A Smörgåsbord of Typos: Exploring International Keyboard Layout Typosquatting

Victor Le Pochat, Tom Van Goethem, Wouter Joosen
imec-DistriNet, KU Leuven
3001 Leuven, Belgium
{firstname.lastname}@cs.kuleuven.be

***Abstract*—Typosquatting is the malicious practice of registering domains that result from typos made when users try to visit popular domains. Previous works have only considered the US English keyboard layout, but of course other layouts are widely used around the world. In this paper, we uncover how typosquatters are also targeting communities that use these other layouts by examining typo domains on non-US English keyboards for 100 000 popular domains. We find that German users are the most targeted, with over 15 000 registered typo domains. Companies such as Equifax and Amazon have defensively registered such domains but are often incomplete; moreover, other major companies ignore them altogether and allow malicious actors to capitalize on their brand. Parking domains or advertising them for sale remains the most popular monetization strategy of squatters on at least 40% of registered domains, but we also see more harmful practices, such as a scam website that spoofs a local newspaper. This proves that domain squatters also consider typos on non-US English keyboards to be valuable, and that companies should be more alert in claiming these domains.**

***Index Terms*—Typosquatting, Web Security, Domain Name System, Internationalization**

I. Introduction

Domain names remain one of the properties of a website that are most visible to end users: they are prominently displayed in the address bar of browsers, shown in the listings of search engine results and generally mentioned in marketing material. They form a major part of a website's and, by extension, a brand's identity, which also makes them a prime target for malicious practices that try to either capitalize on a domain's popularity or impersonate it.

Typosquatting is one such practice, where malicious actors register domains that exploit human error when entering the URL of popular (authoritative) domains. For instance, they might register `faceboik.com`[1], which may be reached by unwitting users when they mistype `facebook.com`, and attempt to monetize it in a variety of ways, such as showing advertisements and links to 'related' websites through parking services [1], redirecting to the authoritative domain with affiliate links that provide the squatter with a commission on all purchases [1]–[3], or serving malware [4], [5].

Previous works have studied the prevalence of typosquatting over time [4] and for a large set of popular domains [5], but when enumerating potential typosquatting domains based on the proximity of keyboard keys, these works only consider the US English (QWERTY) keyboard layout. However, around the world other keyboard layouts are commonly used as well: these rearrange ASCII letters (such as the AZERTY or QWERTZ layouts used in e.g. France and Germany respectively) or swap punctuation symbols for commonly used accented characters (e.g. ñ on Spanish or å on Scandinavian keyboards).

In this paper, we study how the typosquatting phenomenon has expanded to target specific languages and communities, exploiting typos made on non-US English keyboard layouts. We generate candidate squatting domains across 100 000 popular domains, refining our search to domains that we can most reliably attribute to non-US English typosquatting. For those domains that are registered, we determine which countries they target, who owns them and how they are (ab)used.

We see that both brand owners and domain squatters are aware of non-US English typosquatting opportunities. While some companies with targeted domains have made defensive registrations, unfortunately they often miss certain variants. In addition, 6 of the 18 most targeted brands have made no defensive registrations whatsoever. Domain squatters take full advantage of these lapses, mostly monetizing the typo domains through domain parking, but we also observe malicious activity such as scams. Moreover, clusters of sites registered by the same entity as well as parked pages that reference the targeted brands make it clear that domain squatters are specifically targeting non-US English keyboards. This confirms that companies should pay attention to this kind of typosquatting as well, as we see that it is already prevalent today.

In summary, we make the following contributions:

- We study typosquatting that targets non-US English keyboards and the communities that use them, finding 28 943 registered typo domains that mostly target German users.
- We find that 12 out of the 18 most targeted companies have registered at least one domain defensively, but that only one has covered all potential typo domains.
- We classify the registered typo domains, and see that parking is the most popular way to monetize them at 40% of registered domains; only 3% are registered defensively.
- We detect 113 domains that lead to blacklisted websites, and find at least 116 more sites that maliciously redirect to a scam website.

