Now You See It, Now You Don't: A Large-scale Analysis of Early Domain Deletions

Timothy Barron
Stony Brook University

Najmeh Miramirkhani
Stony Brook University

Nick Nikiforakis
Stony Brook University

Abstract

Domain names are a valuable resource on the web. Most domains are available to the public on a first-come, first-serve basis and once domains are purchased, the owners keep them for a period of at least one year before they may choose to renew them. Common wisdom suggests that even if a domain name stops being useful to its owner, the owner will merely wait until the domain organically expires and choose not to renew.

In this paper, contrary to common wisdom, we report on the discovery that domain names are often deleted before their expiration date. This is concerning because this practice offers no advantage for legitimate users, while malicious actors deleting domains may hamper forensic analysis of malicious campaigns, and registrars deleting domains instead of suspending them enable re-registration and continued abuse. Specifically, we present the first systematic analysis of early domain name disappearances from the largest top-level domains (TLDs). We find more than 386,000 cases where domain names were deleted before expiring and we discover individuals with more than 1,000 domains deleted in a single day. Moreover, we identify the specific registrars that choose to delete domain names instead of suspending them. We compare lexical features of these domains, finding significant differences between domains that are deleted early, suspended, and organically expiring. Furthermore, we explore potential reasons for deletion finding over 7,000 domain names squatting more popular domains and more than 14,000 associated with malicious registrants.

1 Introduction

The Domain Name System (DNS) makes the modern web possible by allowing users to navigate by human readable names rather than machine route-able IP addresses. Domain names are often core to the identity of a Web site so it is no surprise that they can sometimes hold significant monetary value depending on popularity, brand recognition, or specific keywords [8].

Domain names are not registered permanently and eventually they expire and become available for anyone to re-register. Prior work has studied security issues revolving around expiring domains. Specifically, in 2016 Lever et al. showed how the new owner can abuse the *residual trust* inherited by a re-registered domain name [34]. Other work has studied the ecosystem around domain name re-registering [32,37]. In most cases this occurs at the end of the domain life-cycle when the domain is expiring, but little attention has been given to domains which are deleted before their expiration date.

Deleting a domain name before expiration is not recommended and not supported by many registrars. Domain names are paid for in full upon registration and registrars generally will not offer refunds prorated or otherwise. Therefore, deleting a domain name wastes the investment that was made in it. A registrant who no longer wishes to use a domain name might as well keep it and choose not to re-register when it normally expires. In spite of this, we discover that domain names deleted prior to their expiration date are surprisingly common. Such deletions by malicious registrants may hamper forensic analysis of their malicious campaigns, while registrars deleting domains instead of suspending them enable re-registration and continued abuse.

For years, TLD zone files have been a popular tool in the security community to find active domain names [12, 30,34,36,37,43]. Zone files are publicly available to researchers who request access and they represent a snapshot of *resolvable* domains, but not all registered domains will appear in zone files. Halvorson et al. observed that 5.5% of registered domains do not appear in zone files, but they specifically refer to domains which are purchased and not assigned name servers by the registrant so they are not yet added to zone files [20]. Alowaisheq et al. discuss suspensions via EPP `client hold` status which remove domains from zone files [11]. In this paper we present evidence of other cases of registered domains disappearing from zone files, making it clear that zone files alone should not be relied upon as a record of domain registrations and de-registrations.

In this paper, we present the first systematic study of domain names deleted prior to expiration. Over three weeks studying live data, we find the surprising result that 6.4% of all dropping domains are actively deleted prior to expiration. We confirm this phenomenon over the long-term using available historic data sets, finding a combined total of over 386,000 prematurely deleted domains. Among other trends, we find that domains names deleted early are longer on average and much more likely to be pronounceable compared to normally expiring domains. We explore potential motivations for deletions finding more than 7,000 domains abusing trademarks and squatting more popular domains, and over 14,000 associated with malicious activity such as phishing and malware. Furthermore, we investigate the participating parties finding that over 100 registrants deleted domains in bulk, and several registrars, including GoDaddy and Domain.com, delete large numbers of domains instead of suspending them. Our results lead us to a discussion of registrar policies regarding domain name deletion, as well as the advantages of publicly available sources for registration information.

2 Background

To enable the translation from domain names to IP addresses for over 300 million domains, DNS utilizes a hierarchical look-up process beginning at the root zone which leads to top level domain (TLDs) zones such as .com, or .net. These TLD zones are maintained by *registries*, such as Verisign, which then delegate the selling of domain names to *registrars*, such as GoDaddy and eNom. Registrars communicate with the registry through the Extensible Provisioning Protocol (EPP) [23] in order to perform a series of domain-name-related operations, such as checking domain availability, registering new domains, and deleting domains. Finally, *registrants* are the users who buy domain names from the registrars.

Domain names are registered for a period of at least one year, and optionally longer for additional cost. Registrants pay the full cost of the domain at the time of registration and control it until the registration period expires, after which the registrant has the option to renew the domain and pay for another period before anyone else has the opportunity to buy it. This process is often automatic, but if the registrant chooses not to renew, then they lose control of the domain name and it becomes publicly available for anyone to register again.

The complete life-cycle of a domain name contains several phases, details of which are not obvious to typical registrants and easy to confuse. Figure 1 illustrates these phases showing the duration of each and indicating when the domain appears in the TLD zone file. These phases revolve around the registration and expiration of domains in an attempt to make the process fair and reduce the risk of accidental expirations.

The first phase after registration is the 5 day *add-grace period* which is the only time when it is possible to receive a refund for a domain name. This can lead to a form of abuse called domain tasting which was studied in detail by Coull et al. [15] so registrars are limited to issuing a maximum number of add-grace deletions [25]. After the first 5 days, refunds are no longer available and the longest phase is the *registered period* during which the domain should be in the zone files or else it cannot be resolved.

Once the domain expires it enters the *auto-renew grace period*. The registrar is in control at this stage and based on their policies they will notify the registrant and often remove the DNS records from the zone file after a certain point. If the registrar does nothing, the domain will automatically renew between the registrar and registry, so if the registrant does not choose to renew then the registrar will typically delete the domain (to avoid having to pay for a domain whose owner does not want) right before the end of the 45 day period (though it is possible for them to delete sooner). The domain then moves to the *redemption grace period* in which the domain is not in the zone file, but the registrar still has an opportunity for 35 days to reclaim the domain name at an increased cost. Finally, if the registrar takes no further action, the domain enters the *pending delete phase*. At this point the domain is still not in the zone file and after 5 days it becomes publicly available for re-registration.

This domain life-cycle is typical for most domain names, but there is also the possibility for a registrant to delete their domain name before the end of the registration period. This is unusual behavior, because the registrant will not be refunded the registration cost. Even if the registrants are no longer using the domain and do not want it, there is no obvious reason to actively request its deletion, as opposed to allowing it to expire. Yet, as we show later, more than 8,000 domains are deleted early every day.

3 Methodology

In this section we describe the data sources, the method for finding early domain deletions, and a discussion of obstacles and limitations of these data sets.

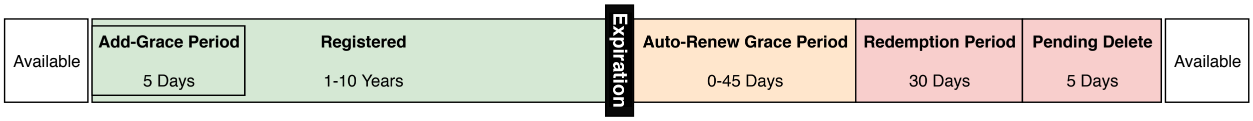

Figure 1: *Stages in the life-cycle of a domain name registration. Green or red indicate the domain is present in or absent from the zone files respectively. Orange means the domain may or may not be in the zone files depending on registrar policies.*

3.1 Data Collection

Zone files. Top level domain (TLD) zone files contain at the very least name server records for all resolvable domain names. These zone files are publicly available for all generic TLDs (gTLDs) once requested from the corresponding registries. We use zone files from ten of the largest TLDs: .com, .net, .org, .info, .biz, .top, .xyz, .loan, .club, and .online. These zone files were collected every day and the deltas between each day were computed to find when domains appeared and disappeared. In most cases, appearances and disappearances correspond to new registrations and expirations respectively. These deltas were indexed to create a searchable database of registration periods for each domain. In practice, we find that zone files are not a perfect representation of registration status, but they are a useful starting point using public data.

