Extortion or Expansion?
An investigation into the costs and consequences of ICANN's gTLD experiments

Shahrooz Pouryousef[1], Muhammad Daniyal Dar[2], Suleman Ahmad[3], Phillipa Gill[1], and Rishab Nithyanand[2]

[1] University of Massachusetts, Amherst MA, USA
{shahrooz,phillipa}@cs.umass.com
[2] University of Iowa, Iowa IA, USA
{rishab-nithyanand,mdar}@uiowa.edu
[3] University of Wisconsin-Madison, Madison WI, USA
{suleman.ahmad}@wisc.edu

Abstract. Since October 2013, the Internet Corporation of Assigned Names and Numbers (ICANN) has introduced over 1K new generic top-level domains (gTLDs) with the intention of enhancing innovation, competition, and consumer choice. While there have been several positive outcomes from this expansion, there have also been many unintended consequences. In this paper we focus on one such consequence: the gTLD expansion has provided new opportunities for malicious actors to leverage the trust placed by consumers in trusted brands by way of typosquatting. We describe gTLDtm (The gTLD typosquatting monitor) – an open source framework which conducts longitudinal Internet-scale measurements to identify when popular domains are victims of typosquatting, which parties are responsible for facilitating typosquatting, and the costs associated with preventing typosquatting. Our analysis of the generated data shows that ICANN's expansion introduces several causes for concern. First, the sheer number of typosquatted domains has increased by several orders of magnitude since the introduction of the new gTLDs. Second, these domains are currently being incentivized and monetarily supported by the online advertiser and tracker ecosystem whose policies they clearly violate. Third, mass registrars are currently seeking to profit from the inability of brands to protect themselves from typosquatting (due to the prohibitively high cost of doing so). Taken as a whole, our work presents tools and analysis to help protect the public and brands from typosquatters.

1 Introduction

With the stated goal of improving the choice of domain names for brand holders, since 2013, ICANN approved the delegation of over 1.2K new generic Top Level Domains (gTLDs). Since its initial expansion, the new gTLD program has been experiencing continuous growth with processes for adding new gTLDs being more codified and streamlined [1]. We provide a brief history of the gTLD expansion in the Appendix of this paper (§5.1). While these new gTLDs have been a boon for organizations seeking to gain relevant domain names for their brands, they also present exciting opportunities for malicious actors. Previous work examined the types of content hosted on the domains using the new gTLDs and found higher incidence rates of malicious content such as malware, in comparison with domain names using the old gTLDs [1,2,3]. The problem is exacerbated by the fact that domain names are a source of trust with sites using HTTPS and certificates linked to them and cyber criminals have exploited this trust placed by users in safe domain names by utilizing visually similar domain names [4] or typos of these safe domain names [5,6,7] to launch attacks – a practice generally referred to as typosquatting. Despite many studies analyzing the incidence rates of typosquatting in the context of the original gTLDs [5,6,7,8,9,10], there has been little attention on typosquatting using the new gTLDs. What remains unknown, specifically, is how ICANN's gTLD expansion has impacted established and trusted brands seeking protection from typosquatting. In this paper, we fill this gap. Our overall objective is to understand how ICANN's gTLD expansion impacts brands trusted by Internet users. To achieve this objective, we develop techniques to reliably identify and monitor typosquatting and understand the challenges and costs facing organizations seeking to protect their brands from typosquatters. More specifically, we make the following contributions.

gTLDtm: The gTLD typosquatting monitor. We develop a framework, called the gTLD typosquatting monitor (gTLDtm), which routinely performs Internet-scale measurements to identify when popular domains are victims of typosquatting, which parties are facilitating the typosquatting – on old and new gTLDs, and what the cost is to prevent typosquatting. gTLDtm is open source and available at https://sparta.cs.uiowa.edu/projects/auditing.html . Periodic dumps of gTLDtm gathered data and inferences are also available for download. The data gathered by this framework forms the basis of the analysis conducted in this paper and will serve many communities seeking to understand the abuse of user trust in established brands online – e.g., studies characterizing typosquatting for fake news and propaganda dissemination, malware distribution, and online scams, amongst many others. The framework may also be used by organizations seeking to identify instances of typosquatting on their brands. During construction of this framework, we also identify several inconsistencies in records maintained by ICANN and gTLD registries.

Characterizing perpetrators and victims of typosquatting. We uncover the mechanics of typosquatting – e.g., types of content and domains that are targeted by typosquatters, the role of advertisers and mass registrars in the typosquatting ecosystem, the extent of knowledge of typosquatters by web intelligence sources such as McAfee [11], as well as the intent behind typosquatting and the cost for a victim to defend against typosquatting. Our characterization explicitly focuses on identifying the differences in these mechanics for each generation of gTLDs. This allows us to understand how the typosquatting ecosystem has changed as a consequence of ICANN's gTLD expansions.

2 The gTLDtm Framework

In order to understand the ecosystem of typosquatting, we construct a measurement framework called the gTLD typosquatting monitor (gTLDtm). gTLDtm consists of several components: a URL curator, typo generator, data generator, typosquatting detector, and a defense cost estimator. The interaction between these components is illustrated in Figure 1 and described in this section.