[1]This domain is defensively registered by Facebook.

II. Background and methods

A. Typosquatting model

Investigations of typosquatting abuse require a model of which domains are most likely to result from a mistyping. Wang et al. [6] defined five kinds of typos:

1) **Missing-dot typos:** the dot following "www" is omitted: e.g. `wwwexample.com`.
2) **Character-omission typos:** one character is omitted: e.g. `exmple.com`.
3) **Character-permutation typos:** consecutive characters are swapped: e.g. `exmaple.com`.
4) **Character-replacement typos:** one character is replaced by an adjacent character on a 'standard' (i.e. US English) keyboard layout: e.g. `exzmple.com`.
5) **Character-insertion typos:** one character is inserted, either the character itself (duplication) or an adjacent character on a 'standard' (i.e. US English) keyboard layout: e.g. `exaample.com` or `exazmple.com` respectively.

These generate domains with a Damerau-Levenshtein distance [7], [8] of one, i.e. the insertion, deletion, substitution or transposition of one character; for the case of adjacent keyboard characters, these domains have also been coined as having a *fat-finger distance* of one [2]. These are the most frequent occurrences of typing errors: domains with more than one modification are less likely to occur [9] and more prone to be false positives.

We construct our specific typosquatting model conservatively: we ignore domains that could have been generated through more 'common' and previously studied techniques, which do not specifically target non-US English keyboard layouts. We therefore consider two kinds of typos:

1) **Character-replacement typos:** one character is replaced by a character that is adjacent on any *non-US English* keyboard layout but not adjacent on a US English keyboard: e.g. `zest.com` for `test.com` on a QWERTZ keyboard.
2) **Character-insertion typos:** one character is inserted that is adjacent on any *non-US English* keyboard layout but not adjacent on a US English keyboard: e.g. `tzest.com`.

Certain keyboard layouts feature adjacent keys that are accented variants of each other, such as `í` and `i` on the Czech QWERTZ layout. Domains that contain such visually resembling characters are used to deceive users, enabling spoofing or phishing in so-called homograph attacks [10]–[13]. However, as these attacks leverage the confusability of similarly looking domains (passively) and not users incorrectly typing the domain (actively), we omit homograph domains resulting from adjacent keys from our set of candidates.

Finally, in order to reduce coincidental collisions with non-squatting domains, we remove two additional classes of candidates: those where the second-level domain is shorter than five characters, and those that are the same as or homographs of a popular domain as we assume them to be non-squatting or a homograph attack respectively.

B. Data collection

1) Keyboard layouts: In order to generate candidate typosquatting domains, we first obtain the set of keyboard layouts defined in version 2.25 of the X Keyboard Configuration Database [14]. We limit ourselves to the 'basic' (default) keyboard layout for each included country, disregarding more obscure layouts that are unlikely to be commonly used and would likely introduce more collisions with benign domains. We extract the mappings from physical keys to characters using the parser and grammar of the `Keyboard Layout Editor` application [15], and list the adjacent characters for each key.

2) Input domains: We generate candidate typosquatting domains for the 100 000 most popular domains, retrieved from the Tranco list of December 22, 2018[2]. This list was proposed by Le Pochat et al. [16] as a replacement for the commonly used Alexa list, as this list has been shown to be both very volatile [17] and vulnerable to large-scale manipulation, and is instead generated by combining four rankings over 30 days.

3) Domain properties: To assess whether our candidate typosquatting domains are registered and how they are used, we collect the following data sets:

DNS records: We request `A`, `NS` and `SOA` records for both the candidate domains and the authoritative domains they are based on. We assume candidate domains to be registered if any record does not return an `NXDOMAIN` response, except for domains where the TLD returns a default record for all unregistered domains (such as `.fm` or `.ws`).

WHOIS records: We obtain registration data by retrieving and parsing WHOIS records with the Ruby Whois library [18]. This data set is incomplete, as WHOIS data is difficult to acquire in bulk (due to rate limits) and process automatically (due to varying formats) [19]. Registrant details may also be obfuscated (out of privacy concerns), outdated (e.g. company name changes) or inconsistent (e.g. slight differences in spelling or format between records for the same entity).

