The Guardians of Name Street: Studying the Defensive Registration Practices of the Fortune 500

Boladji Vinny Adjibi
Georgia Tech
vinny.adjibi@gatech.edu

Athanasios Avgeditis
Georgia Tech
avgetidis@gatech.edu

Manos Antonakakis
Georgia Tech
manos@gatech.edu

Michael Bailey
Georgia Tech
mbailey@gatech.edu

Fabian Monrose
Georgia Tech
fabian@ece.gatech.edu

***Abstract*—Using orthographic, phonetic, and semantic models, we study the prevalence of defensive registrations related to a wide spectrum of transformations of the base domain names of Fortune 500 companies. As part of a large-scale evaluation, we explore several questions aimed at (a) understanding whether there are explainable factors (e.g., the size of the company's security team or its domain name's popularity rank) that correlate with a company's level of engagement regarding defensive registrations; (b) identifying the main actors in the defensive registration ecosystem that Fortune 500 companies rely upon; (c) uncovering the strategies used by these actors, and (d) assessing the efficacy of those strategies from the perspective of queries emanating from a large Internet Service Provider (ISP).**

Overall, we identified 19,523 domain names defensively registered by 447 Fortune 500 companies. These companies engage in defensive registrations sparingly, with almost 200 companies having fewer than ten defensive registrations. By analyzing the registrations, we found many similarities between the types of domain names the companies registered. For instance, they all registered many TLD-squatting domain names. As it turns out, those similarities are due to the companies' reliance on online brand protection (OBP) service providers to protect their brands. Our analysis of the efficacy of the strategies of those OBPs showed that they register domain names that receive most of the potential squatting traffic. Using regression models, we learned from those strategies to provide recommendations for future defensive registrants. Our measurement also revealed many domain names that received high proportions of traffic over long periods of time and could be registered for only 15 USD. To prevent the abusive use of such domain names, we recommend that OBP providers proactively leverage passive DNS data to identify and preemptively register highly queried available domain names.

I. Introduction

Initially designed as a more memorable alternative to using IP addresses when accessing remote computers, domain names have become a cornerstone of the Internet as we know it today. For everyday users, visiting a domain name allows them to access a host of services that are vital to daily life. For the companies providing those services, their domain names represent assets that help drive business closer to the customers. The mindful choice of a domain name, along with proper security measures (e.g., a certificate from a trusted party) and maintaining an online visual presence that is consistent with a company's brand in the physical world, are all ways in which companies build trust with their user base. However, gaining and maintaining that trust is seldom an easy battle, due in part to miscreants who are always on the lookout for ways to undermine trust relationships to perpetrate crimes.

One of the ways in which miscreants abuse this trust is by registering domain names that share commonalities with the target organization's domain name. Users inadvertently access the contents hosted on those sites because of typing mistakes [1], hardware errors [2], or even when dictating the domain name to a virtual assistant [3]. In other cases, these skilled adversaries deliberately trick users with a domain name that is easily confused with the real brand and social engineer users into clicking malicious links [4], [5]. More insidious adversaries rely on DNS manipulation attacks that can lead to stealthy and lasting subdomain takeovers [6]–[10].

Whenever these attacks are successfully perpetrated, the financial hit and loss in reputation can be significant. For example, United Airlines has been in a protracted legal battle with an unsatisfied customer who registered the domain name untied[.]com and shared damaging reviews about the airline [11]. Given the reputation damage and financial impact of such incidents, protection against domain name abuse has become a critical issue for many companies.

One viable option is to defensively register the domain names that would otherwise be used for abuse. According to ICANNWiki, the practice of defensive registrations "refers to registering domain names, often across multiple TLDs and in varied grammatical formats, for the primary purpose of protecting intellectual property or trademark" [12]. Given DNS's importance, it makes sense that the idea of defensively registering domains has been of much interest in recent years.

In the academic literature, researchers have studied instances of defensively registered domain names across various forms of domain abuse [1]–[3], [5], [13]–[17]. Unfortunately, the results from these studies are often contradictory. For example, in a study of homograph squatting, Quinkert et al. [17] reported that 8% of the studied domains were defensively registered, but a subsequent study of a larger set of domain

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.241202
www.ndss-symposium.org

squatting types concluded that only 1.69% of the candidate homograph domains were registered defensively [13], despite the fact that both studies were conducted on the Alexa 500 list at similar points in time. Those discrepancies motivate the need for a more holistic measurement of defensive registrations.

While the prevalence of defensive registration practices has been understudied in the literature, the business case for protecting intellectual property online has not gone unnoticed. Indeed, the online brand protection sector is a booming business [18]. Although protection against domain name abuse is only a subset of the services offered (e.g., alongside protection from social media abusers) by providers in this market, the domain name abuse segment alone can have annual revenue of around 80 million USD for some players [19]. These earnings seem to corroborate the findings of Halvorson et al. [14], who suggested that in 2014, companies spent about 11 million USD to defensively register domain names in the .xxx generic top-level domain (gTLD) alone. Yet, despite such a booming market, malicious actors continue to succeed in domain name squatting abuses. Understanding some of the reasons why they succeed requires an assessment of the *strategies* used by brand protection service providers when shielding their customers against domain squatting, coupled with an in-depth analysis of the *efficacy* of the protection they provide.

To answer those questions, we study the defensive registration practices of the 500 biggest U.S. corporations by revenue. Besides being very popular, the Fortune 500 offer many services such as banking, e-commerce, and accommodation that are pervasive among Internet users. This makes them prime targets for social engineering lures, which provides an incentive for those businesses to protect their online brand against abuse. Moreover, they can afford the financial expenditure that comes with protecting one's brand online, whether in-house or through a third-party like a specialized brand protection company. Therefore, this set of companies is an appropriate seed for an in-depth study of OBPs' strategies.

Our contributions are:

1) A conservative approach for identifying defensive registrations using DNS zone files, WHOIS information, official company reports, and publicly available information related to ownership and relationships between brands.
2) The largest and most comprehensive study of defensive registrations covering TLD-squatting, bitsquatting, typosquatting, homographs, homophones, abbreviations, brand name, and stock name squatting.
3) An assessment of the effectiveness of defensive registrations made by brand protection providers using three years of passive DNS data from a large ISP. This analysis revealed many highly queried domain names that are available for registration for a 15 USD fee.
4) A comparison of the adequacy of six defensive registration strategies by leveraging the predictive power of regression models to understand and ultimately predict which domain names a provider is more likely to register for a new customer. Those models helped us draw insights that can aid future defensive registrants in their choices.

Overall, our results highlight the need for research into new models that provide more comprehensive protection against squatting abuse to help brands curb social engineering and make peoples' online experiences safer.

II. Approach

To date, a number of techniques have been used to identify defensively registered domains. The approaches used are based either on the domain's web content [1], [13], information parsed from WHOIS records [16], the underlying infrastructure (resolving IP addresses, name servers), or a combination thereof [2], [3], [20]. While each approach has shortcomings, we decided to utilize the registrant organization field of the relevant WHOIS records enriched with supplementary data. The need for supplementary data stems from concerns around the utility and trustworthiness of WHOIS records, especially post-GDPR [21]. Specifically, we include an additional check of the organization names against the database of companies provided by the U.S. Security and Exchange Commission.

A. Datasets

Fortune 500: We use the 2023 rankings [22] to collect each company's name, rank, website link, and ticker symbol (for publicly traded companies). We also collected each company's market value, number of employees, and reported revenues as of July 2022 from the Fortune magazine's website. We used a company's ticker symbol to find its full name on the stock market. For the 27 private companies, we use the copyright information on their website as their official name. We used those names to generate semantic transformations such as abbreviation and brand name squatting.

After navigating to each of the website entries, we used the second-level domain (SLD) of the final destination URL as the *base domain* for every company except for Alphabet, Inc. for which we used google[.]com instead of abc[.]xyz.

In total, we gathered 500 base domain names with about a fifth ($\approx$ 20%) having an e2LD with at most four characters. Such short domain names, along with those related to common words are usually omitted from domain squatting abuse studies because they can lead to false positive cases of domain abuse. By applying our conservative approach to identify defensive registrations, we were able to correctly identify defensive registrations for those domain names as well.

DNS Zone files: We use daily snapshots of 1,191 unique gTLDs obtained through the Centralized Zone Data Service (CZDS) program of ICANN since October 2020. We used this data to determine the historical availability dates of the generated domains in a similar fashion to Halvorson et al. [15]. Our study focuses on 383 gTLDs with open registration policies after we filtered out those to which organizational (e.g., .google) or professional (e.g., .archi) restrictions apply.

B. Name Transformations

Domain name abuse can happen at different levels of a domain name, from the top-level domain (TLD) to subdomains [15], [23]. In this work, we consider a wide variety of

TABLE I: Overview of the datasets used in our work. Dates are in the format DD/MM/YYYY.

Label	Description	Size	Time span
Fortune 500	Companies on the Fortune 500 used to seed our analyses	500	01/06/2023 - 31/05/2024
DNS zone files	Historical zone files	429,884	01/10/2020 - 01/12/2023
Passive DNS	Passive DNS records from a recursive server of an ISP	109,180,596	01/10/2020 - 01/12/2023
Name transformations	Number of candidate domain names	146,397,537	NA

transformations that apply to the effective second-level domain (e2LD) and/or the TLD of a base domain name with one exception: we do *not* study combo-squatting mainly because of its unbounded space of possibilities. From the literature, we considered TLD-squatting [15], [24]–[26], typosquatting [27], bitsquatting [28], homophones [3], [13], homographs [17], brand name squatting [5], and abbreviation squatting [29]. Additionally, we introduced stock name squatting, which is the use of a company's ticker symbol as the e2LD of a domain name (e.g., ABNB[.]apartments for Airbnb Inc.).

C. Identifying Defensive Registrations

As shown in Fig. 1, we consider a domain name to be defensively registered if it satisfies any of the following conditions:

Condition ❶ *All* the name servers set for the domain name in the daily zone file are subdomains of the base name. Under this rule, the domain name gogle[.]com, generated from google.com which uses ns[1-4].google[.]com as nameservers is considered defensively registered; *or*

Condition ❷ *All* the name servers set for the domain name in the daily zone file are subdomains of domain names owned by the company or its affiliates. For example, the domain meta[.]tube, a transformation of meta[.]com that uses the nameservers [a-d].ns.facebook[.]com that are owned by the parent company (Meta Platforms) is considered defensively registered; *or*

Condition ❸ The registrant organization field of the WHOIS record points to ownership by the company or its affiliates. For example, the domain mondelez-international[.]com (a transformation of mondelezinternational[.]com) uses the nameservers dns[1-2].cscdns[.]net that are owned by a third party, but the WHOIS record identifies the parent company as owner of the domain name; *or*

Condition ❹ The domain name pointed to by the email address in the WHOIS record for the transformation is owned by the company or its affiliates. For instance, while the registration organization field for the domain name kla[.]net is redacted, that domain uses nameservers ns4[1-2].domaincontrol[.]com (owned by GoDaddy) and has a reference to domainadmins@kla-tencor[.]com as its registrant email address. Given that kla-tencor[.]com is owned by KLA Corporation (as verified through WHOIS), we consider this domain name to be defensively registered.

Conditions 1 and 2 apply when a company self-manages its defensive registrations. On the other hand, conditions 3 and 4 apply when a company delegates the resolution of their defensively registered domain names to third parties. As we show later in Section III-C, many of these third parties are online brand protection service providers.

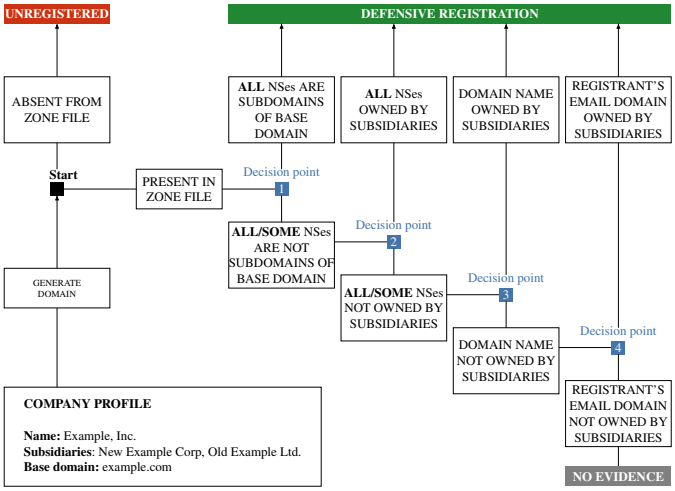

Fig. 1: Approach used to decide if a domain name is defensively registered. We identified subsidiaries using data from the U.S. Security and Exchange Commission and Wikipedia.