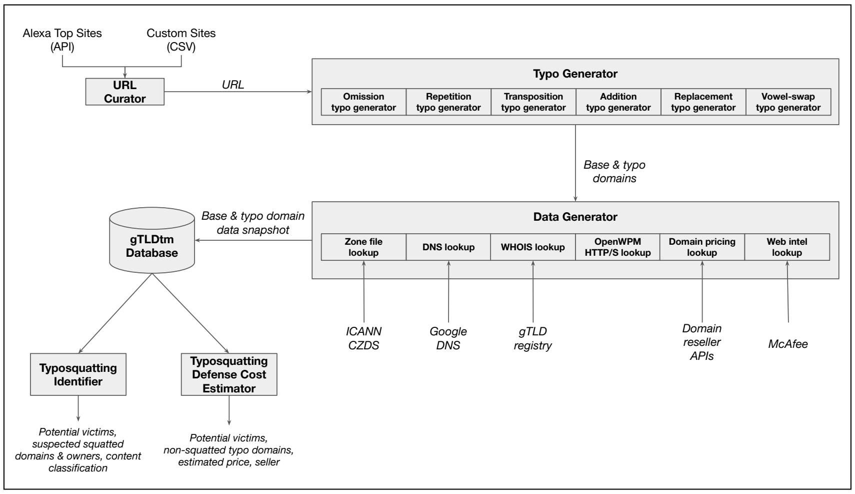


Fig. 1: The gTLDtm architecture.

2.1 URL curation and typosquatting candidate generation

The URL curator periodically fetches a list of URLs whose typos will be monitored by our system. The current implementation grabs a list of 231 most popular URLs in the News, Business, Society, and Shopping categories using the Alexa top sites API. It also has the capability of accepting custom lists of URLs. Given a base domain (obtained by our URL curator), we need to generate domain names likely to be targeted by typosquatters seeking to exploit user trust in the base domain. We do this by leveraging six typosquatting generation techniques: omissions, repetitions, transpositions, additions, replacements, and vowel-swaps. These techniques are applied to the second-level domains (SLDs) only. *For each second-level domain typo generated, we use every possible gTLD to form a typosquatting candidate.* We explain each of the six second-level domain typo generation techniques using the SLD "`icann-example`" as an example: (1) **Omission:** We generate new SLDs by excluding a single character in the base SLD. This method yields `cann-example`, `iann-example`, . . . , and icann-exampl as typo SLDs; (2) **Repetition:** We generate new SLDs by repeating a single character in the base SLD. This method yields iicann-example, iccann-example, . . . , icann-examplee as typo SLDs; (3) **Transposition:** We generate new SLDs by swapping two adjacent characters in the base SLD. This method yields ciann-example, iacnn-example, . . . , icann-exampel as typo SLDs; (4) **Addition:** We generate new SLDs by inserting an additional character at the end of the base SLD. This method yields icann-examplea, . . . , icann-examplez as typo SLDs; (5) **QWERTY- and visually- adjacent replacements [12]:** We generate new SLDs by replacing a single character in the base SLD with one which is adjacent to it on the QWERTY keyboard. This method yields ocann-example, ucann-example, . . . , icann-examplw as candidate typosquatting SLDs. In addition, we generate new SLDs by replacing a single character in the base SLD with one which is visually similar to it (using the sans-serif font). This method yields lcann-example as a typo SLD; and (6) **Vowel-swap [13]:** We generate new SLDs by replacing the vowel in the base SLD with another vowel. This method yields acann-example as a typo SLD. We are currently working on incorporating new typo-generation strategies into our measurement framework.

2.2 Domain intelligence and data gathering

For each base and typo domain, gTLDtm gathers domain intelligence and domain metadata from a variety of Internet authorities. These are described below.

Zone files. ICANN mandates that all open gTLD registries make their up-to-date zone files available to the public via ICANN's CZDS, after the user is able to identify themselves via a physical and IP address [14]. gTLDtm downloads all the zone files made available by the ICANN CZDS repository [15] each day. Given a domain name as input, gTLDtm verifies that it is present in the appropriate zone file. This helps us infer the registration status of a domain.

DNS and WHOIS records. Given a domain name as input, gTLDtm gathers A, AAAA, MX, and NS records by querying Google's public DNS server at 8.8.8.8. Similarly, it also fetches WHOIS records from the corresponding gTLD registry. Data extracted include the *domain registration date*, *registrar*, *organization*, and *contact emails*. This data helps us infer ownership information of a domain.

Web content. Given a domain name, we also attempt to make connections via HTTP and HTTPS to them. We utilize the OpenWPM [16] crawler to visit the domain and gather data associated with the content hosted on it. This includes page content, content sources, cookies, and certificates. This data helps us make inferences about content type and registration intent.

Domain pricing data. Domain resellers are third-party organizations that offer domain name registrations through authorized registrars such as GoDaddy and Namecheap. gTLDtm is registered as a domain reseller with one of the most popular mass registrars – GoDaddy. gTLDtm uses the domain reseller API exposed by this registrar [17,18] to obtain data regarding the availability of the

	pre-2000	2000-12	post-2012
gTLDs	7	15	1.2K
w/ access to zone file	1	7	715
Typosquatting candidates	22K	47.8K	3.9M
Owned domains	8.8K	7.5K	353.4K
w/ WHOIS records	8.6K	6.1K	195.6K
w/ DNS records	7.4K	3.8K	300.9K
w/ Zone file entry	6.7K	198	10.3K
w/ HTTP(S)	625	555	9.6K
w/ TLS certificate	152	437	3.8K
Categorized by Mcafee	579	335	1.4K
Unowned domains	13.2K	40.3K	3.5M
w/ pricing data (randomly sampled)	352	514	29.7K

Table 1: Data gathered by gTLDtm for 231 base domains between 03-10/2019.

input domain name and the associated cost of purchase. This data helps us estimate the cost of registering an typo domain.