Drop Lists. Domain drop lists provide a reliable view of domains which have been de-registered. These domains will become available for re-registration on a specific date and registrars want to advertise as many domains as they can to increase sales on the drop date. These drop lists were collected on a daily basis starting from 1/10/2017 and aggregated from SnapNames, DropCatch, Pool, Namejet and Dynadot, all of which are companies that allow users to re-register valuable domains which were left to expire. These lists cover .com, .net, .org, .info, and .biz.

Historic WHOIS. Whenever a registrant purchases a new domain, their details are added to a WHOIS record which is publicly accessible. Before the recent GDPR legislation, WHOIS records contained personally identifying details (PII) about the owners of each domain, such as their name, address, and phone number. Since the GDPR legislation went into effect, WHOIS records are mostly anonymized although they still provide EPP status, dates of registration/expiration, and the registrar.

WHOIS queries are restrictive (i.e. individual WHOIS servers set limits as to how many queries a client can perform per day) and therefore difficult to obtain in large numbers, especially for past records. For our experiments, we used commercial services to obtain historical WHOIS data taken at the time of registration for all new domains in 2017 and all dropping domains between February and October of 2017 [6, 9, 47].

RDAP. The recent pilots for the new Registration Data Access Protocol (RDAP) have made it practical to collect live information about domains as they disappear. Unlike WHOIS which is just a text protocol, RDAP provides structured data which allow us to straightforwardly extract important domain information, such as a domain's registration and expiration dates. Between 12/19/2018 and 1/27/2019 we collected registration information for all of the domains that were removed from the zone files each day. In cases where RDAP failed we fell back to WHOIS. Between these two methods, only 7.0% of queries failed and 74% of failures were from .xyz while .com had only 3.8% failures. To account for delay in server updates and temporary status changes, we also requeried the same domains one week after disappearance starting with domains that disappeared on 12/31/2018.

Blacklists. We collected hphosts, malc0de, zeustracker, conficker, and malware domains blacklists [1, 3–5, 7] on a daily basis and used the Internet Archive to retrieve snapshots of older versions of the blacklists. We supplemented blacklists by querying for domains using the Google Safe Browsing API (GSB) [2].

Typosquatting, bitsquatting, and combosquatting. Starting from the Alexa top 1 million on 7/1/2017 and 12/31/2018, we generated a set of typo domain names using the typo models described by Wang et al. [46] and a set of bit-flipped domain names as described by Dinaburg [17]. We also compiled a list of 279 popular trademarks modified from those used by Kintis et al. [29] to find combosquatting domains. Each of these lists are used to find squatting domains taking advantage of the trademarks and brand names of others.

3.2 Finding Early Deletions

WHOIS/RDAP 2019 data set. Our primary experiment during the first three weeks of 2019 makes use of zone file deltas from the 10 TLDs mentioned above and public registration information to study domain names as they disappear and become unresolvable on a daily basis. The combination of both RDAP and WHOIS provide EPP status codes [26] and expiration dates, allowing us to investigate these domains and determine the cause of their disappearance.

The two situations that would typically cause a domain name to be removed from the zone files are deletion (including expiration) and suspension. There are five relevant statuses that we track on disappearing domains. A status of `redemption period`, `auto renew period`, or `pending delete` indicates that the domain is in the process of being deleted as shown in Figure 1. The `client hold` and `server hold` statuses indicate suspension by the domains' registrar and registry respectively. In some cases a registrar may set the `server hold` status on a domain that is expiring, but this would not impact our results for domains prior to their expiration date and the domain may also have the `auto renew period` status indicating that this is a deletion and not a suspension. Once we know the reason for the domain's disappearance, we then check its expiration date to determine if it was deleted prematurely. We query the same domains again a week later allowing us to remove inaccuracies due to delays in updates on the registration servers and determine whether the status changes are transient or permanent.

Historical 2017 data set. In order to longitudinally study the phenomenon of early deletions, we also conduct a measurement using historical data from 2017. Even though we initially attempted to rely as much as possible on the indexed zone file deltas, we quickly discovered that zone files alone are not a reliable indicator of a domain's registration. We discovered a large number of cases where domains disappeared and reappeared days, weeks, or months later, and in some cases multiple times without dropping or changing name servers or WHOIS information. Below, we describe our process using these zone file deltas and the additional data sources we used to improve the accuracy of our experiments.

Based on the domain life-cycle shown in Figure 1, we are able to narrow down the list of domains that may have been deleted early. First, if a domain name is newly registered and then de-registered within 365 days then it is a candidate early deletion since the shortest possible registration period is one year.

We ignore any domains that were deleted within the 5 day add-grace period because, as described in Section 2, this falls under domain tasting which is a well known practice and may result in a refund. Domain names may be suspended or deleted by the registrar after 15 days if their owners do not respond to inquiries regarding the accuracy of their WHOIS contact information [24]. While this is 15 days from an inquiry and not necessarily since creation, we conservatively choose to ignore disappearances 15 days after creation when contact information was given to the registrar.

Given our observation of domain names disappearing and reappearing without their registration status changing, we filter these domains further, identifying cases where they could not have been registered or deregistered. As shown in Figure 1, a domain name that is deleted should be absent for *at least* 35 days during the redemption period and pending delete phase following its disappearance from the zone files.

We still observed domains which remained registered according to WHOIS, but were absent for more than 35 days so we also used publicly available drop lists. A domain appearing in a drop list must have been deregistered so we use this to filter out temporary domain disappearances from the zone files. This requirement limits the results of this data set to .com, .net, .org, .info, and .biz, which are covered by our collected drop lists. On the other end of the life-cycle, when we see a domain name appear we do not know if it is newly registered or reappearing from a temporary disappearance. We could similarly require the domain name to be absent for 35 days before appearing, but again the domain may be reappearing after a longer period and the redemption period/pending delete phase do not apply to domains deleted during the add-grace period. Therefore, we require additional information analogous to the drop lists to verify new registrations. We use historic WHOIS records to confirm the zone file appearance date with the listed creation date. This also has the advantage of obtaining registrar information and registrant contact details which we use to characterize early deletions.

4 Results

In the following we present the results with a data-driven approach to explain the phenomenon of early deletions.

4.1 Categorizing Disappearances

Collecting registration information with RDAP and WHOIS allows us to broadly categorize unexpected domain disappearances from the zone files. Based on the EPP statuses of each domain, we assign one of four labels: i) *active* domains which are still registered and have no adverse status, ii) *suspended* domains which were removed by the registrar or registry, iii) *deleted* domains which are on their way to being de-registered (typically due to expiration), and iv) domains for which the queries *failed* to obtain registration information.

On average, we observed 127,318 domains disappearing from the zone files every day. Figure 2 shows that on a typical day about 70% of these disappearing domains were deleted with most of the remainder caused by suspensions by either the registry or registrar.

