# Squatspotting: Towards the Systematic Measurement of Typosquatting Techniques

Wei-Shiang Wung*, Calvin Kranig*, Eric Pauley*, Paul Barford*, Mark Crovella†, Joel Sommers‡
*University of Wisconsin-Madison, Madison, WI USA †Boston University, Boston, MA USA ‡Colgate University, Hamilton, NY USA
wwung@wisc.edu, ckranig@wisc.edu, epauley@cs.wisc.edu, pb@cs.wisc.edu, crovella@cs.bu.edu, jsommers@colgate.edu

***Abstract*—**
**Typosquatting—the practice of registering a domain name similar to another, usually well-known, domain name—is typically intended to drive traffic to a website for malicious or profit-driven purposes. In this paper we assess the current state of typosquatting, both broadly (across a wide variety of techniques) and deeply (using an extensive and novel dataset). Our breadth derives from the application of eight different candidate-generation techniques to a selection of the most popular domain names. Our depth derives from probing the resulting name set via a unique corpus comprising over 3.3B Domain Name System (DNS) records. We find that over 2.3M potential typosquatting names have been registered that resolve to an IP address. We then assess those names using a framework focused on identifying the intent of the domain from the perspectives of DNS and webpage clustering. Using the DNS information, HTTP responses, and Google SafeBrowsing, we classify the candidate typosquatting names as resolved to private IP, malicious, defensive, parked, legitimate, or unknown intents. Our findings provide the largest-scale and most-comprehensive perspective to date on typosquatting, exposing potential risks to users. Further, our methodology provides a blueprint for tracking and classifying typosquatting on an ongoing basis.**

***Index Terms*—DNS, network security, typosquatting**

## I. Introduction

Typosquatting — the practice of registering a domain name similar to another, usually well-known domain name — is typically intended to drive traffic to a website for malicious or profit-driven purposes. Typosquatting campaigns can lead users to provide credentials to adversaries, misroute payments, or download malicious software [1], [2]. In more benign cases, these domains can be reserved by grey-hat actors to serve low-quality ads to errant users, or to advertise the domain for acquisition by other parties (including the original site owner). In response to this phenomenon, the largest and most sophisticated organizations often purchase typosquatting domain names proactively to defend against registration by these adversaries.

Although typosquatting has been studied for over two decades, the community still lacks a comprehensive assessment of the techniques used by malicious actors, as well as a large-scale analysis from the perspectives of Domain Name System (DNS) characteristics and webpage appearance. In this paper, we assess the current state of typosquatting, both broadly (across a wide variety of candidate name generation techniques) and deeply (using an extensive and novel dataset). We introduce a broad taxonomy of typosquatting *techniques* (i.e., the techniques used to generate candidate domains for registration by an adversary), representing a superset and expansion of those considered in prior studies. We also introduce an analysis pipeline for evaluating the deployment of candidate names in the DNS to infer the *intent* (i.e. goals and outcomes from the adversary or defender in reserving and serving records under this domain) of the name registration.

To evaluate real-world typosquatting within our taxonomy, we combine *(i)* existing censuses of active DNS measurement [3] and *(ii)* TLS certificate issuance [4] with *(iii)* automated name generation based on our taxonomy. The three data sources comprise $3.3\,\mathrm{B}$ total records, enabling a comprehensive analysis of typosquatting domains. Both active DNS lookup and publicly available DNS censuses can have gaps – false negatives that would reduce the accuracy of our analysis. Hence, we adopt a strategy of combining the active measurement with DNS censuses where possible to ensure maximum accuracy; our strategy also leverages the precollected DNS censuses to reduce the active measurement load, which is important for maintaining a sustainable pipeline. We study popular domains in the Tranco top 1k [5], [6], using the above datasets and taxonomy to determine candidate typosquatting techniques. We then seek to infer the intent of the parties that host content under each domain, and thereby classify the domain. For this, we actively measure these domain names via the DNS itself, via safe browsing APIs, and through direct browser scraping. We develop a novel classification pipeline that integrates a range of features, including DNS domain-IP connectivity and webpage layouts – features not considered in previous typosquatting classification efforts. In so doing, we offer a complete characterization of typosquatting prevalence and motivations.

In total, we find $2.3\,\mathrm{M}$ potential typosquatting domains that have been registered and resolve to an IP address. In analyzing the share of these domains that actually respond to HTTP requests ($608\,\mathrm{k}$), we use perceptual hashing to cluster and compare website content with known ground truth. We find that the majority of these sites ($311\,\mathrm{k}$) are parking services that either serve ads or solicit purchase offers on the domain. The second largest set of domains ($230\,\mathrm{k}$) consists of legitimate services that are not related to the original popular registrable domains. In contrast, purely malicious actions by typosquatting domains (e.g. SafeBrowsing violations) account for only $0.24\,\%$ of all collected domains. This suggests that typosquatting is largely a gray-hat, profit-driven enterprise, rather than a concerted effort

ISBN 978-3-903176-72-0 © 2025 IFIP

by more sophisticated adversaries to exploit clients.

Our findings provide the most comprehensive perspective to date on typosquatting, which exposes potential risks to users. In addition, our methodology provides a blueprint for tracking and classifying typosquatting on an ongoing basis. We anticipate that continued and expanded monitoring of typosquatting will prove a valuable source of data for organizations to prioritize their defensive efforts.

## II. Background & Related Work

*Domain Typosquatting* has been employed since the early days of the DNS and exploded in use during the 1990s and early 2000s [7], [8]. While early studies were fairly narrow in scope, in 2015 Agten *et al.* [2] performed a comprehensive longitudinal study of typosquatted domains. This work considered five common typosquatting models having a Damerau-Levenshtein distance of 1 from a legitimate name, and generated typosquatting domain candidates. Then, the authors manually classified website screenshots to evaluate domains. While our work also adopts website screenshots in the typosquatting classification, we consider a broader scope of typosquatting models and automate the classification pipeline.