Web intelligence data. gTLDtm also seeks to gather intelligence about a domain name from existing domain categorization services. Given an input domain name, gTLDtm makes a request for the domain category (if available) to the McAfee categorization service [11]. This data helps us make inferences about the content type and registration intent.

All together, the data gathered by gTLDtm can be used to make inferences about the ownership of a domain, the type of content it serves, the intent behind its registration, and the cost associated with its purchase. gTLDtm currently repeats this data gathering once every fortnight. A summary of the data gathered by gTLDtm that was used is shown in Table 1. The data shows that there are numerous inconsistencies in the data made available by gTLD registries – e.g., one would expect every domain with a WHOIS record would have a zone file record, but this is not the case. We note that the registries of the post-2012 gTLDs have been the most inconsistent. To deal with this challenge, we categorize domains which have either a valid WHOIS, DNS, or zone file record to be "owned" and those with no WHOIS, DNS, or zone file record to be "unowned".

2.3 Typosquatting identification and domain cost estimation

At a high-level, we say that a typo domain is being squatted on if the entity owning the base domain does not also own the typo domain.

Identifying domain owners. In order to uncover the owner of an owned domain, we rely on the organization details (i.e., *name* and *email*) reported by the WHOIS record. In rare cases (<200) where a WHOIS record does not exist but a DNS or zone record does (due to inconsistent records), we use the owners of the DNS infrastructure (i.e., NSes) reported by DNS records or zone files.

Recognizing typosquatting. We identify when the owner of a base domain is different from the owner of a typo domain, as different owners imply typosquatting. This process is complicated as simply checking for inequality of strings is insufficient for identifying differences in ownership due to inconsistencies in the domain registration process – e.g., we observed the organization names Name.com, Inc., Name.com, and Name, Inc. in the WHOIS records for domains are all owned by the same mass registrar (name.com). To circumvent this, we use a conservative approach for each (base, typo) domain pair: (1) if both domains list identical organization contact details in their WHOIS records, we conclude that they have the same owners; (2) for remaining domain pairs, we find the longest contiguous subsequence of the organization name for each domain (e.g., Name.com, Name.com, and Name in our previous example) and check if the similarity of the extracted sequence is high (>.50: determined through a manual pilot study involving 200 randomly sampled pairs to have a false-positive rate of .01), we say they have the same owner; (3) any remaining domain pairs are said to have different owners. We note that a similar approach has been leveraged in previous work seeking to identify owners of ASes and their siblings [19]. We do not rely on comparisons of hosting infrastructure due to the possibilities of inaccurate conclusions brought by the widespread use of popular CDNs by popular websites and typosquatters. Similarly, we are currently unable to identify inaccuracies caused by the practice of outsourcing defensive domain registrations to organizations such as MarkMonitor.

Unowned domain cost estimation. To identify the cost of an unowned domain, we randomly sampled unowned typo domains that had SLDs which were up to a Damerau-Levenstein edit-distance of three away from the base domain. Random sampling was performed due to constraints on the number of queries that our reseller API permitted us to make (60 queries/minute). Given the cost distributions for typo domains at a particular edit distance, we extrapolate the estimated cost for purchasing all domains at that edit distance.

3 Results

In total, our method identified 188K typosquatted domains (from 4M candidate domains). Of these, 176K domains were from the post-2012 gTLD era (with 6.8K (pre-2000) and 5.4K (2000-2012) across the other eras respectively). We attribute this large skew towards post-2012 gTLDs to the fact that there are over 1.2K post-2012 gTLDs in comparison to just 22 pre-2012 gTLDs. This has two major consequences: (1) post-2012 gTLDs present more opportunities for typosquatting due to the larger number of typosquatting candidate domains and (2) due to the large number of candidates described in (1), it is increasingly expensive for brands to protect themselves by defensive registrations. We note that although ICANN provides Trademark Clearinghouse (TMCH) [20] which allows brands to perform defensive registrations on new gTLDs before they are open to public registration, the TMCH limits access only to paying members (up to $750 per trademark) and only allows registration of domains which exactly match the brand trademark (e.g., for the organization registered as ICANN Example: `icannexample.money` and `icann-example.money` may be pre-emptively registered with TMCH, but registration of any typos such as

	length	rank	category {shop, news, biz, soc}
Linear regression fit on $risk_{norm}$			
R^2 score	.76***	.70**	NA
Pearson correlation coefficient	.57	-.79	NA
Logistic regression classifier			
Accuracy: 81%			
Log-odds ratios	-3.4	-5.3	{-4.0, -3.1, -3.3, -2.9}
Decision tree classifier			
Accuracy: 97%			
Gini feature importance	.23	.56	{.02, .03, .02, .04}

Table 2: Relationships between base domain characteristics and $risk_{norm}$. ** and *** indicate F-test p-values of $< 10^{-2}$ and $< 10^{-3}$ for our linear regressions.

`icann-examples.money` are not allowed). These consequences are further compounded by the non-uniform release of new gTLDs which prevent a single effort to register all trademarked domains – instead forcing constant monitoring and action.

