

The Landscape of Domain Name Typosquatting: Techniques and Countermeasures

Jeffrey Spaulding
SUNY Buffalo
Buffalo, NY, USA
Email: jjspauld@buffalo.edu

Shambhu Upadhyaya
SUNY Buffalo
Buffalo, NY, USA
Email: shambhu@buffalo.edu

Aziz Mohaisen
SUNY Buffalo
Buffalo, NY, USA
Email: mohaisen@buffalo.edu

Abstract—With more than 294 million registered domain names as of late 2015, the domain name ecosystem has evolved to become a cornerstone for the operation of the Internet. Domain names today serve everyone, from individuals for their online presence to big brands for their business operations. Such ecosystem that facilitated legitimate business and personal uses has also fostered “creative” cases of misuse, including phishing, spam, hit and traffic stealing, online scams, among others. As a first step towards this misuse, the registration of a legitimately-looking domain is often required. For that, domain typosquatting provides a great avenue to cybercriminals to conduct their crimes.

In this paper, we review the landscape of domain name typosquatting, highlighting models and advanced techniques for typosquatted domain names generation, models for their monetization, and the existing literature on countermeasures. We further highlight potential fruitful directions on technical countermeasures that are lacking in the literature.

Keywords. Domain Names, Typosquatting, Defenses.

I. INTRODUCTION

Ever since the process of the domain name registration began in the 1990’s, cybercriminals have seized the opportunity to profit on the backs of others by misusing such process in so many ways [24], [33], [31]. As Internet commerce quickly rose and more companies began registering domain names to get a foothold on the action, certain individuals realized that they could preemptively register these domain names on a first-come first-serve basis. These so called “cybersquatters” would purchase domain names in the hopes of selling them back to companies and trademark owners for a substantial profit. As these popular domain names attracted more users to their websites, it was not long before cybercriminals recognized that people often made mistakes when typing URLs into their browsers—thus sparking a new form of domain name exploitation called *typosquatting*.

In this paper, we survey the landscape of *typosquatting*, which is the deliberate registration of a domain name that uses typographical variants of other target domain names. Typically, these variant domain names are generated in such a way as to exploit common typographical errors made by users that manually type URLs into web browsers. For example, the popular social networking site Facebook was the target of several typosquatters who registered domain names such as www.fagebook.com and www.facewbook.com [29]. Unfortunately for these typosquatters, they were ordered to transfer

over their domain names and pay Facebook up to \$1.34 million in damages.

As we will discuss in subsequent sections, other forms of domain squatting have emerged that not only uses common typographical mistakes, but employs the use of (among others): visually-similar letters [18], similar-sounding words [26] or the exploitation of hardware errors that store domain names [28]. In addition, it has been shown that not only are the most popular domain names targeted by typosquatters, the “long tail” of the popularity distribution has also come under their sights as potential targets for exploitation [30]. By providing a comprehensive treatment of typosquatting, we hope that this paper will catapult research on mitigating this problem.

Organization. In §II we review the anatomy of typosquatting. In §III we review the monetization techniques of typosquatting. In §IV we review the countermeasures to typosquatting. In §V we provide concluding remarks and open directions.

II. TYPOSQUATTING ANATOMY

While typosquatting as a phenomenon is perhaps known for many years, the term itself has been in use for almost two decades. Several studies have been conducted to understand models of typosquatting, including advanced techniques and squatted domains features. In the following, we briefly review the historical background of typosquatting, and follow it by a technical anatomy of models, techniques and features.

A. Historical Background

The term *typosquatter* may have been coined as far back as 1998 by R. C. Cumbow in The New York Law Journal (NYLJ) [17], who was one of the first to write about this new trend of cybersquatting. One of the first large-scale studies on typosquatting was conducted in 2003 by Edelman [15], who located more than 8,800 registered domains that were minor typographical variations of popular domain names. Surprisingly, most of these domain names were traced back to one individual, John Zuccarini, who often redirected users to sexually-explicit content and even used nefarious tactics to “mousetrap” these users from leaving these sites (*e.g.* blocking the ordinary operation of a browser’s Back and Close commands). Some of these typosquatted domain names went so far as to target websites frequently visited by children, such

as `disenystore.com` (a typo on `disneystore.com`) which redirected to a website with sexually-explicit content.