In addition to typosquatting that exploits mistyping errors, a number of other typosquatting strategies have emerged. *Bitsquatting* [9], [10], is based on bit errors occurring in computer memory to redirect Internet traffic. *Homograph-squatting* domains [11] contain unicode characters having visual similarity to standard alphabetic characters. We include both bitsquatting and homograph-squatting within the typosquatting techniques of edit distance 1 and 2 in our taxonomy (Table I). *Sound-squatting* is mainly aimed at voice devices and relies upon poor enunciation or errors in machine translation. Nikiforakis *et al.* [12] developed a methodology for generating soundsquatting domains; we use a similar technique in our study as described in §III-C3. *Combo-squatting*, described in [13], refers to the technique of combining a popular domain name with other words or phrases; that work highlighted the role of combo-squatting in various security threats.

Typosquatting has also been considered from the perspective of *intent*. Szurdi *et al.* [1] showed that many less-popular domain names have become typosquatting targets, studied the monetization strategies of typosquatting domains, and proposed potential solutions for the mitigation of typosquatting. Zeng *et al.* [14] reported the distribution of intent for various typosquatting techniques, which they obtained by manual inspection of the homepages of those domains. There are also several studies focusing on specific typosquatting intents, such as phishing [15] and defensive registration [16]. Perhaps the most well-known intent for typosquatting is to generate revenue through ad delivery via a domain parking service [17]. A recent empirical study of this phenomenon utilizes a list of 82 parking services and DNS-based indicators to identify 60M parked domains [18]. We use similar techniques to classify domains as described in subsection V-B. In addition to intent analysis, Khan et al. [19] attempted to quantify the time effects of typosquatting on Internet users.

Our study also develops new tools for studying typosquatting. Previous studies [1], [20], [21] proposed several tools for automatically identifying typosquatting domains. These tools mainly use domain features such as domain length, DNS and Whois records, and Web attributes such as number of redirections, HTML page size and keywords. While we also develop an automated pipeline to classify typosquatting intents, our framework is designed from the perspective of DNS domain-IP connectivity and landing webpage appearance. To the best of our knowledge, domain-IP connectivity has not been considered in prior studies. Webpage content classification has previously been performed by manual classification [2], [14] but seldom integrated into automated classification tools. We develop a classification pipeline integrating these two aspects, and use it to conduct our large-scale typosquatting analysis.

## III. Definitions & Data Sources

Next, we present the techniques we use to generate potential typosquatting domains, definitions used to classify typosquatted domains, and the data sources used in our study. Note that all domain names in the following discussion are at the level of registrable domains (eTLD+1).

### A. Taxonomy of Typosquatting Techniques

Table I depicts the types of name generation techniques we study; in the table, 'domain label' refers to the label in front of the effective top-level domain (eTLD). In some cases, these techniques target the physical act of typing by users (i.e., edit distance-based), while others target users misremembering domains (e.g., TLD-squatting, combo-squatting) or even mis-hearing domains as spoken (i.e., sound-squatting). Although ED1, TLD-, combo-, and sound-squatting have been considered in prior work to some extent, ED2 and hybrid techniques have not been considered previously.

### B. Typosquatting Domain Intent

While prior work has focused on domain parking and malicious typosquatting domains, we show that name generation techniques can be applied for other purposes as well. Here, we describe eight different domain characteristics that can be used to identify either adversarial or benign intent.

- **Domains linked to private IP:** In this case, a DNS A record translates the domain to a private IPv4 address. As a result, access from the public Internet is impossible. This bears on intent because it is likely indicative of the name being used for internal infrastructure purposes without regard to the domain's potential as a typosquatting domain.
- **Malicious domains:** Malicious behaviors of these typosquatting domains include social engineering, accessing malware, or downloading other unwanted software.
- **Defensive domains:** In this case, the domain is managed by the same entity as the authoritative domain and is used to forward traffic back to the authoritative domain.

TABLE I
TAXONOMY OF TYPOSQUATTING TECHNIQUES. TECHNIQUES ARE DIVIDED INTO BASE APPROACHES AND COMBINATIONS OF OTHER APPROACHES.

| Technique | Description | Example |
|---|---|---|
| Edit Distance 1 (ED1) | The Damerau-Levenshtein distance of domain label is 1 | example.com→**e**example.com |
| Edit Distance 2 (ED2) | The Damerau-Levenshtein distance of domain label is 2 | example.com→**e**example**e**.com |
| TLD-squatting | Suffix (eTLD) is replaced but domain label remains the same | example.com→example.**org** |
| Combo-squatting | Single or multiple strings are appended to domain label | example.com→**test**example.com |
| Sound-squatting | Words are replaced by homophones with similar pronunciations | youtube.com→**u**tube.com |
| ED1+TLD-squatting | The edit distance of domain label is 1 and eTLD is replaced | example.com→**e**example.**org** |
| Combo+TLD-squatting | Strings are appended to domain label and eTLD is replaced | example.com→**test**example.**org** |
| Sound+TLD-squatting | Words are replaced by homophones and eTLD is replaced | youtube.com→**u**tube.**net** |

- **Benign domains:** These domains are managed by the same entity as the authoritative domain and are used for specific purposes.
- **Parked domains:** Domains in this type do not provide active services. Instead, they are managed by some domain parking entity and usually present ads or domain-selling information. Owners may be targeting a sale to the original authoritative domain, or simply serving ads to errant page viewers.
- **Unrelated legitimate domains:** These domains host web pages for legitimate purposes by 3rd party entities irrespective of the service's typosquatting potential.
- **Unknown domains with host IP changed:** We place domains in this category if there is an inconsistency between historical DNS data and those seen during our experimental crawls. Because there has been a recent change in DNS we cannot soundly infer domain intent.
- **Unknown domains without HTTP responses:** At the time of browsing the unknown typosquatting domains, they fail to respond to HTTP requests, and no web pages are returned. Without knowing hosted content or related organizations we are unable to identify domain intent.

