Cybercrime After the Sunrise: A Statistical Analysis of DNS Abuse in New gTLDs

Maciej Korczyński, Maarten Wullink, Samaneh Tajalizadehkhoob, Giovane C. M. Moura
Arman Noroozian, Drew Bagley, and Cristian Hesselman
maciej.korczynski@univ-grenoble-alpes.fr

ABSTRACT

To enhance competition and choice in the domain name system, ICANN introduced the new gTLD program, which added hundreds of new gTLDs (e.g. .nyc, .io) to the root DNS zone. While the program arguably increased the range of domain names available to consumers, it might also have created new opportunities for cybercriminals. To investigate that, we present the first comparative study of abuse in the domains registered under the new gTLD program and legacy gTLDs (18 in total, such as .com, .org). We combine historical datasets from various sources, including DNS zone files, `WHOIS` records, passive and active DNS and HTTP measurements, and 11 reputable abuse feeds to study abuse across gTLDs. We find that the new gTLDs appear to have diverted abuse from the legacy gTLDs: while the *total* number of domains abused for spam remains stable across gTLDs, we observe a growing number of spam domains in new gTLDs which suggests a shift from legacy gTLDs to new gTLDs. Although legacy gTLDs had a *rate* of 56.9 spam domains per 10,000 registrations (Q4 2016), new gTLDs experienced a rate of 526.6 in the same period–which is almost one order of magnitude higher. In this study, we also analyze the relationship between DNS abuse, operator security indicators and the structural properties of new gTLDs. The results indicate that there is an inverse correlation between abuse and stricter registration policies. Our findings suggest that cybercriminals increasingly prefer to register, rather than hack, domain names and some new gTLDs have become a magnet for malicious actors. ICANN is currently using these results to review the existing anti-abuse safeguards, evaluate their joint effects and to introduce more effective safeguards before an upcoming new gTLD rollout.

KEYWORDS

Security Metrics, DNS, Top-Level Domains, Registrars, Cybercrime

1 INTRODUCTION

Starting in 2007, The Internet Corporation for Assigned Names and Numbers (ICANN) introduced the new Generic Top-Level Domain (gTLD) program[1], which enabled hundreds of new gTLDs to enter the domain name system (DNS) since the first delegations. More than 1,900 applications for new gTLDs were filed after the process opened in 2012. To date, more than 1,200 new gTLDs have been delegated to the DNS root zone (e.g.: .nyc, .io). This expansion of the domain name space not only offers a wide range of options for consumers, but also potentially provides new avenues for cybercriminals to abuse domain names. Anticipating potential problems, ICANN has also built safeguards into the program in an attempt to mitigate the prospect of abusive, malicious, and criminal activity in these new gTLDs, such as phishing, spam, and malware distribution[2].

In a previous study, Halvorson *et al.* [11] concluded that speculative and defensive registrations dominate the growth of registrations in new gTLDs. Their work, however, provides very little empirical information about the security of new gTLDs. In this paper, we investigate the following research question: *how do abuse rates in the new gTLDs compare to legacy gTLDs, since the implementation of the new gTLD program?* We take into account the new gTLDs as well the different parts of the industry involved: registries, registrars, and privacy/proxy service providers.

To this end, we combine multiple datasets from various sources including zone files, domain name `WHOIS` records, data obtained through active measurements, and 11 abuse feeds provided to us by 5 reputable organizations. These represent malware, phishing, and spam abuse and cover a three-year period from 2014 to 2016.

Overall, our main contributions can be summarized as follows:

- The research offers a comprehensive descriptive statistical comparison of rates of domain name abuse in new and legacy gTLDs as associated with spam, phishing, and malware distribution (§5.1) to evaluate joint effects of the existing anti-abuse safeguards.
- Using regression modeling we perform inferential statistical analysis to test the correlation between passively and actively measured properties of new gTLDs as predictors of abuse rates (§5.2).
- We analyze proportions of abusive domains across other entities relevant to abuse prevention practices, i.e. registrars and privacy/proxy providers (§5.3 and §5.4).

Our findings reveal surprising, previously unknown trends that are relevant since new gTLDs operate on the basis of different business models and history in comparison to legacy gTLDs. While patterns of abuse vary with respect to abuse type, our analysis suggests that the total number of spam domains in all gTLDs remains relatively constant. Simultaneously, the number of spam domains in new gTLDs is higher (Q4 2016) and growing. We also observe a significant decrease in the number of malicious registrations in

[1] https://gnso.icann.org/en/group-activities/inactive/2007/new-gtld-intro

[2] https://newgtlds.icann.org/en/reviews/cct/dns-abuse

ASIA CCS '18, June 4–8, 2018, Incheon, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5576-6/18/06…$15.00
https://doi.org/10.1145/3196494.3196548

legacy gTLDs (§5.1.7). Therefore, we see a new trend: attackers switching from abusing legacy to new gTLD domain name space. Our analysis of the Spamhaus blacklist also reveals that in the last quarter of 2016, new gTLDs collectively had approximately one order of magnitude higher rate of spam domains per 10,000 registrations compared to legacy gTLDs (§5.1.8).

This research also systematically analyzes how different structural and security-related properties of new gTLD operators influence abuse counts. Our inferential analysis reveals that abuse counts inversely correlate with the restrictiveness of registration policies (§5.2). The analysis of abuse across new gTLDs, registrars, and privacy/proxy service providers reveals discrete entities afflicted with significantly high concentrations of abused domains. We find new gTLDs and registrars with concentrations of blacklisted domains above 50% (§5.1.8 and §5.4). For one registrar, more than 93% of its domains were reported as abusive by SURBL.

ICANN may further expand the number of gTLDs available. Therefore, it is important to understand how miscreants are using the expanded domain name space in their favor. Finally, as the presented state of the art in gTLD abuse is in clear need of improvement, we develop cases for modifying the existing safeguards and proposing new ones. This work is a corollary to our analysis that ICANN is currently using to review the existing anti-abuse safeguards, evaluate their joint effects and to introduce more effective ones before an upcoming new gTLD rollout [28].

2 BACKGROUND

The Internet Domain Name System (DNS) comprises one of the critical services of the Internet, mapping hosts, applications, and services from names to IP addresses [34]. ICANN [21] is the organization responsible for maintaining the Root domain namespace and its expansion with new top-level domains, in particular new gTLDs. ICANN also delegates the responsibility to maintain an authoritative source for registered domain names within a TLD to registry operators (e.g.: Verisign is the registry for .com). Registries, manage themselves and the domain names under their respective TLDs.

Three main entities are involved in the registration of a domain: registries (aforementioned), registrars, and registrants (so-called tripe-R). A registrant is a user or company, which in turn has to contact a registrar to register a domain name. A registrar (e.g.: GoDaddy), if affiliated with the TLD of the registrant's choice, will ask the registry to perform the registration of the requested domain.

In parallel, web hosting providers maintain server infrastructure that is used to host content for the domain. DNS providers operate authoritative DNS servers that resolve domain names to their corresponding IP addresses. Finally, WHOIS Privacy and Proxy service providers conceal certain personal data of domain name registrants.

2.1 Generic TLDs

The first group of generic top-level domains (gTLDs) was defined by RFC 920 [37] in October 1984 and introduced a few months later. The initial group of gTLDs (.gov, .edu, .com, .mil, .org, and .net) were distinct from country-code TLDs (ccTLDs). Until 2012, several gTLDs were approved and further introduced by ICANN, including a set of sponsored gTLDs such as .asia, .jobs, .travel, or .mobi. We refer to all gTLDs introduced before the new gTLD program initiated by ICANN in late 2013 as *legacy* gTLDs. This study analyzes a set of 18 legacy gTLDs (.aero, .asia, .biz, .cat, .com, .coop, .info, .jobs, .mobi, .museum, .name, .net, .org, .post, .pro, .tel, .travel, and .xxx), for which we were able to obtain zone files and WHOIS data, and compare them to *new gTLDs*.

2.2 New gTLDs

ICANN's new gTLD program began in 2012, expanding the root zone by delegating more than 1,200 new gTLDs starting in October 2013 [23]. To obtain a new gTLD, applicants are required to undergo an intensive application and evaluation process [11] that includes screening the applicants technical and financial capabilities for operating a new gTLD. Ultimately, after a new gTLD is assigned to an applicant, it will then be delegated to the root zone. Following initial delegation, each new gTLD registry is required to have a "sunrise" period of at least 30 days, during which trademark holders have an advance opportunity to register domain names corresponding to their marks, before the names are generally available to the public.

New gTLDs can be classified into four broad categories [23][3]:

- *Standard or generic gTLD*: gTLDs that are generally open for public registration, e.g. .movie, .xyz, or .family[4]
- *Geographic gTLD*: gTLDs that cover cities, states, or regions, e.g. .amsterdam or .berlin.
- *Community gTLD*: gTLDs that are restricted to a specific community, such as .thai, .radio or .pharmacy.
- *Brand gTLD*: gTLDs specific to a company or a brand, such as .google or .hitachi.

In our study, we analyze new gTLDs that are intended for public use. Therefore, we exclude the great majority of brand gTLDs for which domains cannot be registered by regular users[5], in particular for malicious purposes. We cover new gTLDs for which registries have submitted their sunrise date information requested by ICANN. In the first quarter of 2014, there were 77 new gTLDs for which the sunrise period ended and domain names were available for public registration. For comparison, by the end of 2016 the group consisted of 522 new gTLDs.

2.3 Safeguards Against DNS Abuse

In preparation for the new gTLD program, ICANN sought advice from different DNS abuse and security experts. As a result of broad discussion with multiple stakeholders such as Anti-Phishing Working Group (APWG), Registry Internet Safety Group (RISG), the Security and Stability Advisory Committee (SSAC), Computer Emergency Response Teams (CERTs), members of the financial, and Internet security communities, ICANN proposed 9 safeguards [16, 19].

The first safeguard mandated that all new gTLD registry operators provide descriptions of the technical back-end services to ensure their technical competence (vetting registry operators). The second safeguard requires all new gTLD registries to implement DNSSEC at the root level. The third safeguard prohibits domain

[3] Note that some gTLDs cross categories. For example, some community gTLDs such as .madrid are also geographic gTLDs [18].

[4] While most of these gTLDs are open to public registration, some may registries impose restrictions on the entities that can register domains.

[5] With a few exceptions such as .allfinanz or .forex brand gTLDs for which the sunrise period has been announced and ended.

wildcarding to ensure that domains resolve for an exact match and do not redirect users for non-existent domain names. The fourth safeguard requires new gTLD registries to remove orphan glue records when it is proved that such records have been used in malicious activity. For the fifth safeguard, operators have to create and maintain "Thick WHOIS" records, i.e. complete WHOIS information from all the registrars on all domains coresponding to a given new gTLD. New gTLD operators are also required to make their zone files available to approved requestors via the Centralized Zone Data Service (CZDS)[6]. The agreement mandates all new gTLD registry operators to document abuse contact details for registries and registrars on their websites. The agreement also obliges operators to respond to security requests to address security threats but do not define specific procedures for doing so. The ninth safeguard proposed to create a framework for a "high security zone verification program", however, due to a lack of consensus this safeguard has never been implemented.