Web pages: To determine the purpose of our candidate typosquatting domains, we crawl the root page for each candidate that has a valid `A` record. By limiting our crawl to one page, we minimize the impact on the servers hosting the websites. We capture the request and response headers, the redirection path and final URL of the response, the HTML source and a screenshot.

Domain blacklists: To detect whether our candidates are known to exhibit malicious behavior, we match them and the domains they redirect to against the blacklists provided by Google Safe Browsing [20] (malware and phishing), PhishTank [21] (phishing), Spamhaus DBL [22] (spam, phishing, malware, botnets) and SURBL [23] (spam, phishing, malware and cracking).

III. Results

A. Distribution of typosquatting domains

For the 100 000 most popular domains, we generated 13 189 391 candidate typosquatting domains, of which we

[2]https://tranco-list.eu/list/M5LN/100000

TABLE I
DISTRIBUTION OF REGISTERED AND CANDIDATE TYPOSQUATTING DOMAINS ACROSS NON-QWERTY LAYOUTS.

	Registered		Candidate	
Keyboard layout	ASCII	IDN	ASCII	IDN
QWERTZ				
Hungary	15 830	58	776 715	737 309
Germany/Austria	15 195	139	771 576	306 684
Albania	15 195	0	771 576	241 679
Czechia/Slovakia	15 008	12	746 860	2 253 636
AZERTY				
France/Belgium	12 895	36	718 088	1 025 854
Senegal (Wolof)	12 895	38	718 088	1 158 234

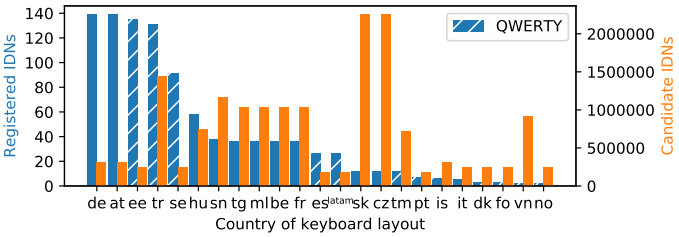

Fig. 1. Distribution of registered and candidate typosquatting IDNs across layouts. QWERTY layouts, where only IDN typo domains exist, are hatched.

found 28 943 to be registered. Table I shows that the majority stems from ASCII domains on non-QWERTY layouts. 290 Internationalized Domain Names (IDNs), which contain non-ASCII characters, have been registered, even though there are not proportionally fewer candidates. However, IDNs are less known and registration policies for IDNs are more restrictive [13]. Whereas candidate ASCII domains may have more coincidental collisions with benign sites, we do expect that nearly all of these IDNs are registered purely for typosquatting, especially as we have removed homograph domains.

Typosquatters can target specific communities by registering variants of local domains that stem from typos made on that country's keyboard layout. We find that German users are targeted the most: Table I and Figure 1 show that their QWERTZ keyboard layout is among the most registered overall. Figure 2 shows the distribution of registered typosquatting domains over TLDs, and while the `.de` TLD is the fifth most popular worldwide [24], we find that it is the second most targeted TLD for typosquatting of non-US English keyboard layouts. Finally, as shown in Figure 3, the number of domains in the `.de` TLD that are typos on the German QWERTZ keyboard is also the highest among all matches of TLD and country, which indicates that squatters know which community they target.

B. Distribution of targeted authoritative domains

Overall, the 28 943 registered typo domains target 14 860 authoritative domains, of which 6 365 were targeted more than once. Figure 4 shows that squatters have a preference for very popular domains, as their typo domains are likely to attract the most visitors, but that the domains are otherwise relatively evenly distributed in popularity, even for more targeted domains. Moreover, Figure 5 shows that shorter typo domains are much more prevalent, even though the possibility of typos is higher for longer domains (leading to more candidate typo domains), with the majority exploiting a replacement typo.