B. Identifying Typo Domains: Models

Prior experiments conducted on the subject of typosquatting typically began their data collection phase by first identifying a set of domain names and then generating a list of possible typo variations on those domain names. Often these experiments used a subset of the top-ranking domain names according to some domain ranking websites, such as Alexa. The rationale of using such domains is that typosquatters will naturally target the most popular domain names to increase the chances of obtaining unsuspecting visitors. Table I summarizes these several approaches of which authoritative domains they studied, the number of possible typosquatted domains they generated, and what percentage of them were active (*i.e.* resolved to an IP address hosting a website). In the following section, we describe the models that generated typos variations of an authoritative domain.

Typo-Generation Models. One of the first and widely cited approaches in this area was introduced by Wang *et al.* [32] where given a target domain (e.g. `www.example.com`), the following five typo-generation models are commonly used:

- 1) **Missing-dot typos:** this typo happens when the dot following “www” is forgotten, e.g., `wwwexample.com`
- 2) **Character-omission typos:** this typo happens when one character in the original domain name is omitted, e.g., `www.exmple.com`
- 3) **Character-permutation typos:** this typo happens when two consecutive characters are swapped in the original domain name, e.g., `www.examlpe.com`
- 4) **Character-substitution typos:** this typo happens when characters are replaced in the original domain name by their adjacent ones on a specific keyboard layout, e.g., `www.ezample.com`, where “x” was replaced by the QWERTY-adjacent character “z”.
- 5) **Character-duplication typos:** this typo happens when characters are mistakenly typed twice (where they appear once in the original domain name), e.g., `www.exaample.com`

While this previous study presented the first attempt to systematically understand techniques for typosquatting that are most prevalent based on certain usage aspects, later studies looked at exhaustively generating typo domains using other methods. For example, a similar approach in 2008 by Banerjee *et al.* [9] suggested the following methods for generating typosquatted domains:

- 6) **1-mod-inplace:** this typo happens when the typosquatter substitutes a character in the original domain name with all possible alphabet letters.
- 7) **1-mod-deflate:** this typo happens when a typosquatter removes one character from the original domain name (or URL)—and unlike [32] where a specific character is considered (e.g., dot), this work systematically considers all possible characters as candidates.

- 8) **1-mod-inflate:** this typo happens when a typosquatter increases the length of a domain name (or URL) by one character. Unlike in [32] characters are added based on distance (e.g., using a keyboard layout), this work considers all characters as potential candidates.

Certain aspects of the techniques proposed in [9] can be viewed as generalization of the techniques proposed in [32]. For example, rather than substituting adjacent characters on a keyboard as shown by Wang *et al.*'s fourth model, Banerjee *et al.* substituted all possible alphabet characters when generating typo domains. In addition, they also experimented with two and three character modifications for their **inplace**, **inflate** and **deflate** schemes thereby generating roughly three million possible typo domain names starting with a corpus of 900 original domain names.

After probing for the existence of a possible typo domain, Banerjee *et al.* observed that approximately 99% of the “phony” typosquatted sites they identified utilized a one-character modification of the popular domain names they targeted. Essentially, these are domain names that have a Damerau-Levenshtein distance [22] of one from the domains they target. The Damerau-Levenshtein distance is the minimum number of operations needed to transform one string into another, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters (a generalization of Hamming distance).

C. Advanced Squatting Techniques

While the two representative studies discussed in §II-B present examples of systematic typosquatting techniques, other techniques that exploit visual, hardware, and sound similarities have been explored as well. In the following, we review those techniques and their use.