The role of safeguards in the new gTLD program is critical since a broadened domain name space creates new opportunities for cybercriminals. The majority of the existing safeguards however, may not directly prevent domain abuse. For example, DNSSEC is intended to increase the security of the Internet by adding authentication to DNS resolution to prevent attacks such as DNS spoofing [9] rather than, for example, preventing legitimate domains from being hacked. We agree that making the zone files of new gTLDs open to security research may indirectly contribute to improving security of new gTLD domain space. It does not, however, prevent miscreants from registering domains for malicious purposes.

As it may be difficult to statistically measure the effects of the existing safeguards individually, we opt for a rigorous approach to assess their joint effects on domain abuse rates.

2.4 Related Work

Numerous studies have looked into discovering, predicting, or explaining abuse across the DNS ecosystem [5, 6, 13, 25, 26, 39, 40, 47]. In addition to those, there are other studies that investigated domain re-registrations patterns and their relation with domain abuse [14, 30, 31, 33]. For example, Lever *et al.* studied the maliciousness of domains before and after re-registration with a focus on when malicious behavior occurs. Their findings showed hundred thousands of expired domains that were maliciously re-registered [31].

When it comes to quantifying the impact of specific factors that influence security of gTLDs, in particular new gTLDs, there exists very little empirical work. Rasmussen and Aaron regularly release APWG Global phishing reports in which they examine phishing datasets collected by APWG and several other supplementary phishing feeds. Recently, they concluded that phishing in the new gTLDs is rising but is not yet as pervasive as it is in the domain space as a whole [2]. Halvorson *et al.* found that new gTLD domains are more than twice as likely as legacy TLDs to appear on a domain blacklist, within their first month of registration [11]. Vissers *et al.* studied large-scale malicious campaigns in the .eu TLD for a period of 14 months and observed that 80% of the malicious registrations are part of just 20 long-running campaigns. Moreover, out of all domains operated by these campaigns, 18% never appeared on any blacklist [53].

Previous literature highlighted the importance of reliable security metrics to estimate abuse rates across network players in the domain ecosystem such as hosting providers or Autonomous Systems [36] and discussed specific factors that can influence this concentration of abuse [35, 43, 48, 49]. For the case of TLD operators, Korczyński *et al.* designed security metrics to measure and benchmark entire TLDs against their market characteristics [27]. They found that next to TLD size, abuse primarily correlates with domain pricing (free versus paid registrations), efforts of intermediaries (measured through the proxy of their DNSSEC deployment rate), and strict registration policies [27].

We build on the existing work in several ways. First, we analyze and compare the distribution of abuse across new and legacy gTLDs. Next, we make the first attempt to develop a comprehensive approach that can statistically quantify the impact of operator security indicators along with the structural properties of new gTLDs on DNS abuse rates.

3 MEASUREMENT DATASETS

In this section we cover six types of datasets used in this research: abuse feeds, WHOIS records, DNS zone files, active web scans, DNS scans, and passive registry data.

3.1 Abuse Feeds

To assess the prevalence of maliciously registered[7] and compromised domains[8] per gTLD and registrar, we use 11 distinct abuse feeds. These represent malware, phishing, and spam abuse and have been generously provided to us by Spamhaus [42], APWG [4], StopBadware [44], SURBL [45], the Secure Domain Foundation (SDF) [10], and CleanMX [7]. All six reputable organizations provide abused domain or URL data feeds employed in operational environments. **Spamhaus** data contains domains with low reputation collected from spam payload URLs, spam senders and sources, known spammers, phishing, virus, and malware-related websites [41]. **APGW** contains black/white listed phishing URLs submitted by accredited users through the eCrime Exchange (eCX) platform. The **StopBadware (SBW)** feed consists of abusive URLs shared by ESET, Fortinet, and Sophos security companies, Google's Safe Browsing appeals results, the StopBadware community, and other contributors [38]. **SURBL ph** is a phishing domain blacklist comprised of data supplied by among others MailSecurity, PhishTank, OITC phishing, PhishLabs, US DHS, NATO [46]. The **SURBL jp** blacklist contains domains analyzed and categorized as spam (e.g. unsolicited) by jwSpamSpy software, traps, and participating mail servers. **SURBL ws** contains mainly spam domains from SpamAssassin, the Anti-Spam SMTP Proxy, as well as information from other data sources including internal and external trap networks. The **SURBL mw** feed contains data from multiple sources that cover malicious domains used to host malware websites, payloads or associated redirectors [46]. The **SDF** contains domains and URLs classified as phishing or malware. The domain names were queried against the SDF's Luminous API which aggregates data from open

[6] https://czds.icann.org/en

[7] Domains registered by miscreants for the purpose of malicious activity

[8] Domains exploited using vulnerable web hosting

source blacklist feeds and registrar suspension lists [10]. Note that unlike the other data feeds the SURBL and SDF feeds cover the 2,5-year study period between July 2014 and December 2016. Finally, **CleanMX** contains three URL blacklists identifying phishing, malware websites, as well as a "portals" category that contain defaced, spamvertized, hacked, and other types of abused websites. Table 1 shows the number of unique blacklisted 2nd-level domain names per feed. In Appendix A, we further discuss the overlap among blacklists.

Note that some of the aforementioned feeds contain data at the URL level while others at the domain level. The distinction is important from an operational level. While some domain names that appear in *URL blacklists* are registered by miscreants for malicious purposes only, the majority of domain names are compromised domains, i.e. they were registered by legitimate users and hacked (see e.g.: phishing survey [1]). From the operational point of view blocking the domain name element of a blacklisted URL might harm legitimate operations. With this in mind, Spamhaus and other data providers maintain *blacklists of domain names* and perform extensive checks to prevent legitimate domain names from being listed. Therefore, the domain blacklists can be used by production systems to, for example, block emails that contain malicious domain names. In this paper, we refer to both domain names that appear in the domain blacklists and as part of blacklisted URLs as "abused domains" or "blacklisted domains".

3.2 WHOIS Data

Most of the abuse feeds used for this study contain no additional domain name attributes such as registrar name or date of registration. We obtained these attributes via WHOIS databases covering the 3-year study period provided by Whois XML API [54] and DomainTools [51]. These databases contain WHOIS information for the domains of the aforementioned 18 legacy gTLDs and for the domain names of the 1,196 new gTLDs that had been delegated during our study period [24].

We extract `<domain, registrar name>` tuples from WHOIS data and use these in conjunction with our abuse feeds to map domain names or the domain element from abused URLs to a sponsoring registrar. The registrar name is used to determine the amount of abuse related to the registrar. We also extract the `<domain, creation date>` tuples and use these to determine if the domain has been maliciously registered or compromised.

3.3 DNS Zone Files

The sizes of gTLDs vary significantly. In order to provide a fair comparison criteria across gTLDs, we need to take into account their size, i.e., the number of domains registered. To do that, we processed daily zone files (containing all domains for each gTLD on a given date) for the 3-year study period. The rate of abuse, i.e. X number of blacklisted domains over the Y number of total registrations, provides a more fair comparison criteria across gTLDs.

To give an idea of this difference, we show in Figure 1 a time series of unique domain names under legacy and new gTLDs. As can be seen, the legacy gTLDs still account for the majority of registrations (160.9M vs 24.5M in Q4 2016).

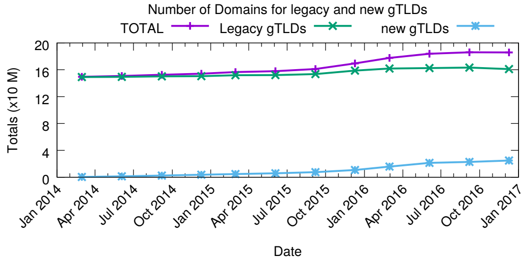


Figure 1: Zone file sizes for legacy and new gTLDs

We also relied upon zone files to determine the number of DNS Security Extensions (DNSSEC)-signed domains for each gTLD. One of the new gTLD program safeguards requires that all new gTLD applicants have a specific plan for DNSSEC deployment [19]. We used this data in our inferential analysis (see §5.2). Using regular expressions we matched `DS` records in the zone files and counted the distinct number of domains with `DS` records. The `DS` record is kept in the parent (TLD) zone and is used to prove the validity of cryptographic DNSSEC chain. Presence of a `DS` record indicates that the domain supports DNSSEC.

3.4 Active Web Scan

Using our web measurement platform, we crawled each new gTLD domain found in the zone files generated on May 2, 2017 (24,2M domains). We crawled these domains to determine how many are active and hosting content (see §4.2 for more details). The number of legacy gTLD domain names proved too voluminous to scan for this study. Therefore, we created a representative sample of 16,7M domain names (from the same date) to scan, using stratified sampling. A domain was considered non-responsive, if fetching www.example.com or example.com respectively, returned an error. If our crawler detects a redirect in either the retrieved HTML code or the HTTP headers then these redirects are followed. Any domain resulting in a crawl chain of more than 5 redirects is also marked as non-responsive.

The crawler is designed to have a minimal impact on the servers that are crawled. For this reason, only the main page is retrieved. The data captured for each domain includes the HTML code, HTTP headers and status codes. To determine if a domain is parked, the HTML code is analyzed using pattern matching to search for strings, which might indicate that the domain is for sale. The crawler also looks for URLs that are linked to known parking service providers.

3.5 Active DNS Scan

During the domain scan process, we also queried the DNS system to retrieve the `A`, `AAAA` and `SOA` records for each domain to detect active domains serving content (see §4.2). The DNS crawler sends queries to a dedicated instance of the unbound DNS resolver to check whether domains resolve. Moreover, the `SOA` record is indicative of whether the primary authoritative name server for the domain is linked to a known parking services provider.

3.6 Passive Data for Registries

In this study, we analyzed new gTLDs whose domain names became available for public registration within the study period. The time

Table 1: Overview of blacklists: unique blacklisted gTLD domain names for the StopBadware Data Sharing Program (SBW DSP), APWG, Spamhaus, SDF, CleanMX, and SURBL datasets for 2014, 2015, 2016.

Year	SBW	APWG	Spamhaus	SDF	CleanMX ph	CleanMX mw	CleanMX pt	SURBL ph	SURBL mw	SURBL ws	SURBL jp
2014	403,347	60,681	1,901,970	41,094	68,523	169,237	205,051	68,208	289,664	1,229,698	1,484,807
2015	501,982	139,538	2,505,407	142,285	98,112	117,140	124,608	134,591	220,073	1,813,858	2,475,745
2016	502,579	83,215	3,944,684	110,687	138,869	149,632	68,413	173,326	106,819	2,023,178	2,442,592

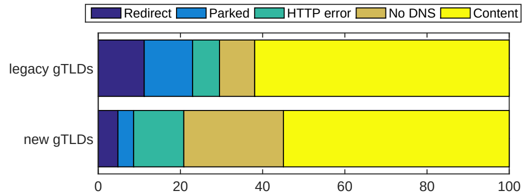


Figure 2: Categorization for all domains in the new TLDs and a random sample of the legacy TLDs.

between the delegation of a new gTLD and the end of its sunrise period might take several months[9]. Consequently, our analysis includes new gTLDs after their respective sunrise periods. This data, provided by ICANN via their public portal [24], contains 522 new gTLDs with sunrise periods ending during the timeframe of the study.