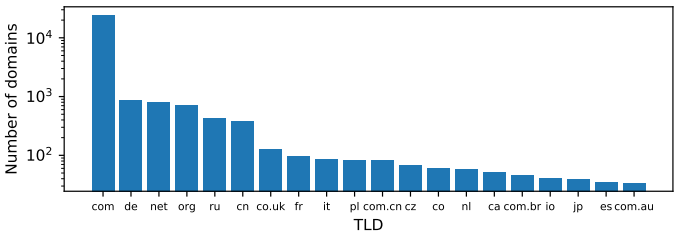

Fig. 2. Distribution of registered typosquatting domains across TLDs.

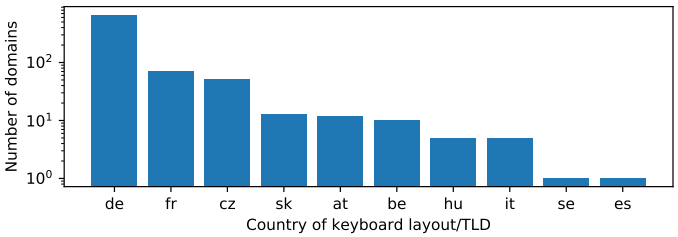

Fig. 3. Distribution of registered typosquatting domains where the TLD and the country of the keyboard layout match.

Table II lists the 18 authoritative domains for which the most typo domains are registered, alongside a (manual) classification of how those typo domains are being used. 12 of the most targeted domains have at least one defensive registration, but only retailmenot.com has succeeded in owning all possible typo domains. 3 others have missed one domain, with equifaxsecurity2017.com being a particularly strange case: the missed domain appears to be seized on behalf of Equifax by MarkMonitor (a brand protection company), but its nameservers were never reconfigured, which allows the

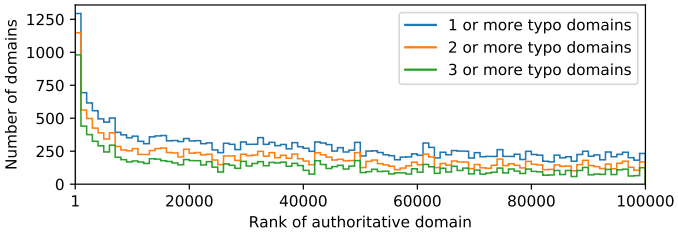

Fig. 4. Distribution of the popularity ranks of targeted authoritative domains.

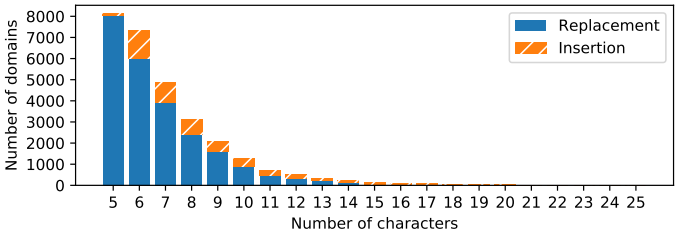

Fig. 5. Distribution of the lengths of candidate domains.

TABLE II
THE 18 AUTHORITATIVE DOMAINS THAT HAVE THE MOST REGISTERED TYPO DOMAINS, AND A CLASSIFICATION OF HOW THOSE TYPO DOMAINS ARE USED. THE MOST COMMON CATEGORY IS HIGHLIGHTED IN BOLD.

