Innovation activity to detect and prevent Typosquatting attacks

Aruna Prem Bianzino
ElevenPaths, Telefónica CyberSecurity Unit
Ronda de Comunicación s/n
Edificio Oeste 1, Planta 1
arunaprem.bianzino@global.11paths.com

Javier Artiga
ElevenPaths, Telefónica CyberSecurity Unit
Ronda de Comunicación s/n
Edificio Oeste 1, Planta 1
javie.artiga@global.11paths.com

***Abstract*—This document describes the innovation activity aiming at detecting and preventing typosquatting. The activity has been carried out by the innovation and laboratory department of ElevenPaths.**

***Index Terms*—Typosquatting, Automatic analysis**

I. What is Typosquatting?

Typosquatting is a form to acquire the online identity of another entity relying on mistakes such as typos made by the internet user when entering a website address into a browser. Typosquatting has the purpose of acquiring the business and/or user base of the target entity, redirecting the mistaken user to a proper web-page faking the original one, eventually asking for sensitive information of the user, leveraging on their trust on the original brand. Figure 1 reports some examples of typosquatting, including explicit typosquatting, domains for sale, domains registered but not yet used, and domains unrelated to the original one (i.e., false positives for typosquatting).

Typosquatting may return revenue to the typosquatter in different ways, including registering a domain before the owner of the corresponding brand and sell it back to the brand owner at a higher price, advertisement subject to a considerable traffic flow, diverting fees, profit margin on resales of products of the original brand, etc. A single typosquatter may earn Millions of dollars per year [1], while brand owners suffer image damage from typosquatting as well as costs in term of money and time to fix the issue once discovered. Such damages have been estimated in $20 billions per year [2], making the issue worth a huge investment in protection.

II. Current Solutions

The current standard solution for typosquatting detection is a reactive solution and consists of the following.

- For each brand to be protected, a list of keywords is generated manually by an analyst. The brand management may also set a level of alert and filters for unwanted notifications. This last option allows removing known false positives in advance.
- A time interval is set as a pace to check new registered domains. New domains are checked against the list of keywords (looking for exact matches and regular expressions. e.g., the newly registered domain branda.com will be detected as suspicious for the keyword *brand*, from the brand "brand").
- Matchings generate alerts, which are manually checked by analysts, and eventually signaled to the brand manager.

The main disadvantages of this practice are:

- The current check introduces delay (the usual time interval between consecutive checks is of 24 hours, and a delay in the order of days exists between the domain is registered and the corresponding update is received). A higher check frequency will not make sense, as the checked domain lists are updated at most every 24 hours, and with much lower frequency for less common TLDs.
- This process requires a lot of manual intervention (starting from the keyword generation, down to the check of the suspicious domains, which include a wide majority of false positives – about 97% in the current commercial solution available to us).
- The list of registered domain and its updates are usually really expensive to access.
- Usually the checks are performed only against the domains registered since the last check. As such, if a new brand is created for a domain owner, or new keywords are added to the search for any reason, they will be checked only against the most recent registered domains.
- This process only returns information about the registration status of a domain, and not about its resolution or other information that may be relevant to the analyst.

III. The Proposed Solutions

In order to overcome these limitations, we proposed the solution depicted in Figure 2. In particular, the domain owners provide the list of official domains, social network official accounts and brands. They may also provide a white-list of terms they do not want to be notified and/or a list of previously detected malicious domains. Finally, the domain owner provides a list of official languages.

A. Dictionary generation

The official web-pages and the content generated by the official social network accounts are crawled for keywords by

Identify applicable funding agency here. If none, delete this.