We also used a list of registry operators, their affiliates, and associated new gTLDs provided to us by ICANN. We mapped gTLDs to related registry operators regardless of which name they were operating under. We used the mapping of parent companies of registry operators and the corresponding new gTLDs in our inferential analysis as a proxy for registration practices.

Relying upon ICANN's categorizations of new generic, community, geographic, and brand gTLD registry applications, we conducted an inferential analysis on registration restrictions. We assigned registration "levels" to new gTLDs, from the least to most restricted groups: 1 generic, 2 geographic, 3 community, and 4 brand. Intuitively, while generic gTLDs are normally unrestricted and open for public registration, registration policies of community or brand gTLDs are strict and may prevent miscreants from malicious registrations.

4 METHODOLOGY

4.1 Security Metrics

To determine the distribution of abusive activities across the gTLDs and registrars, we analyze the occurrence of *unique abused domains*. Previous research has also proposed two complementary security metrics, i.e. the number of *unique fully qualified domain names* (*FQDNs*) and *unique blacklisted URLs* aggregated by TLDs [27]. However, due to space constrains, we do not present our results for such additional metrics and refer the reader to the related work [28].

4.2 Size Estimate of TLDs

In order to have a fair comparison criteria, we normalized the number of reported domains from blacklists (Table 1) by the size of their respective TLD. We calculated the size of each gTLD by counting the number of 2nd-level domains present in a zone file for each gTLD at the end of an observation period. We used zone files as they are the most accurate source to determine gTLD sizes. An alternative would be to use the ICANN monthly reports that summarize domain activity for all registered domains [22]. This would however result in an over counted gTLD size since some registrants register domain names but do not associate them with name servers.

The TLD size can also be used as an explanatory factor for the concentrations of abused domains, as indicated in the previous research [27, 36, 49]. However, it is unclear what portion of the domain names are in use and serve content. Halvorson *et al.* have shown that in 2015 as many as 16% of domain names in new gTLDs with NS records did not resolve [11]. Using our Web and DNS crawling platform, we performed a new scan and classified each domain as belonging to one of five groups: *i) No DNS* domains that do not resolve when queried by our DNS crawler, *ii) Parked* domains that are owned by parking services, advertisement syndicators, and advertisers. We follow the classification methodology outlined by Vissers *et al.* [52], *iii) HTTP Error* domains for which authoritative name servers return valid responses but the corresponding websites do not return an HTTP 200, *vi) Redirect* domains are redirected to a different domain, and *v) Content* domains that serve valid Web content.

Figure 2 shows the categorization results for all domains in the new gTLDs and a random sample of the legacy gTLDs. Interestingly, there is a significant increase in erroneous domains in the new gTLDs ("No DNS" and "HTTP Error" categories) as compared to legacy gTLDs. "No DNS" domains account for about a quarter of all domains (24.2%), whereas domains for which the corresponding websites serve an HTTP error account for another 12.2%.

Note that we use this measurement data in the inferential analysis to adjust measured TLD sizes. Intuitively, only the domains serving content are exposed to certain types of vulnerabilities and can be hacked. On the other hand, parked domains may be used to scam users or to distribute malware. One might therefore expect a positive correlation between the number of parked and maliciously registered domains.

4.3 Size Estimate of Registrars

Since we are interested in comparison between registrars, we calculated their sizes from the WHOIS data by counting the number of distinct domain names linked to each registrar name. Note that the WHOIS data may contain multiple name variants for a single registrar. E.g.:, GoDaddy is listed as a registrar using 52 variations, such as "GODADDY.COM, LLC", "GoDaddy.com, LLC (R91-LROR)" and "GoDaddy.com, Inc.". Therefore, we need an additional entity

[9] E.g. delegation of .zuerich: December 25, 2014 [20], zone file seen for the first time: January 1, 2015, sunrise period termination: June 5, 2017 [24]

resolution step to group together all the different registrar name variants as a single registrar.

We also used the IANA Registrar ID, which is assigned to ICANN accredited registrars [15]. We automatically matched the list of registrar names against names found in the WHOIS data. Then, we manually mapped the remaining registrar variants. To determine the amount of abuse related to a registrar, we mapped each domain name found in an abuse feed to its respective registrar using the WHOIS records with the closest enclosing time-window.

4.4 Compromised Versus Maliciously Registered Domains

Miscreants can both register or compromise and abuse legitimate domains. To distinguish between compromised and maliciously registered domains, we build on three heuristics previously used in domain abuse surveys (e.g. phishing survey by Aaron and Rasmussen [2]). More specifically, we label a domain as maliciously registered if it was involved in criminal activity within a relatively short time after its registration or if it contains a brand name or a misspelled variant of brand name. We refer the reader to Appendix B for more details on the methodology used in our study.

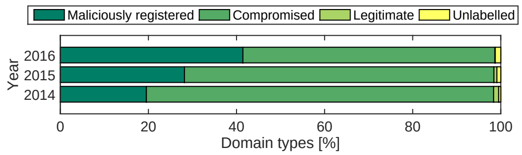


Figure 3: Fraction of maliciously registered, compromised, legitimate, and unlabelled domains for the APWG feed.

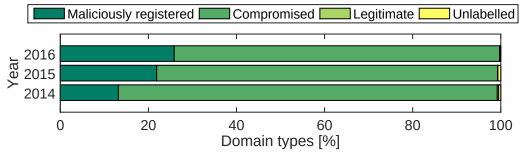


Figure 4: Fraction of maliciously registered, compromised, legitimate, and unlabelled domains for the SBW feed.

Figure 3 and Figure 4 show the categorization of domains blacklisted by APWG and SBW respectively during the study period (2014, 2015, and 2016). Note that up to 1.1% of all domains submitted to the APWG have been pre-filtered based on the maintained list of domains corresponding to legitimate services and labeled as "legitimate". For comparison, we have excluded less than 0.3% of the StopBadware domains. A previous study showed that domains of legitimate services are often misused by miscreants to distribute malware or used in phishing campaigns [27]. However, some may also represent legitimate domains that were incorrectly blacklisted.

The results indicate that 78.8% of abused phishing and 86% of malware domains (listed on URL blacklists in 2014) were compromised by criminals (see Figure 3 and Figure 4). In 2016, those percentages were smaller: 57.2% and 73.9% of phishing and malware domains were labeled as compromised. Although domains listed in URL blacklists are predominantly compromised, their number has been gradually decreasing. Instead, miscreants are registering domain names more often. We find that 19.5%, 28.2%, 41.5% and 13.2%, 21.9%, 25.8% (in 2014, 2015, and 2016) of all phishing and malware domains respectively were presumably maliciously registered by miscreants. This trend suggests a shift in the behavior of miscreants that over time seem to prefer registering rather than compromising legitimate domains.

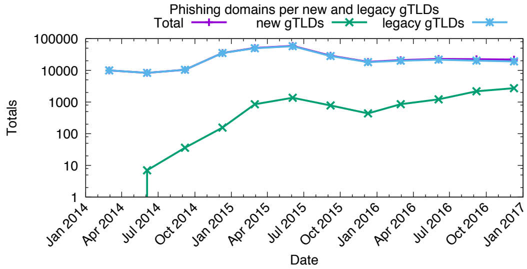


Figure 5: Time series of counts of phishing domains in legacy gTLD, new gTLDs, and all gTLDs (Total) based on the APWG feed (2014-2016, Table 1). Log scale on y axis.

5 RESULTS

5.1 TLD Reputation

5.1.1 Phishing Abuse. Figure 5 plots a time series of the number of phishing domains for new gTLDs, legacy gTLDs, and a "Total" number for our 2014–2016 study period based on data from the APWG feed. We aggregate phishing incidents on a quarterly basis and present counts using a logarithmic scale. We observe that the total number of phishing domains (purple line) overlaps largely with the number of phishing domains in legacy gTLDs. This phenomena is due to the disproportionate market share of domain names registered in legacy gTLDs. While the number of abused domains remains relatively constant in legacy gTLDs, we observe a clear upward trend in the absolute number of phishing domains in new gTLDs. We observe similar trends in SURBL phishing and CleanMX phishing datasets (which have been omitted due to space constraints).

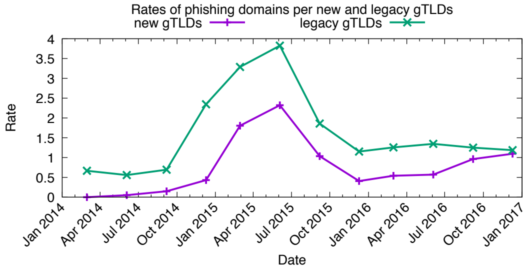


Figure 6: Time series of abuse rates of phishing domains in legacy gTLDs and new gTLDs based on the APWG feed (2014-2016). $Rate = 10,000 * \#blacklisted\ domains/\#all\ domains$.

5.1.2 Normalized Phishing Counts. As previously discussed, reliable reputation metrics must account for market shares (i.e. size) as larger market players may experience a higher amount of domain abuse. Figure 6 shows a time series of abuse rates of phishing domains for legacy gTLDs and new gTLDs based on the APWG

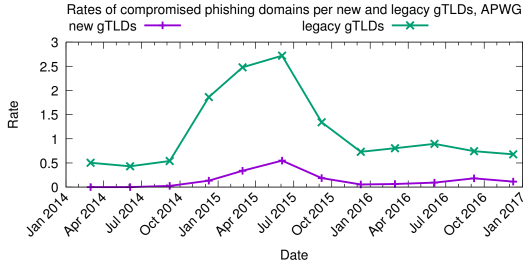


Figure 7: Time series of abuse rates of *compromised* phishing domains in legacy gTLDs and new gTLDs, the APWG feed.

feed (due to space limitation we do not present figures related to abused CleanMX phishing and SURBL phishing domains).

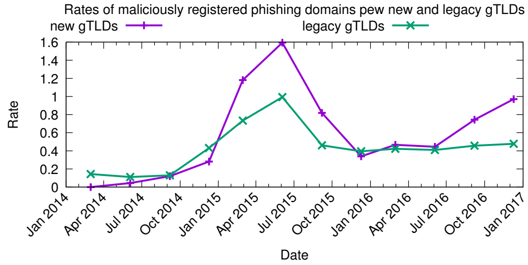


Figure 8: Time series of abuse rates of *maliciously registered* phishing domains in legacy and new gTLDs, the APWG feed.