1) **Homograph Attacks:** Per Holgers *et al.* [18], the homograph attack relies on the visual similarity of letters or strings that might be confused with one another. For example, an attacker can exploit the fact that the lower-case letter ‘L’ (l) is visually confusable with the upper-case letter ‘i’ (I) in sans-serif fonts and register `www.paypai.com` which targets the popular payment site PayPal. The end result, in sans-serif font, looks very similar to the original: `www.paypal.com` vs `www.paypai.com`. Alarming as it may seem, the measurement results of Holgers *et al.* shows that these homograph attacks are rare and not severe in nature. However, these types of attacks may continue to be an attractive choice for would-be cyber-criminals since it can fool most users—as demonstrated in the user study “Why Phishing Works” by Dhamija *et al.* [13], where 90.9% of their participants were fooled by such an attack. In that particular case, the researchers generated a phishing website that was an exact replica of the Bank of the West homepage that was hosted at `www.bankofthevest.com`, with two “v”s instead of a “w” in the domain name.

2) **Bitsquatting:** This unique approach to domain squatting was introduced in 2011 by Artem Dinaburg at the BlackHat Security Conference. This technique relies on random bit-errors to redirect connections intended for popular domains

TABLE I

SUMMARY OF TYPO DOMAIN IDENTIFICATION APPROACHES. †www.MillerSmiles.co.uk is one of the internet’s leading anti-phishing sites, maintaining a massive archive of phishing and identity theft email scams.

Approach	Authoritative Domains	Typo Model(s)	Typo Domains Generated	Active Typo Domains
Wang 2006 [32]	Alexa Top 10,000 Alexa Top 30 MillerSmiles† Top 30 Top 50 Children’s Sites	(1) Missing-Dot (1-5) (1-5) (1-5)	10,000 3,136 3,780 7,094	51% (5,094) 71%(2,233) 42%(1,596) 38%(2,685)
Keats 2007 [19]	Top 2,771 (<i>Various Sources</i>)	(1-5)	1,920,256	7% (127,381)
McAfee Labs 2008 [16]	Top 2,000 (<i>Unknown Source</i>)	<i>Unknown</i>	<i>Unknown</i>	80,000
Banerjee 2008 [9]	Top 900 (<i>Various Sources</i>)	(6-8)	~3 million	35%
Moore 2010 [25]	Alexa Top 3,264	(1-5)	1,910,738	~49%(938,000)
Szurdi 2014 [30]	Alexa Top 1 million	(1-5)	~4.7 million	~20%
Agten 2015 [7]	Alexa Top 500	(1-5)	28,179	61% (17,172)

[14]. To test this theory, Dinaburg conducted an experiment and registered 30 bitsquatted versions of popular domains (e.g. www.microsoft.com) and logged all HTTP requests. Much to his surprise, there were a total of 52,317 bitsquat requests from 12,949 unique IP addresses over an eight-month period. Nikiforakis *et al.* [28] studied Dinaburg’s findings further and conducted one of the first large-scale analysis of the bitsquatting phenomenon. Their results show that new bitsquatting domains are registered daily and that these attackers monetize their domains through the use of ads, abuse of affiliate programs and even malware installations and distribution. While typosquatting relies on humans to make mistakes, bitsquatting on the other hand relies on computers (hardware) to make mistakes.

3) *Soundsquatting*: Discovered by Nikiforakis *et al.* [26] while researching domain squatting, *soundsquatting* takes advantage of the similarity of words with regard to sound and user confusion on which word represents the desired concept. Unlike typosquatting, soundsquatting does not rely on the typographical mistakes made by users—it is based on *homophones*, which are two words that sound the same but spelled differently (e.g. “ate” and “eight”). To verify how much this soundsquatting technique is used in the wild, Nikiforakis *et al.* developed a tool to generate possible soundsquatted domains from a list of target domains. Using the Alexa top 10,000 sites, they were able to generate 8,476 soundsquatted domains where 1,823 (21.5%) of those were already registered. The results presented in [26] indeed show that soundsquatting is a viable threat that should be taken into account when defending against domain squatters.