Target	Cand	Def	Park	Aff	Mal	Ques	Err	Coll
equifaxsecurity2017.com	37	**36**	1	0	0	0	0	0
amazon.com	32	**23**	5	0	1	3	0	0
amazon.de	28	**9**	8	5	0	4	2	0
brazzers.com	27	9	**17**	0	1	0	0	0
teamviewer.com	21	**20**	1	0	0	0	0	0
gmail.com	21	1	**12**	0	3	1	2	2
youjizz.com	19	0	**14**	0	0	3	1	1
zalando.de	18	1	**12**	2	0	1	2	0
xhamster.com	18	0	**16**	0	1	0	1	0
sznews.com	18	0	**9**	0	0	0	4	5
mymms.com	17	**8**	2	0	0	0	3	4
hotmail.com	17	1	**12**	0	1	2	1	0
retailmenot.com	16	**16**	0	0	0	0	0	0
mytoys.de	16	4	1	**6**	0	2	3	0
allstate.com	16	**15**	1	0	0	0	0	0
youtube.com	15	0	**8**	0	4	1	2	0
szhome.com	15	0	**7**	0	0	0	**7**	1
google.com	15	0	**13**	0	1	0	0	1

Cand = Candidates, Def = Defensive registration, Par = Parking/for sale, Aff = Affiliate abuse, Mal = Malicious behavior, Ques = Questionable behavior, Err = Error/empty, Coll = Collision with benign site

previous domain squatters to continue serving a parking page on the domain. This highlights the complexity for companies to combat typosquatting: companies must register all potential typo domains, while malicious actors only need one to be successful. For the remaining domains, the typo domains are mostly used for parking, but harmful behavior is also present through affiliate abuse, malicious redirects to scam sites or other questionable behavior such as displaying adult sites on typo domains of non-adult sites.

Certain typo domains redirect through URL shorteners to their destination. As bit.ly and goo.gl provide analytics data on the number and source of visits (accessed by adding a plus sign to the short URL), we can get insight in how often users of non-US English keyboards mistype domains. A typo domain targeting saturn.de saw 2 393 visits over 16 months, while one targeting zalando.de saw 468 visits over 24 months. Moreover, German users visited both domains the most, which corresponds to the keyboard layout of the typo. This shows that end users are making these typos and are ending up on the squatters' sites. We also see how companies and squatters manage typo domains: Amazon sends visitors of its defensively registered domains through a short URL to the authoritative domain, while a cluster of domains targeting Tipico (a betting service) all go through one short URL that redirects to an affiliate link.

C. Abuse on typosquatting domains

Domain squatters need a way to monetize their domains, e.g. by displaying advertisements or through more malicious practices such as spreading malware. Conversely, the owners of targeted popular domains want to prevent traffic from being diverted to malicious actors, even when end users make typos.

TABLE III
DISTRIBUTION OF TYPOSQUATTING DOMAINS ACCORDING TO THEIR PURPOSE.

Category	Count	%	Category	Count	%
Parking/for sale	11 444	39.5	Defensive	873	3.0
Affiliate abuse	93	0.3	Redirect to authoritative	181	0.6
Malicious	229	0.8	Unclassified	10 202	35.2
Empty/Error	5 921	20.5			

We determine whether the candidate typo domains on non-US English keyboards are subject to abuse, whether they are related to the authoritative domain or whether they are coincidental collisions with unrelated domains. We classify domains through four methods: we identify the most prevalent DNS and WHOIS record values and determine whether these can be attributed to one class; we check DNS records against the list of parking services compiled by Vissers et al. [1]; we compute four image hashes [25] of the screenshot, where visually similar images have similar hash values, and cluster their pairwise Hamming distances using DBSCAN [26] to find groups of websites with (nearly) the same content, of which we manually label the largest and most uniform; and we search certain keywords (e.g. 'parking') in the HTML source.

Table III lists the distribution of candidate typo domains over the categories in our classification. We see that at 39.5% of domains, parking remains the most popular way of monetizing typosquatting domains. More concerning, only 3% is registered defensively by the owner of the authoritative domain, for only 585 distinct authoritative domains, even though brands are at risk of being abused by malicious entities.

By redirecting to the authoritative domain with a tag indicating an affiliate account, a squatter can monetize the typo domain by receiving a commission on all sales made [3]. We consider potential affiliate abuse if a typo domain redirects to the authoritative domain with a non-empty path or query string, and manually verify 93 domains to exhibit abuse, based in part on the affiliate companies listed by Mathur et al. [27]; the count includes those domains that redirect to a specific product with the goal of increasing its sales. For one such cluster of sites, the localized character of the typosquatting is very apparent: 3 domains for amazon.com and 5 for amazon.fr, all typos on the French AZERTY layout, redirect to the Amazon page of the same French book on money creation.