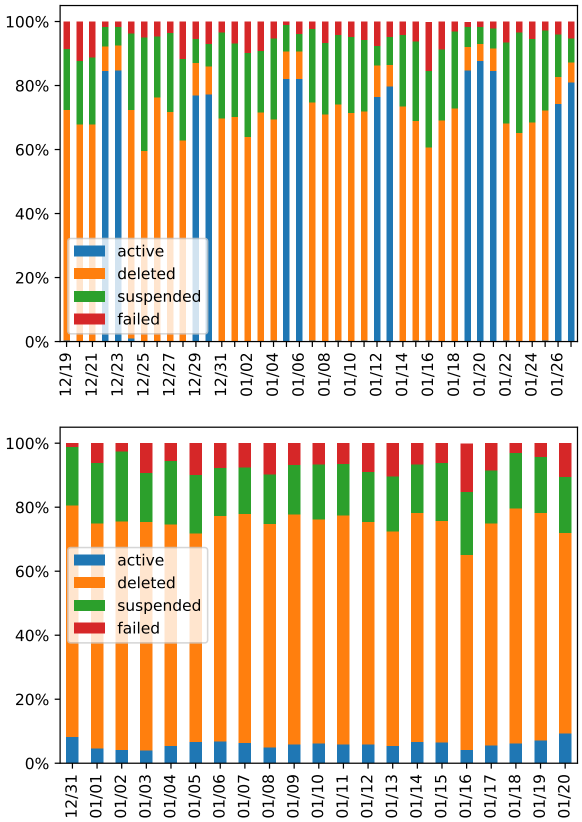

Figure 2: *Percentage of disappearing domains each day with status indicating deleted, suspended, or active. (Top) Queried on the day of disappearance. (Bottom) Queried one week after disappearance.*

There is a notable weekly cycle of days where the majority of the disappearing domains are still listed as active. This occurs on Saturdays and Sundays and on one Monday (1/21) which was a US national holiday. By querying the same domains one week later (bottom of Figure 2) we obtain more consistent results and we no longer see this pattern. This data depends heavily on Verisign's RDAP pilot deployment and our observations suggest that they have some weekend delay in updating registration information on their servers for newly expired and suspended domains.

Ignoring these weekends and looking specifically at domains which change status one week later, we find that by far the most common change on a typical day is domains having their suspension lifted and becoming active again. Table 1 shows how many domains changed status out of the total number of disappearances. Overall we find the results are very stable which lends confidence to the results from RDAP and WHOIS. For the rest of this section we use the second round of results to avoid temporary status changes and server inconsistencies.

Transition			**Percentage of domains**
Active	⟶	*Deleted*	0.0002%
Active	⟶	*Suspended*	0.0004%
Deleted	⟶	*Active*	0.075%
Deleted	⟶	*Suspended*	0.005%
Suspended	⟶	*Deleted*	0.36%
Suspended	⟶	*Active*	5.28%
Failed	⟶	*Not Failed*	1.69%
Not failed	⟶	*Failed*	1.44%

Table 1: *Changes in EPP status one week after disappearance.*

4.2 Early Disappearances

To find domains that disappeared early we compare the date they were removed from the zone files to the expiration date returned by RDAP/WHOIS. Figure 3 shows the breakdown of status for domains that disappeared before their expiration date. This ignores domains with failed queries since we do not have their expiration dates or status. Since domains disappearing after expiration is the normal expected behavior, it is unsurprising that when we look at these cases, we find that most are caused by suspensions. However, a surprising 22.1% (173,292 total) of the early disappearances are related to domain names that were deleted before their expiration date. This corresponds to 6.4% of all disappearing domains and an average of 8,252 early deletions per day.

We also observe 5,061 domains which still have an active status after both queries despite their removal from the zone files. 99.96% of these occur before expiration so we know these are not expired domains with inaccurate status. Without an entry in the zone files these domains cannot be resolved, but they still appear to be registered and they do not have a client or server hold. Registrants do not typically have the ability to directly remove a domain from the zone files while it is still registered, but this may be a process controlled by the registrar or registry. Because of this behavior, we conclude that zone files alone cannot be used to indicate whether a domain is registered or not. Therefore systems that rely on zone files [12, 30, 34, 36, 37, 43] are bound to be missing domains that are temporarily suspended and ones that are registered but are missing.

To find early deletions over a longer period of time we use zone files, drop lists and historical WHOIS as described in Section 3. We classify 212,864 domains as early deletions over 13 months which is about 0.5% of all disappearing domains during the same time period. Figure 4 shows the number of early deletions over time which is fairly steady on typical days, but there are multiple outliers, the largest being 3/16/2017 with more than 20 times the average number of deletions. These cases are discussed in more detail later in this section. We also observe a slight increase over time, as the number of

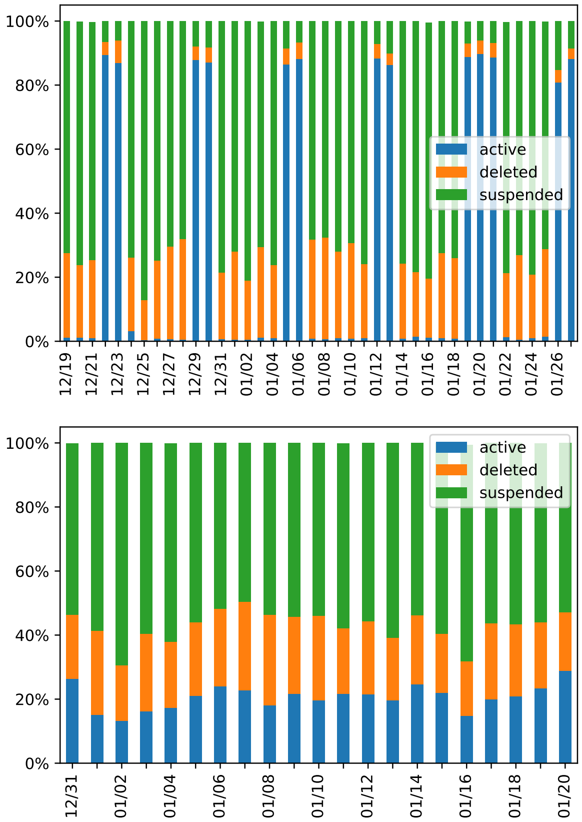

Figure 3: *Percentage of domains which disappeared before their expiration date each day with status indicating deletion, suspension, or that the domain is still active. (Top) Queried on the day of disappearance. (Bottom) Queried one week after disappearance.*

monitored, newly-registered domains grows. These data sets cover fewer TLDs than our RDAP/WHOIS measurement, but based on the percentage of disappearing domains, it is still significantly fewer early deletions. The main reason is that this method can only capture early deletions that occur within one year of registration, missing domains which were registered for a longer period.

This data set provides another vantage point from which we observe domains disappearing and reappearing while still registered. Without using drop lists or WHOIS to filter domain changes in the zone files, we found 5,907,109 cases of domains disappearing before expiration. We do not have access to domain status when these domains disappeared, but based on our above findings where we observed about 20,000 suspended domains per day we expect this to be the primary cause of early disappearances. However, most of these suspended domains names will remain suspended and will not reappear in the zone files for the lifespan of the domain.

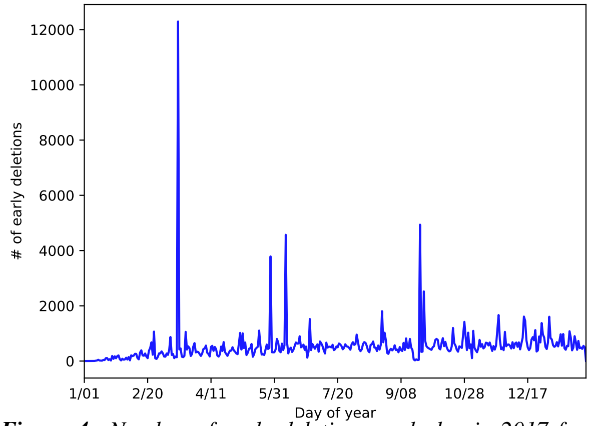

Figure 4: *Number of early deletions each day in 2017 from .com, .net, .org, .info, and .biz.*

4.3 Duration of Registration

To understand potential patterns in early deletions, we start by examining the lifespan of deleted domains. Figure 5 shows the distributions of registered duration and days remaining. We measure the duration of registrations as the number of days from a domain's creation to the observed date of deletion. The remaining time is the number of days prior to expiration that we observe deletion. The median remaining time was 322 days so domains are often deleted long before expiration, but only 4% of domains had more than 1 year remaining. Interestingly, deleted domains are not all short-lived. In fact, 81% of deleted domains were registered for over one year with a median of 458 days. This makes it clear that the primary cause for the difference in number of observed early deletions in our two data sets is the limitation of domains in zone files for under one year in our historical data set.