4) *Typosquatting Cross-site Scripting (TXSS)*: In a study conducted by Nikiforakis *et al.* [27] that examined malicious JavaScript inclusions, they identified a new type of vulnerability that occurs when a web developer mistypes the address of a JavaScript library in their HTML pages or JavaScript code. This simple mistake allows an attacker to register the mistyped domain and easily compromise the site that includes the script. To further explore the impact of this type of attack, the researchers registered a typo variation of a popular JavaScript inclusion domain (googlesyndication.com vs. googlesyndicatio.com) and observed its traffic:

163,188 unique visitors over the course of 15 days. Nikiforakis *et al.* argue that the damage of TXSS is much greater than that of typosquatting, since every user visiting the page containing the typo will be exposed to malicious code hosted on the attacker’s site.

D. Features of Typosquatted Domains

In the following, we review features of typosquatted domain names as confirmed by measurements and their evolution over time, including length of domain names (§II-D1), popularity of domain names (§II-D2), popularity of top-level domain (TLD) (§II-D3), and domain landing behavior (§II-D4).

1) *Domain Name Length*: One of the features of domain names investigated for its correlation with typosquatting is their length. For example, while investigating if domain name length affects the chances of being typosquatted, Banerjee *et al.* [9] observed that more than 10% of all possible “phony” typosquatted sites registered on the Internet have URLs with less than 10 characters. This fulfills their expectation that typosquatters target domains with shorter names, since popular sites often have short names.

However, in a contradictory study by Moore and Edelman [25], the authors show that no matter the length of the popular domain, typo domains within the Damerau-Levenshtein distance of one or adjacent-keyboard distance of one from popular domains were overwhelmingly confirmed as typosquatted. Naturally, we can expect that as the length of domain names increases the probability of it being typosquatted increases since the number of possible typo variations increases. This concept is solidified in the results of the 2015 study by Agten *et al.* [7], which concluded that typosquatters have started targeting longer authoritative domains in the years following 2009, due to the fact that most short typosquatting domains were already in use.

2) *Domain Name Popularity*: Another feature of domains names that has been investigated for its correlation with typosquatting is their popularity. It is naturally expected that typosquatters will target the most popular domain names to maximize the return on their investment (e.g., the number of visits by unsuspecting users). The results of Banerjee *et al.* [9] initially suggest that the percentage of active typosquatting

domains for a given authoritative domain decreases significantly with the declining popularity. This is in contrast to the results presented by Szurdi *et al.* [30], who performed a comprehensive study of typosquatting domain registrations within the .com TLD—the largest TLD in the domain name ecosystem. They concluded that 95% of typo domains target the “long tail” of the popularity distribution. The longitudinal study by Agten *et al.* [7] also confirms this trend, suggesting a shift in trends and behaviors of typosquatters.

3) *Effect of the Top-Level Domain:* The popularity of a TLD has been also investigated as a feature for its correlation with typosquatted domain names. For example, since the .com TLD was introduced as one of the first TLDs when the Domain Name System (DNS) was first implemented in January 1985 [2], it makes up a large portion of the total number of registered domain names (As of June 30, 2015, the total number of registered domain names was 294 million, out of which 117.9 million domain names were registered under .com, making up roughly 40% of the total domain names (<http://bit.ly/1VKiMr3>)). As such, a majority of the existing studies conducted on typosquatting have only considered domain names in the .com TLD. In their results, Banerjee *et al.* [9] observed that for nearly a quarter of all initial .com URLs, at least 50% of all possible phony sites exist; confirming that a domain name ending with .com has a high chance of being typosquatted. Interestingly, the results of Agten *et al.* [7] finds that certain country-code TLDs (.uk, .jp, etc.) affect the number of typosquatted domains they contain due to either an unconventional domain dispute policy or domain cost (e.g., cheaper domain names are more likely to be typosquatted).