Here, abuse rates are presented on a linear scale. For example, in the second quarter of 2015 the domain abuse rate for legacy gTLDs is equal to 3.82503. This means that, on average, legacy gTLDs had 3.8 blacklisted phishing domains per 10,000 registered domains. Our results suggest phishing abuse rates in legacy and new gTLDs to be converging towards similar values over time and were almost equal the end of 2016.

5.1.3 Phishing: Compromised vs Maliciously Registered. Up to this point, our descriptive statistical analysis of phishing abuse rates in the new and legacy gTLDs has conflated compromised and maliciously registered domains. Now, we compare abuse rates for these two types, separately.

Figure 7 plots abuse rates for compromised phishing domains within legacy gTLDs and new gTLDs, based on the APWG feed over time. The curves corresponding to all blacklisted phishing domains and compromised phishing domains of legacy gTLDs (cf. Figure 6 and Figure 7) follow a similar pattern due to a disproportionate concentration of compromised domains in legacy gTLDs.

Figure 8 on the other hand, shows abuse rates for maliciously registered phishing domains in the legacy and new gTLDs in APWG feed over time. When comparing the rates of all blacklisted domains of new gTLDs with rates of maliciously registered domains (cf. Figure 6 and Figure 8), we conclude that (despite higher relative concentrations of compromised domains in legacy gTLDs) miscreants more frequently choose to maliciously register domain names using one of the new gTLDs.

Moreover, we observe relatively higher rates of maliciously registered domains in new gTLDs in the first three quarters of 2015. We find 616 abused new gTLD domains. We observe 182 and 111 abused .work and .xyz domains, respectively. Manual inspection indicates that the majority of .work domains were registered by the same person: 150 domains were registered on the same day using the same registrant information, the same registrar, and the domain names were composed of similar strings.

Attackers often seem able to maliciously register strings containing trademarked words. Manual analysis of maliciously registered domains in the fourth quarter of 2015 revealed 88 abused .top domains 75 out of which contain the words: Apple, iCloud, iPhone, their combinations, or misspelled variants of these strings suggesting that they may have been all used in the same phishing campaign against users of Apple Inc. products.

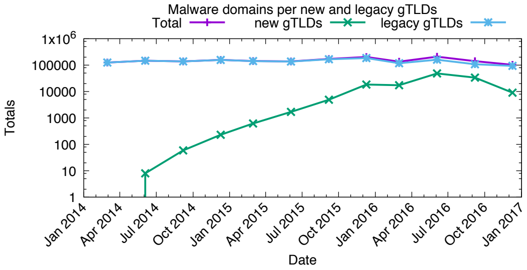


Figure 9: Time series of counts of malware domains in legacy gTLD, new gTLDs, and all gTLDs (Total), the SBW feed.

5.1.4 Malware Reputation. Having examined phishing abuse, we now analyze the malware related activity.

Figure 9 presents a time series of the number of malware domains in legacy gTLD, new gTLDs, and a "Total" based on the StopBadware feed between 2014 and 2016. Similar to phishing abuse, the total number of malware incidents in all gTLDs is mainly driven by incidents in legacy gTLDs (88.6%). We observe that the number of abused malware domains in legacy gTLDs remains relatively constant, whereas a growing trend in the number of malware domains in new gTLDs is clearly visible. SURBL mw and CleanMX malware datasets (not presented due to space limitation) confirm this observed trend.

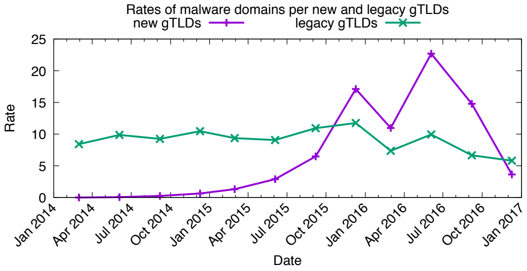


Figure 10: Time series of abuse rates of malware domains in legacy gTLDs and new gTLDs, the SBW feed.

5.1.5 Normalized Malware Counts. We now account for gTLD market shares by constructing a time series of abuse rates of malware domains in legacy and new gTLDs based on the StopBadware feed (see Figure 10). As before, the abuse rates are presented on a linear scale. Here, we observed an exponential growth of malware domain abuse rates in the new gTLDs up to the first quarter of

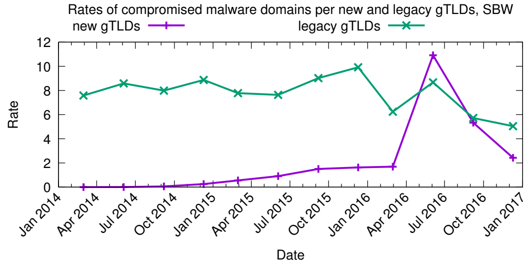


Figure 11: Time series of abuse rates of *compromised* malware domains in legacy and new gTLDs, the SBW feed.

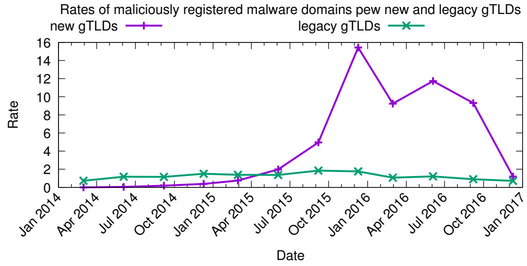


Figure 12: Time series of abuse rates of *maliciously registered* malware domains in legacy and new gTLDs, SBW.

2016. Differences between malware abuse rates in legacy and new gTLDs is the most prominent in the second quarter of 2016. While legacy gTLDs collectively had a malware-domains-per-10,000 rate of 9.9, the new gTLDs experienced a rate of 22.7. In absolute terms, malware domains in new gTLDs constitute 23% of all gTLD domains blacklisted by StopBadware during this period. SURBL and CleanMX malware datasets confirm the growing trend in terms of the malware rates in new gTLDs in comparison to legacy gTLDs.

5.1.6 Malware: Compromised vs Maliciously Registered. To distill factors that drive higher abuse rates in new gTLDs, in our analysis, we will differentiate between maliciously registered and compromised domains as we did for phishing abuse. Figure 11 and Figure 12 plot time series of abuse rates of compromised and maliciously registered malware domains, respectively, in legacy and new gTLDs. The results suggest that similar to phishing, malware abuse rates in legacy gTLDs are mainly driven by compromised domains (cf. Figure 10 and Figure 11). As expected, the malware abuse rates for new gTLDs are driven by maliciously registered domains (cf. Figure 10 and Figure 12).

Manual analysis of maliciously registered domains reveals distinctive common patterns in domain names. For example, we find 9,376 .link domains of which 9,256 were created in the first quarter of 2016 and 9,253 were registered through the Alpnames Limited registrar. 8,381 of all .link domains were registered using two registrar names only. Moreover, 8,205 and 1,027 were composed of 5 and 6 randomly generated characters, respectively. We created a user account with Alpnames Limited and tested bulk domain registration options. In fact, it is possible to randomly generate up to 2,000 domains at once from the selection of 27 new gTLDs using different patterns like letters, time, cities, zip codes, etc.

Finally, note that the registries of the most abused new gTLDs such as .win, .loan, .top, and .link compete on price, and their registration prices were occasionally below US $1, which was lower than the registration fee of a .com domain. Prior work has also found anecdotal evidence that indeed the price is likely to be one of the main driving factors of domain abuse [3, 32].

5.1.7 Spam Reputation. The results of the spam activity in the new and legacy gTLDs reveal very surprising trends. Due to space limitation, we only present our analysis of the Spamhaus feed. Note that Spamhaus provides *domain* rather than *URL* blacklists, which means that the great majority of listed domains are maliciously registered. Figure 13 presents a time series for the number of spam domains observed in legacy gTLDs, new gTLDs, and the total number of spam domains. While we observed an upward trend in the number of *phishing* and *malware* domains in new gTLDs, in contrast the absolute number of malicious *spam* domains in new gTLDs was actually higher than in legacy gTLDs. Note that the total number of spam incidents in all gTLDs is relatively constant and in the Q4 2016 is mainly driven by incidents in new gTLDs (58.8%). Figure 19 and Figure 20 (see Appendix C), presenting spam domains in legacy and new gTLDs for SURBL ws and SURBL jp spam datasets confirm this observed trend. The results suggest an alarming trend that miscreants seem to be switching from abusing legacy to new gTLDs when it comes to spam domains.

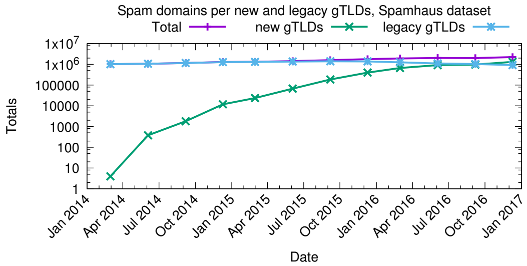


Figure 13: Time series of counts of blacklisted domains in legacy gTLD, new gTLDs, and all gTLDs (Total), Spamhaus.

5.1.8 Normalized Spam Counts. Figure 14 plots a time series of spam domain abuse rates for legacy gTLDs and new gTLDs based on the Spamhaus feed. As expected, the difference between spam abuse rates in legacy and new gTLDs is quite prominent. While legacy gTLDs collectively had a spam-domains-per-10,000 rate of 56.9, in the last quarter of 2016, the new gTLDs experienced a rate of 526.6–which is almost one order of magnitude higher. When comparing abuse rates based on our SURBL jp and SURBL ws spam feeds in the same period we observed a spam-domains-per-10,000 rates of 46.6 and 26 for legacy gTLDs, whereas for new gTLDs the spam-domains-per-10,000 rates are 286.3, and 265.2, respectively.

Table 2 (see the Appendix section) lists the top 10 new gTLDs with the highest relative concentrations of blacklisted domains for selected feeds in the fourth quarter of 2016. For example, spam-domains-per-10,000 registration rates calculated using the Spamhaus feed for .science, .stream, and .study are equal to 5,154, 4,756 and 3,343, respectively. In other words, as many as 51.5%, 47.6% and 33.4% of all domains in the corresponding zones were abused by cybercriminals and blacklisted by Spamhaus. Note that our results clearly indicate that the problem is not caused by just a few abused new gTLDs. As many as 15 most abused new gTLDs had spam-domains-per-10,000 registration rates calculated using Spamhaus

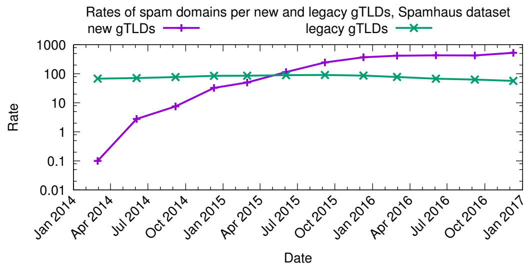


Figure 14: Time series of abuse rates of blacklisted domains in legacy gTLDs and new gTLDs, the Spamhaus feed.

feed higher than 1,000 at the end of 2016. Does this problem affect all new gTLDs? No. Our analysis of Spamhaus and SURBL blacklists reveals that approximately 32% and 36% of all new gTLDs available for registration did not experience a single incident in Q4 2016.

