Poster: 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

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`, 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, across 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).

II. Background and methods

We have studied how the typosquatting phenomenon has expanded to target specific languages and communities, exploiting typos made on non-US English keyboard layouts [6]. 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.

A. Typosquatting model

Investigations of typosquatting abuse require a model of which domains are most likely to result from a mistyping. Wang et al. [7] defined five kinds of typos: omitting the dot following "www", omitting one character, swapping consecutive characters, replacing one character by an adjacent character, and inserting the same or an adjacent character. These are the most frequent occurrences of typing errors: domains with more than one modification are less likely to occur [8] 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`.

We omit visually resembling 'homograph' domains [9]–[12] generated on layouts with adjacent accented variants (e.g. `í` and `i` on the Czech QWERTZ layout), as these leverage the confusability of similarly looking domains (passively) and not users incorrectly typing the domain (actively). In order to further reduce coincidental collisions with non-squatting domains, we also remove those candidates 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: We generate domains for the 'basic' variant of all country-based keyboard layouts defined in version 2.25 of the X Keyboard Configuration Database [13].

2) Input domains: We generate candidate typosquatting domains for the 100 000 most popular domains, retrieved from the Tranco list [14] of December 22, 2018[1].

3) Domain properties: We collect DNS records and WHOIS records; crawl their Web pages; and match them against four well-known blacklists.

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

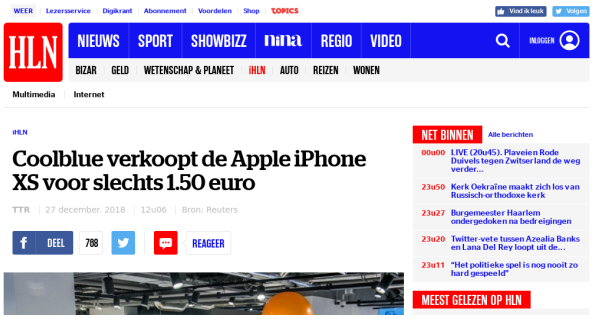

Fig. 1. Fake website spoofing a local newspaper, found on typosquatting domains that link victims to a scam page claiming to sell cheap iPhones.

III. Summary of results

We see that both brand owners and domain squatters are aware of non-US English typosquatting opportunities. For the 100 000 most popular domains, we generated 13 189 391 candidate typosquatting domains, of which we found 28 943 to be registered, mostly targeting German users with over 15 000 registered domains. These domains target 14 860 authoritative domains, more often popular and short domains. While some targeted companies, such as Equifax and Amazon, have made defensive registrations, they often miss certain variants: only one of the 18 most targeted brands (retailmenot.com) has covered all potential typo domains. In addition, 6 of them have made no defensive registrations whatsoever.

Table I lists the distribution of how candidate typo domains are being (ab)used. We see that at 39.5% of domains, parking or advertising them for sale remains the most popular way of monetizing typosquatting domains. More concerning, only 3% is registered defensively by the owner of the authoritative domain. We also observe malicious activity: 113 domains are blacklisted due to spam, phishing, malware or unwanted software; 93 domains abuse affiliate links [3]; and 116 domains redirect to a scam website for cheap iPhones that spoofs a local newspaper (Figure 1). Moreover, 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.

We found several instances where the localized character of the typosquatting is very apparent. The sites in one cluster of typo domains on the French AZERTY layout of Amazon (3 for amazon.com and 5 for amazon.fr) all redirect to the Amazon page of the same French book on money creation. Moreover, 'related links' shown on parked pages sometimes refer to the authoritative domain and its content: for example, `googöe.se` has '`Google.SE`' as its only related link, with `ö` being adjacent to `l` on a Swedish QWERTY keyboard. This serves as evidence that malicious actors recognize and actively exploit typos made by international users.

Overall, we see that companies have acknowledged the legitimate threat of typosquatting on non-US English keyboards to their brands by defensively registering typo domains.

TABLE I
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	Redirects to authoritative	181	0.6
Malicious	229	0.8	Unclassified	10 202	35.2
Empty/Error	5 921	20.5			

However, because they often fail at covering them all, end users become vulnerable to harmful practices as malicious actors also consider such domains valuable, mostly monetizing them through parking services, confirming that companies should pay more attention to this kind of typosquatting as well, as we see that it is already prevalent today.

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] V. Le Pochat, T. Van Goethem, and W. Joosen, "A smörgåsbord of typos: Exploring international keyboard layout typosquatting," in *4th International Workshop on Traffic Measurements for Cybersecurity*, 2019.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.