We also found that domains have a bias towards being deleted early in their overall lifespan. This can be observed in Figure 6 which shows the distribution of remaining time as a percentage of the total number of days from creation to expiration. We also observe that there are multiple sharp jumps in each of these graphs which likely correspond to bulk activity relating to a few registrants that had many domains which were registered on the same date and deleted on a specific following date.

4.4 Trademark Abuse

One explanation for domain deletions could be an association with a brand that requests for the domain to be taken down. For serious disputes, trademark holders can file UDRP (Uniform Domain-Name Dispute-Resolution Policy) requests, but this can cost thousands of dollars [48]. In 2017 we found records of only 1,557 total filed disputes covering 3,487 domains using a third-party

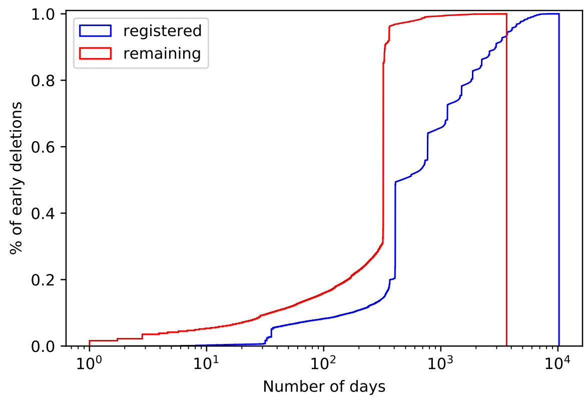

Figure 5: *Cumulative distribution of duration of registration and days remaining before expiration for domains deleted early. Number of days on log scale.*

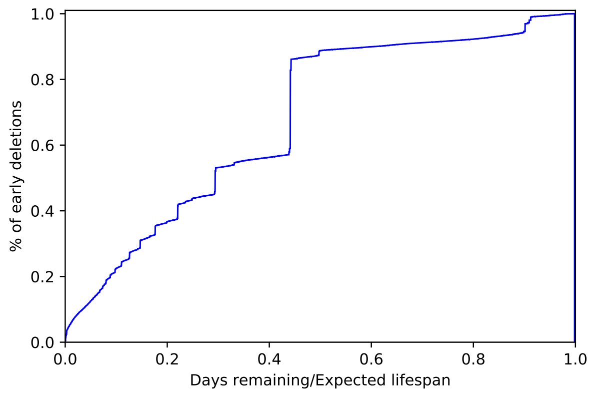

Figure 6: *Cumulative distribution of percentage of paid lifespan remaining before deletion.*

tool that crawls and indexes UDRP cases [28]. In most cases where the ruling is in the complainant's favor, the domain is transferred. This gives the trademark holder the most control as they can now sinkhole or redirect it to their primary site and they can choose to continue renewing it, so it stays under their control indefinitely. In only 3% of these cases, the decision was to delete the domain name instead of transferring it. As such, we argue that formal disputes have a minimal impact on early deletions. More often, the first step in dealing with a squatting domain is sending a cease and desist to the offending domain registrant demanding that the domain name be deleted, transferred, or abandoned. In this scenario, a registrant who is concerned about legal trouble may choose to delete the domain name right away to satisfy the trademark holder. To estimate the number of early deletions that may have been caused by this situation before an official complaint, we looked for typosquatting, bitsquatting, and combosquatting domains.

Typosquatting domains target popular Web sites and capitalize on users making typos of those sites. Bitsquatting it very similar, but it involves registering domains with an ASCII encoding one bit different from the target to capitalize on random bitflips during the resolution process. We generated a list of typos and bitflips from the Alexa top 1 million on 12/31/2018 for the RDAP/WHOIS data set and from 7/1/2017 for the historic data set as described in Section 3.1. From these lists we found 4,751 typosquatting/bitsquatting early deletions. We also explored combosquatting which is a broader category of domains which contain popular trademarks along with additional keywords meant to confuse and deceive users reading the URL. As described in Section 3, we obtained the list of combosquatting trademarks compiled by Kintis et al. [29] which we extended with trademarks that we observed during our manual analysis of early domain deletions. From the resulting list of 279 trademarks, we found 2,612 combosquatting early deletions in our combined data sets.

By combining the discovered typosquatting, bitsquatting, and combosquatting results, we found a combined total of 7,322 (1.9%) early deletions abusing trademarks. It is clear that this represents a level of malicious activity although we cannot differentiate whether these early deletions were due to cease-and-desist orders or due to malicious registrants attempting to cover their tracks by deleting domains after they abuse them. Only 134 of these squatting domains were present in our collected blacklists which amounts to less than 2% coverage by popular blacklisting approaches.

4.5 Domain Features

We analyze features of domains which are deleted early and compare them to popular domains, domains which expire normally, and suspended domains. Since the domain names we are studying are offline by the time we observe them, we are limited to lexical features of the names themselves. The simplest metrics are length, and whether the domain contains at least one number or hyphen. We also look at Shannon entropy which can be an indicator of names created by Domain Generation Algorithms (DGA) which create a large number of domains where botnets know to contact command and control servers. Plohmann et. al. [41] discuss various types of DGAs and show that most have high entropy. The domain names are segmented into the most likely words based on known frequencies of words and pairs of words in the English language [40]. From this we get number of words, and check those words against a blacklist of adult keywords [16]. We filter out extracted words which do not appear in a dictionary, but since many domain names contain made-up brand names or newer slang, we also

Metric	Alexa Top 1M	Normal Expirations	Suspensions	Early Deletions
Length	10.55	12.14	11.59	12.85
Entropy	2.825	2.937	2.893	3.005
Number of words	1.377	1.534	1.368	1.719
Percent containing numbers	6.46%	14.04%	15.99%	9.87%
Percent containing hyphens	10.05%	7.44%	11.20%	8.43%
Percent containing adult keywords	3.129%	2.633%	2.562%	2.649%
Percent unpronounceable	14.89%	21.64%	24.58%	15.02%

Table 2: *Comparison of lexical domain features between popular, expired, suspended, and deleted domain names.*

measure pronounceability based on frequency of letter bi-grams in the English language. Average pronounceability scores were very similar across all domains, but we set a threshold to label domains as pronounceable or not which we applied equally to each category. Using only English words is a potential limitation, but in many cases pronounceability may work well across languages and we exclude from this analysis Internationalized Domain Names which begin with `xn--`. Similar metrics are commonly used in part for domain name appraisals which combine many features of a domain to estimate its monetary value [10].

From the extracted English words we used WordNet Domains [13] to label domain names with more general topics. Wordnet Domains has 181 possible topics, many of which are closely related, so we manually aggregated similar topics to more appropriately group domain names. For example, banking, commerce, and finance are distinct topics which we include under the *Economy* label along with eight other sub-topics. Since domain names can have multiple words and each word may have multiple meanings each mapping to different topics with WordNet Domains, we take the most common topic from the list of all related topics for that name. If a domain name has no words or the words do not have a mapped topic, then we label it as *Unknown*. A breakdown of the most represented topics among early deleted domain names can be seen in Table 3. A full table with all 28 topics and a statistical comparison between different categories of domains appears in Appendix B.