To conclude, while the number of abused domains in legacy gTLDs seem to remain relatively constant over time (or are decrasing), new gTLDs that underwent rigorous security analysis by ICANN are much more frequently affected by phishing, malware, and especially spam activities. Despite the new safeguards a number of new gTLDs are more susceptible to DNS abuse in comparison to legacy gTLDs. Given these observations, we systematically analyze the potential factors driving DNS abuse in new gTLDs.

5.2 Inferential Analysis of Abuse in New gTLDs

Previous work used regression analysis to study the impact of factors that influence the variation of abuse counts across networks of different intermediaries such as hosting providers [49] or TLDs [27]. Examples of such factors or more specifically intermediary properties are size, pricing, domain popularity index, or security effort [27, 49, 50]. In this section, we aim to analyze and quantify the relationship between the collected new gTLD properties (independent variables), and abuse counts (dependent variable), at the level of gTLDs. In other words, we use regression analysis to examine the amount of variance that gTLD properties can collectively explain, out of the total observed variance in the abuse counts.

Our regression models in Table 3 are built using the datasets explained in §3.1. We model the number of abused domains as a dependent variable (i.e. blacklisted domains or domain name elements of blacklisted URLs) using negative binomial[10] generalized linear model (GLM) with a Log link function. Depending on the model, we use the total number of abused domains or treat maliciously registered and compromised domains separately (details follow later). The independent variables in the models are the following properties of new gTLDs: *"new gTLD size"*: number of domains in TLD, *"Parked"*: number of parked domains, *"No DNS"*: number of domains that do not resolve, *"HTTP Error"*: number of domains for which corresponding websites return an HTTP error, *"DNSSEC"*: number of DNSSEC-signed domains, *"Type"*: an integer corresponding to the type of new gTLD, from least to most restricted group: 1 generic, 2 geographic, 3 community, and 4 brand, *"Registry"*: name of the registry operator that the TLD is operating under. Given that we find anecdotal evidence that pricing is one of the main driving factors of abuse, one limitation of this work is that we do not include it in our analysis due to a lack of historical data.

Table 3 in the Appendix section contains the summary of the regression models, i.e., the estimated coefficients, and their significance levels together with the goodness-of-fit measures such as the maximum Log likelihood, θ values and minimum Akaike information criterion (AIC) value (for more details, we refer the reader to the relevant literature). Note that the presented models are chosen from a stepwise addition of the variables into a baseline model with a single explanatory variable. Each column of the table contains a regression model for one of the abuse feeds with the count of abuse being the dependent variable.

The results in Table 3 are very consistent among all the analyzed abuse feeds. While all types of abuse show a positive and statistically significant correlation between the new gTLD size and abuse counts, the coefficients are very weak. We suspect that this is because the majority of abused domains in the new gTLDs are maliciously registered rather than compromised.

As expected, two variables indicating the number of domains that do not serve valid Web content to their users, i.e. "No DNS" and "HTTP Error" show a weak negative significant relationship with abuse counts. That means, the more domains labelled as "No DNS" and/or "HTTP Error", the less abused domains. Those two variables also correspond to the count of compromised domains rather than maliciously registered counts.

Moreover, the number of parked domains in new gTLDs plays a weak positive and statistically significant role in explaining the variance in phishing and malware domains. The more parked domains in a new gTLD, the more abused domains. This is to be expected as landing pages of parked domains may serve malware on a large scale. Note that the coefficients are very small. For example, if we hold the other independent variables constant and increase the number of parked domains by one unit (which is the equivalent to multiplying the number of parked domains of a gTLD by 10 since it is in the log_{10} scale), the number of phishing domains in APWG is multiplied by $e^{0.0003} = 1.0003$.

Previous research found a negative significant relation between the DNSSEC deployment and the count of phishing domains [27]. The authors used DNSSEC deployment as a proxy for the security efforts of both ccTLDs and gTLDs. In our analysis, we test the relationship between the number of DNSSEC-signed domains and abuse counts using various types of blacklists for new gTLDs. Note that ICANN requires each new gTLD to demonstrate a plan for DNSSEC deployment to ensure integrity and utility of registry information. Therefore, in our analysis, the number of DNSSEC-signed domains cannot serve as a proxy for registry efforts and obviously it does not prevent malicious registrations. One may suspect that attackers could be interested in deploying DNSSEC and signing their maliciously registered domains. Although it is not clear if that is the case, we indeed observe a weak but positive and statistically significant correlation between the number of DNSSEC-signed domains and the number of abused domains.

The regression results consistently show a negative correlation between the "Type" variable reflecting strict registrations and the count of phishing domains. In fact, in comparison to other variables, the obtained coefficients indicate the strongest statistically significant negative correlation for APWG, CleanMX phishing, and

[10] We choose negative binomial over Poisson due to the over-dispersion (unequal mean and variance) in our data.

SURBL phishing datasets: −0.54, −0.4, and −0.76, respectively (see Table 3). Note that for all other considered datasets, in particular malware, we also observe negative but not statistically significant correlations. When we consider separately maliciously registered and compromised domains (models not presented due to space limitation) the "Type" of new gTLD plays a significant role in explaining phishing abuse counts only for malicious registrations. Again, the results are intuitive. For example, it is much easier to register domains in the .top *standard* gTLD than it is for the .pharmacy *community* gTLD, for which the registration policy restricts the sale of domains to legitimate pharmacies only.

We also considered other models that contain "Registry" as a fixed effect to capture systematic differences in the policies of registries across new gTLDs such as pricing, bulk registration options, etc. Interestingly, our results indicate that none of the registry operators have a statistically significant effect on the abuse counts.

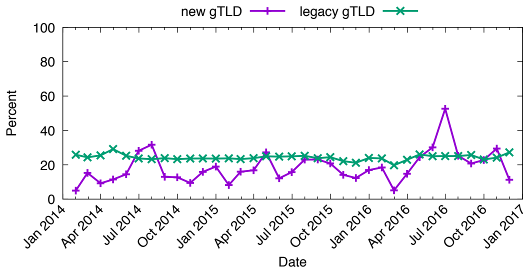


Figure 15: Usage percentage of Privacy and Proxy services for newly registered domains

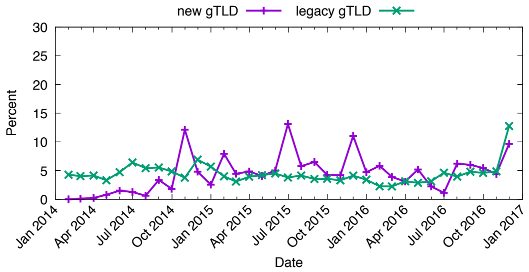


Figure 16: Percentage of abusive newly registered domains using Privacy and Proxy services

5.3 Privacy and Proxy Services

In this section, we present the results of an analysis to determine if there is a difference in the usage of WHOIS Privacy and Proxy services for abused domains in legacy gTLDs and new gTLDs. WHOIS Privacy and Proxy services are designed to conceal certain personal data of domain name registrants who use them. In practice, this works by replacing the registrant information in `WHOIS` with the information of the WHOIS Privacy and Proxy service.

There are many legitimate reasons why someone may want to conceal possession of a domain name. The usage of a WHOIS Privacy and Proxy services by itself alone is, therefore not a reliable single indicator of malicious activity. A previous study by National Physical Laboratories [29] did however find that a significant portion of abusive domains use Privacy and Proxy services.

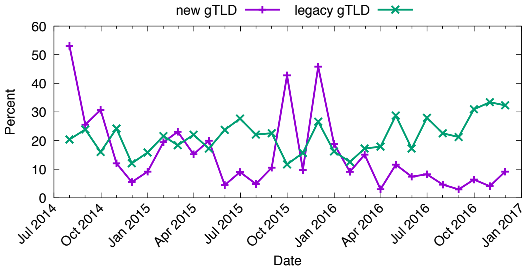


Figure 17: Usage of Privacy and Proxy services for abusive domains, reported by SURBL

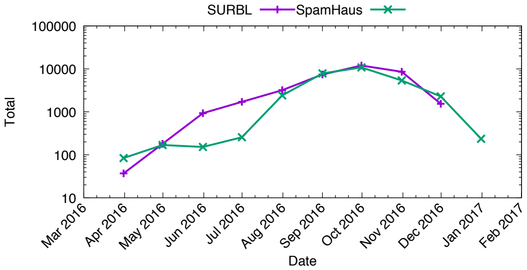


Figure 18: Abusive domains managed by Nanjing Imperiosus Technology

There are numerous WHOIS Privacy and Proxy services available, which can be used by domain owners. In Appendix D, we describe the methodology used in this study to identify commonly used WHOIS Privacy and Proxy services.

To get an indication of how common WHOIS Privacy and Proxy service usage is, we aggregated all domains from the `WHOIS` data by their create date. This shows us the number of newly added domains per month for legacy and new gTLDs. After checking how many of these domains were using a Privacy and Proxy service when the domain was registered, we calculated what percentage of the total number of newly registered domains is using a Privacy and Proxy service (see Figure 15). We find that for legacy gTLDs the usage is stable with a mean of 24%, and a standard deviation of 1.6. For new gTLDs the usage is generally below that of legacy gTLDs with a mean of 18% and a standard deviation of 9.3, which is visualized by the larger spikes and the increase to above the level of legacy gTLDs near the end of the study period.

Figure 16 shows the percentage of all newly created domains using Privacy and Proxy service, that have been reported to the Spamhaus or SURBL blacklist on or after the registration date. We have chosen to use Spamhaus and SURBL for this figure because these blacklists mainly contain maliciously registered domains. Here again, just as seen in Figure 15, we find that the variability for the new gTLDs is higher than compared to the legacy gTLDs.

For each blacklist used in this study we analyzed the proportion of domains that were using a Privacy and Proxy service at the time the domain was found to be abusive and included in the blacklist. Here again, we make a distinction between legacy and new gTLD domains.

For all SURBL feeds combined in 2016 the mean usage per month of privacy and proxy services by abusive domains in new gTLD observed is 5,874, with a standard deviation of 1,984, while for legacy gTLDs the mean usage per month is 21,744 with a standard deviation of 9,475. For Spamhaus the 2016 new gTLDs mean usage per month is 8,951 with a standard deviation of 2,892, while for legacy gTLDs the mean usage per month is 16,569 with a standard deviation of 3,843.

In the SURBL data we find 2 large peaks (see Figure 17) of abusive new gTLD domains using Privacy and Proxy services. Both of these peaks are driven by the .xyz, .click and .link new gTLDs. We attempted to find peaks in new registration that correspond to the two peaks seen in Figure 17. In the 7-15 day period leading up to a peak we do see an increase in the number of new registrations for the .xyz, .click and .link new gTLDs with the same registrar. However, we do not find strong evidence that the malicious registrations belong to a single or multiple campaigns using WHOIS Privacy and Proxy services.