### C. Data Sources

Our study focuses on typosquatting domains that are similar to domains in the Tranco top 1k list. We set a minimum length restriction to the target authoritative domains for the following two reasons: (1) the registration cost of short typosquatting domains is high, and (2) we observe that typosquatting operations to short authoritative domains often lead to very different and unrelated legitimate domains. To reduce the likelihood of bias caused by short unrelated domains, we only consider domains at least 6 characters in length, resulting in 721 domains. We then collect resolved potential typosquatting domains from Rapid7 Sonar, Certificate Transparency (CT) logs, and via brute-force generation (described below). Note that we exclude domains that do not successfully resolve from our study, as no servers host these domains at the time of data collection. All of the datasets we use were collected in September 2024.

*1) Rapid7 Sonar:* Rapid7's Project Sonar [3] actively collects DNS records for Internet and security research. Starting with scanning all public IPv4 addresses, a list of domain names is collected using various proprietary techniques. Based on the most comprehensive list available, Project Sonar actively requests records of DNS resolution from DNS resolvers. We focus on the dataset of A records for the analysis of resolvable typosquatting domains.

*2) Certificate Transparency Logs:* Certificate Transparency (CT) logs [4] list domain names having issued TLS certificates, which provides a list of candidate domains. Our work collects full CT logs scanned by Censys through September 2024 and filters only resolvable domains, as determined using public resolvers.

*3) Brute Force Domain Generation:* To achieve the most complete corpus of potential typosquatting domains, we generate candidate typosquatting domains by brute force (exhaustive search over all valid possibilities in ASCII character set) and then resolve them through zdns [22]. Due to the unbounded size of the namespace of combo-squatting domains, we only use brute force generation for the other six techniques. We generate potential typosquatting domains with edit distance 1 and edit distance 2 through four operations on the domain labels, including insertion, deletion, substitution, and transpose. To generate TLD-squatting domains, we iterate all valid eTLDs using the Public Suffix List [23] and concatenate them with target domain labels. Sound-squatting domains are generated by replacing words in domain labels with homophones in two steps. First, the domain labels of the authoritative domains are divided into lists of meaningful English words, if possible through ChatGPT. Then, we collect 2,178 homophone pairs from homophone.com, Wikipedia [24], and the conversion of digits to words from 0 to 100. Candidate sound-squatting domains are generated by searching for words in the domain label word lists and replacing them with homophones.

We generate typosquatting domains that incorporate hybrid techniques by combining the domain labels of the previously generated typosquatting domains and all valid eTLDs in the Public Suffix List. However, the size of the set of generated domains will be the product of the namespace sizes of two selected techniques (e.g. there are over 2 trillion potential domains when combining ED1+TLD-squatting), which presents a challenge for both storage and resolution. Hence we choose instead to narrow the namespace while including as many resolved typosquatting domains as possible. For ED1+TLD-squatting, we generate

typosquatting domains with the combination of domain labels from the resolved ED1 typosquatting domains and eTLDs from the resolved TLD-squatting domains. With a similar technique, sound+TLD-squatting domains are generated by combining all sound-squatting domain labels and eTLDs from the resolved TLD-squatting domains.

All A records of the resolved typosquatting domains generated by brute force are collected through zdns and Cloudflare's public DNS resolver. For the generated domains in hybrid typosquatting models, the restricted namespace still leads to a hit rate $> 1\%$ in DNS resolution, showing that our strategy to narrow the domain name space is effective.

## IV. Methodology

We analyze typosquatting domains in two stages: (1) we collect typosquatting domains and verify that they resolve in the DNS (subsection IV-A), and (2) we classify them based on their DNS characteristics and website screenshots (from subsection IV-B to subsection IV-E).

### A. Typosquatting Domain Collection

To perform a comprehensive analysis of typosquatting domains, all eTLD+1 domains and resolved IPv4 addresses from Rapid7 Sonar, CT logs and brute-force generation are considered. Due to the different collection practices of these three data sources, a series of data transformations are required to combine them. Rapid7 Sonar aims to collect DNS records from all known domain names, so its raw data contains both A and CNAME records. We recursively resolve domains associated with CNAME records to the corresponding IPv4 addresses. In addition, to expose DNS configuration characteristics at the eTLD+1 level, fully qualified domain names in the A records of Rapid7 Sonar and CT logs are converted to eTLD+1 domains by removing subdomains. Next, unique pairs of eTLD+1 domains and resolved IPv4 addresses from all A records in these three data sources are utilized in the analysis of typosquatting domains.

Using the 721 authoritative domains of the Tranco top 1k, we collect all potential typosquatting domains by applying the eight typosquatting techniques in subsection III-A. For the combo-related squatting techniques, candidate domains are searched from the existing Rapid7 and CT logs datasets due to the unbounded namespace. For the other six typosquatting techniques, candidate typosquatting domains are collected both by brute force generation and also by a search-based approach applied to the Rapid7 and CT logs. In Rapid7, non-ASCII characters may be included in domain names; additionally, it contains fully qualified domain names (FQDNs) whose eTLD+1 domains have no DNS records. Hence, some typosquatting domains matching the six techniques may exist in the Rapid7 or CT logs that are not discovered by brute force generation. To enhance domain coverage, we consider all potential typosquatting domains from each of the three data sources.