Topic	% of Early Deletions
Unknown	26.71%
Economy	8.79%
Science/technology	8.49%
Play/sports	6.16%
Health/medicine	5.64%
Architecture	4.24%
Geography	4.09%
Politics/goverment	3.95%
Art	3.54%
Travel/transport	3.46%
Person	3.07%
Writing/language	2.32%
Other	19.56%

Table 3: *Top domain topics in early deleted domain names.*

Table 2 compares each metric between domains in the Alexa top 1 million, domains which disappear normally after their expiration, domains which were suspended, and domains deleted early between 12/31/18 and 1/20/19. One difference is that early deletions are the longest in both number of characters and number of words. Early deletions also have fewer numbers and hyphens than suspended domain names and are almost as likely to be pronounceable as popular domains. Shorter domain names are generally considered more valuable, but the pronounceability and number of words suggests that early deleted domain names may be well formed. If the domain is considered valuable we expect it is less likely a registrant chooses to delete it, but one explanation for the length could be DGA names that are built from a dictionary of words. Three examples of such wordlist-based DGAs are Matsnu, Suppobox, and Gozi which were studied by Plohmann et. al. [41] in 2016. These are less likely to be suspended or blacklisted because they do not look like random strings of numbers and letters. At the same time, the paper points out that these DGAs are more likely to collide with existing domain names than other methods. To counteract this issue, modern wordlist-based DGAs may require longer names with more words than would be common among popular domains in order to generate a consistent variety of unused domain names.

We wish to determine if the observed differences between early deletions and the three other categories in Table 2 are statistically significant or if they could have occured randomly as a sample of all domain names. We use the Welch's t-test to test the null hypothesis that early deletions are sampled from the same population as the other groups. Due to our large sample sizes, we obtained very small p-values for almost all tests leading us to reject the null hypothesis in these cases and conclude that early deletions are drawn from distinct populations. Appendix A includes more details of the statistical analysis including Cohen's *d* effect size, *t*-statistic, and *p*-value for each test. The fact that early deletions have statistically significant differences means that it is very unlikely that domain names are being deleted indiscriminantly. Rather, some reasoning went into which domains were deleted, or the entities that deleted domain names had some pattern as to the domains they held.

4.6 Registrant Patterns

While domain owners can choose to use a WHOIS privacy service to hide personal information, such as email addresses, we found that in our 2017 data set, after filtering out anonymous addresses, 73% of early deletions had real registrant email addresses. Out of 58,773 registrant email addresses, 97% of them deleted less than 10 domains before expiration, but 100 registrants deleted more than 100 domains early. We clustered the email addresses using DBSCAN and Levenshtein distance to find registrants who are likely the same person registering under different accounts, but this had only a small impact on our analysis. After clustering we had 56,279 registrants, 105 of whom deleted more than 100 domains early. These bulk deletions may contribute to the spikes seen in Figure 4. We count the number of deletions each day by individual registrants and find that the spike on 9/23/19 was caused by a few bulk registrants, one of whom deleted 952 domains on that day.

We argue that these types of mass deletions suggest malicious use. Bulk registrants are often domain speculators which are not necessarily malicious, but hope to sell the domain for a profit and collect small amounts of advertisement money from parked pages in the meantime [44]. Therefore, they have no incentive to delete their domains. Contrastingly, spam domains and DGA C&C domains are also commonly registered in bulk and tend to be short lived [21, 45].

4.7 Registrar Patterns

We take advantage of the registrar listed by RDAP/WHOIS to find the registrars which either enable or otherwise actively delete domains. Table 4 shows the top 10 registrars in terms of number of early deletions. Godaddy has by far the most early deleted domain names, but according to ICANN transaction reports [27] it is also the largest registrar in the world by number of registered domains. To address this, we also ranked registrars normalized by their total number of .com domains reported by ICANN, excluding registrars with less than 100 domains. This reveals some additional smaller registrars which had notable portions of their domains deleted. Dropcatch.com is at the top of this list which is interesting because they sell expired and re-registered domains which Miramirkhani, et al. showed are rarely used for legitimate web pages [37]. The ranking of registrars by total deletions is similar for 2017 with six of the top ten appearing on both lists. Notably, the registrar Web Drive deleted 27% of their total domains early over the course of our measurement.

While we discovered that many domains are suspended by the registrar with a `client hold` status, some registrars may delete abusive domain names instead, which could explain high numbers of deletions from certain registrars. In particular, we found that Godaddy only suspended 412 during the same time period which is only 0.0009% of their total domains. Godaddy and Wild West Domains are both among the lowest of all registrars in terms of percentage of domains suspended. While these registrars deleting domains could explain a large number of early deletions, it is surprising that registrars choose deletion over `client hold`, thereby allowing the domain to be re-registered.

We investigated the largest 15 registrars as well as all of those shown in Table 4 to determine how easy it is for registrants to delete their own domain names. We found that only GoDaddy, Google, Hetzner Online, and RegistryGate refer to or provide an option for users to delete domains [18, 19, 22, 42]. This is further evidence that the other registrars are deleting domain names early, but we cannot rule out the possibility that these registrars would allow deletions if requested through customer support.

Figure 4 exhibited multiple significant outliers which we investigate further using WHOIS listings for the registrar and registrant email addresses. As with our analysis of registrants (Section 4.6), for each outlier, we group together early deletions of domains belonging to the same registrar. Three of these outliers are dominated by single registrars:

- **3/16**: The largest outlier by far, Domain.com had 12,014 early deletions while the second most was 1&1 Internet with only 119. Several Hotmail email addresses were tied to hundreds of these domains.
- **5/28**: Cronon had 3,576 early deletions, again significantly more than Godaddy having the second most at 182, while no registrants deleted more than 16.
- **6/9**: 1&1 Internet had 4,267 early deletions with GoDaddy trailing at 207 with no dominant registrant emails.

The evidence strongly suggests that certain registrars are responsible for these deletions. While registrars are welcome to take action against abusive domain names before they expire, it is surprising that they choose to delete the domain names instead of placing them on `client hold`. Once a domain is deleted, a malicious actor can re-register the same domain and regain control of the infrastructure associated with it. In fact, prior stud-

Registrar	# Early Deletions	Registrar	% of Total Domains
GoDaddy.com, LLC	125,059	DropCatch.com, LLC	7.89
Tucows Domains Inc.	6,084	Ednit Software Private Limited	2.37
Cronon AG	4,896	NetTuner Corp. dba Webmasters.com	1.72
Wild West Domains, LLC	4,713	Vautron Rechenzentrum AG	0.97
Google, Inc.	3,599	Deutsche Telekom AG	0.90
Key-Systems GmbH	2,712	Metaregistrar BV	0.79
Name.com, Inc.	2,545	Cronon AG	0.78
CSC Corporate Domains, Inc.	2,502	RegistryGate GmbH	0.77
RegistryGate GmbH	2,160	Hetzner Online GmbH	0.74
PSI-USA, Inc. dba Domain Robot	1,821	HTTP.NET Internet GmbH	0.68

Table 4: *Registrars with the most early deletions.*

ies have shown that malicious domains are *more* likely to be re-registered [34, 37]. Therefore, by deleting suspicious domains (instead of placing them on hold) registrars are merely inconveniencing malicious actors who can use a new registrar to re-register their domains and resume their malicious campaigns.

4.8 Blacklisted Domains

A possible motive for registrants deleting domains is for them to cover their tracks after malicious use. For example, associated IP addresses and WHOIS information cannot be obtained for deleted domains which may stump later forensic investigations of abusive domains.