3.1 Characteristics of typosquatting victims

Our 231 base domains were found to have 188K typosquatted domains. We now analyze the characteristics of the base domains which make them vulnerable to being typosquatted on. We refer to the number of typosquatting candidates for a base domain as $risk_{potential}$, the number of typosquatted domains for a base domain as $risk_{realized}$, and their ratio as $risk_{norm}$. To explore the relationships between characteristics of the base domains (i.e., length, rank, and category) and risk outcomes, we rely on two approaches: (1) linear regressions and correlations to measure the dependence and statistical significance of the variables and (2) using interpretable machine learning models on base domain characteristics and domain risk to measure the predictive nature of each characteristic. Our intuition with the latter approach is that if an interpretable classifier (e.g., logistic regression or decision tree classifier) can achieve a reasonable high classification success rate, then interpreting their feature importance will yield domain characteristics that are predictive of the likelihood of a domain being typosquatted on. For our classification task, models were built to predict the level of normalized risk associated with the domain (each level was associated with a quartile from the distribution of all risks). In order to interpret the logistic regression model, we computed the estimated weights for each feature and their corresponding log-odds ratio. If the log-odds ratio of a feature f is x, it means that a unit increase in f changes the odds of our outcome variable y by a factor of e^x when all other features remain the same. Therefore, higher values are indicative of more predictive features. These log-odds for the length and rank features are shown in Table 2. In order to interpret the decision tree model, we computed the Gini importance score for each feature. At a high-level, the Gini importance counts the number of times a feature is used as a splitting variable in proportion with the fraction of samples it splits. We expect higher scores to represent more important features.

Our results are shown in Table 2. Here we see that there are statistically significant relationships between base domain lengths and ranks with the associated $risk_{norm}$. Our 10-fold cross-validated interpretable classifier models, whose task was to classify a base domain into its correct $risk_{norm}$ quartile, also found these characteristics strongly predictive of the quartile range of $risk_{norm}$. Interestingly, our analysis showed that the category of the base domain was not predictive of its $risk_{norm}$.

Takeaway. A domain's normalized typosquatting risk ($risk_{norm}$) is predictable using off-the-shelf interpretable classifiers. When considering individual features, the rank of the domain is the most predictive feature, while the domain category contains little predictive information. This suggests that higher ranked domains are the most common target for typosquatters.

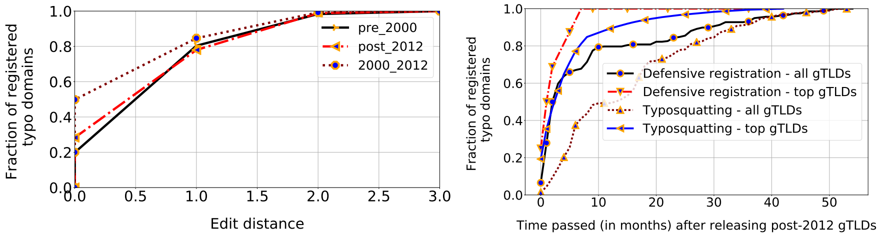
3.2 Characteristics of typosquatted domains

We now analyze characteristics of typosquatted domains which use different era gTLDs with a specific focus on how they are selected, used, monetized, and understood by the web.

gTLD era	Content-3rd	Parked-3rd	Parked-Orig	Redirect-3rd	Redirect-Orig	Sale	Unused
pre-2000	26.91	12.43	0.68	0.82	5.61	45.41	8.14
2000-2012	23.12	8.35	0.41	0.61	6.46	55.41	5.65
post-2012	37.23	12.82	0.72	0.77	3.74	38.74	6.20

Table 3: Typosquatted domain intent by gTLD era (as a percentage of all non-error pages). A suffix of '-3rd' indicates that the inferred intent was associated with a third-party and '-orig' indicates that the intent was associated with the original base domain.

How are typosquatted domains selected by squatters? As shown by our six typosquatting candidate generation methods, there are millions of targets for typosquatters to select with each having relatively short edit distances (less than 3) from a base domain. To understand the predictive nature of the edit-distance, gTLD era, and typo generation method on the `is_domain_squatted_on` variable, we use interpretable logistic regression and decision tree classifiers to find the most predictive features of typosquatted domains. We convert each of our inputs into binary features (e.g., `is_pre2000_gTLD`, `is_post_2012_gTLD`, etc.) and use a 10-fold cross-validation evaluation. Our classifiers had accuracies of 62% and 69% in identifying typosquatted domains from all candidates, respectively. Our analysis of the predictiveness of each feature finds that domains with lower edit distances from the base and using the 'omission' typo generation method are most likely to be squatted on. Figure 2a shows number of registered typo domains as a function of edit distance from base domains in each gTLD era. As it is clear, most of the typo domains (%80) have a short edit distance (less than 2) from a base domain. Amongst the different eras of gTLDs, pre-2000 gTLDs are most likely to be squatted on (followed closely by post-2012 gTLDs, while 2000-2012 era gTLDs are not predictive of squatting). We note that our analysis tool may be leveraged for brands to identify which domains need to be targeted for pre-emptive defensive registration.



(a) CDF of typosquatting domain registrations as a function of edit distance from base domain.

(b) CDF of domain registrations as a function of time since release of the post-2012 gTLD.

Fig. 2: CDF of domain registrations as a function of edit distance and time of gTLD release.

How are typosquatted domains being used by squatters? In order to understand how typosquatted domains are being used, we relied on a three-step process: similarity computation, clustering, and tagging. First, we compute the semantic pairwise-similarity of the textual content in each `html` page fetched by our framework's OpenWPM crawling module. To make this process scalable, we rely on Jenks natural breaks optimization to find ideal clusters based on the one-dimensional parameter: file size. The intuition here is that the similarity between files belonging to different Jenks clusters will be low owing to the large differences in their file sizes. We then compute the similarity matrix such that the similarity of files in different Jenks clusters is set to zero and only intra-Jenks-cluster similarities are computed. Using this similarity matrix, we use k-means clustering to identify clusters of similar pages. k was determined by iterating through all possible values and selecting the candidate value with the highest silhouette score. Our clusters achieved a silhouette score of 0.48 with $k = 54$ clusters. Finally, we manually inspected and tagged 25 randomly sampled pages from each cluster to verify similarity. One of nine tags was then assigned to each cluster: content-original, content-third-party, parked-original, parked-third-party, redirect-original, redirect-third-party, sale, unused, and error.

