear to be associated with spamming.

We also analyzed two other _time-of-registration_ features: (1) the degree to which spammers tend to use distinctive nameservers to host their domains, and (2) whether newly registered spammer domains contain common English words. We did not find much discriminative power for either of these features.

Other than a couple of particularly abuse-prone registrars—which by themselves do not account for a significant portion of spam domains—none of the _time-of-registration_ features that we examined by themselves serve as a “smoking gun”. Nevertheless, many features exhibit different behavior for spammer domains versus non-spammer domains, suggesting that an apt application of machine learning may enable the development of an accurate time-of-registration detector that can enable us to nip spammer domain registrations in the bud.

Acknowledgments

This work was supported in part by the U.S. Army Research Office under MURI grant W911NF-09-1-0553, by the Office of Naval Research under MURI grant number N000140911042, and by the National Science Foundation under grants 0831535, 1111723, and 1237265. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES


Figure 13: Cumulative percentage of (brand-new) domains that appear in the same registration spike and have at least one English subword in common with another domain in the spike.