113 typosquatting domains lead users to domains that are listed on at least one of our studied blacklists, with Table IV showing that these are mostly used for unwanted software and spam. However, we find that at least 116 additional sites across multiple parking services engage in malicious redirects [1], [28], taking victims to a page spoofing a local newspaper with a scam involving cheap iPhones (Figure 6). As we crawled each typosquatting domain only once, as parking services only redirect intermittently [1], and as the domain serving the scam page is not blacklisted, we expect the number of typosquatting domains that lead users to malicious content to be even higher.

3 291 (11.4%) domains use a WHOIS privacy/proxy service to conceal the identity of their owner. Malicious actors as

TABLE IV
NUMBER OF TYPOSQUATTING DOMAINS THAT LEAD USERS TO A DOMAIN THAT APPEARS ON A DOMAIN BLACKLIST. A DASH INDICATES THAT THE BLACKLIST DOES NOT CONSIDER THE GIVEN CATEGORY.

Blacklist	Spam	Phishing	Malware	Unwanted software
Google Safe Browsing [20]	–	8	5	59
Spamhaus [22]	22	0	0	–
SURBL [23]	19	2	0	–

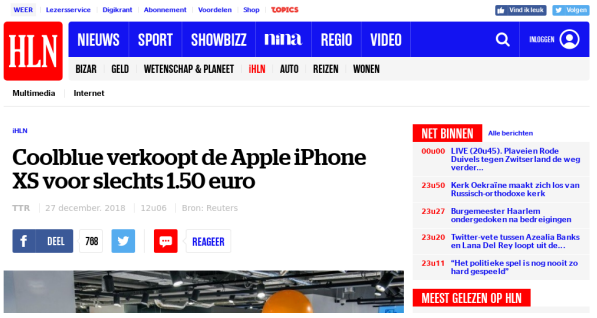

Fig. 6. Fake website spoofing a local newspaper, found on typosquatting domains that link victims to a scam page claiming to sell cheap iPhones.

well as typosquatters tend to use such services more than average [29], but as individuals may have legitimate reasons to protect their personal identity, using a privacy/proxy service does not imply maliciousness [30].

The 'related links' shown on parked pages make it clear that domain squatters are well aware of the potential traffic from certain countries: users of e.g. ParkingCrew can configure keywords [31], and several domains parked there refer to the authoritative domain and its content, whereas other parked pages tend to show a default set of links. For example, `googöe.se` has '`Google.SE`' as its only related link, with `ö` being adjacent to `l` on a Swedish QWERTY keyboard. The localized and highly targeted examples that we discussed, together with the high proportion of squatted domains, serve as evidence that malicious actors recognize and actively exploit typos made by international users.

IV. Related work

Typosquatting leverages typing errors made when humans try to visit popular websites. Edelman [32] first reported on the issue in 2003, finding thousands of sexually explicit typosquatting domains likely linked to one individual.

Wang et al. [6] developed models for automatically generating potential typosquatting domains, based on one-character modifications of popular domains using only adjacent keys. They used these models to crawl and analyze active typosquatting domains, finding that they were concentrated with a few large domain parking services. Banerjee et al. [9] found that typosquatting domains can have insertions, deletions or substitutions of an arbitrary number of (non-)adjacent characters, but that squatters prefer one-character modifications of shorter domains, as the probability of typos resulting in traffic to these domains is higher.

Moore and Edelman [2] analyzed typosquatting domains and their revenue sources, highlighting the role of advertising platforms in providing typosquatters with a way to monetize the domains. They found the phenomenon to be concentrated at a few large squatters and ad platforms. Vissers et al. [1] found typosquatting domains to be unevenly distributed over parking services.