To gauge the level of malicious activity we rely on blacklists. Blacklists cannot possibly cover all malicious domains, but with the domains deleted and the web sites down, they are the best available method to estimate the number of malicious early deletions. We referenced multiple sources to find blacklisted deleted domains as described in Section 3.1. We found 402 (0.23%) blacklisted domain names that were deleted early in the 2019 data set and another 1,107 (0.52%) in the 2017 data set. We can expand the number of potentially malicious domains by grouping early deletions by owner. We know from Section 4.6 that some registrants were associated with large numbers of deleted domains. If one of those domains was blacklisted then we may suspect that the registrants' other deleted domains may also be malicious. Using the clusters of registrant emails from the previous section, we mark a registrant as malicious if they deleted at least one blacklisted domain. The number of 2017 domains deleted by these malicious registrants is 9,782 (4.60%), significantly more than was found with blacklists alone. Despite the unavailability of registrant email addresses, we are able to apply a similar analysis to the 402 blacklisted domains in the 2019 data set. We grouped deleted domain names by registrar and date, then clustered very similar domain names using Levenshtein distance and DBSCAN to find groups of domains that were likely created by the same registrant. This resulted in 8,577 clusters containing two or more domains and and average cluster size of 4.5. Then, with the same approach of labeling a group as malicious if at least one of its domains was blacklisted, we extend the number of of potentially malicious domains to 5,028 (2.90%). While associating domains in this way does not guarantee that they are malicious, we apply this method because shared ownership is a concrete connection between domain names. Malicious domains are often registered in bulk, and this is within the capabilities of registrars who ultimately control early deletions.

Even though the percentage of domains found in blacklists may appear small, to get a rough comparison of coverage we check domains that were suspended between 12/31/18 and 1/20/19 and find that only 2,163 (0.47%) appear in the same set of blacklists. We expect that most suspended domains were malicious, and yet only a small percentage of them appear in blacklists. The percentage of blacklisted early deletions is similar and our estimates for potentially malicious domains after association are 6-10 times more than the coverage for suspended domains. Therefore, we are confident that many of these domains were malicious but did not make their way onto blacklists before their deletion.

We suggest that deleting a domain name before it would normally expire is suspicious enough that it may be a signal worth considering when blacklisting/suspending domains. This signal can be used to tie this information to the registrant and scrutinize other domains that they register. Moreover, since malicious domains are often re-registered after their deletion [34, 37] or after sink-holing systems let them expire [11], knowing that a domain was deleted early can predict future abusive behavior. In fact, this is one reason why the owner of a blacklisted domain may choose to delete it, i.e., to make the association with their other live domains less conspicuous.

5 Discussion

Registrars deleting domains. Through two distinct data sets we quantify disappearing domains and identify hundreds of thousands of domain names which were deleted before expiration. By analyzing registrar patterns over time we conclude that at least 10% of deletions may be initiated by registrars. For registrars that are deleting domains to deal with abuse, we recommend that they instead suspend them with `client hold` and/or disable the registrant's account, but maintain control of the domain so that it cannot be re-registered for malicious use. As we mentioned in Section 4.7, most registrars already make it difficult for registrants to delete domain names or at least warn against it which is appropriate for the majority of users. For these registrars we recommend that they go a step further and screen requests to delete domains to determine if the domain names were abused which may be used to find other abusive domains from the same owner.

Registrants deleting domains. For cases where the registrant initiated the deletion, explanations are not obvious, but we present the following possibilities: i) a negative association with another name, ii) attempts to hide malicious activity, or iii) due to ignorance or indifference regarding the domain life-cycle. Since most registrars do not make deleting domains easy, and even the ones that do warn against it, we believe that category three is an unlikely cause. We found many domains fitting the first category and due to our thorough approach to checking multiple types of domain squatting we expect that we found the majority of cases in this category. For the second category, we identify as many malicious domains as possible using blacklists and association by registrant. Due to the limited coverage of blacklists and the difficulty of finding malicious activity after the domain has been deleted, we argue that without another likely explanation, unexplained cases are likely to fall into the second category.

Registration information. Another lesson from this study is that registration data including status, registrar, and dates should be maintained as a publicly available resource. Public access to zone files has been very successful in aiding security research and applications [12,30,34,36,37,43], but it is not enough to identify all registered domain names, nor does it cover all stages of the domain life-cycle making cases like early deletion and dropcatching more difficult to monitor. In the past, query limits on WHOIS were a reasonable precaution to prevent mass collection of registrants' personal information, but with recent changes to WHOIS privacy, largely driven by the EU's GDPR, this is no longer necessary. Now there is an opportunity to make generic, non-personal domain registration information publicly available. RDAP is a step in the right direction particularly with unified data structure and the addition of authorization, but as it is still in pilot and specific zones are managed by different parties, it remains to be seen whether this will continue to be a public resource.

Limitations. A limitation of our study is the inability to determine for each domain name whether its deletion was initiated by the domain owner or the registrar. In Section 4.7, we are only able to use indirect evidence to estimate that at least 9.33% were deleted by registrars, and cannot guarantee that the remainder were all initiated by registrants. Nevertheless, we hope that this work motivates security-conscious registrars who do have this visibility for their customers' domains to further explore premature deletions.

For an analysis of domain names after they have already been deleted, we are also limited in our ability to determine what the domain name was used for while it was still active. Our primary tool is analysis of lexical features, but in many cases this is not enough to identify whether a domain name was registered and used for malicious purposes. We were able to find that 1.9% of early deletions were domain squatting, and based only on blacklists we estimated another 3.84% were likely malicious. We acknowledge that this leaves a majority of cases unexplained, and future work should aim to fill this gap. However, even these limited findings draw attention to the phenomenon of early deletions and warrant a closer look into a practice that has been ignored up to this point.

6 Related Work

A domain name may enter the pending delete state of the domain life-cycle as a result of early deletion or expiration. While re-registration of expired domains, and the security implication of this practice have received extensive attention [31–34, 37], to the best of our knowledge, no one has exclusively looked into the unusual behavior of early deletion. Since our work is the first systematic study on this phenomenon, we review prior work on deleted domains which are closest to our work.

Lauinger et al. [32] showed that there is an intense competition between dropcatch registrars in registering desirable deleted domains. Some of these registrars maintain large numbers of registrar accreditations to be able to submit more requests and in this way increase their chance of catching available domains. Only three large dropcatch registrars control 75% of registrars which translates to millions of dollars in accreditation fees. In more recent work [31], they took a closer look at when expired domain names are re-registered. Using a model to infer the deletion time of domains, they showed that 9.5% of deleted domains are re-registered less than one second after they became available. Miramirkhani et al. found that the domains which are shorter, older, have more residual traffic, and malicious history are more likely to be re-registered and the majority of registrations are for speculative or malicious purposes [37]. They also reported that premature deletions are relatively uncommon compared to the usual yearly life-cycle. In this paper we investigated the reasons and motivations behind early deletions because, although it is a smaller percentage of cases, it creates an opportunity for malicious registrants to avoid detection from DNS monitoring systems.

In related work, potential and actual abuse of the reputation of deleted domains are discussed. Nikiforakis et al. showed that popular websites contain remote JavaScript inclusions that are pointing to expired domains which can be re-registered to perform code injection attacks [39]. Later, Moore et al. investigated the domains of US banks and reported that 33% of these domains are re-registered to host advertisement, distribute malware, or to carry out search engine optimization (SEO) activities [38]. A recent study by Vissers et al. showed that re-registration of expired domain names can result in hijacking thousands of domain names through their name servers [43]. Lever et al. explored the domains that were deleted in a span of six years and for those that were abused for malicious intentions, they examined whether the malicious activity occurred before or after domain re-registration. They found hundreds of thousands of re-registrations occurred with the intention of abusing negative or positive residual trust of the original domains [34]. Recent studies [14, 35] have shown a similar use-after-free problem exists for IP addresses as well. Specifically, Borgolte et. al. [14] demonstrated that stale DNS records pointing to cloud IP addresses could be abused to hijack domains' traffic and create new SSL certificates under an attacker's control.