It is prudent to note that because many registrars shifted toward privacy-protected WHOIS records to comply with GDPR rules, among others [21], WHOIS records are currently less useful for attribution than before. Given that, we adopted a conservative approach to using the records by first filtering out clearly redacted or anonymous registrations. We extracted the registrant organizations from the remaining records (see Table VI in the appendix). We matched that against ten years of data from the U.S. Securities and Exchange Commission listing that contains the names of businesses operating in the U.S. that filed a report with the agency between 2014 and 2023 [30]. We used ten years because it corresponds to the maximum length of time for which a domain name can be registered without renewal according to ICANN [31]. We addressed inconsistencies between the names extracted from the U.S. Securities and Exchange Commission files and the WHOIS records by stripping special characters and corporate endings (e.g., Inc.) from the names. Through that matching process, we identified 1,770 U.S. entities among the registrants in our dataset, which we manually searched online to establish their relationship to the Fortune 500. We used the resulting mapping to determine if a Fortune 500 owns a domain name.

TABLE II: Number of candidate domains by transformation type. Gray lines indicate semantic transformations. Note that some candidates can fall under multiple types.

Transformation	Description	Example	Generated	Registered	Defensive (%)
Typosquatting	The insertion, replacement, transposition or omission of characters	capitaline.com	125,779,498	367,375	7,710 (0.02%)
Homograph Squatting	Replace characters with visually similar ones	capital0ne.com	20,100,606	1,601	224 (13.99%)
Bitsquatting	Flip a bit in the e2LD's binary representation	capitanone[.]com	6,343,246	49,113	671 (1.37%)
Abbreviation squatting	Abbreviate words from the company's name	**C**apital **O**ne **F**inancial (conef.com)	394,107	18,167	332 (1.83%)
Brand name squatting	Concatenate words in the company's name	capitalonefinancialcorp.com	278,058	4,352	1,728 (39.71%)
TLDSquatting	Replace the TLD with another valid one	capitalone.net	191,500	22,969	8,782 (38.23%)
Stock name squatting	Use the ticker symbol as an e2LD	cof.com	155,498	22,348	615 (2.75%)
Homophone squatting	Replace constituent words with their homophones	capitalwon.com	78,515	1,407	33 (2.35%)
Overall			**146,397,537**	**402,934**	**19,523 (4.84%)**

While using such a strict approach helps us identify defensive registrations with high confidence, it also limits our ability to identify every defensive registration. For instance, the domain name wal-mart[.]com, has both the registrant organization and email address fields of the WHOIS record redacted, and the nameservers to which the domain name points are all owned by third parties (Akamai and UltraDNS). As such, we can not consider this domain name as being defensively registered under any of our rules, even though we identified several other domain names defensively registered by Walmart (e.g., wal-mart[.]net) that used subdomains of wal-mart[.]com as their primary nameservers. While those observations suggest that the domain name is probably owned by Walmart, we do not count wal-mart[.]com as being defensively registered as we have no direct evidence from WHOIS records to bolster the claim that Walmart owns that domain name. Consequently, we take a conservative approach and label wal-mart[.]com as unknown (see Fig. 1). We discuss the overall limitations of our approach in Section IV.

We applied this methodology to the 146,397,537 domain names that we generated. Table II summarizes the number of domain names generated, registered, and defensively registered by transformation type. It shows that semantic transformations (in light gray) have a much larger registration rate than their non-semantic counterparts. We explore some reasons that could explain this behavior in the remaining sections.

III. Evaluation

A. The Clientele

Overall, we found 19,523 domain names defensively registered by 447 Fortune 500 companies. Most of them engage in defensive registrations sparingly, with almost 200 companies having fewer than ten defensive registrations each (see Fig. 2).

Level of engagement: To account for the disparities in the size of the squatting spaces, we normalized the count of defensive registrations by the total number of candidate transformations generated for each company. Consistent with our previous findings, we observed in Fig. 3 that 218 companies register less than 0.005% of all the domain names in their squatting space. We consider those companies to adopt a *timid* approach to defensive registrations while the others appear to be *resolute*. We obtained the cutoff of 0.005% using Satopaa et al.'s [32] approximation for finding the elbow of a curve. The spread of values in the region of the distribution representing

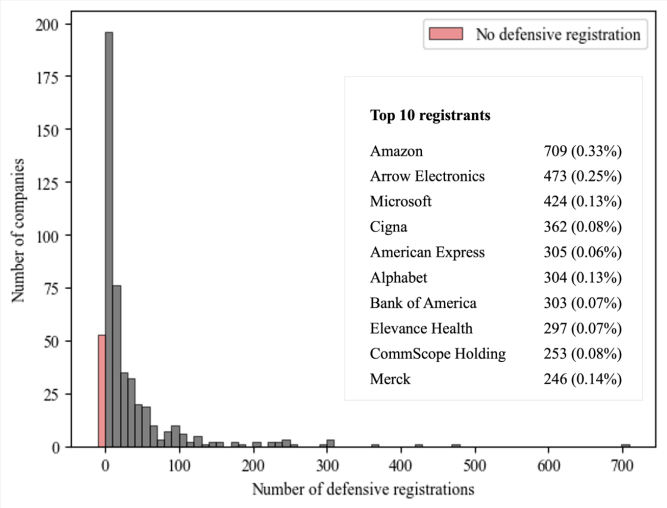

Fig. 2: Number of defensively registered domain names for the Fortune 500 companies. Each bin has a size of 10.

the resolute companies suggests that certain companies are more aggressive than others in their approach to defensive registrations. Next, we attempt to identify whether there are factors that correlate with their level of engagement.

Factors influencing defensive registrations: We examined a set of high-level properties to verify whether they could explain the disparity between the timid and resolute groups of companies. This list of non-exhaustive properties aimed to represent the financials as well as the threat profile of each of the companies. In particular, we examined:

- **Domain length**: Length of the base domain name's effective second-level domain (e2LD).
- **Domain popularity**: We took the inverse of the median Tranco rank [33] for the base domain name during the period of our measurement.
- **OSINT evidence of abuse**: We counted, for each company, the number of unique domain names generated by a transformation of the base domain name that appears on any of the following threat intelligence feeds.
  - *Phishtank*: a repository of crowdsourced phishing URLs. We only used the list's manually verified URLs.
  - *Artists Against 419*: a list of URLs related to bank fraud from which we retained the manually verified URLs.

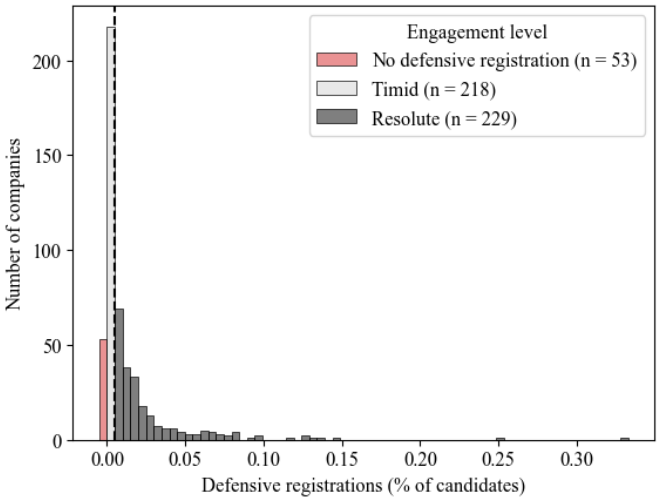

Fig. 3: Proportion of defensively registered domain names for the Fortune 500 companies. Each bin has a size of 0.005.

Fig. 4: Spearman-ranked correlation between certain properties and a company's level of engagement in defensive registrations at a significance level of $p = 0.05$.

  - *OpenPhish*: an open-source repository of phishing links that are vetted by the organization.
  - *Spamhaus Domain Blocklist*: a list of domains with a poor reputation (e.g., previously used maliciously).

  In addition to the transformations mentioned earlier, we also counted instances of combo-squatting [4]. For short domain names (< 5 characters), we only considered instances of OSINT TLD-squatting abuse.
- **Assets**: The estimate, in millions of USDs, of the financial assets of the company according to the Fortune Magazine.
- **Number of security analysts employed**: Using the 2022 employment and wage statistics data [34], we estimated the ratio of security analysts to every employee in an industry. We multiplied each company's number of employees by the estimated ratio for their industry.

Using those features, we conducted a Spearman-ranked correlation test that measures the monotonicity of the relationship between two variables. The results of the analysis, conducted at a confidence level of 95%, are presented in Fig. 4. It suggests that across the two groups of companies, domain popularity exerts the highest correlation with the number of defensive registrations made by a company. This finding confirms some of Pouryousef et al.'s [16] earlier observations that domain popularity was a strong indicator of the risk of abuse by typosquatters. Similarly, the evidence of abuse (OSINT hits) and the number of security analysts employed by each company are slightly correlated with the aggressiveness of a company when it comes to defensive registrations. Intuitively, these results make sense because we expect highly abused brands to invest more in protecting their online identities.

The remaining two factors (i.e., domain length and financial assets) set the two groups apart. For timid companies, the domain length is negatively correlated with the propensity to defensive registrations (-0.33). This means that the shorter the domain name of a company, the more defensive registrations they do relative to their available set. Those results seem to echo results from Banerjee et al. [35] that concluded that shorter domain names are more likely to be abused. However, this does not seem to matter to the 229 resolute companies where no statistically significant correlation was obtained with that feature. Unlike the timid group, their level of engagement positively correlates (0.25) with their financial assets.

Takeway: Overall, our analysis shows that although the majority of Fortune 500 companies participate in defensive registration practices, companies facing a higher risk of domain name abuse tend to engage more than their counterparts. For resolute companies who do engage aggressively, financial means seem to play a non-negligible role. The lingering question at this point is *what* domain names do they register?

B. The Catalog

Figure 5 shows the defensively registered domain names by transformation. The data shows that registrations from both groups of companies span all the transformations studied, with homophones being the least represented operation of all. TLD-squatting, on the other hand, is very common and accounts for close to 2/5 of all defensive registrations. It is by far the largest transformation type registered by both timid and resolute companies. These results confirm the observations of Halvorson et al. [15] that hinted at companies' willingness to defensively register their domain names in the new gTLDs.

The other semantic transformations (brand name, stock name, and abbreviation squattings) are also largely represented, especially for the timid companies where their proportion is significantly higher than for the resolute companies. Together with transformations that modify a delimiter (insertion or removal of a dash between two words) or the grammatical form of a word (pluralize or singularize), this set of operations seems to be the favorite choice of timid companies. These transformations are highlighted in Fig. 5 with a bold font on the y-axis. Next, we provide some explanations for their prevalence among the registrations of timid companies.

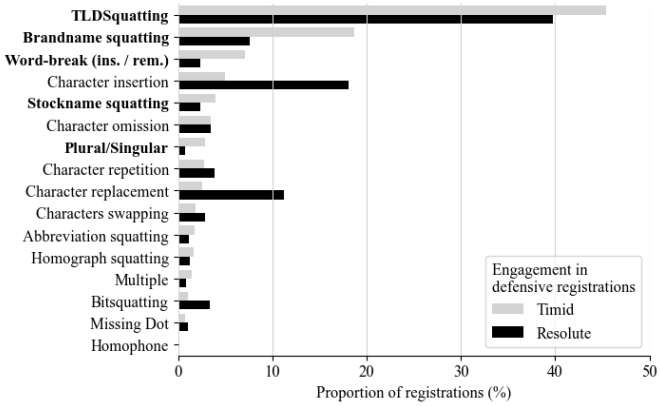

Fig. 5: Proportion of companies defensively registering each type of transformation by level of engagement.