Additionally, the TLD portion of a domain name may also be a target for exploitation. For example, one .com domain may have a malicious .org counterpart unbeknownst to the original registrant of the .com domain. A noteworthy example of this was mentioned in [12], where unsuspecting viewers inadvertently typed `www.whitehouse.com` instead of `www.whitehouse.gov` and got exposed to questionable contents instead of the official White House website. Banerjee *et al.* [9] further studied this effect and observed that domains under the .com TLD are impersonated primarily in .biz, .net and .org domains, and that domains not registered in the .com TLD extension are impersonated primarily in .com, .net and .org domains.

4) *Probability Models for Domain Landing:* The 2015 study by Khan *et al.* [20] introduced a novel approach for detecting typosquatting domains called the *conditional probability model*, which requires a vantage point at the network level to examine DNS and HTTP traffic records. This model identifies domains that have a high proportion of visitors leaving soon after landing on a site (domain name), followed by a visit to a more popular site (domain name) with a similar name. Specifically, they generated pairs of domains (d_1, d_2) such that each visit was performed within 33 seconds of each other and the Damerau-Levenshtein edit distance between the two domains is one. When dealing with lexically-similar domain pairs, where one of the two domains is unlikely a typo

of another, e.g., `nhl.com` and `nfl.com`, the advantage of applying the conditional probability model is that it does not correlate such domain pairs. In the results reported by Khan *et al.*, a request for `nhl.com` is only followed by a load of `nfl.com` .08% of the time where the reverse rate is even lower at $< 0.01\%$. However, they also reported that visits to the site `eba.com` are followed by visits to `ebay.com` 90% of the time, thus indicating that visits to `eba.com` are likely to be typos.

III. MONETIZATION STRATEGIES

The main drive of typosquatting is monetary in the first place, thus typosquatters employ various techniques to capitalize on their typosquatted domain names and generate revenues. In the following section, we review the various techniques that typosquatters employ to profit from deliberate registrations of typo domain names, including domain name parking (§III-A), domain name ransoming (§III-B), affiliate marketing (§III-C), hit stealing (§III-D), and scams (§III-E).

A. Domain Parking

The results of the 2006 study by Wang *et al.* [32] revealed that a large percentage of typo domains they observed were “parked”, where there was no real content on these pages except for advertisements that were generated by domain parking services. For example, Moore and Edelman’s 2010 study [25] highlighted the case of the typosquatted site `www.expendia.com`, which led to a web page that contained a list of sponsored links to travel-related websites. The popular travel site `expedia.com`, the most likely target, was at the top of the list followed by sponsored links to competitors such as `Orbitz.com` and `CheapTickets.com`. In the most recent study on typosquatting conducted by Agten *et al.* [7], domain parking continues to be the most popular scheme chosen by typosquatters.

Domain parking is not limited to benign applications, as show in the previous studies and more recently in [23], but may also include malicious behaviors and activities. For example, Alrwais *et al.* [8] explored the dark side of domain parking, and showed that parked domain names can be actually used for click fraud, traffic stealing, and spam delivery, all of which generate more than 40% of the revenue for some parking services.

B. Selling and Ransoming Domain Names

In addition to being “parked” with advertisements, a typosquatted site may have no content other than being advertised as for sale. In the extreme case, these typo domain names may be held for ransom—as in the Zuccarini case highlighted by the 2003 study by Edelman [15]. Edelman found that the vast majority of the typosquatted domain names acquired by the infamous cyber-criminal John Zuccarini were often redirected to websites with sexually-explicit content. For the owners of the authoritative domain names that Zuccarini’s typosquatted sites targeted, Edelman argued that having redirections to sexually-explicit content only increased their willingness to pay. Furthermore, Edelman has also pointed out that

Zuccarini may have profited from another source of revenue: affiliate marketing programs which we review next.