Our results, broken down by gTLD era, are shown in Table 3. Here we notice that approximately 75% of all typosquatted domains identified by our framework were either hosting third-party content (i.e., content not provided by the base domain) or listed for sale. On average, less than 4% of all typosquatted domains were parked by or redirected to their base domains. Broken down by gTLD era, we see that the typosquatted domains using post-2012 gTLDs are indeed more likely to host content from parties unrelated to the base domain. While we do not currently study the nature of the differences in content in this study, it is clear that this often results in negative impact for users and brands. For example, post-2012 gTLD typos of the `cbsnews` base domain were frequently used to spread political misinformation during the 2016 US Presidential elections – simultaneously harming public discourse and brand reputation.

How are typosquatted domains monetized? While our analysis of the domain intent yields some insights into how typosquatted domains are being used, we also seek to understand how the advertising and tracking ecosystem fuels the typosquatting economy. To this end, we analyzed the incidence rates of different advertising and tracking services using the Easylist and Easyprivacy filter lists [21]. We notice several interesting trends here. First, 67% of all the post-2012 gTLD typosquatters hosting third-party content served ads or hosted trackers in comparison to 53% of the other typosquatted domains. Interestingly, the ad and tracker networks participating in the typosquatting ecosystem vary by the gTLD era. Over 1.6K unique networks were observed in the post-2012 gTLD typosquatted domains in comparison to 1.2K and 384 in the pre-2000 and 2000-2012 eras gTLD typosquatted domains. We identified 103 unique domains serving ads only in the post-2012 gTLD typosquatted domains, including vertamedia, adsnative, and others. We note that the top 20 ad providers for the base domains were all observed in large fractions of typosquatted domains. This suggests the absence of enforcing policies that are meant to prevent the monetization of harmful practices such as typosquatting – e.g., Google's adsense (which was the most prevalent advertising service in our typosquatted domains) policy prohibits using their program to place ads on sites which have 'misrepresentative content' including content which 'misrepresents, misstates, or conceals information about you, your content or the primary purpose of your web destination' or 'falsely implies having an affiliation with, or endorsement by, another individual, organization, product, or service' [22].

How quickly do brands perform defensive registrations? Using the "creation date" entry in each typosquatting candidate domain WHOIS record and knowledge of the release dates for each gTLD (gTLD's delegation date based on ICANN), we seek to understand the amount of time that passes between the availability of a typosquatting candidate domain (using a post-2012 gTLD) and its registration by brands and typosquatters. Figure 2b shows the domains registered by typosquatters and organizations with post-2012 gTLDs as a function of time since release of the post-2012 gTLDs. We find that in the cases where brands do make defensive registrations to prevent typosquatting, a majority occur within the first year of the domains availability (85% of the time when considering all post-2012 gTLDs and 98% of the time when considering only the most popular post-2012 gTLDs observed in our dataset of registered typosquatting candidates (i.e., `app`, `media`, `mobi`, `xxx`, and `agency`)). Typosquatters are rarely left behind. In fact 30% and 98% of all typosquatted domains using the most popular gTLDs are registered within the first month and year of their public availability, respectively. When considering all post-2012 gTLDs however, we observe that there is no landrush – only 45% are registered within the first year of their availability. Our results show that brands are generally able to outpace typosquatters in registering typosquatting candidate domains. Despite this, our previous results show that typosquatting is extremely common. This points to a barrier in either resources or interest in pre-emptive defensive registrations by brands.

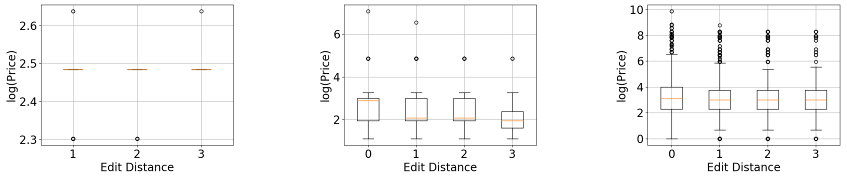
How are typosquatted domains viewed by the web? Web intelligence services such as OpenDNS [23], VirusTotal [24], and McAfee's domain categorizer [11] play a crucial role in protecting users from deceptive online practices. Our measurements of their coverage of typosquatted domains yielded underwhelming results. In total, only 6.6%, 4.5%, and 0.4% of all pre-2000, 2000-2012, and post-2012 gTLD typosquatted domains were found to be categorized. Besides the overall poor coverage of typosquatted domains, these results also suggest that web intelligence services have not yet begun covering domains utilizing new gTLDs to the same extent of those using older gTLDs – leaving users of their services vulnerable to deception from them.

3.3 Cost of brand protection

We now focus on understanding the costs associated with defensive registration of typosquatting candidates by brands.