TLDsquatting: We observed a median number of six TLD-squatting registrations in the 383 gTLDs. How companies choose in which gTLDs to protect their brands is an open question. We hypothesize that those decisions are largely based on the relevance of the gTLDs to their business activities. To test this conjecture, we leveraged Google Cloud Natural Language's API [36] to categorize the gTLDs based on their description on TLD-List [37] or GoDaddy's website. After manually mapping each high-level category to the closest sector, we observed in Fig. 6 that a gTLD category receives more attention from companies whose sector of activity directly maps to the said category. For instance, registrations in gTLDs associated with food and cuisine are more frequent with accommodation businesses than in any other sector. The decision to prioritize such business-related gTLDs is a rational one (e.g., it is easier to trick a user into accessing airbnb[.]apartments than airbnb[.]surgery).

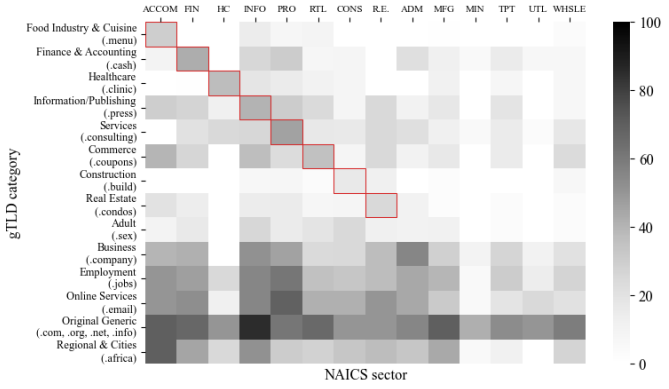

Fig. 6: Use of gTLDs per sector (Accommodation, Finance, Healthcare, Information, Professional, Construction, Real Estate, Administrative, Manufacturing, Mining, Transportation, Utilities, Wholesale).

Additionally, the gTLDs with a more general connotation receive comparable interest across all sectors. This is particularly true in the case of the original gTLDs where more than half of the companies in each sector have TLD-squatting defensive registrations. Interestingly, although the adult category also receives some attention, it is much less so than what the study by Halvorson et al. [14] would suggest.

In total, our findings corroborate Pouryousef et al.'s assertion that the gTLD expansion places financial burdens on companies concerned about their online reputation [16]. Specifically, our results show that when selecting which gTLDs to register their brands in, companies tend to prioritize those associated with their business activities. The large number of remaining gTLDs with vague connotations puts added pressure on these companies, and because it's unreasonable to expect them to register every possible transformation of their domain name, we echo Korczyski et al.'s concern that ICANN should take additional measures to make the new gTLDs safer [38].

Brand, stock, and abbreviation squatting: These transformations are similar to the TLD-squatting transformation in that they are intimately related to the brand's identity in the physical world. Thus, it is logical to expect users to take the bait easily if an attacker decides to leverage those.

Delimiter modification and grammatical mistakes: The tendency among the timid companies to prefer these transformations also makes sense. This family of transformations was found to be common in a study of package confusion [39] with more than 600 instances in the RubyGems ecosystem alone. The ease with which such transformations have fooled developers explains why companies might attend to them first.

Takeway: We found that Fortune 500 companies' defensive registrations span a wide gamut of transformations. However, when they shyly engage in such practices, companies tend to focus on those transformations that seem to have a high potential for confusion among users. This pattern makes us question whether those choices are made independently or if certain entities are responsible for those shared behaviors among the companies. We investigate that question next.

C. The Guardians of Name Street

Almost 3/4 (73.64%) of the defensive registrations that we identified were matched by condition ❸ of our methodology. This suggests that the majority of defensive registrations use at least one name server managed by a third party. Besides large DNS and CDN service providers (e.g., Vercara [formerly Neustar], Microsoft, Amazon, etc.), those third-party name servers appeared to be owned by domain name registrars that offer online brand protection (OBP) services. Given this insight, we measured which domain name registrars were used by Fortune 500 companies for their defensive registrations.

20 of the 47 unique registrars we identified advertise OBP services. As a whole, they account for the majority of defensive registrations that we observed in terms of domain names (99.07%) as well as Fortune 500 companies serviced (436 out of 447). The two largest of those companies are MarkMonitor and CSC CD with five times more defensive registrations than any other provider. Out of the ten registrars shown in Fig. 7, all but RegistrarSEC offer OBP services. According to the

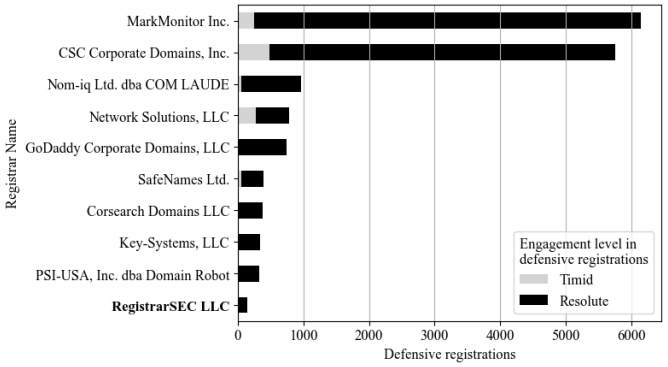

Fig. 7: Top 10 registrars of record for the domain names defensively registered by the Fortune 500 companies.

information available on the registrar's website, it is a wholly-owned subsidiary of Meta Platforms who used it for most of its defensive registrations. It is plausible that this is an instance of a company offsetting the costs of defensive registrations by using its self-managed registrar. The other Fortune 500 companies that own a registrar (Alphabet Inc. [1] and Amazon Technologies Inc.) do not exhibit such behaviors.

In our subsequent analysis, we assess how relevant the providers' defensive registrations for the Fortune 500 were.

1) Experimental Design: Our evaluation of the efficacy of the providers' strategies focused on comparing the domains they defensively registered with available domain names.

Passive DNS: We use passive DNS data from the recursive DNS resolvers of a large US telecommunication provider that serves over 40 million Internet devices spread across the US. The data from the ISP only consists of the number of DNS requests of a certain type (e.g., A) that a domain name receives daily and the number of distinct IPs generating those queries. The data contains no IP addresses or any other personally identifiable information. Moreover, because the ISP's customers agree to data collection for analytics and inferences (as part of their service agreement), there are no ethical concerns related to our use of the data. We consider the data from the ISP to be adequately representative because of its high cardinality and substantial overlap with the Tranco Top 1M list [33] leading to a rank-biased overlap (RBO) score of 36.2%. The RBO is much higher than those between major popularity lists (e.g., Alexa and Umbrella) that ranged from 4.5% to 33% according to Pochat et al. [33]. Overall, 321,652 domain names appeared simultaneously on the top 1M list from Tranco and the list computed from the passive DNS data on November 8th, 2023. The overlap difference between the two lists is not surprising. It is merely because our data targets US residents while Tranco assesses popularity on a global stage. Regardless, for the scope of our study, this passive DNS data is adequate, as the Fortune 500 list only includes US-headquartered corporations. From here on, we refer to this dataset as the *volumetric data*. That dataset spans a period of just over 3 years, from October 2020 to December 2023.

Providers: For a concise yet meaningful analysis of the performance of the providers, we restricted our measurement to the domain names in the ten gTLDs in which we found the most defensive registrations. After discarding all the instances where a provider had fewer than ten registrations for a customer in those gTLDs, we were left with six providers with at least five *resolute* customers as summarized in Table III.

TABLE III: No. of customers serviced and domain names defensively registered by provider.

Provider	Customers	Domains	
		Median (%)	Total
CSC Corporate Domains, Inc.	76	23.5 (0.40%)	3,236
MarkMonitor Inc.	64	22.0 (0.42%)	3,094
Network Solutions, LLC	9	12.0 (0.23%)	349
GoDaddy Corporate Domains, LLC	7	19.5 (0.70%)	425
Nom-iq Ltd. dba COM LAUDE	7	21.0 (0.38%)	200
SafeNames Ltd.	5	27.0 (0.57%)	227
Overall	**166**	**21.0 (0.40%)**	**7,486**

Measurement methodology: For the period between October 2020 and December 2023, we compare the number of type A DNS queries received by defensive registrations with those received by domain names that were available for registration at any point during the measurement period. To determine whether a squatting domain name is available for registration, we follow the guidance of Affinito et al. [41]. That work showed that when domain names expire, they can remain on the zone files for the auto-renew grace period (45 days), after which they are moved to a redemption grace period, followed by a pending delete period (usually 30-35 days in total). It is not until then that they become available for registration. To be sure that we cover this entire window, we only consider a domain name to be available on a specific date if it is absent from the zone files for more than 90 consecutive days. From this set of available domain names, we select those that are at least five characters long and do not appear in the English dictionary. Even after following the methodology recently proposed by Affinito et al. [41] to find domain name availability via the zone files, we identified several cases where although the domain names (e.g, wak-mart[.]com, c0mcast[.]net, and googlew[.com]) did not have zone file records at the time they received the DNS requests, the domains were actually registered according to WHOIS data. Thus, we query the VirusTotal [42] historical WHOIS database to filter out domain names that were registered but absent from the zone files for more than 90 consecutive days.

We also used the zone files to determine the days on which a domain name was defensively registered in the past. For the period from October 2020 to December 2023, we consider a domain name to be defensively registered on a given date if and only if its name servers on that day are exactly those in the zone files (snapshot of November 8th, 2023) we used to build our dataset of defensive registrations.

[1]Note that Alphabet sold Google Domains to Squarespace in late 2023 [40]. Still, none of Alphabet's defensive registrations had Squarespace as registrar.

Close inspection of the data revealed several cases where defensive registrations received a comparable amount of traffic as the base domain name. For the most part, it appeared that some domains that were defensively registered in the past were later used by the company for official purposes. For instance, the domain name econocophillips.com is used to host an internal application of Conoco Phillips. To filter such cases of non-security-related defensive registrations, we used the Censys scanner data [43]. First, we identified all the resolvable domain names that map to a service other than a simple HTTP redirection to the base. For the ones that do not map to any service, we used the Censys' certificate data and considered them re-purposed if a valid SSL certificate is associated with them with the organization mapping to the parent company. We do so because those domain names could be in use for internal services behind a firewall. In total, we identified 432 repurposed registrations hosting various services (e.g., SMTP, HTTPS, FTP), which we excluded from our measurements.

2) Results: The distribution of the proportion of query traffic captured by the defensively registered domains is given in Fig. 8. The median share of traffic received by defensively registered domains ranges from 78% (Network Solutions) to 94% (SafeNames). The findings suggest that, on average, almost 3/4 of queries are captured by the providers' defensive registrations. For two of the outliers we observed (i.e., Ulta Beauty [ulta[.]com] and Target Corporation [target[.]com]), we hypothesized that their poor performances could be because of the shortness of their domain names, which might make them more attractive targets to adversaries. Unfortunately, our OSINT data does not contain evidence of abuse to those customers, so we could not validate that conjecture.

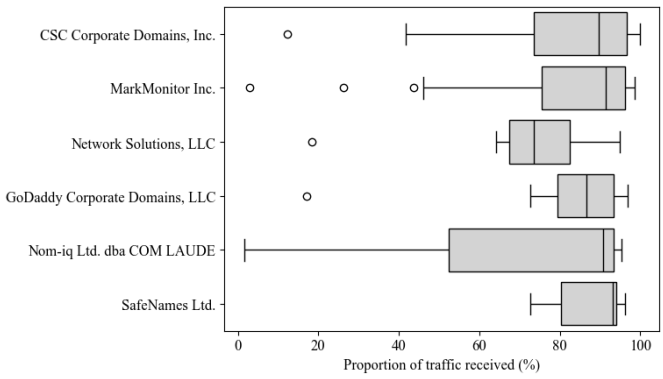

Fig. 8: Distribution of query volume captured by the defensive registrations. The missed queries go to domain names that were available according to zone files and WHOIS records.

Figure 9 suggests that the providers registered domain names that received many daily queries. One of CSC Corporate Domains' picks (goldmansachs[.]info) received an amount of traffic that is proportional to 20% of that received by the base domain name. The fact that many defensive registrations appear in the top-right quadrant of the plots shows that the OBP service providers pick several high-value domain names. SafeNames' high defensive performance can be traced back to the fact that most of their registrations are in the right quadrant.