Szurdi et al. [5] found that typosquatters increasingly target less popular domains through a study of the entire `.com` zone. They presented a tool that uses domain features from sources such as DNS or WHOIS records to identify and categorize likely typosquatting domains. Agten et al. [4] studied changes in typosquatters' behavior over time by tracking domains that target 500 popular domains over seven months, finding that they regularly change monetization strategies and are quick to claim expired registrations, as well as a trend towards longer domains and a concentration of certain hosters and TLDs.

Spaulding et al. [33] reviewed the typosquatting landscape, listing the models used to generate deceptive domains, the features that suggest a higher probability of typosquatting, the monetization strategies of the squatters and potential countermeasures. They also compared the effectiveness of typosquatting techniques through a user study [34]. Tahir et al. [35] studied why typosquatting is effective from a human-centric perspective, predicting the likelihood of typos based on domain composition, hand anatomy and keyboard layouts.

Domain squatters have also been found to exploit other types of errors or create the perception of dealing with a legitimate party through credible domain names. Homograph attacks leverage visually similar domains, e.g. through the addition of diacritical marks [10]–[13]. Bitsquatting [36], [37] leverages hardware errors that cause bit flips and subsequently character changes, requiring no human input. Soundsquatting [38] leverages domains constructed with words that sound similarly to those in popular domains. Combosquatting [39] leverages the combination of the intended domain with words or other characters. "AbbrevSquatting" [40] leverages alternative abbreviations of organization names or other phrases.

Nikiforakis et al. [37] analyzed the overlap of bitsquatting and typosquatting domains for three keyboard layouts (QWERTY, AZERTY and QWERTZ), but did not further study typosquatting in general on all three layouts. The `dnstwist` library [41] supports the same layouts when generating potential typosquatting domains based on adjacent keys, however without the country-specific accented characters. All other studies of typosquatting only consider the 'standard' US English QWERTY keyboard. We are the first to systematically analyze typosquatting for other keyboard layouts, as these allow for attacks that target specific countries.

V. Conclusion

Domain squatters abuse human errors made when typing popular domains by registering the resulting domains and monetizing them in a variety of ways. In this paper, we

highlight how this practice has moved beyond the 'standard' US English keyboard, to target other communities that use other keyboard layouts, in particular German users. Through a comprehensive analysis of 28 943 potential typo domains, we see that companies have acknowledged the legitimate threat of typosquatting on non-US English keyboards to their brands by defensively registering typo domains, but that they often fail at covering them all. Unfortunately, this leaves end users vulnerable to harmful practices as malicious actors also consider such domains valuable, mostly monetizing them through parking services, while also revealing the localized characteristics of the typosquatting abuse.

Acknowledgment

This research is partially funded by the Research Fund KU Leuven. Victor Le Pochat holds a PhD Fellowship of the Research Foundation - Flanders (FWO).

References

[1] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking sensors: Analyzing and detecting parked domains," in *22nd Annual Network and Distributed System Security Symposium*, 2015.

[2] T. Moore and B. Edelman, "Measuring the perpetrators and funders of typosquatting," in *14th International Conference on Financial Cryptography and Data Security*, 2010, pp. 175–191.

[3] N. Chachra, S. Savage, and G. M. Voelker, "Affiliate crookies: Characterizing affiliate marketing abuse," in *2015 Internet Measurement Conference*, 2015, pp. 41–47.

[4] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven months' worth of mistakes: A longitudinal study of typosquatting abuse," in *22nd Annual Network and Distributed System Security Symposium*, 2015.

[5] 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*, 2014, pp. 191–206.

[6] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels, "Strider typo-patrol: Discovery and analysis of systematic typo-squatting," in *2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet*, 2006, pp. 31–36.

[7] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964.

[8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics-Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[9] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan, "Cyber-fraud is one typo away," in *27th International Conference on Computer Communications*, 2008, pp. 1939–1947.

[10] E. Gabrilovich and A. Gontmakher, "The homograph attack," *Communications of the ACM*, vol. 45, no. 2, p. 128, Feb. 2002.