C. Affiliate Marketing

These programs are set up by companies to allow third parties to collect commissions on sales or referral fees for redirecting customers to their websites [11]. Such redirection can be for legitimate [10] or illicit applications [21]. For example, the “Amazon Associates” program was one of the first online affiliate marketing programs that was launched in 1996 [6]. When “Associates” (*i.e.* affiliates) create URL links and potential customers click through those links and buy products from Amazon, the Associates earn referral fees. Typically, these URL links contain unique identifiers to determine which affiliate has forwarded visitors. As Agten *et al.* point out, many typosquatters abuse such affiliate programs when they redirect visitors to the intended site, collecting referral fees from the authoritative owner for a visit that should have been theirs in the first place [7].

D. Hit Stealing

Not only do typosquatters redirect visitors to their intended websites (for monetary gain), but they can also forward them to websites of competitors. Essentially, these typo domain registrations “steal” traffic meant for authoritative domains. The study by Agten *et al.* [7] found that this behavior was mostly associated with adult sites (and for spam and click fraud as shown by Alrwais *et al.* [8]). However, some non-adult sites steal hits from their competitors in situations involving Internet marketing companies who draw traffic to the sites of their customers.

E. Scams

In this scenario, unsuspecting visitors may fall victim to a scheme that tricks them into divulging personally identifiable information (PII). As reported in [4], the typosquatted sites `Wikipedia.com` and `Twitter.com` emulated the real sites (`Wikipedia.org` and `Twitter.com`) and displayed advertisements for contests offering Apple iPads and MacBooks as prizes. Ultimately, users were prompted to enter their credit card number and other sensitive information as part of the contest to claim their prizes.

IV. COUNTERMEASURES

Countermeasures to typosquatting involve technical and policy-based aspects. Technical aspects to typosquatting, as indicated in the surveyed work in this study, take the identification of typosquatted domains as a first step. Then, the policy-based approach employs legal frameworks to resolve disputes between registrants in case of typosquatted domain names. While the technical aspects are treated at length in §II, in the following we review the policy-based aspects to countermeasures domain name typosquatting.

In November of 1998, the United States Department of Commerce identified a private, non-profit organization called the Internet Corporation for Assigned Names and Numbers (ICANN) as the new entity to oversee the domain name

registration system [3]. As the Internet rapidly expanded and the number of domain names being registered spiked, cybersquatters and typosquatters alike were quickly snatching up available domain names. In response, the ICANN introduced the Uniform Domain Name Dispute Resolution Policy (UDRP) in late 1999 which states that a domain registrant is required to submit to a mandatory administrative proceeding in the event that a third party (a “complainant”) disputes such a domain [5]. As Moore and Edelman [25] point out, while the majority of UDRP arbitration proceedings are successful for the complainants, the filing fees can range from \$1,300 to \$4,000—since December 1, 2002 (in use as of March 2016), the world intellectual property organization (WIPO), one of the main arbitration organizations assigned by ICANN, charges a tiered fee structure; \$1,500 for up to five domains, and \$2,000 for up to 10 domains in a single complaint reviewed by one panelist, and \$4,000 and \$5,000, respectively, with three panelists (<http://bit.ly/1MHwR1c>). While this might not be a lot of money for big companies, it might discourage smaller companies from filing a complaint, especially if targeted by large number of typosquatters/typosquatted domains.

Since the only remedies available to a complainant in the UDRP are the cancellation or the transfer of the domain name, another alternative became available through legal means: The Anti-cybersquatting Consumer Protection Act (ACPA). As noted in *Shields v. Zuccarini* [1]: “On November 29, 1999, the ACPA became law, making it illegal for a person to register or to use with the “bad faith” intent to profit from an Internet domain name that is “identical or confusingly similar” to the distinctive or famous trademark or Internet domain name of another person or company.”

As mentioned earlier, the typosquatters in the Facebook case were found to have violated the ACPA and were ordered to surrender their domain names as well as pay Facebook, netting them a total of \$2.8 million in damages [29]. While both the UDRP and ACPA can have successful outcomes for the authoritative domain name owners who decide to take the policy intervention route, eliminating the opportunity via defensive registration is perhaps the best strategy.