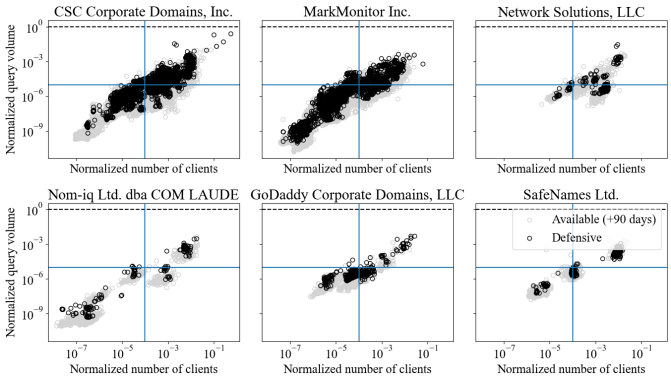

Fig. 9: Queries received by the defensive (black) and available (lightgrey) domain names. The axes are normalized for each day by the amount of lookups the base domain name received.

At a 95% confidence level, SafeNames' defensive registrations yielded a higher amount of traffic (total query volume) over more days than the available domain names for all their customers. This analysis was conducted through a Mann-Whitney U test under the null hypothesis (H_0) that the distributions are equivalent or else that the defensive registrations were greater in their measures (H_1). The results were similar for the remaining 4 providers showcasing the overall aptness of the choices made by the providers. When we switched the criteria to be the median number of distinct IPs that queried each domain name, we found that the defensively registered domain names outperformed the available domain names in at least 22% – for Network Solutions – of instances. SafeNames Ltd. once again had the best return with 4 out of 5 of their customers greatly benefiting from their registrations.

Such a high-level assessment of the efficacy of defensive registrations suggests that the providers registrations are in general reasonable. However, drilling down into specific types of operations depicts a different picture. In fact, we found many domain names that, while being available, received more than a thousand unique queries for at least one hundred days.

As seen in Table IV, the most queried domain name received 95 million requests for more than three years. However, it was never registered during the three years of our study according to both zone files and historical WHOIS. Similarly, many other domain names received thousands of requests while they were available for registration, posing as prime candidates for defensive registrations that the providers missed during parts or the whole period of our measurements. Using VirusTotal's historical data, we found that 5 of the 15 domain names had been used maliciously. For those domain names, we were curious if scanners were responsible for all their queries.

Under the hypothesis that scanner-generated traffic would be uniformly distributed over time, we compared the query volume distribution of those five domain names to the uniform distribution using a Kolmogorov-Smirnov statistical test [44]. At a 99% confidence interval, we found that none of the

TABLE IV: Top 15 available domain names with significant queries by clients with the providers who should have registered them. These domain names did *not* have zone file records *nor* did they appear in historical WHOIS during the observation period. We mark previously abused domains with 🐞. To limit potential data abuse, we replaced the available domain names with their originating root domain names.

Root Domain		Operation Type	Provider Responsible	Days Queried	Days Available	Query Volume
blackrock.com		replacement	CSC CD.	1,000	1,187	95,480,075
pacificlife.com		wrong-tld	CSC CD	915	1,187	269,647
bestbuy.com		replacement	MarkMonitor	696	1,187	164,885
wellsfargo.com		replacement	MarkMonitor	103	1,187	49,858
cisco.com	🐞	repetition	MarkMonitor	201	334	42,892
honeywell.com		transposition	MarkMonitor	134	239	7,616
pacificlife.net		transposition	CSC CD	372	1,187	6,589
caterpillar.com		replacement	COM LAUDE	649	1,187	5,748
bankofamerica.com	🐞	omission	CSC CD	835	1,187	5,663
pacificlife.com		transposition	CSC CD	434	1,187	5,230
bankofamerica.com	🐞	addition	CSC CD	189	838	5,036
broadcom.com		replacement	CSC CD	133	1,187	4,111
adobe.com	🐞	repetition	COM LAUDE	507	1,187	3,905
discover.com		replacement	CSC CD.	319	1,083	3,167
adobe.com	🐞	addition	COM LAUDE	245	618	2,962

domain names' daily, weekly, or monthly traffic distributions were uniform. Similarly, our visual inspection of the moving average of the number of IPs querying those domains revealed no sign of regularity (see Fig. 16 in the appendix). Those results indicate probable user activity, underscoring the need for the providers to defensively register such domain names.

D. Rating the Guardians

At this point, it is only natural to wonder what might explain such oversights by the providers. Indeed, although it may be difficult to understand why some transformations received such high traffic volumes, many appear to be obvious candidates for defensive registration. Do the providers miss out on these opportunities because they consider certain classes of transformations less valuable than others? To explore that question, we attempted to mimic the observed strategy employed by each provider and then measured which of the learned strategies would lead to better protection for the set of 53 Fortune 500 companies that do not participate in defensive registration practices (see Section III-A).

1) Understanding the providers' strategies: With a large pool of available domain names to choose from, it is unclear how each provider selects which domain names to register for their customer. Our next goal is to reproduce the selection processes of each provider using machine learning.

Assumptions and problem definition: While we acknowledge that defensive registrations can be done either reactively [45], [46] or proactively, we simplify our analysis by assuming that the ones we observed were done proactively. This assumption aligns with the documented preference of the providers for proactive registrations [47]–[50], which we confirmed by analyzing the historical WHOIS records of the OBP providers' defensive registrations. Our analysis revealed that in 85.83% of cases, the OBP registrar was the first to register the domain names in that set. This analysis suggests that the OBPs that we studied engage mainly in intentional, proactive registrations rather than reactive ones.

As such, we consider that the OBP providers in our study pick which domain names to defensively register from the pool of available candidates. We aim to understand which domain names a given provider would register for a customer, C_i, with a fixed budget covering no more than n domains. We liken the problem to a recommendation system in which we learn an OBP provider's preferences from past registrations and evaluate our understanding of their strategy using future registrations. We selected a provider's past registrations in two ways: by using all the registrations for a *random* sample of 80% of their customers or *temporally* by using the first 80% domain names they registered according to the WHOIS records. Using those past registrations, we trained a logistic regression model that predicts, for a given company, an ordered list of domain names unseen at training time the provider would defensively register for them. We used the remaining 20% of the data to evaluate our models' performance.

Evaluation metrics: We consider the model to be effective as long as its predictions largely intersect with the ones that the provider actually made. We measured this likeness using the metrics described below.

Recall represents the proportion of domain names in the top-n predictions of the model that were part of the defensive registrations that the provider made for the company.

Similarity provides a numerical measure of the likeness between the top-n predicted domain names, and those registered by the provider. It is measured across three equally weighted factors: the transformation (insertion, replacement, deletion), the gTLD, and the high-level categories under which the domain name falls. Those categories include 1) *bitsquatting*; 2) *fat-finger mistakes* on a QWERTY keyboard; 3) *other typing mistakes* including omissions, repetitions, and transpositions; 4) *homophones and pseudo-homophones*; 5) *homographs and pseudo-homographs*; *delimiter modification* referring to the insertion or removal of a dash between two words; 6) *grammatical mistakes* including the use of plurals or singular; 7) *semantic transformations*, and 8) *other* domain names that do not fall under any of the previous categories.

Recall@K/Similarity@K are variants that account for the fact that only a small fraction of candidate domain names are ever defensively registered. For various values of K, representing a percentage of the predicted domain names, we evaluate each of the metrics of interest. Similar metrics are used in traditional recommender systems study [51].

Differences in strategies: Understanding OBP service providers' strategies assumes that they have a consistent pattern of defensive registrations. To assess the validity of that assumption, we performed, for each provider, a Principal Component Analysis reduction on a matrix that represents the proportion of domain names from a group of high-level categories that the provider registered for each of their customers. The results showed that the first principal component amounted to an explained variance ratio of at least 0.712, suggesting a homogeneous pattern of registrations.

However, this homogeneity does not transfer over to other providers based on a Kruskal test between providers. The

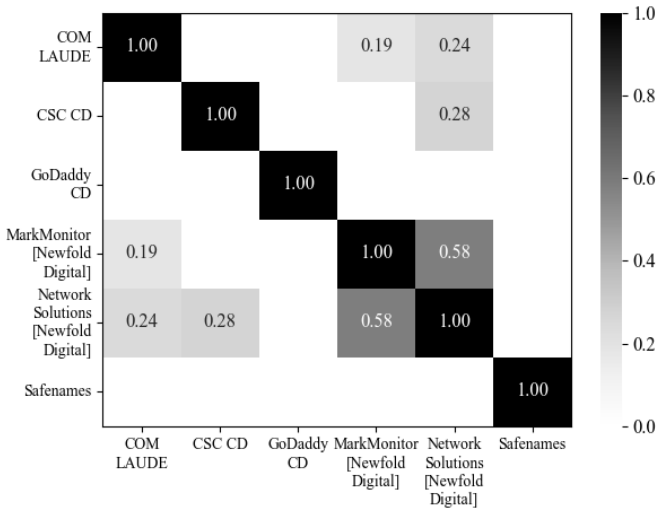

Fig. 10: Kruskal test on the registration patterns of each provider. Each column shows the associated p value for every statistically significant result at the 95% confidence level.

results obtained at a significance level of 95% show that registration patterns are different from one provider to another (see Fig. 10). Interestingly, the highest estimated similarity with a p value of 0.58 was between MarkMonitor and Network Solutions, both of which Newfold Digital acquired recently [52], [53]. Com Laude's similarity with those two providers resulted in p values of 0.19 and 0.24 for MarkMonitor and Network Solutions, respectively. In trying to understand why that might be the case, we found disclosures in official French documents [54] showing that two founders of Com Laude worked for a company called Net Searchers which was renamed to Register.com in 2005 and then subsequently renamed to CSC three years later. The fact that Newfold Digital now owns Register.com might explain the weaker commonality between Com Laude and the two providers. Conceivably, the frequent mergers and acquisitions [18] in the online brand protection field might have resulted in a set of common practices.

Model and features description: For each transformation, we trained a logistic regression model with a boolean dependent variable indicating whether the target registrar defensively registered the domain name for their customer. Our independent variables consisted of several features summarized in Table VIII (in the appendix) that we extracted from each candidate domain name. Those features were inspired by works on domain name squatting [16], [55]–[60], package confusion [39] and psychology [61], [62]. At inference time, we used the model's probability score to rank the candidate domain names. Before combining the predictions from the transformation-specific models into a unified sorted list for each provider, we scaled the predicted probabilities using a prior probability estimated as the proportion of the provider's defensive registrations that emanate from said transformation.

Results: Overall, we observed similar performance in both settings (random/temporal) so in what follows, we restrict the discussion to the random split as it better aligns with our use of the models in Section III-D2 to predict a provider's registration for new companies based on their choices for existing customers. Interested readers are referred to Section A-D of the appendix for the results of the temporal setting.

Using a ten-fold cross-validation analysis for each provider, we found that under our strictest measure (i.e., recall), our models correctly predicted 80% of the registrations of almost every provider while going through less than 10% of the available space of candidates. On average, registering 10% of the available space (about 782 domains) would cost approximately 8K USD for each company. We present the results in Fig. 11.

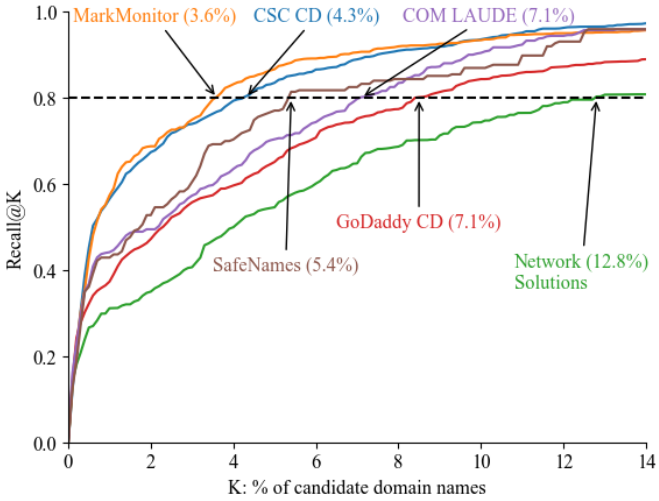

Fig. 11: Recall@K for the six providers that we studied.

As the plot shows, our model performed worst on Network Solutions, requiring us to go through 12.8% of the predictions for an 80% recall. In examining why, we found that Network Solutions registered 227 domains for one customer (Xcel Energy) while their median number of registrations per customer overall was 15. When the company with the most registrations was in the testing set, the model did not have enough data to learn those specific patterns and thus performed poorly.