[11] T. Holgers, D. E. Watson, and S. D. Gribble, "Cutting through the confusion: A measurement study of homograph attacks," in *USENIX Annual Technical Conference*, 2006, pp. 261–266.

[12] B. Liu, C. Lu, Z. Li, Y. Liu, H. Duan, S. Hao, and Z. Zhang, "A reexamination of internationalized domain names: The good, the bad and the ugly," in *48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2018, pp. 654–665.

[13] V. Le Pochat, T. Van Goethem, and W. Joosen, "Funny accents: Exploring genuine interest in internationalized domain names," in *20th Passive and Active Measurement Conference*, 2019, pp. 178–194.

[14] S. V. Udaltsov *et al.*, "X keyboard configuration database," Version 2.25, Oct. 2018. [Online]. Available: https://www.freedesktop.org/wiki/Software/XKeyboardConfig/

[15] S. Xenitellis, "Keyboard layout editor," Oct. 2008. [Online]. Available: https://github.com/simos/keyboardlayouteditor

[16] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *26th Annual Network and Distributed System Security Symposium*, 2019.

[17] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of Internet top lists," in *2018 Internet Measurement Conference*, 2018, pp. 478–493.

[18] S. Carletti. Ruby Whois. [Online]. Available: https://whoisrb.org/

[19] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul, "Who is .com?: Learning to parse WHOIS records," in *2015 Internet Measurement Conference*, 2015, pp. 369–380.

[20] Google. Safe browsing. [Online]. Available: https://safebrowsing.google.com/

[21] OpenDNS. PhishTank. [Online]. Available: https://www.phishtank.com

[22] Spamhaus Project. The domain block list. [Online]. Available: https://www.spamhaus.org/dbl/

[23] SURBL. SURBL URI reputation data. [Online]. Available: http://www.surbl.org/

[24] W3Techs. Usage of top level domains for websites. [Online]. Available: https://w3techs.com/technologies/overview/top_level_domain/all

[25] J. Buchner. Imagehash. [Online]. Available: https://github.com/JohannesBuchner/imagehash

[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[27] A. Mathur, A. Narayanan, and M. Chetty, "Endorsements on social media: An empirical study of affiliate marketing disclosures on YouTube and Pinterest," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 119:1–119:26, Nov. 2018.

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

[29] R. Clayton and T. Mansfield, "A study of Whois privacy and proxy service abuse," in *13th Annual Workshop on the Economics of Information Security*, 2014.

[30] 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," in *Asia Conf. on Computer and Communications Security*, 2018, pp. 609–623.

[31] ParkingCrew. How can I set keywords? [Online]. Available: https://www.parkingcrew.com/faq.php?mode=detail&detail=9

[32] B. Edelman, "Large-scale registration of domains with typographical errors," Berkman Center for Internet & Society - Harvard Law School, Tech. Rep., Sep. 2003. [Online]. Available: http://cyber.law.harvard.edu/people/edelman/typo-domains

[33] J. Spaulding, S. Upadhyaya, and A. Mohaisen, "The landscape of domain name typosquatting: Techniques and countermeasures," in *11th International Conference on Availability, Reliability and Security*, 2016, pp. 284–289.

[34] J. Spaulding, D. Nyang, and A. Mohaisen, "Understanding the effectiveness of typosquatting techniques," in *5th Workshop on Hot Topics in Web Systems and Technologies*, 2017.

[35] 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 *27th International Conference on Computer Communications*, 2018, pp. 2618–2626.

[36] A. Dinaburg, "Bitsquatting: DNS hijacking without exploitation," Raytheon Company, White Paper #2011-307, 2011.

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

[38] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *17th International Conference on Information Security*, 2014, pp. 291–308.

[39] 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 *2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 569–586.

[40] 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 *2018 International Conference on Computational Science*, 2018, pp. 221–233.

[41] M. Ulikowski, "dnstwist." [Online]. Available: https://github.com/elceef/dnstwist