Defensive registration is a tactic where companies and trademark owners will deliberately register typo variations of their own domains, keeping it out of the hands of typosquatters and thus redirecting users to the proper domain. Despite this simple strategy, the results of Agten *et al.* shows that only 156 of the Alexa top 500 have defensive domain registrations, meaning that 344 domains (68.8%) have no defensive registrations whatsoever [7].

In line with defensive registration efforts, various registries offer domain name suggestion and trademark clearinghouse services to reduce the risks associated with typosquatting and typosquatted domain name registration by speculators. For example, ICANN specifies the structure and pricing of trademark clearinghouse, which can be deployed by any interested registry (e.g., Neustar, Nominet, Verisign, etc.)¹. Furthermore,

¹<http://bit.ly/1Sn770J>

Verisign provides name suggestion services that may include, among their suggestions, typosquatted domains².

V. CONCLUSIONS AND OPEN DIRECTIONS

In this paper we reviewed the landscape of domain name typosquatting and identified techniques used for typosquatting, methods used for their monetization, and countermeasures, including policy-based approaches.

While the current state-of-the-art highlights the problem, detection techniques, and policy-based approaches, less work is done on the technical front towards defending against this threat. To this end, we foresee a great opportunity in pursuing technical solutions to typosquatting utilizing features and inartistic characteristics of typosquatted domain names. In particular, we are currently pursuing three directions to realize informed solutions to the problem at hand:

End-user feedback. The nature of a domain name is often inferred from certain side channels, as by Khan *et al.* [20], or the type of contents it delivers as per Agten *et al.* [7]. However, none of the prior work considered end-users' feedback on the nature of those domain names, and whether they are domains of interest to them. Such ground truth is valuable, and could highlight new trends and features of typosquatted domain names that are not obvious, or captured by the algorithmic models we know so far. On the other hand, we envision a system that utilizes the well-known features of typosquatted domain names (blacklists of them or new features discovered using users' feedback) to inform users about the risks associated with domain names they are about to visit. Such feedback can be delivered to users in a usable way in the browser.

DNS-level filtering. While the user-centric approach to the problem provides the highest fidelity capturing the users' intent, it does not scale well. To this end, we also foresee a complementary solution that outsources all computations and decisions to the network. For example, based on a fine-grained ground truth of typosquatted domain names, one solution to prevent users from landing on those domain names and exposing themselves to attackers is to implement the blocking of such typosquatted domain names at the DNS level. One realization of such approach to implement the defense using a blacklist as a middlebox in the network. While the approach scales well, and is agnostic to the behavior of users and interaction with the defense system (unlike the end-user feedback based solution), it is also agnostic to the users' intent, and may block domains of interest to users. Furthermore, the system would rely on a blacklist that need to be actively and frequently updated, which comes at cost.

Up-to-date view via measurements. Many of the studies in the literature concerning the features of typosquatted domain names, their correlation with domain properties, and models for generating them are outdated. Furthermore, some of the recent studies concerning a partial set of those feature refute well-established belief in the earlier studies on this topic. To this end, we foresee a fruitful direction in revising those studies

in light of recent datasets (and previously not studied TLDs). In particular, as the use of new generic TLDs (gTLDs) is on the rise, we will extend our measurements to those TLDs in the pursuit for new features for finer understanding of the threat of domain name typosquatting and its evolution.