The result is that we identify 2,305,556 candidate typosquatting domains derived from the authoritative target domains. The DNS resolutions of these candidate domains are verified using the Rapid7 dataset as well as via active DNS measurement for domains from brute-force generation and CT logs via zdns in September 2024. This ensures that all the candidate typosquatting domains considered in this work resolve to host IPs from the DNS infrastructure.

### B. DNS Graph Structure

DNS A records not only demonstrate the existence of possible typosquatting domains, but may also indicate how domain names and associated IPv4 addresses relate to one another more broadly. Inspired by [25], we convert the A records of the combined dataset to a bipartite graph with domain names and IPv4 addresses as the two types of vertices. Each edge connecting a domain name and an IPv4 address in the DNS graph represents an existing A record in the DNS infrastructure. By running the decomposition algorithm in [25], the DNS graph is decomposed into millions of fully connected components.

Connected components from the DNS graph provide a context in which to consider domain names and IPv4 addresses that relate to one another. As discussed in [25], it is often the case that if some domain names in a component are known to be parked domains, or are otherwise classified in some way, we can consider other names in the same component to have *similar properties by association*. We make use of connected components to identify related malicious, defensive, and parked typosquatting domains within our classification pipeline. In particular, we identify typosquatting domains as *potentially malicious* if they reside in the same component as known malicious domains. We use this inference to flag those potentially malicious domains for further analysis. Similarly, typosquatting domains that reside within components with known defensive or parked domains are classified as such.

### C. Domain Nameserver Collection

In addition to the A records, the nameservers of the typosquatting domains are also important references from a DNS perspective. Nameservers refer to authoritative DNS servers responsible for managing A records for a list of domains. Many entities such as Google and Amazon AWS managed their top-ranked authoritative and defensive domains with their own nameservers. Prior studies of domain parking [18] have shown that nameserver delegation is also a common approach to parking unused domains. Therefore, the nameservers of these typosquatting domains become additional indicators to identify the intent of the typosquatting domains.

We collect the nameservers of the 2.3M typosquatting domains through Cloudflare's public DNS resolver in October 2024. Recall that all these typosquatting domains have previously been resolved in the DNS and have associated A records from at least one data source. After experimenting with other means of collecting NS records for these domains, we found that using the standard `dig` tool with its `+trace` option to request A records resulted in finding 94% of NS records for the typosquatting domain candidates in the DNS lookup processes. Most of the responses of the remaining 6% of typosquatting

domains indicate that they are no longer resolvable to A records using the public DNS resolver. This may be caused by the nature of the churn in DNS (consistent with the observation in [25]) or by the different collection methods of the data sources.

### D. Website Screenshot Clustering

An additional means for typosquatting domain classification relies on whether a server associated with a typosquatting domain supports HTTP, and if so, on the contents of the landing page for that host. Unrelated legitimate domains typically host active web pages for their own services. However, some servers associated with domain parking may not support the HTTP protocol at all, leading to unresolved parked domains. In addition, based on prior studies [18], [26], parked domains with active HTTP service often host web pages with highly repetitive layouts such as ads and domain-selling information. Thus, support for HTTP along with any web page screenshots can serve as key indicators for typosquatting domain analysis.

Using Selenium and Chromedriver, we automatically browse all typosquatting domains and collect webpage screenshots in Chrome headless mode. For each typosquatting domain, 20 seconds are allotted to load the page before timeout. In the process we collect the server IP addresses, HTTP status codes, and destination URLs. Since the full size of a screenshot image is relatively large, we store the 128-byte perceptual hash values of the screenshots instead [27]. Unlike cryptographic hashing algorithms, perceptual hashing is designed to capture image similarity using fixed-length hash values. All the browsing results are collected in October 2024.

We use webpage screenshots to classify typosquatting domains through image clustering. To identify similar web pages, we apply hierarchical clustering with bit-wise comparison as the metric to measure the distance of two 128-byte screenshot hashes. To discover the non-virtual hosted servers, we set a bit-wise distance threshold = 150 to identify domains hosting visually identical webpages. To identify parked domains with similar website layouts, we found that setting a threshold = 300 performs well to group similar webpage layouts with context differences and small pop-up windows into the same clusters.

### E. Typosquatting Domain Classification

Potential typosquatting domains can be registered for various purposes. To clarify the registered intent or property of candidate typosquatting domains at a large scale, we develop a pipeline to automate the classification processes to the extent possible. Note that typosquatting domains already labeled as one of the categories in the previous steps will no longer be considered in subsequent classification decisions.

*1) Domains linked to private IP address:* By inspecting the resolved IPv4 addresses in the DNS A records, we find and label typosquatting domains linked to private IPs. Since these domains are publicly accessible, the registered intents and usages cannot be further classified.

*2) Malicious domains:* The process of identifying malicious domains consists of two steps. First, we select Google's SafeBrowsing as a reference to identify malicious domains that have been discovered and reported. These domains are labeled as malicious (SafeBrowsing). Second, we explore other potential malicious typosquatting domains in the same DNS connected components via our graph-based analysis as the malicious domains identified by SafeBrowsing. The idea is that for a non-virtual hosted server with known malicious domains, the server provides identical web content regardless of domain names. Hence, if hosts within a connected component are identified as non-virtual hosted through web page screenshots, then we classify the other typosquatting domains in the same connected components into malicious domains (non-VH CC). This technique enables the identification of malicious domains that are not in the existing SafeBrowsing database.