The analysis of the use of WHOIS Privacy and Proxy service leads us to conclude that the usage of a WHOIS Privacy and Proxy services by itself is not a reliable indicator of malicious activity. Apart from the peaks, the usage of Privacy and Proxy services for abusive domains is not that high (see e.g. Figure 17). The usage of Privacy and Proxy seems to be higher in legacy gTLDs.

5.4 Registrar Reputation

Here we present the distribution of abused domains across ICANN accredited registrars. For each registrar, we find how many (#Incidents) can be attributed to the registrar and the total number of domains sponsored by that registrar (#Domains). We then calculate what proportion (Percentage) of all domains managed by the registrar is reported as abusive by a blacklist (see e.g. Table 4 in the Appendix section). An outlier with a relatively high rate compared to its peers may be caused by registrar-specific policies or operational practices.

Note, sinkholing of confiscated abusive domains or preventive registration of botnet C&C infrastructure domains is a common practice and special registrars have been created for this purpose e.g. “Afilias Special Projects” or “Verisign Security and Stability”. These registrars have high numbers of abuse and have been filtered out during the analysis because they are not regular registrars.

Our analysis reveals that “Nanjing Imperiosus Technology Co. Ltd.” is an outlier: over 93% of its domains are reported as abusive by SURBL (35,502, with a total number of 38,025 under its management) and 78% by Spamhaus (see Table 4). Figure 18 shows that both blacklists have marked domains managed by this registrar as abusive starting from early 2016. Starting from November 2016 we see a sharp decline in domains reported by Spamhaus and SURBL. This can be explained by the fact that ICANN has terminated the registrar accreditation [12] for this registrar, as it was determined that the registrar was in breach of the RAA [17]. Termination of the RAA had an effect on the amount of abuse linked to this registrar.

Alpnames Limited is another registrar that suffers from a high volume of abusive new gTLD domains reported by both Spamhaus and SURBL. The SURBL feed shows 2 distinctive peaks with a high number of abuse reports in 2016. After more detailed analysis, we find that these peaks correspond with 103,758 reports of abusive domains in the .top gTLD in August 2016. In October 2016, we find another peak, which is caused by 120,669 reports of abusive domains in the .science gTLD. In 2016 Alpnames did have promotions for domains using the .science gTLD for US $1 or less. We did not find corresponding peaks in the size of the .top and .science zone files, indicating the abusive domains have been registered over a longer period of time.

6 NEW ANTI-ABUSE SAFEGUARDS

Our results indicate that the implementation of the 9 anti-abuse safeguards have not effectively prevented domain name abuse in new gTLDs in comparison to legacy TLDs. Our findings, therefore, beg the question of whether more effective safeguards could be implemented by ICANN before the upcoming new gTLD rollout.

Our regression and descriptive analysis suggest that lesser strict registration policies, low registration pricing, and the possibility of bulk domain name registration lower barriers to abuse.

In addition, we observe that some of the more specific safeguards (e.g. DNSSEC deployment and prohibition of wildcarding) do not to raise barriers enough to prevent abuse. Yet, we cannot, for example, expect registries and registrars to raise registration prices to reduce abuse levels as this might be in conflict with their economic incentives. Alternatively, registries and registrars with disproportionately higher concentrations of abused resources could be penalized while those with relatively lower concentrations could be financially rewarded, e.g., through lowered ICANN fees, to align incentives towards raising abuse barriers. This would also incentivize intermediaries to develop their own anti-abuse best practices while balancing their anti-abuse policies against their economic incentives and allow for self-regulation.

Our analysis of domain abuse across new gTLDs revealed that some distinct entities are (or have been) afflicted with significantly hight concentrations of abused resources. We observed large concentrations of blacklisted domains associated with Nanjing Imperiosus Technology in early 2016. ICANN has terminated its registrar accreditation in this case in early 2017. Yet at the time of writing this paper, registry operators of the most abused new gTLD (e.g. .science, .stream or .racing) still remain ICANN-accredited. Accreditation terminations may be effective penalizing factors.

That being said, existing safeguards mostly concentrate on individual complaints (e.g. removing orphan glue records) rather than on security reputation metrics. An alternative more effective path forward could be to introduce continuous monitoring of abuse rates (including that of domain resellers) and employing enforcement mechanisms such as immediate accreditation termination if the concentrations of abused domains are persistent and exceed certain levels. Note that all above-mentioned proposals are currently under consideration by the ICANN community for upcoming new gTLDs rollout [8].

7 CONCLUSIONS

Since its inception, the new gTLD program has led to more than 1,200 strings being delegated in the root DNS zone, which greatly expanded the domain name space and increased consumer choice. We presented in this paper the first comprehensive study comparing the rates of malicious and abusive behavior in the new and legacy gTLDs. To that end, we employed datasets from many sources, including zone files, domain WHOIS information, data obtained through our active measurements, and heterogeneous blacklists representing malware, phishing, and spam.

While the number of abused domains in legacy gTLDs seem to stay relatively constant over time (or in some cases decreasing), new gTLDs that underwent rigorous application and evaluation process by ICANN are more frequently affected by phishing, malware, and especially spam activities.

The systematic investigation of the relation between structural and security-related properties of new gTLD operators, and abuse counts has shown that the number of domains in the new gTLDs, number of parked, and DNSSEC-signed domains play a statistically significant but weak role in explaining the differences in abuse counts among different new gTLDs. Low domain registration prices, unrestrictive registration practices, a variety of other registration options such as WHOIS privacy, registration in bulk and finally the increased availability of domain names decrease barriers to abuse, and seem to make some new gTLDs very attractive for miscreants.

Taken together, our findings indicate that the existing safeguards do not prevent domain name abuse and therefore we further develop cases for modifying the existing safeguards and proposing new ones, which we extensively discussed with the ICANN community.

Acknowledgements. ICANN funded the initial research upon which this further analysis was based. We thank ICANN, DomainTools, Whois XML API, Spamhaus, SURBL, StopBadware, CleanMX, SDF, APWG for providing access to their data, Roland van Rijswijk for his help in obtaining additional DNS data and anonymous reviewers for their valuable comments.

REFERENCES

[1] G. Aaron and R. Rasmussen. 2015. APWG Global Phishing Survey: Trends and Domain Name Use in 2H2014. http://docs.apwg.org/reports/APWG_Global_Phishing_Report_2H_2014.pdf. (2015).

[2] G. Aaron and R. Rasmussen. 2016. Global Phishing Survey: Trends and Domain Name Use in 2016. http://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf. (2016).

[3] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. 2015. Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse. In *Proc. of NDSS.*

[4] APWG. 2017. APWG: Cross-industry Global Group Supporting Tackling the Phishing Menace. http://antiphishing.org. (2017).

[5] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. 2011. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis.. In *Proc. of NDSS'11.*

[6] D. Chiba, T. Yagi, M. Akiyama, T. Shibahara, T. Mori, and S. Goto. 2017. DomainProfiler: toward accurate and early discovery of domain names abused in future. *International Journal of Information Security* (2017), 1–20.

[7] CleanMX. 2017. Spam-Filter Anti-Spam Virenschutz. http://clean-mx.de. (2017).

[8] Consumer Trust Competition and Consumer Choice (CCT). 2017. *New Sections.* Technical Report. https://www.icann.org/en/system/files/files/cct-rt-draft-recs-new-sections-27nov17-en.pdf

[9] D. Dagon, N Provos, C. P. Lee, and W. Lee. 2008. Corrupted DNS Resolution Paths: The Rise of a Malicious Resolution Authority. In *Proc. of NDSS.*

[10] The Secure Domain Foundation. 2017. https://securedomain.org/. (2017).

[11] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker. 2015. From .Academy to .Zone: An Analysis of the New TLD Land Rush. In *IMC.*

[12] S. Hansmann. 2017. ICANN: Notice of Termination of Accreditation Agreement. https://www.icann.org/uploads/compliance_notice/attachment/895/serad-to-hansmann-4jan17.pdf. (2017).

[13] S. Hao, N. Feamster, and R. Pandrangi. 2011. Monitoring the initial DNS behavior of malicious domains. In *Proc. of the IMC.* ACM, 269–278.

[14] S. Hao, M. Thomas, N. Paxson, V.and Feamster, C. Kreibich, C. Grier, and S. Hollenbeck. 2013. Understanding the Domain Registration Behavior of Spammers. In *Proc. of IMC'13.* ACM, 63–76.

[15] IANA. 2017. IANA: Registrar IDs. https://www.iana.org/assignments/registrar-ids/registrar-ids.xhtml. (2017).

[16] ICANN. 2009. New gTLD Program Explanatory Memorandum: Mitigating Malicious Conduct. https://archive.icann.org/en/topics/new-gtlds/mitigating-malicious-conduct-04oct09-en.pdf. (October 2009).

[17] ICANN. 2013. Registrar Accreditation Agreement. (2013). https://www.icann.org/resources/pages/approved-with-specs-2013-09-17-en#whois-accuracy

[18] ICANN. 2015. .madrid. https://icannwiki.org/.madrid. (March 2015).

[19] ICANN. 2016. New gTLD Program Safeguards Against DNS Abuse. https://newgtlds.icann.org/en/reviews/dns-abuse/safeguards-against-dns-abuse-18jul16-en.pdf. (2016).

[20] ICANN. 2017. ICANN: .zuerich TLD. https://icannwiki.org/.zuerich. (2017).

[21] ICANN. 2017. Internet Corporation for Assigned Names and Numbers (ICANN). https://www.icann.org. (2017).

[22] ICANN. 2017. Monthly Registry Reports. https://www.icann.org/resources/pages/registry-reports. (2017).

[23] ICANN. 2017. New gTLD Program. icannwiki.com/New_gTLD_Program. (2017).

[24] ICANN. 2017. TLD Startup Information. https://newgtlds.icann.org/en/program-status/sunrise-claims-periods. (Retrieved on February 2017).

[25] I. Khalil, T. Yu, and B. Guan. 2016. Discovering malicious domains through passive DNS data graph analysis. In *Proc. of the ASIACCS.* ACM, 663–674.

[26] M. Korczyński, M. Król, and M. van Eeten. 2016. Zone Poisoning: The How and Where of Non-Secure DNS Dynamic Updates. In *Proc. of the IMC.* ACM, 271–278.

[27] M. Korczyński, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. van Eeten. 2017. Reputation Metrics Design to Improve Intermediary Incentives for Security of TLDs. In *Proc. of IEEE Euro SP.*

[28] M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, and C. Hesselman. 2017. *Statistical Analysis of DNS Abuse in gTLDs Final Report.* Technical Report. https://www.icann.org/en/system/files/files/sadag-final-09aug17-en.pdf

[29] National Physical Laboratory. 2013. A Study of Whois Privacy and Proxy Service Abuse. gnso.icann.org/en/issues/whois/pp-abuse-study-20sep13-en.pdf. (2013).