REFERENCES

- [1] —. Shields v. Zuccarini, 254 F. 3d 476 - Court of Appeals, 3rd Circuit. <http://1.usa.gov/1Sn9zon>, 2001.
- [2] —. History Behind .COM. <http://bit.ly/1UHdba0>, 2015.
- [3] —. Registrar Accreditation: History of the Shared Registry System. <http://bit.ly/1NWexTL>, 2015.
- [4] —. Typosquatting: How Spelling Errors Could Lead to Scams. <http://bit.ly/1fjuhO7>, 2015.
- [5] —. Uniform Domain Name Dispute Resolution Policy. <http://bit.ly/1tHVeEn>, 2015.
- [6] —. What is the Amazon Associates program? <http://amzn.to/1HymVJt>, 2015.
- [7] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse. In *NDSS*, 2015.
- [8] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang. Understanding the dark side of domain parking. In *USENIX Security*, 2014.
- [9] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. Cyber-Fraud is One Typo Away. In *IEEE INFOCOM*, 2008.
- [10] H. Burema and Y. Makino. System and method for tracking affiliates and merchants. <http://goo.gl/KWkkgY>, Aug. 7 2001. USPTO 09/922,953.
- [11] N. Chachra, S. Savage, and G. M. Voelker. Affiliate crookies: Characterizing affiliate marketing abuse. In *ACM IMC*, 2015.
- [12] C. G. Clark. The Truth in Domain Names Act of 2003 and a Preventative Measure to Combat Typosquatting. *Cornell Law Review*, 2004.
- [13] R. Dhamija, J. D. Tygar, and M. Hearst. Why Phishing Works. In *ACM CHI*, 2006.
- [14] A. Dinaburg. Bitsquatting: DNS Hijacking without Exploitation. <http://dynaburg.org>, 2011.
- [15] B. Edelman. Large-Scale Registration of Domains with Typographical Errors. <http://bit.ly/1IEGvql>, 2003.
- [16] B. Edelman. Typosquatting: Unintended Adventures in Browsing. *McAfee Security Journal*, 2008.
- [17] D. B. Gilwit. The Latest Cybersquatting Trend: Typosquatters, Their Changing Tactics, and How to Prevent Public Deception and Trademark Infringement. *Wash. UJL & Pol'y*, 2003.
- [18] T. Holgers, D. E. Watson, and S. D. Gribble. Cutting through the Confusion: A Measurement Study of Homograph Attacks Homographs and Confusability. In *USENIX ATC*, 2006.
- [19] S. Keats. What's In A Name: The State of Typo-Squatting 2007. <http://bit.ly/1mqonpl>, 2007.
- [20] M. T. Khan, X. Huo, Z. Li, and C. Kanich. Every Second Counts: Quantifying the Negative Externalities of Cybercrime via Typosquatting. *IEEE Security and Privacy*, 2015.
- [21] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, et al. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE Security and Privacy*, 2011.
- [22] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
- [23] L. Metcalf and J. Spring. Domain parking: Not as malicious as expected. Technical report, DTIC, 2014.
- [24] A. Mohaisen. Towards automatic and lightweight detection and classification of malicious web contents. In *HotWeb*, 2015.
- [25] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In *FC*, 2010.
- [26] N. Nikiforakis, M. Balduzzi, and L. Desmet. Soundsquatting: Uncovering the use of homophones in domain squatting. In *ISC*, 2014.
- [27] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. You Are What You Include: Large-scale Evaluation of Remote JavaScript Inclusions. 2012.
- [28] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, and F. Piessens. Bitsquatting: exploiting bit-flips for fun, or profit? In *WWW*, 2013.
- [29] C. Roth, M. Dunham, and J. Watson. Cybersquatting; typosquatting Facebook's \$2.8 million in damages and domain names. <http://bit.ly/1SnahSF>, 2013.

²<http://bit.ly/22yW0Xq>

- [30] J. Szurdi, B. Kocso, G. Cseh, M. Felegyhazi, and C. Kanich. The Long Tail of Typosquatting Domain Names. In *USENIX Security*, 2014.
- [31] M. Thomas and A. Mohaisen. Kindred domains: detecting and clustering botnet domains using DNS traffic. In *WWW, Companion Volume*, 2014.
- [32] Y. Wang, D. Beck, and J. Wang. Strider typo-patrol: discovery and analysis of systematic typo-squatting. *USENIX SRUTI*, 2006.
- [33] A. G. West and A. Mohaisen. Metadata-driven threat classification of network endpoints appearing in malware. In *DIMVA*, 2014.