As expected, under the similarity metric, the model's performance is significantly better (see Fig. 12). By predicting domain names that resemble the actual categories from which the providers drew, we improved our performance on Network Solutions (examining only 4.6% of the space to achieve a score of 80% on the strict measure). Success was best for MarkMonitor (0.8%), closely followed by CSC Corporate Domains (0.9%), with the smaller providers trailing behind (SafeNames Ltd. - 2.1%, Nom-iq Ltd dba COM LAUDE - 2.4%, GoDaddy Corporate Domains - 2.7%).

The fact that the models for MarkMonitor and CSC Corporate Domains achieve a higher result than the other providers underscores the utility of having sufficient data to train and test the models on. This justifies our data-driven approach to understanding the strategies of the providers. To further illustrate that point, we trained a new model using *only* domain

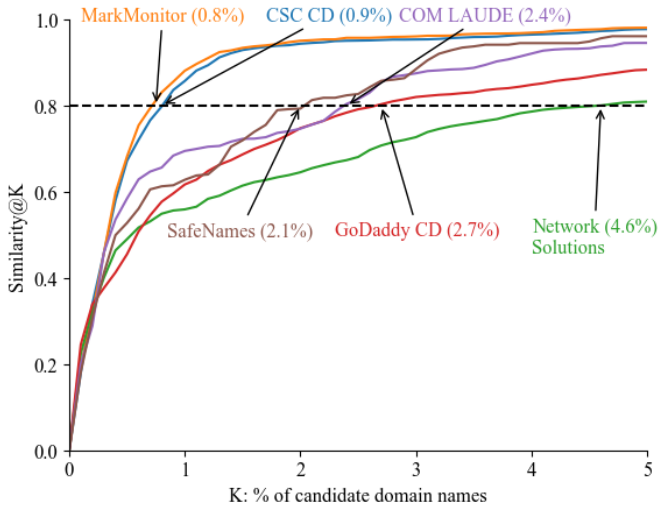

Fig. 12: Similarity@K for the six providers that we studied.

names seen in OSINT data. The resulting model performed almost 8 times worse compared to the others.

2) Applying the strategies to new customers: Having successfully captured the preferences of the providers using features derived from the scientific literature, we now examine the aptness of the providers' ranking of domain names to capture the worthiest transformations as seen in the volumetric data. In doing so, we observed one provider consistently outperforming the others with an intuitive approach that we summarized.

Experimental design: Using the trained models for the six providers, we predicted the ranking of the transformations for the 53 Fortune 500 for which we did not identify any defensive registrations. We then evaluated how those rankings compare with the volumetric data. Our ranking of domain names using the volumetric data leveraged three variables: 1) *sparsity* referring to the median number of distinct IPs that queried the domain name, 2) *popularity* denoting the total number of queries received by the domain name, and *regularity* measured as the proportion of days during which the domain name received at least one query. We compare the rankings from each of those metrics with the models' predictions using the discounted cumulative gain (DCG) used in the evaluation of recommendation systems [63] to reward early suggestions of highly valuable items and expressed as

$$DCG_d = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)} \quad (1)$$

where rel_i represents the relevance of the domain name. For simplicity, we estimate the relevance as the quartile in which the observed value falls relative to the maximum observed value for a domain name. For example, if the maximum query volume for the transformations of a domain name is 100, the intervals $(1, 25), (26, 50), (51, 75), (76, 100)$ correspond to a relevance of 1, 2, 3, and 4 respectively.

Results: Analyzing the results of this comparison, we found that no provider has a perfect strategy for any of the 53 new customers (DCG = 1). In Fig. 13, we see that the comparison based on the *regularity* criteria led to the highest scores among all the providers with values going up to 0.73 compared to a maximum of 0.54 (*popularity*) and 0.53 (*sparsity*) for the other criteria. This suggests that the providers' selections are best approximated with a ranking based on how many days each domain name was queried.

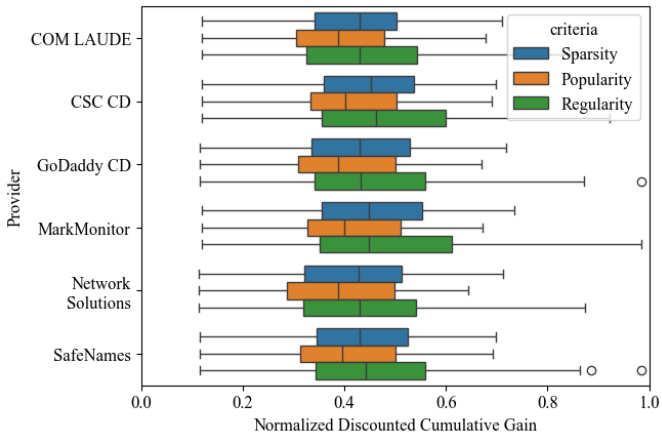

Fig. 13: Normalized Discounted Cumulative Gains per provider for each criteria considered.

The plot also shows that the scores differ from one provider to the other. To assess whether those differences are significant, we ran a Mann-Whitney U test under the null hypothesis that the distributions of DCG scores of two providers are equivalent on set criteria. When we reject the null hypothesis at the 95% confidence level, we accept the alternative hypothesis that the first provider's scores are greater than the other's. In such cases, we attribute a win (+1) to the first provider and a loss (-1) to the second. The results presented in Table V show that MarkMonitor Inc. has the best overall performance, performing better than at least one provider on each criterion. The provider Network Solutions, on the other hand, does worse than at least one provider for every criterion. SafeNames is the only small provider with a non-negative score overall.

TABLE V: Pugh matrix comparing the performance of the providers based on the considered criteria.

Provider	Popularity	Regularity	Sparsity	Total
MarkMonitor Inc.	+2	+1	+3	+6
CSC Corporate Domains, Inc.	+1	0	+2	+3
SafeNames Ltd.	0	0	0	0
GoDaddy Corporate Domains, LLC	0	0	-1	-1
Nom-iq Ltd. dba COM LAUDE	-1	0	-2	-3
Network Solutions, LLC	-2	-1	-2	-6

Lessons learned from the winning strategy: Since they outperformed the other providers, we decided to analyze the strategies employed by MarkMonitor. To be succinct, we focus on two classes of operations that we found the providers missing many valuable domain names from: character insertion and replacement. By taking a look at the model's weight on those operations for MarkMonitor Inc., we observe that this provider

prefers registering domain names in the same gTLD as the root domain name, and those in the legacy gTLDs (com, net, org). When the transformations are in the .info gTLD, MarkMonitor seems less willing to register them.

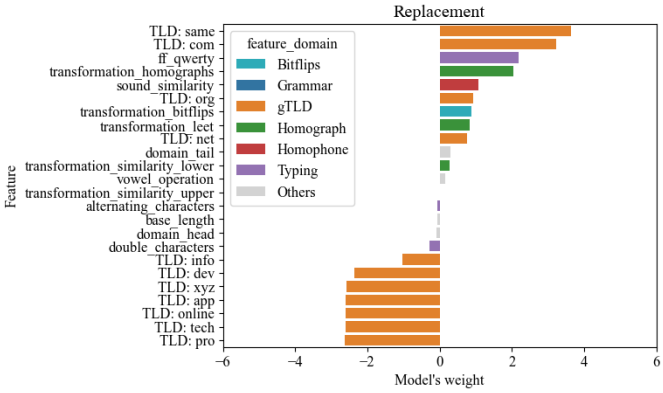

Fig. 14: Features' weight of the replacement model for MarkMonitor Inc.

Specific to the addition operation, we see in Fig. 15, that the provider places more emphasis on easy-to-make grammatical mistakes. The models also weigh highly the domain names that sound similar to the base domain name. This is true for both the replacement and addition transformations. One notable difference is that typing mistakes are weighted more for replacement mistakes than for addition. This makes sense because the set of fat-finger mistakes is much higher for addition than for replacement mistakes. We also observe a significant insistence on homographic transformations for both operations. Overall, those strategies could help improve the performance of the other providers.

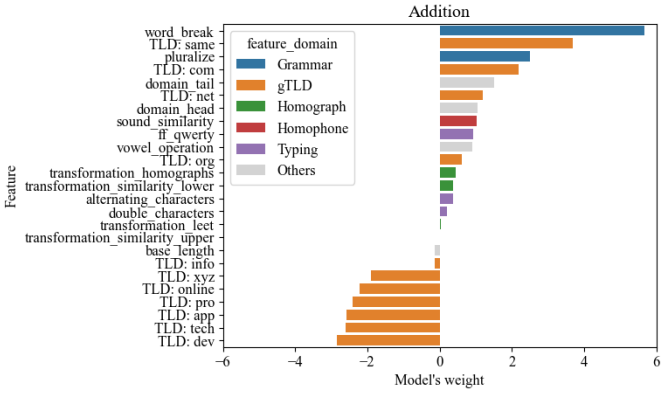

Fig. 15: Features' weight of the addition model for MarkMonitor.

E. Recommendations

During our assessment of the efficacy of the providers' strategies, we often found that they failed to defensively register brand-infringing domain names that received significant amounts of traffic during the observation period. Unfortunately, the longer these providers overlook the available domain names, the higher the chances that a malicious third party will register them. This opportunistic registration could happen by mere happenstance or, in the case of a motivated adversary, through careful analysis of any widely available passive DNS data (e.g., Circl.lu [64], WhoisXML [65], Farsight DBDNS2 [66]) to identify the most valuable transformations of their target brand. Under those circumstances, the affected company can only hope that the monitoring services of their provider (e.g., CSC's 3D domain security and enforcement [67] service) can quickly identify the ongoing case of abuse. In such cases, they can request the transfer of the domain name through the Unified domain name Dispute Resolution Policy (UDRP) by filing a complaint with one of the ICANN-accredited resolution providers (e.g., WIPO). Unfortunately, the process is both lengthy and expensive: the policy requires ten business days only to apply the decision [68], and WIPO charges a minimum of 1,500 USD just for filing a complaint [69]. In contrast, a standard registration of the domain names in Table IV can be completed within minutes, at a median price of 14.99 USD using GoDaddy.

An obvious solution is for providers to leverage passive DNS data to identify heavily queried available domain names and register them before miscreants have a chance to do so. This would improve the defensive posture of their customers while significantly reducing their costs. As a case in point, our analysis of 696 complaints filed by Phillips Morris International — a Fortune 500 company serviced by MarkMonitor — revealed that the company spent anywhere between 1.046 and 2.788 million USD [2] just to register their complaints with WIPO. These costs do not include other legal fees. For other heavy-hitter complainants, like the Danish Lego Group, the costs for protecting their brand is staggering – well over 4.459 million for 1,104 complaints. Providers could lessen their customers' expenditures on domain name disputes if they spot heavily queried available domain names using passive DNS.

IV. Limitations

Our examination of defensive registration practices for the Fortune 500 comes with a few limitations. First, despite the large pool of domain names we studied (spanning more than 146 million transformations), one could argue that we do miss important subclasses that are worth studying. For instance, it is reasonable to expect that very popular domain names like facebook[.]com might have more defensive registrations than other domains (i.e., meta[.]com) that are related to the same brand. However, past research [17] suggests that as little as 23 out of the top 10,000 sites have any defensive registrations. Those findings might be explained by the preponderance of content distribution providers (e.g., akamai[.]net) and websites operated by non-profit entities (e.g., wikipedia[.]org) among the most popular domains: on one hand, the typical Internet user does not directly interact with such domains, and on the other, the cost of defensive registration might be prohibitively

[2]The actual cost depends on how many panelists the complainant requested.

expensive for those sites. The same does not apply to the 500 domain names we studied as they are all meant to be accessed by users, and the companies they belong to can afford to protect their brands online. Those assertions would also hold for successful businesses outside the U.S., like those on the Global Fortune 500. Using such an international list of companies would certainly allow one to get a broader view of defensive registration practices. However, since we assessed the effectiveness of the providers' strategies using passive DNS data collected from U.S. residential customers, we focused on businesses with good visibility in the U.S., such as the Fortune 500, which are all headquartered there. We leave the analysis of global patterns of defensive registrations to future work.