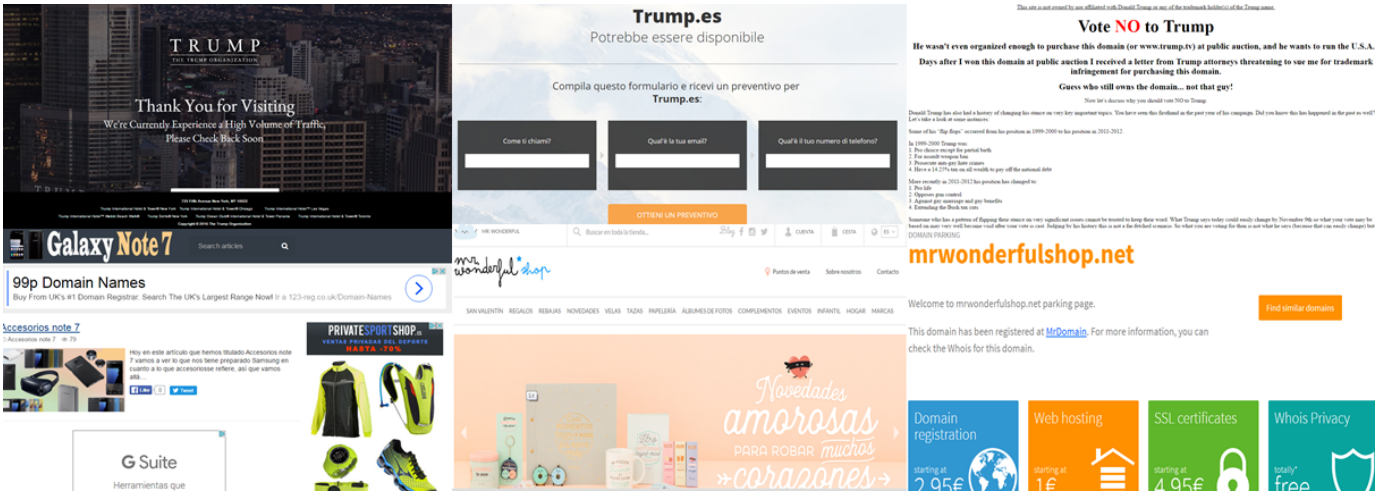

Fig. 1. Examples of typosquatting. On top left the legitimate web-page of Donald Trump, i.e. trump.com. On the top center a domain registered and for sale, i.e., trump.es. on the top right an explicit typosquatting page, i.e., trump.org. On the bottom left a web-page unrelated to the main one, i.e., galaxynote7.es. on the bottom center a legitimate web-page, i.e., mrwonderful.es, and on the bottom right a similar domain registered but not yet used, i.e., mrwonderful.net.

a proper module. In particular, a list of most frequent words is periodically generated. This list is filtered removing common and not interesting words (e.g., "Facebook", "web", "blog", "YouTube", etc.) and the resulting word set, referred to as "Related words" in the architecture, is used to:

- Find matching among the top trending topics. Specific parameter may be set for geographical and depth restriction of this search.
- Combine keywords among them. A maximum length for combinations may be set in terms of keywords or characters.

The obtained list is combined with TLDs and added to the current domain dictionary.

At the same time, official domains are edited accounting for three kind of variations:

- **Phonetic typos**, which include changing a phoneme for another homophonic one, leet, and close Unicode characters, e.g., "f" for "ph", "e" for {"è", "é", "ë", "ê", "3", etc.}, etc.
- **Semantic typos**, which include replacing a signifier with another similar one, e.g., a TLD for another, or "www" for {"www-", "vvv", etc.}.
- **Keyboard typos**, which include applying typos to the original domain/brand. This includes standard typos ("fatty finger" typos, inclusions, omissions, permutations, etc.), and more (detailed below).

In particular, the Keyboard Typos include the following variations:

- **Addition** of characters at the end of the domain name (e.g., branda.com)
- **Bitsquatting**: replacement of a character for another random one (e.g., brant.com)
- **Hyphenation**: an hyphenation is added to the domain (e.g., bra-nd.com)
- **Insertion**: a random character is inserted in the domain name (e.g., bramnd.com)
- **Omission**: a character is deleted from the domain name (e.g., rand.com)
- **Repetition** of a character in the domain name (e.g.., braand.com)
- **Replacement** of a character in the domain name for another close in the (qwerty) keyboard (e.g., brsnd.com)
- **Subdomain**: a point is added in the domain name (e.g., br.and.com)
- **Transposition**: swap of two consecutive characters in the domain name (e.g., barnd.com)
- **Vowel-swap**: swap of a vowel with another one in the domain name (e.g., brend.com)

The dictionary of domains to be checked is hence composed by the combination of (i) the ones generated combining the keywords obtained from the official domains and social networks account, (ii), the ones generated looking for keywords among trending-topics, and (iii) the ones generated applying typos to the official domains.