[30] T. Lauinger, Ka. Onarlioglu, A. Chaabane, W. Robertson, and E. Kirda. 2016. WHOIS Lost in Translation:(Mis) Understanding Domain Name Expiration and Re-Registration. In *Proc. of the IMC.* ACM, 247–253.

[31] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. 2016. Domain-Z: 28 registrations later measuring the exploitation of residual trust in domains. In *Proc. of the IEEE S&P.* IEEE, 691–706.

[32] H. Liu, K. Levchenko, M. Felegyhazi, C. Kreibich, G. Maier, G. Voelker, and S. Savage. 2011. On the Effects of Registrar-level Intervention. *USENIX LEET* (2011).

[33] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul. 2015. Who is. com?: Learning to parse whois records. In *Proc. of the IMC.* ACM, 369–380.

[34] P.V. Mockapetris. 1987. Domain names - concepts and facilities. RFC 1034. (1987).

[35] A. Noroozian, M. Ciere, M. Korczyński, S. Tajalizadehkhoob, and M. Eeten. 2017. Inferring the Security Performance of Providers from Noisy and Heterogenous Abuse Datasets. In *WEIS 2017.*

[36] A. Noroozian, M. Korczyński, S. Tajalizadehkhoob, and M. van Eeten. 2015. Developing Security Reputation Metrics for Hosting Providers. In *USENIX CSET.*

[37] J. Postel and J.K. Reynolds. 1984. *Domain requirements.* RFC 920. RFC Editor.

[38] SBW. 2017. StopBadware: DSP. www.stopbadware.org/data-sharing. (2017).

[39] H. Shulman and M. Waidner. 2015. Towards security of internet naming infrastructure. In *Proc. of the ESORICS.* Springer, 3–22.

[40] K. Soska and N. Christin. 2014. Automatically detecting vulnerable websites before they turn malicious. In *Proc. USENIX Security.*

[41] Spamhaus. 2017. The Domain Block List. https://www.spamhaus.org/dbl. (2017).

[42] Spamhaus. 2017. The Spamhaus Project. www.spamhaus.org. (2017).

[43] B. Stone-Gross, C. Kruegel, K. Almeroth, A. Moser, and E. Kirda. 2009. FIRE: FInding Rogue nEtworks. In *Proc. of the ACSAC.* IEEE Computer Society, 231–240.

[44] StopBadware. 2017. StopBadware: A Nonprofit Anti-malware Organization. https://www.stopbadware.org. (2017).

[45] SURBL. 2017. SURBL - URI reputation data. http://www.surbl.org. (2017).

[46] SURBL. 2017. SURBL Lists. http://www.surbl.org/lists. (2017).

[47] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich. 2014. The Long" Taile" of Typosquatting Domain Names.. In *Proc. of USENIX Security.*

[48] S. Tajalizadehkhoob, R. Böhme, C. Gañán, M. Korczyński, and M. van Eeten. 2018. Rotten Apples or Bad Harvest? What We Are Measuring When We Are Measuring Abuse. *IEEE TOIT* (2018). https://arxiv.org/abs/1702.01624

[49] S. Tajalizadehkhoob, C. Gañán, A. Noroozian, and M. van Eeten. 2017. The Role of Hosting Providers in Fighting Command and Control Infrastructure of Financial Malware. In *Proc. of the ASIACCS.* ACM.

[50] S. Tajalizadehkhoob, T. Van Goethem, M. Korczyński, A. Noroozian, R. Böhme, T. Moore, W. Joosen, and M. van Eeten. 2017. Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting. In *Proc. of the ACM CCS.*

[51] Domain Tools. 2017. DomainTools: Domain Whois Lookup, Whois API & DNS Data Research. http://www.domaintools.com. (2017).

[52] T. Vissers, W. Joosen, and N. Nikiforakis. 2015. Parking Sensors: Analyzing and Detecting Parked Domains.. In *Proc. of NDSS.*

[53] T. Vissers, J. Spooren, P. Agten, D. Jumpertz, P. Janssen, M. Van Wesemael, F. Piessens, W. Joosen, and L. Desmet. 2017. Exploring the Ecosystem of Malicious Domain Registrations in the .eu TLD. In *Proc. of the RAID.* Springer, 472–493.

[54] WhoisXML. 2017. Whois XML API. https://www.whoisxmlapi.com/. (2017).

APPENDIX

A OVERLAP AMONG BLACKLISTS

To determine the overlap among our blacklsits, we present their pairwise intersections as a matrix in Figure 21, after extracting unique domain names from each data feed. Note that darker shades of grey represent larger overlaps among compared feeds. For example, the overlap between Spamhaus and SURBL ws indicates that they have 2,257,450 domain names in common within the observation period. This overlap constitutes 37% of the Spamhaus feed. In comparison, 2,257,450 domain names represent 64% of the SURBL ws feed. This is to be expected as both blacklists contain the same type of abuse, i.e. spam. The rightmost column indicates the absolute number and the percentage of samples that each blacklist has in common with all other feeds combined. For instance, the overlap between Spamhaus and all other blacklists is equal to 3,054,837 and indicates that as many as 51% of all domains blacklisted by Spamhaus are blacklisted by at least one other organization. Combined, these blacklists provide a comprehensive overview of domain name abuse for various criminal purposes.

B METHOD TO DISTINGUISH BETWEEN COMPROMISED AND MALICIOUSLY REGISTERED DOMAINS

We flag a domain name as malicious if it is blacklisted within 3 months after its registration. Aaron and Rasmussen have recently examined the delay between the time when phishing domains were initially registered and when they were ultimately used in attacks [2]. Their analysis indicates that miscreants tend to age the malicious domains they register to ensure a higher reputation score from security organizations. They concluded that the great majority of the domains used for phishing were maliciously registered within three months before they were used in an attack. To estimate the time between original registration and blacklisting, we analyze domain WHOIS information and extract the domain *creation date*. According to the Registrar Accreditation Agreement (RAA) [17], the creation date of the domain registration cannot be changed as long as the domain does not expire.

Furthermore, Aaron and Rasmussen identified 783 unique organizations used as phishing targets in 2015 and 679 in 2016, among which the most popular ones were PayPal, Yahoo!, Apple, and Taobao [2]. We used this information to create a list of keywords that the attackers may incorporate in maliciously registered domain names. As the great majority of phishing attacks target the most popular organizations, we extracted 300 keywords of the most popular domains according to their Alexa ranking and we labelled each blacklisted domain as maliciously registered if it contained an extracted string or its misspelled version. For example, 0paypalpayment.com would be labelled as malicious as it contains the string "paypal". To test if the domain contains a misspelled keyword, we first remove all digits from a domain name and split the resulting string into words with the "–" character. We compute the Levenshtein edit distances between the predefined keywords and a set of words derived from a domain name. If any Levenshtein edit distance is smaller than 2, we label the domain as maliciously registered.

Note that from the categorization process we exclude a list of 11,075 domains of legitimate services that tend to be misused by miscreants. These represent a separate, third group of domains that are neither maliciously registered nor hacked (i.e. third-party domains). For example, bit.ly – a domain used by a legitimate URL shortener service – could be used by an attacker to create a malicious URL (e.g. bit.ly/dcsahy) that may further be used to redirect a legitimate user to a phishing website. In fact, previous research shows that miscreants extensively abuse a variety of services with good reputations, affecting not only the reputation of those services, but of entire TLDs [27]. The list is composed of the 10,000 most popular domains according to their Alexa ranking and our own, manually maintained lists of domains of legitimate services (332 domains of URL shorteners and 840 domains of free hosting providers).

The reader should bear in mind that the categorization used in this study is limited due to a lack of the historical ground-truth data and a reasonable way of labeling blacklisted domains as maliciously registered or compromised. Future work could collect a large number of features related to registration information of a domain, lexical patterns of a blacklisted URL, etc. Based on manual analysis of the content of a website corresponding to a blacklisted but not yet suspended domain, it could be manually labeled as maliciously registered, compromised, or a third-party domain. Finally, future work could use the labeled domains as a ground truth, evaluate classification results of the proposed method and carefully assess importance of each pre-selected feature.

C SPAM DOMAINS IN LEGACY GTLD AND NEW GTLDS BASED ON THE SURBL FEEDS.

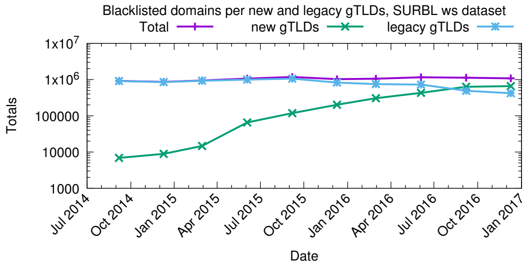


Figure 19: Time series of counts of spam domains in legacy, new, and all gTLDs (Total) based on the SURBL ws feed.

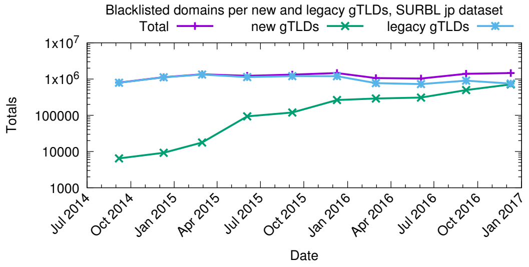
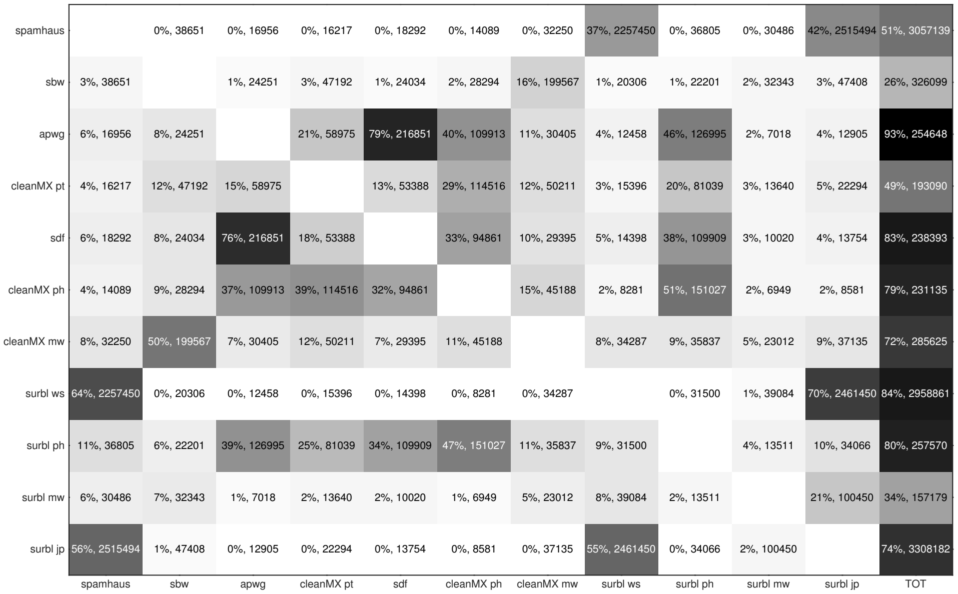