Second, by being conservative about what we consider as a defensive registration, we could have missed certain defensive registrations. First, our approach does not attempt to label domain names that are not in the zone files yet are registered (like wakmart[.]com). Unfortunately, identifying all those instances would require access to the WHOIS records for every generated transformation which is impracticable. We do not have an estimate of such unresolvable false negatives. For the domain names that are in the zone files, we estimated that around 490 might be defensively registered similar to wal-mart[.]com (see §II-C). This figure corresponds to the number of domains with anonymous WHOIS records that are registered through any of the OBPs that we identified who offer no domain name retail business (e.g., CSC Corporate Domains, Inc.). The domain names registered with registrars with a retail offer (e.g., Network Solutions) could have been the work of an opportunistic third party. It is possible that we count such third-party registrations as defensive if the owner sets all the nameservers to those owned by the corresponding Fortune 500 company. This seems an unlikely adversarial approach since the name servers will not respond to queries for the domain names. Alternatively, an adversary can pretend to be a Fortune 500 company by faking the information in the WHOIS records. We found no evidence of this in our dataset. More specifically, upon navigating to the web pages hosted on the defensive registrations we identified, they all redirected to a domain name owned by the Fortune 500 company.

V. Related Work

This paper builds upon years of research on domain name abuse discussing the topic of defensive registrations. Those works independently studied typosquatting [1], [27], [56], bitsquatting [2], sound-squatting [3], homographs [17], [59], [70] and semantic transformations [5], [15], [29]. We not only consider all of those transformations but also introduce a new type, i.e., stock name squatting. The breadth of types allows us to conduct an analysis of defensive registrations spanning more than 146 million potential domain names across 383 gTLDs. This is the largest study to date of this kind, covering orders of magnitude more domains than recent work [71].

Given the scale of our inquiry, it was imperative that we devised a methodology that built trust and confidence in our findings. Therefore, to address growing concerns about the use of cloud DNS infrastructure by miscreants and legitimate companies alike [72]; as well as the increasing number of anonymized or potentially faked WHOIS records [21], we designed a conservative approach to identify defensive registrations. By leveraging name servers' details and WHOIS records when needed, our approach can be likened to that of Pouryousef et al. [16] that considered a domain name to be defensively registered if its WHOIS records or those of its name servers suggest so. However, we take additional constraints into consideration (i.e., steps ❶ and ❹ in Section II-C) such as using non-anonymous organization email addresses like in [73], and we also match the WHOIS records against ten years of SEC data. We could not use Sebastián et al.'s recent attribution technique [74] since it uses the privacy policy of a website and most of the defensive registrations had no web content. Overall, our careful design choices allowed us to identify 447 distinct companies with at least one defensive registration — which is far more than prior work.

Most closely related is the concurrent work of Benjamin et al. [71] that studied 36,027 defensive registrations associated with 370 brand names. Benjamin et al. [71] found that the vast majority of these registrations were instances of TLD squatting followed by combo-squatting, and that most defensive registrars do not take maximum advantage of the claims phases for new generic TLDs. Specifically, they found that 68.7% of the domain name registrations took place after the end of the sunrise period. Coincidentally, while they found a moderately positive relationship between attack frequency and defensive registrations per brand, they suggest that "*other elements such as specific industry sectors may also be influential*" and recommend that "*future research, including regression modeling, could uncover a more comprehensive list of variables contributing to defensive registrations.*" This is precisely the type of analysis we perform in this paper.

For example, by meticulously analyzing the observed defensive registrations, we found that a company is more likely to register domain names in gTLDs related to their sector of activity. In addition to filling this gap in their work, we conducted the first systematic assessment of OBPs' strategies based on passive DNS data. This analysis led to the worrisome finding that the providers exposed their customers to online brand abuse by not registering many highly queried domain names that were available for an extended period. To assert whether those missed opportunities were due to oversights in the providers' decisions, we trained, for the first time in the literature, machine-learning models that replicate their strategies. Our models predict, for a given company, the sorted list of available domain names a provider would likely register.

Along similar lines, ranking systems for defensive registrations were also proposed elsewhere. Ahmad et al. [75], for example, trained a recurrent neural network using n-grams extracted from a passive DNS trace to estimate the probability that users will visit a given typosquat. Likewise, Tahir et al. [57] proposed a metric dubbed the "Hardness Quotient" that gauges how strenuous users will find it typing a specific domain name. Their estimation was based on insights on

how the human hand's anatomy can increase the risks of typing mistakes. Collectively, these works suggest that their models can be used to assist with defensive registrations but did not compare the performance of their approach against others. By contrast, we compare the strategies used by six large providers that protect their customers' brands using multiple transformation classes, and assess the effectiveness thereof by examining how often those defensively registered domains are queried by users. Overall, the breadth of our study allowed us to understand the strengths and weaknesses of each strategy, leading to actionable recommendations (e.g., that domain names generated by hyphenation or grammatical mistakes in the same gTLD should be registered with priority).

VI. Conclusion

We analyzed the defensive registration practices of Fortune 500 companies. Through our longitudinal study of several classes of domain name transformations, we found that most Fortune 500 companies engage in defensive registrations. For the most part, those registrations only consist of simple grammatical or semantic mistakes, yet they receive a lot of traffic. We find that most Fortune 500 companies turn to a few brand protection service providers to protect their online presence. By leveraging several works from the academic literature, we successfully modeled those providers' strategies. While the major players exhibit a propensity for different strategies, our evaluation of the effectiveness of their approaches showed that they each offer a reasonable level of protection to the brands they protect. Nonetheless, they can improve their protections by using passive DNS data to secure highly queried available domain names before a third party can abuse them.

Future work could focus on understanding why such domain names receive so much traffic and what security issues they lead to beyond the scope of our study. Similarly, a promising line of research is to propose models to better guide practices that improve trust and safety online.

Acknowledgment

The authors express their gratitude to the anonymous reviewers for their detailed feedback that helped improve the paper. We also thank Boo Fullwood and Tillson Galloway for their comments and suggestions on an earlier draft of this paper.

References

[1] P. Agten, P. F. Wouter, and N. Nikiforakis, "Seven months' worth of mistakes: A longitudinal study of typosquatting abuse," in *Netw. Distrib. Syst. Secur. Symp.*, Feb. 2015, pp. 1–13.

[2] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting: Exploiting bit-flips for fun, or profit?" in *Proc. 22nd Int. Conf. World Wide Web*. New York, NY, USA: ACM, May 2013, pp. 989–998.

[3] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Inf. Secur.* Springer Int. Publishing, 2014, pp. 291–308.

[4] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis, "Hiding in plain sight: A longitudinal study of combosquatting abuse," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: ACM, 2017, pp. 569–586.

[5] N. Kumar, S. Ghewari, H. Tupsamudre, M. Shukla, and S. Lodha, "When diversity meets hostility: A study of domain squatting abuse in online banking," in *APWG Symp. Electron. Crime Res.*, Dec. 2021, pp. 1–15.

[6] D. Liu, S. Hao, and H. Wang, "All your DNS records point to us: Understanding the security threats of dangling dns records," in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur.* ACM, 2016, pp. 1414–1425.

[7] T. Vissers, T. Barron, T. Van Goethem, W. Joosen, and N. Nikiforakis, "The wolf of name street: Hijacking domains through their nameservers," in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur.* ACM, 2017, pp. 957–970.

[8] B. Liu, C. Lu, H. Duan, Y. Liu, Z. Li, S. Hao, and M. Yang, "Who is answering my queries: Understanding and characterizing interception of the DNS resolution path," in *USENIX Secur. Symp.*, Aug. 2018, pp. 1113–1128.

[9] M. Squarcina, M. Tempesta, L. Veronese, S. Calzavara, and M. Maffei, "Can I take your subdomain? Exploring same-site attacks in the modern web," in *USENIX Secur. Symp.*, Aug. 2021, pp. 2917–2934.

[10] T. Dai, P. Jeitner, H. Shulman, and M. Waidner, "The hijackers guide to the galaxy: Off-Path taking over internet resources," in *30th USENIX Secur. Symp.*, Aug. 2021, pp. 3147–3164.

[11] C. Hellen. (2017, Jun.) United Airlines: Meet the man waging a 20-year war against company that ignored his complaint. Last accessed on Feb. 21, 2024. [Online]. Available: https://www.independent.co.uk/travel/news-and-advice/united-airlines-flights-20-years-untied-parody-complaint-website-jeremy-cooperstock-court-injunction-paxex-a7811921.html

[12] ICANN Wiki. (2022, Jan.) Defensive registration. Last accessed on Oct. 04, 2024. [Online]. Available: https://icannwiki.org/Defensive_Registration

[13] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A comprehensive measurement study of domain-squatting abuse," in *IEEE Int. Conf. Commun.*, 2019, pp. 1–6.

[14] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker, "XXXtortion? Inferring registration intent in the .xxx TLD," in *Proc. 23rd Int. Conf. World Wide Web*. New York, NY, USA: ACM, 2014, pp. 901–912.

[15] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker, "From .academy to .zone: An analysis of the new TLD land rush," in *Proc. Internet Meas. Conf.* New York, NY, USA: ACM, 2015, pp. 381–394.

[16] S. Pouryousef, M. D. Dar, S. Ahmad, P. Gill, and R. Nithyanand, "Extortion or expansion? An investigation into the costs and consequences of ICANN's gTLD experiments," in *Passive Active Meas.*, vol. 12048. Springer International Publishing, 2020, pp. 141–157.

[17] F. Quinkert, T. Lauinger, W. Robertson, E. Kirda, and T. Holz, "It's not what it looks like: Measuring attacks and defensive registrations of homograph domains," in *IEEE Conf. Commun. Netw. Secur.*, 2019, pp. 259–267.

[18] C. Abrahams. (2023, Jan.) 2023 review of the online brand protection market. Last accessed Nov. 01, 2023. [Online]. Available: https://circleid.com/posts/20230110-2023-review-of-the-online-brand-protection-market

[19] U.S. Securities and Exchange Commision. (2022, Sep.) Newfold Digital signs agreement to acquire MarkMonitor from Clarivate. Last accessed Dec. 11, 2023. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1764046/000110465922099073/tm2225609d2_ex99-1.htm

[20] T. Liu, Y. Zhang, J. Shi, Y. Jing, Q. Li, and L. Guo, "Towards quantifying visual similarity of domain names for combating typosquatting abuse," in *IEEE Mil. Commun. Conf.*, Nov. 2016, pp. 770–775.

[21] C. Lu, B. Liu, Y. Zhang, Z. Li, F. Zhang, H. Duan, Y. Liu, J. Q. Chen, J. Liang, Z. Zhang *et al.*, "From WHOIS to WHOWAS: A large-scale measurement study of domain registration privacy under the GDPR," in *Netw. Distrib. Syst. Secur. Symp.*, Feb. 2021, pp. 1–18.

[22] Fortune Media IP Limited. (2023, May) Fortune 500 2023. Last accessed on Nov. 08, 2023. [Online]. Available: https://fortune.com/ranking/fortune500/

[23] K. Du, H. Yang, Z. Li, H. Duan, S. Hao, B. Liu, Y. Ye, M. Liu, X. Su, G. Liu, Z. Geng, Z. Zhang, and J. Liang, "TL;DR hazard: A comprehensive study of levelsquatting scams," in *Secur. Privacy in Commun. Netw.*, vol. 305, no. 1. Springer International Publishing, Dec. 2019, pp. 3–25.

[24] O. R. Gutierrez, "Get off my URL: Congress outlaws cybersquatting in the wild west of the internet comment," *Santa Clara Comput. High-Technol. Law J.*, vol. 17, pp. 139–167, 2001.

[25] Y. Zeng, X. Chen, T. Zang, and H. Tsang, "Winding path: Characterizing the malicious redirection in squatting domain names," in *Passive and Active Meas.* Springer International Publishing, 2021, pp. 93–107.

[26] J. Reynolds, D. Kumar, Z. Ma, R. Subramanian, M. Wu, M. Shelton, J. Mason, E. Stark, and M. Bailey, "Measuring identity confusion with uniform resource locators," in *Proc. 2020 Conf. Human Factors Comput. Syst.* ACM, 2020, pp. 1–12.

[27] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels, "Strider typo-patrol: Discovery and analysis of systematic typo-squatting," *2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet*, vol. 6, no. 2, pp. 31–36, Jul. 2006.