*3) Benign and defensive domains:* Many owners of top-ranked authoritative domains manage domain names and services via server clusters and DNS nameservers. To identify benign and defensive connected components of typosquatting domains, we consider that two conditions need to be satisfied: (1) the Autonomous System (AS) organizations of the IP addresses from the connected components of authoritative and typosquatting domains should match, and (2) the authoritative and all typosquatting domains in the connected components should be managed by nameservers with the same registered domains. Next, we check whether browsing the typosquatting domains will be redirected to similar authoritative domains. If not, it indicates that the benign domains are registered for other purposes; otherwise, these domains forward traffic to the authoritative domains and are likely to be registered for defensive purposes.

*4) Parked domains:* Domain owners have monetization incentives to park unused domains, and delegate authority to parking services through DNS configuration. From the list of 82 parking services released by [18], the DNS indicators can be categorized into three types — NS, A/AAAA, and CNAME records. Domains parked with a delegation of nameservers from some parking services can be identified by the nameserver information collected from subsection IV-C. Meanwhile, parked domains configured by A or CNAME records can be identified through connected components in subsection IV-B. These parked typosquatting domains identified by the DNS characteristics are labeled as parked (known parking services).

However, the parking service list and the corresponding DNS indicators was published over two years ago. We hypothesize that there may be previously unidentified indicators of parking services, and thus part of the parked domains are not identified using purely DNS-related characteristics. Since the browsing results of parked domains show that the resolved web pages are highly duplicated and irrelevant to the typosquatting domains, unknown parked typosquatting domains may also have similar webpage layouts. Through the screenshot clustering of all successfully resolved webpages, unknown typosquatting domains with webpages similar to known parked domains are

filtered and labeled as parked (by a similar screenshot).

In addition to known parking services, some domain registrars also provide domain parking before owners decide the usage of their domain assets. To further identify other parked typosquatting domains not in use, we utilize the characteristics of duplicate or near-duplicate web pages. We therefore manually filter the web page screenshots irrelevant to the typosquatting domains from ∼300 large clusters with more than 50 typosquatting domains. The parked domains identified by this strategy are labeled parked (by large image clusters).

*5) Unrelated legitimate domains:* Based on the destination URLs from the browsing results, the remaining unknown domains with HTTP responses and webpages are classified into unrelated legitimate domains or typosquatting domains redirected to similar authoritative domains. The former are legitimate domains running unrelated services yet have domain names similar to top-ranked authoritative domains (e.g., amazon.com and amazon-rainforest-tours.org). The latter are potentially defensive domains. However, authoritative domain entities may manage these domains with different server clusters or nameservers.

*6) Unknown domains linked to a new IP address:* If the typosquatting domains are not classified into one of the above categories and the server IPs from the browsing results are not in the combined A record dataset and not managed by the same AS organization, they are labeled as unknown domains linked to a new IP. The characteristic typically highlights the instability of the typosquatting domains and the frequent changes to the DNS configurations.

*7) Unknown domains without HTTP responses:* For those unknown typosquatting domains failing to respond to HTTP requests, we do not have sufficient information to understand their registration intent. Therefore, they are labeled as unknown (no HTTP response).

## V. Results

We now present the results of analyses of typosquatting domain collection and classification based on intent. We report on the distribution of collected typosquatting domains and present a summary visualization of classification results. Additionally, we examine the distribution of typosquatting domains across different techniques, analyze parked and unknown domains without HTTP responses, and identify key parking services and nameservers. We conclude with a validation of the classification pipeline.

### A. Typosquatting Domains from the Eight Techniques

Considering the 721 authoritative domains from the Tranco top 1k, we identified 2,305,556 distinct typosquatting domains from our three data sources. Table II presents the distribution of typosquatting domains generated by the eight typosquatting techniques. From techniques with single type (1)-(5), combo-squatting (4) generates the most typosquatting domains followed by edit distance 2 (2), due to the large namespaces of these methods. On the other hand, sound-squatting (5) generates the least typosquatting domains since not all top-ranked domains include homophone words for replacement. For typosquatting domains generated by hybrid techniques (6)-(8), the namespace of typosquatting domains significantly expands due to the factor of suffix replacement. Therefore, typosquatting domains in (6)-(8) contribute to over 75% of all 2.3M typosquatting domains.

TABLE II
Distribution of Domains from Typosquatting Techniques

| Domain Typosquatting Technique | Resolved Domains | Resolution Rate | Unique Domains |
|---|---|---|---|
| (1) ED1 | 46,511 | 2.02% | 38,601 |
| (2) ED2 | 315,207 | 13.67% | 288,582 |
| (3) TLD-squatting | 72,072 | 3.13% | 72,072 |
| (4) Combo-squatting | 342,419 | 14.85% | 307,920 |
| (5) Sound-squatting | 52 | 2.26e-3% | 15 |
| (6) ED1+TLD-squatting | 1,135,578 | 49.25% | 948,226 |
| (7) Combo+TLD-squatting | 612,227 | 26.55% | 425,285 |
| (8) Sound+TLD-squatting | 3,439 | 0.15% | 2,961 |

Some typosquatting domains can be generated by multiple techniques. For example, character insertion at the beginning or the end of a domain label satisfies both (1) edit distance 1 and (4) combo-squatting. Part of (5) sound-squatting domains can also be generated by character insertion, removal, or replacement with (1) edit distance 1 or (2) edit distance 2. Similar phenomenon also applies to typosquatting domains generated by hybrid techniques. To quantify the phenomenon, Table II presents a column of the unique typosquatting domains generated by each technique. Note that even if a domain may match multiple typosquatting techniques, starting from the analysis in subsection V-B, each domain is assigned with only one label based on the priority in Table II.