What is the cost of complete protection from typosquatters? To measure the monetary resources required to register typosquatting candidate domains, we registered as domain resellers on GoDaddy domain registrar which have access to 385 of the all 1230 currently open gTLDs. Since the total number of unregistered typosquatting candidate domains is over $4M and our reseller API are rate limited to 60 queries/minute, we randomly sampled domains with edit distances of less than three from the base domain. In total we received 33K responses to our queries – 352, 514, and 29.7K for queries on candidates using pre-2000, 2000-2012, and post-2012 gTLDs, respectively.

Figure 3 illustrates the cost for each of our queried domains, broken down by gTLDs and edit-distance from the base SLD. Comparing across all gTLD eras, we see that the typo domains with post-2012 gTLDs are generally more expensive than all other eras – regardless of the edit distance from the base domain. Comparing within eras, our results show that typo domains with exact matches of the base domains are also significantly more expensive than higher edit distance domains – i.e., edit-distance 0 domains with post-2012 gTLDs cost $138 on average while edit-distance 1 and edit-distance 2 domains average $95 and $96, respectively. The median of cost of queried domains, broken down by gTLDs and edit-distance 0 from the base SLD for 2000-2012 and post-2012 gTLDs is $17.99 and $21.99, respectively. We also note that GoDaddy advertises these exact match domains as "premium". This suggests that there is knowledge of trademark value of the domain and the increased price and lack of restric-



(a) pre-2000 gTLDs (b) 2000 - 2012 gTLDs (c) post-2012 gTLDs

Fig. 3: Distribution of unowned typosquatting domain prices within 3 edit distances of base domains, broken down by gTLD era.

tions on domain purchase suggests that there is a willful effort to profit off of typosquatting.

From our analysis so far, we can estimate the cost that a brand needs to pay in order to protect itself from typosquatting as a result of the 2012 gTLD expansion. To get the lower bound, we only consider the cost of purchasing domains with open gTLDs (643). To only purchase domains with exactly identical SLDs, a brand would require $63K. Our earlier results suggesting that the majority of typosquatting occurs at an edit-distance of one away from the base SLD indicate that $63K is far from sufficient for meaningful protection from typosquatting. Considering that the average cost of a domain with a post-2012 gTLD and edit-distance of one is $95 and there are hundreds of possible typos with each individual gTLD, it is safe to say that it is not feasible or reasonable to expect brands to be able to protect their domains from typosquatters. Our most conservative estimates show the cost of typosquatting protection against edit-distance 0-1 and open post-2012 gTLD typosquatting to be in the millions of dollars (exact values depend on the length of the base domain SLD).

4 Related Work

ICANNs gTLD expansion. ICANN's gTLD expansion has been the subject of much research over the past several years. Previous research has focused on the economics of the gTLD expansion from the perspective of registries purchasing the new gTLDs. Halvorson et al. [1] found that only a half of the new gTLD-owning registries had recovered their $185K registration costs two years after the expansion. In other work, Halvorson et al. [3] performed specific measurements of the xxx gTLD and found that the gTLD was primarily used for defensive registration with only 4% of the listed domains actually hosting content. In more recent work, the focus has been on how domains with new gTLDs increase security vulnerabilities. Korczy'ski, et al. [2] conducted an investigation of the abuse rates observed in domains using the pre-2012 and post-2012 gTLDs. They found that the incidence rate of spam-domains in the post-2012 gTLD domains was a whole order of magnitude higher than in the pre-2012 gTLD domains. Further, the authors showed an upward trend in the number of spam domains in using the post-2012 gTLDs. Osterweil et al. [25] quantified Man in the Middle (MitM) attacks on web browsing caused due to internal namespace WPAD query leakage. They found that almost all leaked queries are for new gTLD domains and 10% of these highly-vulnerable domains have been registered.

Typosquatting on the web. The incidence of typosquatting on the Internet has been extensively discussed in previous literature. However, the focus has generally been on the pre-2012 gTLDs or on the general behaviours of typosquatters. Agten et al. [5] conducted a longitudinal study on the Alexa top 500 websites and showed that 95% of these websites were actively targeted by typosquatters and that only a handful pursued measures to protect themselves through pre-emptive registrations of candidate domains. Khan et al. [7] demonstrated methods to quantify the harm of typosquatting on the Internet by using time lost for users and visitors lost to brands as their primary metrics. Nikiforakis et al. [26] found a "Typosquatting Cross-site Scripting" (TXSS) vulnerability that exploited typosquatted domains. Wang et al. [27] proposed Strider – a system designed for detecting and discovering large-scale and systematic typosquatters by monitoring neighboring domains. Banerjee et al. [9,10] analyzed phony sites and their network layer behavior, e.g., number of http redirections. While the relationship of domain parking services and malicious domains and parking services has been analyzed in other researches such as [28,29], these papers do not specifically target domain names registered with new released gTLDs.

5 Discussion

Taken in completeness, our study shows that typosquatting incidence rates continue to remain high and that the sheer number of typosquatted domains has significantly increased since ICANN's 2012 gTLD expansion. In fact, typosquatting candidate domains using post-2012 gTLD are already being used by third-parties for content hosting and being monetized at higher rates than any previous gTLD era. Further, our findings highlight a simultaneous failure of multiple entities in the typosquatting ecosystem: (1) advertisers and trackers have failed to enforce their own policies regarding acceptable publishers, therefore presenting monetary incentives for typosquatters and (2) mass registrars, rather than protecting trademarked domains, are themselves seeking to monetize both trademarked and typo domains. These failures have a cost not only to the brands for whom it is unreasonably expensive to defend against typosquatting, but also to the public whose trust in them is more easily exploited by malicious entities – e.g., the 2016 US Presidential election showed that fake news was spread via websites spoofing major media outlets [30]. Finally, our work also shows the cost for brands to protect their own trademarks from typosquatters to be unreasonably high. Taken together, our study suggests that the gTLD expansion has in fact resulted in an ecosystem which facilitates extortion of trusted brands and organizations. We are currently expanding gTLDtm to automatically identify occurrences of typosquatting for the purpose of mis- and dis-information during the 2020 US

Presidential election and also seeking to build tools to enable brands to identify which domains should be targeted for pre-emptive registration.