[28] D. Artem, "Bitsquatiing: DNS hijacking without exploitation," in *Proc. BlackHat Secur.*, Jul. 2011. [Online]. Available: https://media.blackhat.com/bh-us-11/Dinaburg/BH_US_11_Dinaburg_Bitsquatting_WP.pdf

[29] P. Lv, J. Ya, T. Liu, J. Shi, B. Fang, and Z. Gu, "You have more abbreviations than you know: A study of AbbrevSquatting abuse," in *Comput. Sci.*, vol. 10860. Springer International Publishing, 2018, pp. 221–233.

[30] Security and Exchange Commission. (2023, Nov.) EDGAR full text search. Last accessed on Feb. 14, 2024. [Online]. Available: https://www.sec.gov/edgar/search/

[31] ICANN. (2018, Dec.) FAQs for registrants: Domain name renewals and expiration. Last accessed on Feb. 14, 2024. [Online]. Available: https://www.icann.org/resources/pages/domain-name-renewal-expiration-faqs-2018-12-07-en

[32] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st Int. Conf. Distrib. Comput. Syst. Workshops*, Jun. 2011, pp. 166–171.

[33] V. L. Pochat, T. van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," *Proc. 2019 Netw. Distrib. Syst. Secur. Symp.*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:56367624

[34] Bureau of Labor Statistics. (2022, May) Occupational employment and wage statistics query system. Last accessed on Jun. 23, 2023. [Online]. Available: https://data.bls.gov/oes//#/home

[35] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan, "Cyber-fraud is one typo away," in *27th Conf. Comput. Commun.*, 2008, pp. 1939–1947.

[36] Google Cloud. (2023, Oct.) Content categories. Last accessed on Feb. 14, 2024. [Online]. Available: https://cloud.google.com/natural-language/docs/categories

[37] TLD-List. (2023, Dec.) Compare Prices of All Top-Level Domains. Last accessed on Jan. 13, 2024. [Online]. Available: https://tld-list.com

[38] M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, A. Noroozian, D. Bagley, and C. Hesselman, "Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs," *Asia Conf. Comput. Commun. Secur.*, pp. 609–623, 2018.

[39] S. Neupane, G. Holmes, E. Wyss, D. Davidson, and L. D. Carli, "Beyond typosquatting: An in-depth look at package confusion," in *32nd USENIX Secur. Symp.* Anaheim, CA: USENIX Association, Aug. 2023, pp. 3439–3456.

[40] Google Domains. (2023, Sep.) About the Squarespace purchase of Google Domains registrations. Last accessed on Nov. 22, 2024. [Online]. Available: https://support.google.com/domains/answer/13689670?hl=en

[41] A. Affinito, R. Sommese, G. Akiwate, S. Savage, K. Claffy, G. M. Voelker, A. Botta, and M. Jonker, "Domain name lifetimes: Baseline and threats," in *6th Netw. Traffic Meas. Anal. Conf.* International Federation for Information Processing (IFIP), 2022, pp. 1–9.

[42] VirusTotal. (2024) Virustotal-free online virus, malware and URL scanner. Last accessed Aug. 23, 2024. [Online]. Available: https://www.virustotal.com/en

[43] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by Internet-wide scanning," in *22nd ACM Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 542–553.

[44] K. An, "Sulla determinazione empirica di una legge didistribuzione," *Giorn Dell'inst Ital Degli Att*, vol. 4, pp. 89–91, 1933.

[45] Clearinghouse. (2014, Dec.) What is the Trademark Clearinghouse? — www.trademark-clearinghouse.com. Last accessed Jan. 05, 2024. [Online]. Available: https://trademark-clearinghouse.com/content/what-trademark-clearinghouse

[46] V. Holstein-Childress, "Lex Cyberus: The UDRP as a gatekeeper to judicial resolution of competing rights to domain names," *Penn State Law Rev.*, vol. 109, pp. 565–607, 2005.

[47] L. Shane. (2023, Aug.) NFT domains: considerations for corporations. Last accessed on Oct. 07, 2024. [Online]. Available: https://www.markmonitor.com/blog/nft-domains-corporate-considerations/

[48] P. Lodico. (2024, May) Defense is the best offense: Domain names' role in brand protection. Last accessed on Oct. 07, 2024. [Online]. Available: https://gcd.com/posts/defense-is-the-best-offense-domain-names-role-in-brand-protection/

[49] CSC Corporate Domains. (2024, Apr.) How CSC's proactive and personal approach supports Breville's market expansion. Last accessed on Oct. 07, 2024. [Online]. Available: https://cscdbs.com/blog/how-cscs-proactive-and-personal-approach-supports-brevilles-market-expansion

[50] Com Laude. (2023, Jan.) The Com Laude casebook: Two sides of the same coin. Last accessed on Oct. 07, 2024. [Online]. Available: https://comlaude.com/domain-management-brand-protection-2/

[51] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," in *Proc. 2nd ACM Conf. Electron. Commerce*. New York, NY, USA: ACM, 2000, pp. 158–167.

[52] E. Susan Datz and D. Kori. (2011, Aug.) Web.com to acquire Network Solutions. Last accessed on Feb. 14, 2024. [Online]. Available: https://newfold.com/newsroom/webcom-acquire-network-solutions

[53] C. Nicole, W. Amy Bourke, and D. Mark. (2022, Sep.) Newfold Digital signs agreement to acquire MarkMonitor from Clarivate. Last accessed on Feb. 14, 2024. [Online]. Available: https://newfold.com/newsroom/newfold-digital-signs-agreement-to-acquire-markmonitor-from-clar

[54] Figaro Emploi. (2021, Apr.) Corporation Service Company France (75009) : siret, siren, TVA, bilan gratuit... Last accessed on Feb. 14, 2024. [Online]. Available: https://entreprises.lefigaro.fr/corporation-service-company-france-75/entreprise-434129805

[55] M. S. Starr and K. K. Fleming, "A rose by any other name is not the same: The role of orthographic knowledge in homophone confusion errors," *J. Exp. Psychol.: Learning, Memory, and Cognition*, vol. 27, no. 3, pp. 744–760, 2001.

[56] T. Moore and B. Edelman, "Measuring the perpetrators and funders of typosquatting," in *Financial Cryptography and Data Security*, R. Sion, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 175–191.

[57] R. Tahir, A. Raza, F. Ahmad, J. Kazi, F. Zaffar, C. Kanich, and M. Caesar, "It's all in the name: Why some URLs are more vulnerable to typosquatting," in *IEEE Conf. Comput. Commun.*, 2018, pp. 2618–2626.

[58] DNSTwist. (2023) DNS Twist. Last accessed Jan. 05, 2024. [Online]. Available: https://github.com/elceef/dnstwist

[59] T. Holgers, D. E. Watson, and S. D. Gribble, "Cutting through the confusion: A measurement study of homograph attacks," in *USENIX Annu. Tech. Conf.* USENIX Association, May 2006, pp. 261–266.

[60] D. Chiba, A. A. Hasegawa, T. Koide, Y. Sawabe, S. Goto, and M. Akiyama, "DomainScouter: Understanding the risks of deceptive IDNs," in *Int. Symp. Res. Attacks, Intrusions and Defenses*. Chaoyang District, Beijing: USENIX Association, Sep. 2019, pp. 413–426.

[61] M. Perea, J. A. Duñabeitia, and M. Carreiras, "R34d1ng w0rd5 w1th numb3r5," *J. Exp. Psychol.: Human Perception Perform.*, vol. 34, no. 1, pp. 237–241, 2008.

[62] I. C. Simpson, P. Mousikou, J. M. Montoya, and S. Defior, "A letter visual-similarity matrix for latin-based alphabets," *Behav. Res. Methods*, vol. 45, no. 2, pp. 431–439, 2013.

[63] G. Dupret, "Discounted cumulative gain and user decision models," in *String Process. Inf. Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 2–13.

[64] CIRCL.LU. (2023, Oct.) Circl Passive DNS. Last accessed on Jun. 25, 2024. [Online]. Available: https://www.circl.lu/services/passive-dns/

[65] WhoisXML. (2024, Jun.) DNS database download — The largest passive DNS coverage — WhoisXML API. Last accessed on Jun. 25, 2024. [Online]. Available: https://dns-history.whoisxmlapi.com/database

[66] DomainsTools. (2024, Mar.) Introducing DNSDB 2.0 — Passive DNS — DomainTools. Last accessed on Jun. 25, 2024. [Online]. Available: https://www.domaintools.com/products/farsight-dnsdb/

[67] Corporation Service Company. Internet brand monitoring — CSC. Last accessed Jan. 05, 2024. [Online]. Available: https://www.cscdbs.com/en/brand-protection/brand-monitoring-services/internet-monitoring/

[68] ICANN. (2024, Feb.) Uniform domain name dispute resolution policy. Last accessed on Jun. 26, 2024. [Online]. Available: https://www.icann.org/resources/pages/udrp-rules-2024-02-21-en

[69] WIPO. (2015, Jul.) Schedule of fees in WIPO domain name dispute resolution proceedings. Last accessed on Jun. 25, 2024. [Online]. Available: https://www.wipo.int/amc/en/domains/fees/

[70] V. Le Pochat, T. Van Goethem, and W. Joosen, "A Smörgåsbord of typos: Exploring international keyboard layout typosquatting," in *IEEE Secur. Privacy Workshops*, May 2019, pp. 187–192.
[71] B. C. Benjamin, J. Bayer, S. Fernandez, A. Duda, and M. Korczynski, "Shielding brands: An in-depth analysis of defensive domain registration practices against cyber-squatting," in *Netw. Traffic Meas. Anal. Conf.*, May 2024, pp. 1–11.
[72] P. Foremski, O. Gasser, and G. C. M. Moura, "DNS observatory: The big picture of the DNS," in *Proc. Internet Meas. Conf.* New York, NY, USA: ACM, 2019, pp. 87–100.
[73] F. Cangialosi, T. Chung, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, and C. Wilson, "Measurement and analysis of private key sharing in the HTTPS ecosystem," in *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: ACM, 2016, pp. 628–640.
[74] S. Sebastián, R.-G. Diugan, J. Caballero, I. Sanchez-Rola, and L. Bilge, "Domain and website attribution beyond WHOIS," in *Proc. 39th Ann. Comput. Secur. Appl. Conf.* New York, NY, USA: ACM, 2023, pp. 124–137.
[75] I. Ahmad, M. A. Parvez, and A. Iqbal, "TypoWriter: A tool to prevent typosquatting," in *IEEE Ann. Comput. Softw. and Appl. Conf.*, vol. 1, Jul. 2019, pp. 423–432.
[76] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, "Targeted online password guessing: An underestimated threat," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: ACM, 2016, pp. 1242–1254.

APPENDIX A
ADDITIONAL INFORMATION

A. Query Success

The results from WHOIS queries are given in Table VI. Unsuccessful queries arise due to misconfigurations (e.g., the registrar WHOIS server was either missing or unreachable), cases where the domain name had expired or was in the redemption period, or when a TLD did not have a WHOIS service (e.g., .es)

TABLE VI: WHOIS Query Results

Category	Queried	Successfully queried (%)
Base domain name	500	500 (100.0%)
Email domains	3,713	3,083 (83.00%)
Name servers	15,849	15,750 (99.38%)
Transformations	402,934	402,916 (99.99%)
Overall	**419,638**	**418,892 (99.82%)**

B. Traffic distribution of historically abused domain names

Figure 16 shows the moving average of the number of IPs querying each of the five previously abusive domain names in Table IV for a window size a week. The plots show the number of clients querying each of the domain names remains inconsistent over time which demonstrates that scanners are not solely responsible for the traffic those domains received.

C. Features used to train the models

We summarized the features we used to train the logistic regression models in Table VIII.

D. Evaluation based on a temporal training approach

Under the temporal setting, we used the earliest 80% of a provider's registrations to train the model and evaluated the models' performance using their 20% latest registrations. In Fig. 17, we show the proportion of observed defensive registrations each provider had at a given point in time and highlight the cutoff we used for each of them.