### B. Domain Classification

Through the classification pipeline in subsection IV-E, all 2.3M typosquatting domains are classified into one of the categories according to their intents. The classification results are presented in Figure 1. Through the classification process, we have the following observations:

*a) Private IPs:* Among 374k (16.25%) typosquatting domains resolved to private IP addresses from the DNS A records, the nameservers of 42k (1.82%) domains indicate that they are managed by a parking service, ParkingCrew. There are other parking services managing very few typosquatting domains linked to private IPs as well.

*b) Malicious Domains:* Over 96% of 3.7k (0.16%) malicious typosquatting domains are identified by SafeBrowsing due to social engineering issues. This indicates that typosquatting domains have been commonly used to steal sensitive personal information.

*c) Virtual Hosting:* Through website screenshot clustering, 107 connected components with known malicious domains are identified as non-virtual hosted clusters. As a result, an additional 2k (0.09%) typosquatting domains in these connected components are also labeled as malicious, and most of them are associated with social engineering.

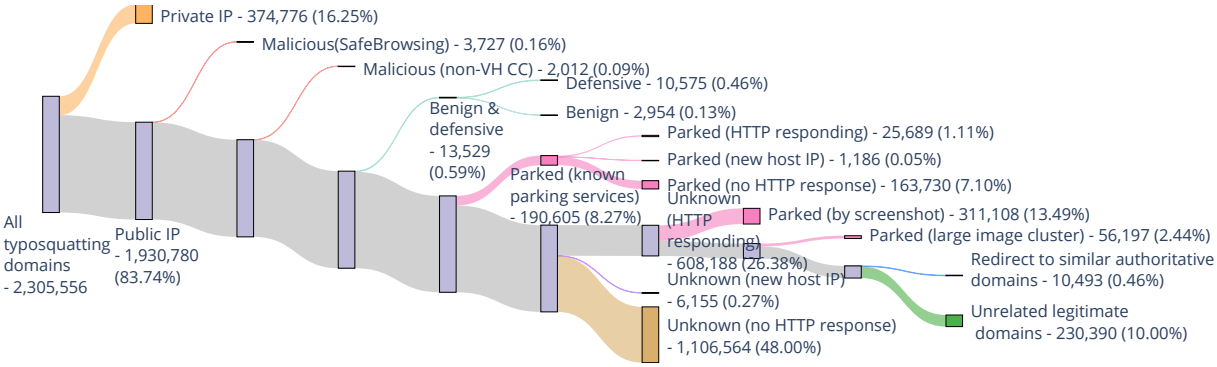


Fig. 1. The Sankey Diagram of Typosquatting Domain Classification

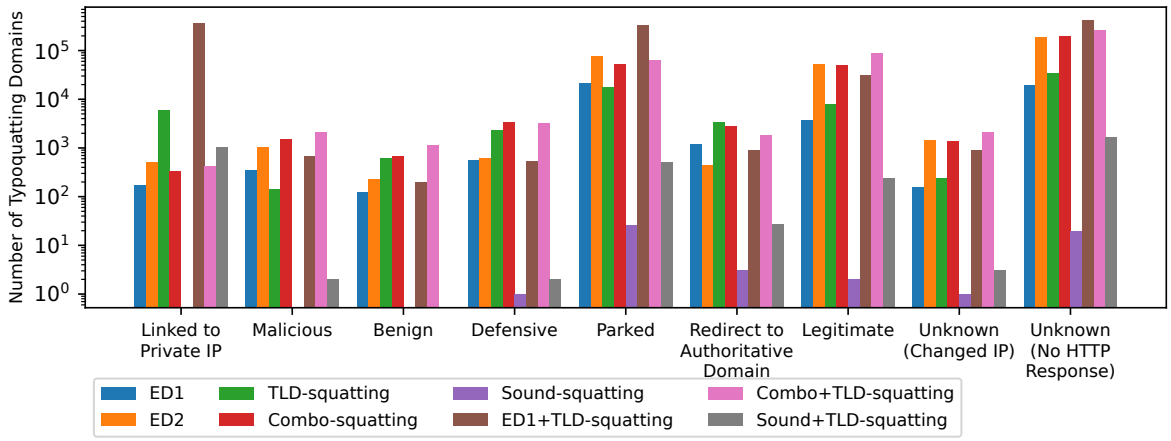


Fig. 2. Intent Distribution of 8 Types of Typosquatting Domains

*d) Parking Services:* From the parked domains identified by known parking services, over 85.9% (163k out of 191k) domains do not resolve to webpages, indicating that only a small portion of these domains actively generate revenue from pages with ads or domain selling information.

*e) Unresolved Domains:* Over 1.1M (48.00%) typosquatting domains do not resolve to any web pages and cannot be classified based on the DNS characteristics. Because of the unknown registration intents for approximately half of the typosquatting domains, we will further analyze these unknown typosquatting domains as well as the parked domains without HTTP responses in subsection V-C.

*f) New Parking Domains:* About 311k (13.49%) domains are identified as parked because their website layouts are similar to known parked domains. New infrastructures or settings are likely to be used to manage these parked domains, but are not included in the list collected by prior researchers. In addition to specialized parking services, we observe that many domain registrars also provide domain parking from parked domains identified by large image clusters. These domains are typically newly registered but not yet in use. When browsing these typosquatting domains, the pages sometimes show that this domain has been registered from the service and forward visitors to search for other available domains.

*g) Defense Registrations:* From the typosquatting domains redirected to similar authoritative domains, we observe that some domains are actually defensive but their IP AS organization and nameservers are mismatched with the authoritative domains, and thus cannot be identified by the classification pipeline in subsubsection IV-E3. For example, T-Mobile is a brand of Deutsche Telekom, but their domains may be deployed on IP clusters with different AS organization titles. This indicates that our identification process of benign and defensive domains is conservative and unable to identify defensive domains with complicated deployment strategies.