References

1. Tristan Halvorson, Matthew F Der, Ian Foster, Stefan Savage, Lawrence K Saul, and Geoffrey M Voelker. From. academy to. zone: An analysis of the new TLD land rush. In *Proceedings of the 2015 Internet Measurement Conference*, pages 381–394. ACM, 2015.
2. Maciej Korczynski, Maarten Wullink, Samaneh Tajalizadehkhoob, Giovane Moura, Arman Noroozian, Drew Bagley, and Cristian Hesselman. Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 609–623. ACM, 2018.
3. Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. Xxxtortion?: inferring registration intent in the. xxx tld. In *Proceedings of the 23rd international conference on World wide web*, pages 901–912. ACM, 2014.
4. Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.
5. Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015.
6. Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long" taile" of typosquatting domain names. In *USENIX Security Symposium*, pages 191–206, 2014.
7. Mohammad Taha Khan, Xiang Huo, Zhou Li, and Chris Kanich. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting. In *Security and Privacy (SP), 2015 IEEE Symposium on*, pages 135–150. IEEE, 2015.
8. Nick Nikiforakis, Steven Van Acker, Wannes Meert, Lieven Desmet, Frank Piessens, and Wouter Joosen. Bitsquatting: Exploiting bit-flips for fun, or profit? In *Proceedings of the 22nd international conference on World Wide Web*, pages 989–998. ACM, 2013.
9. Anirban Banerjee, Dhiman Barman, Michalis Faloutsos, and Laxmi N Bhuyan. Cyber-fraud is one typo away. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pages 1939–1947. IEEE, 2008.
10. Anirban Banerjee, Md Sazzadur Rahman, and Michalis Faloutsos. Sut: Quantifying and mitigating url typosquatting. *Computer Networks*, 55(13):3001–3014, 2011.
11. McAfee. https://www.mcafee.com/en-us/index.html, 2019. [Online; accessed 20-October-2019].
12. Tobias Holgers, David E Watson, and Steven D Gribble. Cutting through the confusion: A measurement study of homograph attacks. In *USENIX Annual Technical Conference, General Track*, pages 261–266, 2006.
13. Brody Stout and Keith McDowell. System and method for combating cybersquatting, January 3 2013. US Patent App. 13/612,603.
14. ICANN Centralized Zone Data Service. https://www.icann.org/resources/pages/zfa-2013-06-28-en, 2019. [Online; accessed 20-July-2019].

15. ICANN-CZDS. https://czds.icann.org/home, 2019. [Online; accessed 20-October-2019].
16. Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1388–1401. ACM, 2016.
17. GoDaddy. https://www.godaddy.com/, 2018. [Online; accessed 20-August-2018].
18. NameCheap. https://www.namecheap.com/, 2018. [Online; accessed 20-August-2018].
19. Rishab Nithyanand, Oleksii Starov, Phillipa Gill, Adva Zair, and Michael Schapira. Measuring and mitigating as-level adversaries against tor. In *23rd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016*, 2016.
20. Trade Mark Clearing House. https://www.trademark-clearinghouse.com/, 2019. [Online; accessed 29-October-2019].
21. EasyList. https://easylist.to/, 2018. [Online; accessed 20-August-2018].
22. Google AdSense. https://www.google.com/adsense/, 2019. [Online; accessed 20-October-2019].
23. OpenDNS. www.opendns.com, 2018. [Online; accessed 20-August-2018].
24. Virustotal. www.virustotal.com, 2018. [Online; accessed 20-August-2018].
25. Qi Alfred Chen, Eric Osterweil, Matthew Thomas, and Z Morley Mao. MitM attack by name collision: Cause analysis and vulnerability assessment in the new gTLD era. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 675–690. IEEE, 2016.
26. Nick Nikiforakis, Luca Invernizzi, Alexandros Kapravelos, Steven Van Acker, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. You are what you include: large-scale evaluation of remote javascript inclusions. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 736–747. ACM, 2012.
27. Yi-Min Wang, Doug Beck, Jeffrey Wang, Chad Verbowski, and Brad Daniels. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. *SRUTI*, 6:31–36, 2006.
28. Thomas Vissers, Wouter Joosen, and Nick Nikiforakis. Parking sensors: Analyzing and detecting parked domains. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*, pages 53–53. Internet Society, 2015.
29. Daniel Plohmann, Khaled Yakdan, Michael Klatt, Johannes Bader, and Elmar Gerhards-Padilla. A comprehensive measurement study of domain generating malware. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 263–278, 2016.
30. The Media Trust. https://mediatrust.com/media-center/real-fake-news-spoofed-domains-are-targeting-major-media-outlets, 2018. [Online; accessed 20-August-2019].
31. Domain Name Stat. Domain name registration's statistics. https://domainnamestat.com/.
32. ICANN. About the program | icann new gtlds. https://newgtlds.icann.org/en/about/program.
33. ICANN. gTLD applicant guidebook. June 2012.
34. Herbert Burkert, John Coates, Robert Faris, Urs Gasser, Jack Goldsmith, Colin Maclay, Laura Miyakawa, John Palfrey, and Jonathan Zittrain. Accountability and transparency at icann: An independent review. October 2010.
35. Association National of Advertisers. Icann generic top level domain developments | ana. http://www.ana.net/content/show/id/icann.