Figure 20: Time series of counts of spam domains in legacy, new, and all gTLDs (Total) based on the SURBL jp feed.



	spamhaus	sbw	apwg	cleanMX pt	sdf	cleanMX ph	cleanMX mw	surbl ws	surbl ph	surbl mw	surbl jp	TOT
spamhaus		0%, 38651	0%, 16956	0%, 16217	0%, 18292	0%, 14089	0%, 32250	37%, 2257450	0%, 36805	0%, 30486	42%, 2515494	51%, 3057139
sbw	3%, 38651		1%, 24251	3%, 47192	1%, 24034	2%, 28294	16%, 199567	1%, 20306	1%, 22201	2%, 32343	3%, 47408	26%, 326099
apwg	6%, 16956	8%, 24251		21%, 58975	79%, 216851	40%, 109913	11%, 30405	4%, 12458	46%, 126995	2%, 7018	4%, 12905	93%, 254648
cleanMX pt	4%, 16217	12%, 47192	15%, 58975		13%, 53388	29%, 114516	12%, 50211	3%, 15396	20%, 81039	3%, 13640	5%, 22294	49%, 193090
sdf	6%, 18292	8%, 24034	76%, 216851	18%, 53388		33%, 94861	10%, 29395	5%, 14398	38%, 109909	3%, 10020	4%, 13754	83%, 238393
cleanMX ph	4%, 14089	9%, 28294	37%, 109913	39%, 114516	32%, 94861		15%, 45188	2%, 8281	51%, 151027	2%, 6949	2%, 8581	79%, 231135
cleanMX mw	8%, 32250	50%, 199567	7%, 30405	12%, 50211	7%, 29395	11%, 45188		8%, 34287	9%, 35837	5%, 23012	9%, 37135	72%, 285625
surbl ws	64%, 2257450	0%, 20306	0%, 12458	0%, 15396	0%, 14398	0%, 8281	0%, 34287		0%, 31500	1%, 39084	70%, 2461450	84%, 2958861
surbl ph	11%, 36805	6%, 22201	39%, 126995	25%, 81039	34%, 109909	47%, 151027	11%, 35837	9%, 31500		4%, 13511	10%, 34066	80%, 257570
surbl mw	6%, 30486	7%, 32343	1%, 7018	2%, 13640	2%, 10020	1%, 6949	5%, 23012	8%, 39084	2%, 13511		21%, 100450	34%, 157179
surbl jp	56%, 2515494	1%, 47408	0%, 12905	0%, 22294	0%, 13754	0%, 8581	0%, 37135	55%, 2461450	0%, 34066	2%, 100450		74%, 3308182

Figure 21: Pairwise overlap of feeds with unique domains (2014-2016)

Table 2: Top 10 new gTLDs with the highest relative concentration of blacklisted domains for StopBadware SDP, APWG, Spamhaus, SDF, and SURBL datasets (fourth quarter of 2016). $Rate = 10,000 * \#blacklisted\ domains/\#all\ domains$.

StopBadware			**APWG**			**Spamhaus**			**SDF**		
TLD	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**
TOYS	32	78	LIMITED	31	66	SCIENCE	117,782	5,154	SUPPORT	510	294
TRADE	221	15	SUPPORT	43	24	STREAM	18,543	4,756	TECH	4,409	158
TAfTAR	1	11	CENTER	72	22	STUDY	1,118	3,343	ONLINE	4,179	83
WANG	1,086	11	CREDITCARD	1	13	DOWNLOAD	16,399	2,016	LIMITED	15	32
JUEGOS	1	9	SERVICES	24	10	CLICK	20,713	1,814	REVIEW	161	24
TOP	3,830	8	ONLINE	417	8	TOP	736,339	1,705	CLAIMS	3	19
MOE	5	8	MOE	5	8	GDN	45,547	1,602	PRESS	91	19
CAB	3	7	HOST	32	7	TRADE	23,581	1,521	FURNITURE	4	18
PICS	10	7	LEASE	1	6	REVIEW	9415	1,318	WEBSITE	298	15
TATTOO	2	7	REPORT	3	6	ACCOUNTANT	6,722	1,279	CREDITCARD	1	13

SURBL ph			**SURBL mw**			**SURBL ws**			**SURBL jp**		
TLD	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**	**TLD**	**# Domains**	**Rate**
LIMITED	51	109	FOOTBALL	7	16	RACING	51,443	3,812	SCIENCE	152,719	6,683
SUPPORT	82	46	TOP	5,066	11	DOWNLOAD	21,515	2,645	CLICK	27,871	2,441
CENTER	93	29	RIP	1	5	ACCOUNTANT	10,543	2,007	GDN	50,940	1,792
SERVICES	61	25	BID	200	3	REVIEW	12,615	1,766	STREAM	6,033	1,547
CRICKET	57	22	DENTIST	1	3	GDN	49,427	1,739	LINK	39,764	1,238
ONLINE	903	16	LGBT	1	3	FAITH	5,540	1,301	REVIEW	8,705	1,219
WEBSITE	318	14	ACCOUNTANT	11	2	TRADE	19,330	1,247	CRICKET	2,468	993
REPORT	7	14	CAB	1	2	CLICK	13,270	1,162	TRADE	14,535	937
HOST	65	13	SUPPORT	5	2	STREAM	4,406	1,130	FAITH	3,130	735
CREDITCARD	1	13	POKER	1	2	DATE	1,3851	999	TOP	285,488	661

D METHOD TO IDENTIFY WHOIS PRIVACY AND PROXY SERVICES

To identify the most commonly used WHOIS Privacy and Proxy services we used the following methodology: *i)* Using the WHOIS data, we aggregated all distinct domains by "registrant name" and "registrant organization" attributes and created a list with the top 5,000 registrants. *ii)* A keyword search on the top 5,000 "registrant name" and "registrant organization" attributes, trying to match any registrant with keywords such as: "privacy", "proxy", "protect", "private", "whois" etc. *iii)* A manual inspection of the suspect "registrant name" and "registrant organization" attributes to decide if the registrant is a Privacy and Proxy service (when this was not immediately clear from the name itself we used an Internet search to find additional information). Using the above described method we identified 570 "registrant name" and "registrant organizations" attribute combinations used by WHOIS Privacy and Proxy services.

Each blacklist abuse incident contains metadata such as the date when the domain was added to the blacklist. We used this date to identify the correct historical WHOIS record for an abused domain. By comparing the "registrant name" and "registrant organization" attributes from the domain WHOIS record to the list of known WHOIS Privacy and Proxy services, we are able to correctly identify abusive domains that were using a WHOIS Privacy and Proxy service at the time the domain was added to a blacklist.

Table 3: Negative Binomial GLM for count of abused domains per new gTLD

	Dependent variable:						
	apwg	sbw	cmx ph	cmx pt	cmx mw	surbl ph	surbl mw
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New gTLD size	0.00002^{***}	0.00001^{***}	0.00002^{***}	0.00003^{***}	0.00001^{***}	0.00002^{***}	0.00002^{***}
	(0.00001)	(0.00000)	(0.00001)	(0.00001)	(0.00000)	(0.00001)	(0.00001)
Parked	0.0003^{***}	0.0001^{***}	0.0002^{***}	0.00003	0.0001^{***}	0.0002^{***}	0.00001
	(0.00004)	(0.00003)	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00004)
DNSSEC	0.00001^{***}	0.00002^{***}	0.00002^{***}	0.00002^{***}	0.00001^{***}	0.00002^{***}	0.00002^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
No DNS	-0.00004^{***}	-0.00003^{***}	-0.00005^{***}	-0.00005^{***}	-0.00002^{***}	-0.00004^{***}	-0.00004^{***}
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00000)	(0.00001)	(0.00001)
HTTP Error	−0.00002	-0.00004^{***}	-0.0001^{***}	-0.00003^{*}	-0.00004^{***}	-0.0001^{***}	-0.0001^{***}
	(0.00002)	(0.00001)	(0.00001)	(0.00002)	(0.00001)	(0.00002)	(0.00002)
Type	-0.540^{**}	−0.150	-0.400^{**}	−0.120	−0.190	-0.760^{***}	−0.170
	(0.220)	(0.120)	(0.180)	(0.170)	(0.160)	(0.190)	(0.220)
Constant	-0.630^{**}	-0.390^{**}	-0.960^{***}	-1.200^{***}	-1.600^{***}	0.330	-2.200^{***}
	(0.280)	(0.170)	(0.230)	(0.230)	(0.220)	(0.230)	(0.290)
Observations	521	521	521	521	521	521	521
Log Likelihood	−566.000	−792.000	−508.000	−546.000	−392.000	−786.000	−284.000
θ	0.140^{***}	0.330^{***}	0.240^{***}	0.200^{***}	0.470^{***}	0.190^{***}	0.240^{***}
	(0.017)	(0.035)	(0.034)	(0.024)	(0.087)	(0.019)	(0.051)
AIC	1,149.000	1,600.000	1,031.000	1,109.000	800.000	1,588.000	583.000
Note:							*p<0.1; **p<0.05; ***p<0.01 Standard errors in brackets

Table 4: SURBL top10 percentage between blacklisted new and legacy gTLD domains (#Incidents) and total number of registrar gTLD domains (#Domains).

#	new gTLD registrar	#Domains	#Incidents	Percent	Legacy gTLD registrar	#Domains	#Incidents	Percent
1	Nanjing Imperiosus Technology	38,025	35,502	93.36	HOAPDI INC.	141	126	89.36
2	Intracom Middle East FZE	20,640	11,255	54.53	asia registry r2-asia (700000)	1,379	598	43.36
3	Dot Holding Inc.	153	76	49.67	Nanjing Imperiosus Technology	35,309	10,834	30.68
4	Alpnames Limited	3,028,011	751,748	24.83	Paknic (Private) Limited	10,525	3,083	29.29
5	Todaynic.com, Inc.	329,399	69,404	21.07	OwnRegistrar, Inc.	22,188	5,238	23.61
6	Web Werks India Pvt. Ltd	785	146	18.6	Eranet International Limited	6,109	1,339	21.92
7	GMO Internet, Inc. d/b/a Onamae	1,734,775	295,641	17.04	BR domain Inc. dba namegear.co	847	158	18.65
8	TLD Registrar Solutions Ltd.	163,988	24,700	15.06	Netlynx Inc.	17,612	3,030	17.2
9	Xiamen Nawang Technology, Ltd	282,925	42,089	14.88	AFRIREGISTER S.A.	1,551	266	17.15
10	Instra Corporation Pty Ltd.	77,642	6,200	7.99	GMO Internet, Inc. d/b/a Onamae	7,306,312	1,177,886	16.12