Another perspective of the classification results is the intents of typosquatting domains generated by different techniques. From Figure 2, we observe that typosquatting domains generated from several techniques are more concentrated in certain categories. From the 375k typosquatting domains linked to private IPs, over 95% typosquatting domains are generated with ED1+TLD-squatting, followed by TLD-squatting domains. For domains with malicious purposes, typosquatting domains generated by combo-squatting-related techniques are the most common cases. This shows that instead of pretending to be the authoritative domains, typosquatting domains that look like they are being registered for special services by the authoritative entities are preferable for malicious purposes. On the other hand, for both benign and defensive domains

TABLE III
TOP 10 PARKING SERVICES OF THE MOST TYPOSQUATTING DOMAINS WITHOUT HTTP RESPONSES

| Parking Service | # Parked Domains without HTTP Responses | Percentage of the Parked Domains Managed by the Service |
|---|---|---|
| (1) Sedo | 23,952 | 94.88% |
| (2) GoDaddy CashParking | 23,060 | 99.85% |
| (3) dan.com | 19,872 | 96.79% |
| (4) ParkingCrew | 16,218 | 99.56% |
| (5) BODIS | 14,969 | 80.32% |
| (6) Afternic | 12,027 | 99.93% |
| (7) Unknown survey-smiles.com | 9,801 | 97.27% |
| (8) Skenzo | 9,467 | 79.22% |
| (9) ParkLogic | 8,936 | 85.36% |
| (10) Above | 7,334 | 54.57% |

TABLE IV
TOP 10 NAMESERVER ETLD+1 DOMAINS OF THE MOST UNKNOWN TYPOSQUATTING DOMAINS WITHOUT HTTP RESPONSES

| Nameserver eTLD+1 Domains | # Unknown Domains without HTTP Responses | Percentage of the Typosquatting Domains Managed by the Nameserver |
|---|---|---|
| (1) domaincontrol.com | 221,378 | 87.20% |
| (2) dns.ws | 149,511 | 78.15% |
| (3) cloudflare.com | 68,026 | 62.58% |
| (4) nic.vg | 41,854 | 100% |
| (5) lanic.net.la | 41,739 | 100% |
| (6) registrar-servers.com | 26,483 | 80.76% |
| (7) namebrightdns.com | 24,695 | 95.47% |
| (8) markmonitor.com | 21,344 | 79.32% |
| (9) dnsnode.net | 18,029 | 99.99% |
| (10) googledomains.com | 16,930 | 70.72% |

registered by authoritative entities, combo-squatting-related and TLD-squatting are the top 3 typosquatting techniques covering most domains while the numbers of edit-distance-based typosquatting domains are relatively small. For the parked domains, while a large number of typosquatting domains from each typosquatting technique are labeled and classified into this category, ED1+TLD-squatting covers the most typosquatting domains in the parking category. Among all the intent categories, one thing in common is that typosquatting domains generated by combining multiple techniques are popular and should be considered in typosquatting domain analysis.

### C. Domains without HTTP Responses

From the classification results in Figure 1, over 163k parked domains and 1.1M unknown domains do not have successful HTTP responses. In this section, we drill down on the details of these typosquatting domains and analyze them from different perspectives. For parked domains managed by known parking services, while hosting webpages that show ads is a common monetization strategy, over 80% of the parked domains identified by the DNS indicators of known parking services do not respond to HTTP requests. To understand the phenomenon, Table III lists the top 10 parking services of the most typosquatting domains without HTTP responses, which cover 89% of the parked domains without HTTP responses. In addition, the table presents the percentage of parked domains without HTTP responses over the total typosquatting domains managed by each service. Among these parking services, many have more than 90% parked typosquatting domains that do not respond to HTTP requests, indicating that only a small portion of the typosquatting domains are hosting web pages and generating revenue for these parking services.

The unknown typosquatting domains without HTTP responses, on the other hand, account for 48% (1.1M out of 2.3M) yet have even less information associated with them. DNS characteristics are the only attributes that can be used to analyze these unknown domains. Therefore, except for the 31k domains without nameserver information in October 2024, the other unknown domains are classified based on the nameservers managing the corresponding DNS A records. Table IV presents the top 10 eTLD+1 domains of the nameservers associated with the most unknown typosquatting domains without HTTP responses, which cover $\sim$56% of the unknown typosquatting domains. Similarly, the last column of the table shows the percentage of unknown typosquatting domains over the total typosquatting domains managed by each nameserver. The fractions of unknown domains without HTTP responses in these nameservers are relatively high, some even achieve 100%. According to WHOIS and Google search information, "domaincontrol.com" is the nameserver used by GoDaddy (an Internet domain registry), and many entities in the top 10 nameserver eTLD+1 domains provide domain registration as part of their business operations. It is likely that these typosquatting domains are just registered with these services but are not yet in use. Note that besides domain registration, Cloudflare, MarkMonitor and Google provide customers with website hosting services as well. A considerable portion of typosquatting domains managed by their nameservers host active web pages and respond to HTTP requests, and thus the percentages of unknown domains without HTTP responses are not as high as the others.

### D. Method Validation

To validate the typosquatting analysis in this study, the techniques to collect candidate domains and classify domain intents need to be verified. Our methods of brute force domain generation pass a careful code review, and the $2.3\,\mathrm{M}$ potential typosquatting domains selected from the three datasets are verified by manually inspecting random samples. All the A records collected from public DNS resolvers are checked to satisfy the RFC standards through an automated testing pipeline. In addition, the connected components utilized in identified domains with similar intents by association are verified to be consistent with the original DNS datasets, and the image-

based clustering approach for landing webpages is verified by random manual sampling. Domain nameserver is adopted by our classification pipeline instead of WHOIS records because WHOIS records of less than 50% of candidate typosquatting domains are found in the database, and the information of many WHOIS records are private. On the other hand, DNS domain nameservers are public data and can be widely used to identify the intents of typosquatting domains.