36. Jon Leibowitz, Thomas Rosch, Edith Ramirez, and Julie Brill. Consumer protection concerns regarding new gTLDs. December 2011.
37. ICANN. New gTLD auction proceeds | icann new gtlds. https://newgtlds.icann.org/en/applicants/auctions/proceeds.
38. ICANN. Base registry agreement. July 2017.

Appendix

5.1 ICANN and gTLD Expansions

In this section, we provide a high-level overview of how gTLDs have been expanded over the years and the role that ICANN plays in regulating these expansions. Since 1998, the Internet Corporation for Assigned Names and Numbers (ICANN), has been responsible for administering the Internet Domain Name System (DNS). This role has included the authority for establishing new top-level domains (TLDs). TLDs have historically been classified into: (1) TLDs reserved for countries and territories (country-code TLDs or ccTLDs), (2) a TLD reserved for Internet infrastructure (infrastructure TLD: `.arpa`), and (3) TLDs that may be used for other purposes (generic TLDs or gTLDs).

gTLD expansion between 1984 and 2012. Between 1984 and 2000, the number of gTLDs increased from five to seven with `.net` and `.int` added to the "core" set (`.com`, `.edu`, `.gov`, `.mil`, and `.org`). Of these seven, three TLDs – `.com`, `.net`, and `.org` – have always been open to public registration with the other TLDs being reserved for use by specific organizations such as universities (`.edu`) and government entities (`.gov`). Starting in 1998, ICANN began considering a more "open" gTLD program which would allow private entities to act as registries and manage new gTLDs. Following a public call for proposals in August 2000 and a two-month period for public comment, ICANN announced seven new gTLDs in November 2000 (`.aero`, `.biz`, `.coop`, `.info`, `.museum`, `.name`, and `.pro`). The process was repeated again in 2004, resulting in the introduction of six new gTLDs (`.asia`, `.cat`, `.jobs`, `.mobi`, `.tel`, and `.travel`). Between 2004 and 2012, only two other gTLDs – `.xxx` and `.post` – were added. By the end of 2012, the Internet had 22 gTLDs – of which 15 were open to public registration. As of August 2013, the 15 additions to the 7 core gTLDs accounted for 3% of all domain registrations while the 7 core gTLDs accounted for 51% of all domain registrations on the Internet (ccTLD domain registrations accounted for 35%) [31].

The 2012-2013 gTLD expansion. In 2008, citing the success of the previous gTLD expansions in 2000 and 2004, ICANN approved new policies to facilitate the large-scale creation of new gTLDs with the stated goal of "enhancing innovation, competition, and consumer choice" [32]. Following the creation and multiple revisions of a guide for the application process of new gTLDs, in 2011 steps were taken to enable the registration of new gTLDs. These guidelines are still applicable today. In order to register a new gTLD, a registry needs to demonstrate capabilities to handle technical, operational, and business operations related to the handling of registrar relationships and submit a $185K application and evaluation fee [33]. Applications for new gTLDs were opened in 2012 following criticism and protest from Internet societies, including Harvard's Berkman Center for Internet & Society [34], the Association of National Advertisers [35], and the United States Federal Trade Commission [36] which primarily cited the lack of transparency in the evaluation process, potential for trademark infringement and other generally malicious conduct. By 2013, over 1,900 applications were received of which 1,543 were granted and 1,208 are still active today. Contested gTLD registration applications were resolved by a bidding process. As of July 2016, the ICANN netted a profit of $233M from the bidding process alone [37]. As of August 2018, the 1,208 active new gTLDs accounted for 9% of all domain registrations on the Internet [31]. We note that statistics regarding the registration of new gTLD domains have not been updated on the ICANN website since 2015 and are only available through other third-party services.

Registry responsibilities and guidelines. Following the delegation of a gTLD, a registry is required to perform certain responsibilities related to maintenance of the gTLD. A full specification of these requirements is available online [38]. We summarize the requirements that are relevant to our study below.

- *WHOIS services.* Registries are required to maintain a fully responsive and searchable WHOIS service available via port 43 and through a web-based interface.
- *Zone files.* Registries are required to provide public access to their *current* zone files via the Centralized Zone Data Access (CZDA) provider [14]. In order for a member of the public to gain access to the zone file, they need to provide "information sufficient to correctly identify and locate" themselves. These may include an organization name and address, IP address, *etc. There is no specified time within which a registry is required to provide a response.*
- *Protected domains.* All registries owning and operating an *open* gTLD are subject to a *sunrise* period of 30 days. During this period, domains may only be registered by organizations registered with ICANNs Trade Mark Clearing House (TMCH). Following this period, all domains are open for public registration – regardless of their trademark status and any trademark disputes are to be resolved using ICANN services. All costs associated with disputes, trademark verification, and TMCH registration are to be paid by the trade mark holder. Further, the TMCH will only accept domains as trademarked if the following criteria are met (examples are demonstrated with the organization "ICANN Example"): (1) exact match rule — `icannexample.org` is a valid trademark domain, (2) hyphen for spaces/special characters rule — `icann-example.org` is a valid trademark domain. *All other domain variations, including plurals are considered invalid* (*e.g.,* `icann-examples.org`).

We note that we were unable to find documents relating to how compliance with these responsibilities were to be monitored or enforced.