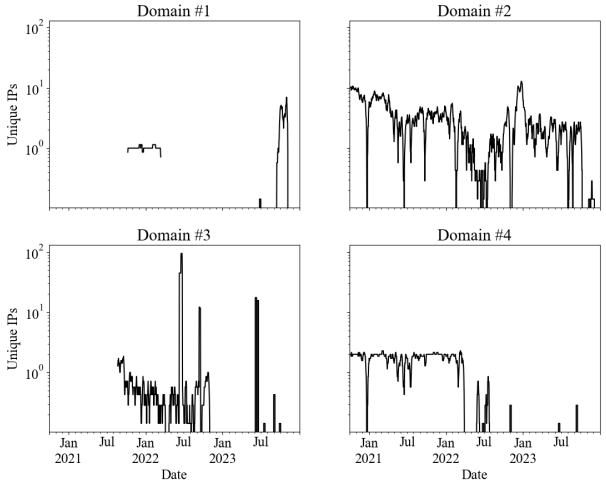

Fig. 16: Moving average of the number of unique IPs querying the previously abusive domain names of Table IV. Empty periods correspond to when a domain name was unavailable.

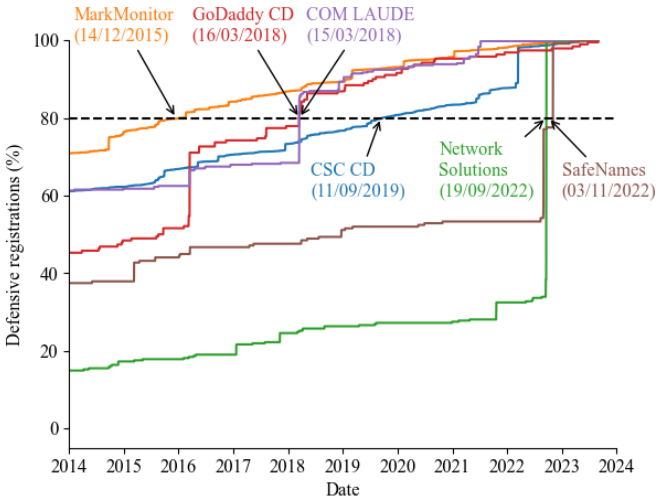

Fig. 17: Time series of the number of defensive registrations per provider and the date (format: DD/MM/YYYY) at which they registered 80% of their defensive registrations.

Table VII shows, for each provider and, under the two settings we considered, the value of k at which the model reached 80% for either of the metrics we used (recall or similarity). It shows that in most cases, the model trained temporally performed at least as well as the one trained randomly, and sometimes significantly better. Only in the recall for GoDaddy Corporate Domains did the temporal model perform worse than the random one, with a fivefold deterioration.

After inspecting the data for GoDaddy Corporate Domains, it appeared that 40% (26 of 65) of the domains in the testing data were related to transformations that either never appeared (69.23%) or only appeared once (30.77%) in the

TABLE VII: Value of k (%) at which the various models achieve 80% under the recall or similarity metric. **Rnd** and **Tmp** respectively stand for the random and temporal settings.

Provider	Recall		Similarity	
	Rnd	Tmp	Rnd	Tmp
CSC Corporate Domains, Inc.	4.3	3.0	0.9	0.6
MarkMonitor Inc.	3.6	2.6	0.8	0.4
Network Solutions, LLC	12.8	12.4	4.6	4.5
GoDaddy Corporate Domains, LLC	8.5	**41.2**	2.7	0.6
Nom-iq Ltd. dba COM LAUDE	7.1	1.6	2.4	0.3
SafeNames Ltd.	5.4	2.4	2.1	0.6

training data. For instance, the training data did not contain any TLDSquatting in the dev, xyz, or app gTLDs, while ten such transformations appeared in the testing data. The absence of relevant data in the training set led the model to rank those transformations poorly, leading to a low recall value. Nonetheless, when we used the similarity metric that measures how similar the transformations predicted are to the observed registrations, the model's performance were much stronger. This suggests that even when the model did not learn specific patterns from the data, it could still predict the relevant high-level classes of transformations.

APPENDIX B
ARTIFACT APPENDIX

We share our code to run many of the experiments described in the paper, including training and evaluating machine learning models to mimic the strategies of online brand protection service providers. Given the potential for abuse by malicious actors, we decided against sharing any data related to which domain names the Fortune 500 companies defensively registered and how much traffic those domains received. The reviewers agreed with that decision as it minimizes the ethical risks related to our study. Thus, instead of the actual data, we generated synthetic data using 50 fictitious companies to demonstrate the functionality and reusability of our artifacts. Below, we describe how to access and use those artifacts.

A. Description & Requirements

1) How to access: You can download the source code and supporting data for our artifacts at https://doi.org/10.5281/zenodo.14188149.

2) Hardware dependencies: While this artifact requires no specific hardware configuration, we recommend a computer with at least 4 GB of RAM and 1 GB of available disk space.

3) Software dependencies: A version of Python greater than 3.6 is required to run our artifacts. We advise creating a virtual environment using the `venv` module or Anaconda. In our local tests, we used Python 3.8 with the `venv` module on Ubuntu.

4) Benchmarks: None

B. Artifact Installation & Configuration

Uncompress the artifact package after downloading it. This creates a directory `defreg-artifacts` from which you can install the library dependencies by running the following commands on Unix-like systems with the `venv` module installed. Navigate to the repository and install the library dependencies by running the following commands on a Unix-like system with the `venv` module installed.

```
$ python3 -m venv .venv
$ source .venv/bin/activate
$ pip install -r requirements.txt
$ python -m nltk.downloader averaged_perceptron_tagger_eng
```

We refer users with a different setup to the `README.md` file for specific instructions on how to set up their environments.

C. Experiment Workflow

In this repository, only experiments E1 and E3 can be executed independently. The other experiments (E2, E4, and E5) are strictly meant to be executed sequentially.

D. Major Claims

- (C1): We can generate transformations across multiple transformation classes. We show this in the experiment (E1), the results of which we reported in Table II.
- (C2): We devised a conservative approach to identify defensively registered domain names using information from CZDS zone files and live WHOIS records. We demonstrate this with experiment (E2) that can be used to obtain the results we presented in Section III-A.
- (C3): We found correlations between some factors and the percentage of domain names a company registers defensively. We replicate this through experiment (E3), whose results we reported in Section III-A and Fig. 4.
- (C4): We trained logistic regression models that accurately predict the preferences of six brand protection service providers. Experiment (E4) showcases those results we discussed in Section III-D1 (see Fig. 11 and 12).
- (C5): Using three measures of the effectiveness of future defensive registrations estimated using passive DNS data, we compared the inferred strategies of six online brand service protection service providers. We show this in experiment (E5), which can be used to obtain results similar to those in Section III-D2, especially Fig. 13.

E. Evaluation

1) Experiment (E1): [Generate transformations] [1 human-minutes + 5 compute-seconds]: This experiment generates all the studied transformations for the Fortune 500 company Capital One used as an example in Table II.

[Preparation] None

[Execution] In your terminal, type the following command:
`$ ./scripts/e1.sh` on Unix-like systems or
`$ ./scripts/e1.bat` on Windows.

[Results] After you execute the code, the script outputs the number of distinct domain names generated by each transformation. It also saves the generated transformations to `results/capitalone-transformations.csv` that contains the transformations. Run the command below to check that all the domains of Table II appear in the list:

```
grep -e "capitaline.com" \
    -e "capital0ne.com" \
    -e "capitanone.com" -e "conef.com" \
```

TABLE VIII: Features extracted from the domain names to train the transformation-specific models for the providers.

Feature	Description (References)	Type (Boolean / Numeric)	Transformation type							
			TLD	Typo	Bit	Hphn.	Hgr.	Brd.	Abbr.	Stk.
Position of change	Whether the transformation happens at the head or tail of the base domain name [76]	B		✓	✓		✓			
Base length	Length of the base domain's e2LD [16]	N	✓	✓	✓	✓	✓	✓	✓	✓
Domain length	Length of the transformation e2LD [57]	N								✓
Relative length	The length relative to the base domain name [57]	N				✓				
Length ratio	Factor of increase/decrease in domain length [56]	N						✓	✓	
Dictionary word	The word is a dictionary word	B	✓					✓	✓	✓
Edit distance	Damerau-Levenshtein distance to base name	N								✓
Substring of base domain	Domain name is contained in base domain name	B								✓
Sound similarity	Phonetic similarity to the base domain name [39], [55]	N		✓	✓	✓				
Fat-finger error	Likely fat-finger error on a QWERTY keyboard [56]	B		✓	✓					
Double characters	No. of repetition w.r.t. to base domain name [57]	N		✓	✓	✓				
Alternating characters	No. of alternating characters w.r.t. to base domain name (e.g., gogle.com) [57]	N		✓	✓	✓				
Vowel operation	Vowel swapping / insertion / deletion [58]	B		✓	✓					
Plural / Singular	A singular word was pluralized or a plural word was singularized [39]	B		✓	✓					
Word break	Two words were split using a dash or a dash splitting two words is removed [39]	B		✓	✓					
Homographic transform	Swapping two ASCII homographs [59]	B		✓	✓		✓			
Leet transformation	Swapping two leet equivalent characters [61]	B		✓	✓					
Visual similarity	The visual similarity of two characters involved in the transformation [62]	N		✓	✓		✓			
SSIM	Structural Similarity Index Measure of two swapped characters [60]	N					✓			

```
-e "capitalonefinancialcorp.com" \
-e "capitalone.net" -e "cof.com" \
-e "capitalwon.com" \
results/capitalone-transformations.csv
```

2) Experiment (E2): [Identify defensive registrations] [1 human-minutes + 2 compute-minutes]: This experiment generates the studied transformations for 50 fictitious companies and identifies the defensively registered domain names based on our methodology, which we applied to synthetic WHOIS and zone file data.

[Preparation] None

[Execution] Using a terminal, run the following command:

`$ ./scripts/e2.sh` on Unix-like systems or

`$ ./scripts/e2.bat` on Windows.

[Results] After execution, the script should report 6,256 defensive registrations identified for 42 companies. It also saves the generated data frame to the `results/domain-names-with-attribution.csv.gz` file.

3) Experiment (E3): [Correlation with defensive registrations] [1 human-minutes + 2 compute-seconds]: this experiment performs a Spearman-ranked correlation analysis on the Fortune 500 companies using multiple features.

[Preparation] None

[Execution] Using a terminal, run the following command:

`$ ./scripts/e3.sh` on Unix-like systems or

`$ ./scripts/e3.bat` on Windows.

[Results] The output file `results/plots/fig4.png` generated by this script should be identical to Fig. 4.

4) Experiment (E4): [Train and evaluate the providers' models] [5 human minutes + 6 compute hours]: This experiment trains various logistic regression models that mimic the strategies of fictional OBP providers in a ten-fold cross-validation process. It also estimates the recall@k and similarity@k metrics for each of the providers. This process is long and can be shortened by 1) reducing the number n of iterations in the cross-validation process and/or 2) increasing the step size of k used for estimating the metrics (set to 0.01 as during our experiments).

[Preparation] None

[Execution] Run the following command to train and evaluate the models of the providers in a 10-fold cross-validation approach.

`$ ./scripts/e4.sh` on Unix-like systems or

`$ ./scripts/e4.bat` on Windows.

[Results] The script ends with generating two plots `fig11.png` and `fig12.png` with a similar appearance to Fig. 11 and 12 of the paper. It also dumps the predictions, the hard metrics, and the metrics@k from the models respectively in the `providers-models-predictions.csv`, the `providers-estimated-metrics.csv` and the `recall-similarity_at_k.csv` files. All the files are saved under the `results/` directory.

5) Experiment (E5): [Compare the providers' strategies using passive DNS data] [1 human-minutes + 35 compute-minutes]: This experiment trains various models for the providers and, using passive DNS data, compares their performance in protecting new customers.

[Preparation] You need a valid VirusTotal API key with enough quota to use the code that collects historical WHOIS records. Since the domain names in our package are fictional, we provided dummy historical WHOIS records to support the code's execution.

[Execution] Run the following command from your terminal with the Python environment activated.

`$ ./scripts/e5.sh` on Unix-like systems or

`$ ./scripts/e5.bat` on Windows.

[Results] The result of this execution is the generation of `fig13.png` in the `results/plots` folder, a graph that mimics Fig. 13 in our paper. The script also dumps the models' predictions to the `predictions-for-new-businesses.csv.gz` file and the computed statistics (Normalized Discounted Cumulative Gain) to the `statistics-for-new-businesses.csv.gz` file.