The domain accessibility and threat extents determine the step orders in the classification pipeline. For example, for candidate domains connecting to private IPs and with nameservers managed by parking services, these domains are classified as *linked to private IP* because they cannot be publicly accessed when we collect. For typosquatting domains matching both the criteria of malicious and parked domains, since the threat extent of malicious domains is more severe, the malicious domains should be filtered and labeled before the following analysis. From our results, benign and defensive typosquatting domains do not overlap with domains in the malicious or parked categories. Hence, we consider the sequence of steps in the classification reasonable and feasible.

## VI. Conclusion

Through a broad analysis of typosquatting domain selection techniques and associated content hosting, we demonstrate insights on the benign and malicious intents behind typosquatting registrations. Typosquatting domains are used for malicious content hosting and profit-driven parking, but also by organizations to defend against malicious actors. In other cases, the relationship to a popular domain is purely coincidental, a property that can only be inferred through the large-scale site analysis we've proposed here. We anticipate that ongoing monitoring of typosquatting techniques and properties will inform organizational defensive practices.

## Acknowledgements

The authors thank the Rapid7 Research Team for providing data and feedback on this project. This material is based upon work supported by the National Science Foundation under Grant No. CNS2312709, CNS2312710, CNS-2312711, CNS-2319367, CNS-2319368, and CNS-2319369. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

[1] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich, "The long {"Taile"} of typosquatting domain names," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 191–206.

[2] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven months' worth of mistakes: A longitudinal study of typosquatting abuse." in *NDSS*, 2015.

[3] "Rapid7 Project Sonar," https://www.rapid7.com/research/project-sonar/, 2021.

[4] B. Laurie, "Certificate transparency: Public, verifiable, append-only logs," *Queue*, vol. 12, no. 8, p. 10–19, aug 2014. [Online]. Available: https://doi.org/10.1145/2668152.2668154

[5] (2024) Tranco top site ranking list. [Online]. Available: https://tranco-list.eu/

[6] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," *arXiv preprint arXiv:1806.01156*, 2018.

[7] J. Golinveaux, "What's in a domain name: Is cybersquatting trademark dilution," *USFL Rev.*, vol. 33, p. 641, 1998.

[8] B. Edelman, "Large-scale registration of domains with typographical errors," *2003a. Harvard University.,"Domain Name Typosquatter Still Generating Millions."," 2003b. Harvard University*, 2003.

[9] A. Dinaburg, "Bitsquatting: Dns hijacking without exploitation," *Proceedings of BlackHat Security*, 2011.

[10] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting: Exploiting bit-flips for fun, or profit?" in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 989–998.

[11] E. Gabrilovich and A. Gontmakher, "The homograph attack," *Commun. ACM*, vol. 45, no. 2, p. 128, Feb. 2002. [Online]. Available: https://doi.org/10.1145/503124.503156

[12] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Information Security: 17th International Conference, ISC 2014, Hong Kong, China, October 12-14, 2014. Proceedings 17*. Springer, 2014, pp. 291–308.

[13] 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 *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 569–586.

[14] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A comprehensive measurement study of domain-squatting abuse," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[15] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 429–442.

[16] B. C. Benjamin, J. Bayer, S. Fernandez, A. Duda, and M. Korczyński, "Shielding brands: An in-depth analysis of defensive domain registration practices against cyber-squatting," in *2024 8th Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2024, pp. 1–11.

[17] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang, "Understanding the dark side of domain parking," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 207–222.

[18] J. Zirngibl, S. Deusch, P. Sattler, J. Aulbach, G. Carle, and M. Jonker, "Domain parking: Largely present, rarely considered!" in *TMA*, 2022.

[19] M. T. Khan, X. Huo, Z. Li, and C. Kanich, "Every second counts: Quantifying the negative externalities of cybercrime via typosquatting," in *2015 IEEE Symposium on Security and Privacy*. IEEE, 2015, pp. 135–150.

[20] A. Banerjee, M. S. Rahman, and M. Faloutsos, "Sut: Quantifying and mitigating url typosquatting," *Computer Networks*, vol. 55, no. 13, pp. 3001–3014, 2011.

[21] 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 and Active Measurement: 21st International Conference, PAM 2020, Eugene, Oregon, USA, March 30–31, 2020, Proceedings 21*. Springer, 2020, pp. 141–157.

[22] L. Izhikevich, G. Akiwate, B. Berger, S. Drakontaidis, A. Ascheman, P. Pearce, D. Adrian, and Z. Durumeric, "Zdns: a fast dns toolkit for internet measurement," in *Proceedings of the 22nd ACM Internet Measurement Conference*, 2022, pp. 33–43.

[23] (2024) Public suffix list. [Online]. Available: https://publicsuffix.org/

[24] Wikipedia contributors, "Lists of common misspellings/homophones — wikipedia, the free encyclopedia," http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/Homophones, 2024, accessed: 2024-10-29.

[25] A. Anderson, A. S. Mondal, P. Barford, M. Crovella, and J. Sommers, "An elemental decomposition of dns name-to-ip graphs," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, 2024, pp. 1661–1670.

[26] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking sensors: Analyzing and detecting parked domains." in *NDSS*. Citeseer, 2015.

[27] J. Buchner, "imagehash: A python perceptual image hashing module," https://github.com/JohannesBuchner/imagehash, 2024, accessed: 2024-11-15.