7 Conclusion

In this paper we presented the first systematic analysis of early domain name disappearances. Using historical zone files, drop lists, and WHOIS for all of 2017 and collecting live RDAP/WHOIS between 12/19/18 and 1/27/19, we uncovered the surprising phenomenon of domain names deleted before their expected expiration date, with thousands of cases every day. We showed that domains deleted early are longer, contain fewer numbers, and are much more likely pronounceable than domains expiring normally. We found thousands of cases of squatting domain names which may have been deleted due to complaints from trademark holders, and domains deleted by malicious registrants potentially to cover digital tracks of abusive activity. We showed that many registrants deleted domains in bulk and a few registrars such as GoDaddy seem to delete domain names rather than suspending them.

We demonstrate issues that arise when relying on zone files to study the global state of domain names and advocate for public access to anonymous registration information to aid automated security tools and forensics. Finally, we recommend that registrars scrutinize registrants who delete domain names and utilize suspensions rather than deletions to prevent malicious domains from being re-registered.

Acknowledgements: We thank our shepherd Katharina Krombholz and the anonymous reviewers for their helpful feedback. This work was supported by the Office of Naval Research (ONR) under grant N00014-16-1-2264 and by the National Science Foundation (NSF) under grants CMMI-1842020, CNS-1617593, and CNS-1617902.

8 Availability

To motivate further research into the unexpected disappearances of domain names and their effect on security, we are releasing our compiled list of 386K domains that disappeared from zone files before their expected expiration date (together with their associated metadata). The compiled dataset can be downloaded from: **https://github.com/timothy-barron/now-you-see-it**.

References

[1] Cert.at Conficker. https://www.cert.at/static/conficker/all_domains.txt.

[2] Google Safe Browsing API. https://developers.google.com/safe-browsing/.

[3] hpHosts. https://hosts-file.net/download/hosts.zip.

[4] Malc0de. https://malc0de.com/bl/BOOT.

[5] Malware Domains. https://mirror1.malwaredomains.com/files/domains.zip.

[6] WHOISDataCenter. https://whoisdatacenter.com/.

[7] ZeuS Tracker. https://zeustracker.abuse.ch/blocklist.php?download=domainblocklist.

[8] Sex.com now officially the most expensive domain name in the world. https://techcrunch.com/2011/02/22/sex-com-now-officially-the-most-expensive-domain-name-in-the-world/, 2011.

[9] Domainiq (a domain intelligence service). https://www.domainiq.com/, 2017.

[10] Estibot appriasal tool. http://www.estibot.com, 2017.

[11] Alowaisheq, E., Wang, P., Alrwais, S., Liao, X., Wang, X., Alowaisheq, T., Mi, X., Tang, S., and Liu, B. Cracking Wall of Confinement: Understanding and Analyzing Malicious Domain Takedowns. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)* (2019).

[12] Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., and Feamster, N. Building a dynamic reputation system for dns. In *USENIX security symposium* (2010), pp. 273–290.

[13] Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources* (2004), Association for Computational Linguistics, pp. 101–108.

[14] Borgolte, K., Fiebig, T., Hao, S., Kruegel, C., and Vigna, G. Cloud strife: mitigating the security risks of domain-validated certificates. In *Proc. Internet Society Symposium on Network and Distributed System Security (NDSS)* (2018).

[15] Coull, S. E., White, A. M., Yen, T.-F., Monrose, F., and Reiter, M. K. Understanding domain registration abuses. In *IFIP International Information Security Conference* (2010), Springer, pp. 68–79.

[16] David Cohn. Parental Controls Bad Word List:NSFW. `https://www.webseoanddesign.com/parental-controls-bad-word-listnsfw/`.

[17] Dinaburg, A. Bitsquatting: Dns hijacking without exploitation. *Proceedings of BlackHat Security* (2011).

[18] GoDaddy. Cancel my domain. `https://www.godaddy.com/help/cancel-my-domain-412`.

[19] Google Domains Help. Delete a domain. `https://support.google.com/domains/answer/6202231?hl=en`.

[20] Halvorson, T., Der, M. F., Foster, I., Savage, S., Saul, L. K., and Voelker, G. M. From. academy to. zone: An analysis of the new tld land rush. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference* (2015), ACM, pp. 381–394.

[21] Hao, S., Thomas, M., Paxson, V., Feamster, N., Kreibich, C., Grier, C., and Hollenbeck, S. Understanding the domain registration behavior of spammers. In *Proceedings of the 2013 conference on Internet measurement conference* (2013), ACM, pp. 63–76.

[22] Hetzner Online. How do I delete a domain from my Managed Server? `https://hetzner.co.za/help-centre/products-and-services/how-do-i-delete-a-hosting-package-from-my-server/`.

[23] Hollenbeck, S. Extensible provisioning protocol (epp). `https://tools.ietf.org/html/rfc5730`, 2009.

[24] ICANN. Registrar advisory concerning the "15-day period" in whois accuracy requirements. `https://www.icann.org/news/advisory-2003-04-03-en`, 2003.

[25] ICANN. AGP (Add Grace Period) Limits Policy. `https://www.icann.org/resources/pages/agp-policy-2008-12-17-en`, 2008.

[26] ICANN. EPP Status Codes. `https://www.icann.org/resources/pages/epp-status-codes-2014-06-16-en`, 2014.

[27] ICANN. Monthly registry reports. `https://www.icann.org/resources/pages/registry-reports`, 2019.

[28] Intelium. Trademark247. `https://www.trademark247.com/`, 2019.

[29] Kintis, P., Miramirkhani, N., Lever, C., Chen, Y., Romero-Gomez, R., Pitropakis, N., Nikiforakis, N., and Antonakakis, M. Hiding in plain sight: A longitudinal study of combosquatting abuse. In *Proceedings of 24th ACM Conference on Computer and Communications Security (CCS)* (2017).

[30] Kountouras, A., Kintis, P., Lever, C., Chen, Y., Nadji, Y., Dagon, D., Antonakakis, M., and Joffe, R. Enabling network security through active dns datasets. In *International Symposium on Research in Attacks, Intrusions, and Defenses* (2016), Springer, pp. 188–208.

[31] Lauinger, T., Buyukkayhan, A. S., Chaabane, A., Robertson, W., and Kirda, E. From deletion to re-registration in zero seconds: Domain registrar behaviour during the drop. In *Proceedings of the 2018 Internet Measurement Conference* (2018), IMC'18.

[32] Lauinger, T., Chaabane, A., Buyukkayhan, A., Onarlioglu, K., and Robertson, W. Game of Registrars: An empirical analysis of post-expiration domain name takeovers. In *Proceedings of the USENIX Security Symposium* (August 2017).

[33] Lauinger, T., Onarlioglu, K., Chaabane, A., Robertson, W., and Kirda, E. Whois lost in translation:(mis) understanding domain name expiration and re-registration. In *Proceedings of the 2016 ACM on Internet Measurement Conference* (2016), ACM, pp. 247–253.

[34] Lever, C., Walls, R., Nadji, Y., Dagon, D., McDaniel, P., and Antonakakis, M. Domain-z: 28 registrations later measuring the exploitation of residual trust in domains. In *Security and Privacy (SP), 2016 IEEE Symposium on* (2016), IEEE, pp. 691–706.

[35] Liu, D., Hao, S., and Wang, H. All your dns records point to us: Understanding the security threats of dangling dns records. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016), ACM, pp. 1414–1425.

[36] McGrath, D. K., and Gupta, M. Behind phishing: An examination of phisher modi operandi. *LEET 8* (2008), 4.

[37] Miramirkhani, N., Barron, T., Ferdman, M., and Nikiforakis, N. Panning for gold.com: Understanding the dynamics of domain dropcatching. In *Proceedings of the Web Conference* (April 2018).

[38] Moore, T., and Clayton, R. The ghosts of banking past: Empirical analysis of closed bank websites. In *International Conference on Financial Cryptography and Data Security* (2014), Springer, pp. 33–48.

[39] Nikiforakis, N., Invernizzi, L., Kapravelos, A., Van Acker, S., Joosen, W., Kruegel, C., Piessens, F., and Vigna, G. You are what you include: large-scale evaluation of remote javascript inclusions. In *Proceedings of the 2012 ACM conference on Computer and communications security* (2012), ACM, pp. 736–747.

[40] Peter Norvig. Natural Language Corpus Data: Beautiful Data. `http://norvig.com/ngrams/`.

[41] Plohmann, D., Yakdan, K., Klatt, M., Bader, J., and Gerhards-Padilla, E. A comprehensive measurement study of domain generating malware. In *Proceedings of the USENIX Security Symposium* (2016), pp. 263–278.

[42] RegistryGate. Registration and management conditions for domain names. `http://www.registrygate.com/fileadmin/user_upload/RyG_KS_ICANN_Registration-Agreement_Registrant_ENGLISH.pdf`.

[43] VISSERS, T., BARRON, T., VAN GOETHEM, T., JOOSEN, W., AND NIKIFORAKIS, N. The wolf of name street: Hijacking domains through their nameservers. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), ACM, pp. 957–970.

[44] VISSERS, T., JOOSEN, W., AND NIKIFORAKIS, N. Parking sensors: Analyzing and detecting parked domains. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS)* (2015).

[45] VISSERS, T., SPOOREN, J., AGTEN, P., JUMPERTZ, D., JANSSEN, P., VAN WESEMAEL, M., PIESSENS, F., JOOSEN, W., AND DESMET, L. Exploring the ecosystem of malicious domain registrations in the. eu tld. In *International Symposium on Research in Attacks, Intrusions, and Defenses* (2017), Springer, pp. 472–493.

[46] WANG, Y.-M., BECK, D., WANG, J., VERBOWSKI, C., AND DANIELS, B. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. *SRUTI 6* (2006), 31–36.

[47] SecurityTrails: The World's Largest Repository of historical DNS data. `https://securitytrails.com/dns-trails`.

[48] WIPO. Schedule of Fees under the UDRP. `https://www.wipo.int/amc/en/domains/fees/`, 2002.

A Statistical Analysis of Domain Features

In order to test our domain features presented in section 4.5, the Welch's *t*-test was chosen for its robustness on large sample sizes and to handle unequal variances. For each feature in Table 2, we test the difference between sample means under the null hypothesis that each sample is drawn from the same population. Since our samples are very large (>172K), the distribution of sample means is closely approximated by the normal distribution. Table 5 shows the Cohen's *d* effect size, *t*-statistic, and *p*-value for each test, comparing the feature means of early deletions against the three other categories of domains. In three cases we have large enough *p*-values that these differences are not considered significant. Two of these are tests for the feature *percent containing adult keywords* which is very similar between early deletions, normal expirations and suspensions. The third is *percent unpronounceable* where early deletions and Alexa domains are similar despite the difference with the other two categories. In all other tests we obtain very small *p*-values leading us to reject the null hypothesis. The effect sizes indicate that the most significant difference between early deletions and normal expirations is in the mean *number of words* and *percent unpronounceable*. Overall, the effect sizes coincide with our observations in section 4.5 of the most notable differences between domain categories.

B Domain Name Topics

In section 4.5 we presented the top domain topics among domains deleted early. We extend this to the Alexa top 1M popular domains, and domains that were suspended or expired between 12/31/18 and 1/20/19. Early deleted domain names are much more likely to have a mapped topic with only 27% unknown while the other categories have between 33% and 40% unknown. This intuitively follows our observation that domain names deleted early tend to be longer and contain more words, making it more likely that we find a match in WordNet Domains. Table 6 compares these categories with the full list of 28 topics. For a more fair comparison, this table only represents domain names that are not unknown. There are a few topics that are ranked in different orders between columns, but overall the distribution of topics appears similar. To support this observation we applied a Chi-square test comparing early deletions to each of the other three categories. The effect size and *p*-value are shown at the bottom of Table 2 below the column being compared to early deletions. The effect size used is $\varphi = \sqrt{\frac{\chi^2}{N}}$. Because the samples are so large, each test produced a *p*-value very near zero, but the effect sizes are small enough that it is not clear that there is a meaningful difference.

Feature	Alexa Top 1M			Normal Expirations			Suspensions		
	Cohen's *d*	*t*-statistic	*p*-value	Cohen's *d*	*t*-statistic	*p*-value	Cohen's *d*	*t*-statistic	*p*-value
Length	0.53	174.79	0.00	0.13	51.05	0.00	0.23	82.11	0.00
Entropy	0.38	155.43	0.00	0.14	58.15	0.00	0.23	84.06	0.00
Number of words	0.33	112.54	0.00	0.15	60.40	0.00	0.29	102.85	0.00
Percent containing numbers	0.16	51.76	0.00	-0.12	-54.45	0.00	-0.18	-68.07	0.00
Percent containing hyphens	-0.04	-15.85	0.00	0.04	14.28	0.00	-0.09	-33.87	0.00
Percent containing adult keywords	-0.03	-11.21	0.00	0.00	0.3984	0.69	0.01	1.918	0.06
Percent unpronounceable	0.00	0.6791	0.50	-0.16	-72.26	0.00	-0.23	-89.43	0.00

Table 5: *Statistical comparison of domain features of early deletions against popular, expired, and suspended domains.*

Topics	Early Deletions	Alexa Top 1M	Normal Expirations	Suspensions
Economy/commerce/banking	11.99%	9.90%	12.19%	10.90%
Science/technology	11.58%	13.53%	12.32%	13.14%
Play/sports	8.40%	9.25%	7.91%	7.70%
Health/medicine	7.69%	6.89%	7.61%	7.53%
Architecture	5.79%	6.16%	6.37%	5.57%
Geography	5.58%	5.41%	5.48%	5.79%
Politics/government	5.39%	3.92%	4.48%	4.51%
Art	4.83%	4.52%	5.50%	5.48%
Travel/transport	4.72%	5.43%	5.04%	4.88%
Person	4.19%	3.95%	4.07%	4.13%
Writing/language	3.17%	4.27%	3.22%	3.30%
History/humanity	2.64%	2.41%	2.35%	2.47%
Psychology	2.56%	1.76%	2.14%	2.26%
Media/telecommunication	2.39%	3.64%	2.30%	2.42%
Religion	2.31%	1.55%	1.72%	1.99%
Earth/environment	2.21%	2.11%	2.25%	2.31%
Time period	2.12%	1.93%	1.96%	2.10%
Food	2.05%	1.90%	2.27%	2.12%
Education	2.01%	2.16%	1.55%	1.62%
Quality	1.77%	1.89%	1.96%	1.80%
Animals	1.61%	1.69%	1.70%	2.00%
Administration	1.27%	1.45%	1.17%	1.30%
Metrology	1.27%	1.50%	1.48%	1.87%
Fashion	0.92%	0.95%	1.34%	1.05%
Sexuality	0.54%	0.77%	0.55%	0.62%
Number	0.51%	0.58%	0.59%	0.66%
Color	0.48%	0.46%	0.47%	0.48%
Paranormal	0.01%	0.02%	0.01%	0.01%
Chi-square test		φ, *p*-value 0.07, 0.00	φ, *p*-value 0.03, 0.00	φ, *p*-value 0.05, 0.00

Table 6: *Comparison of domain name topics between popular, expired, suspended, and deleted domain names.*