The obtained dictionary is filtered removing known false positives (domains matching words in the white-list), and added to the list of analyzed domains in the Database (DB).

B. Domain status and priority

Once in the DB, domains are checked periodically. In order for the solution to scale properly, states are defined for domains (e.g., "Suspicious", "Parking", "Non-suspicious", etc.) and associated to priorities for check (e.g., "High", "Low", etc.), resulting in different check frequency (e.g., daily, weekly etc.). States, priorities and frequencies can be set according to preferences and resources. Indicators of a higher risk status for domains (resulting in higher priority and checking frequency) should be:

- The domain matches a **pattern** common to other previously detected domains.

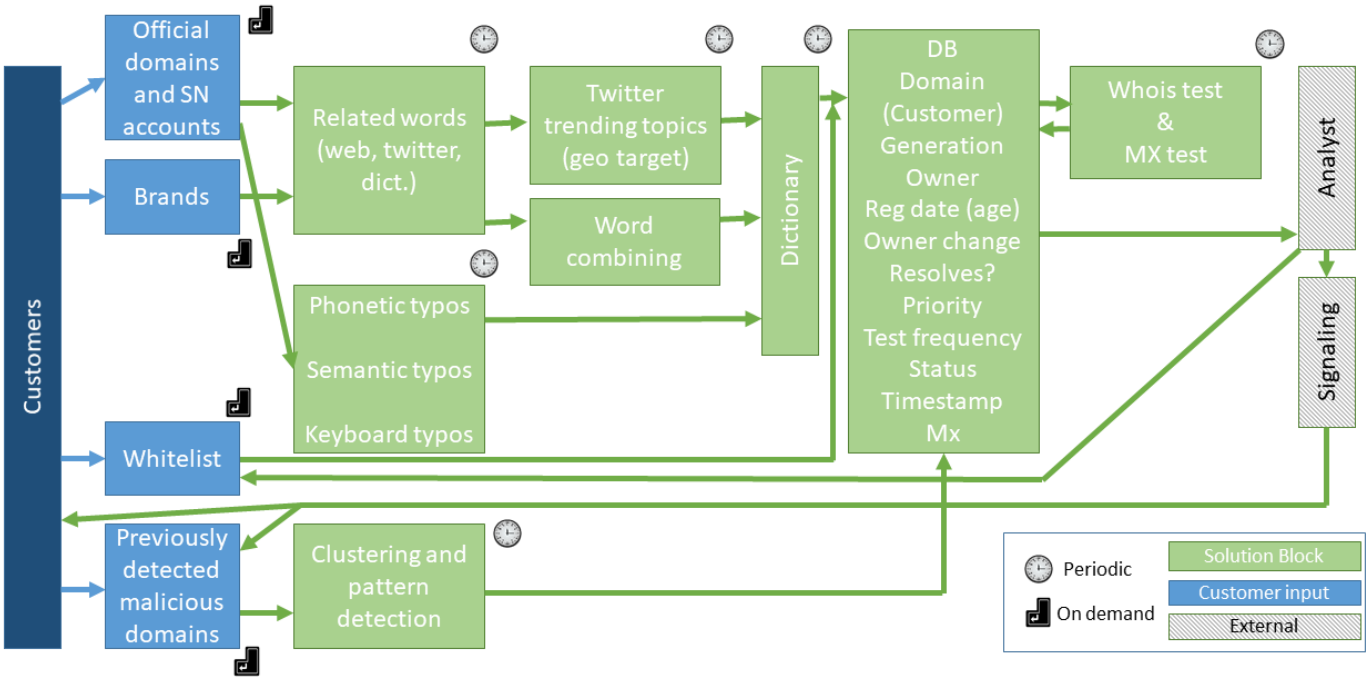

Fig. 2. Architecture of the proposed solution.

- Some **properties** of the domain have recently changed (e.g., registration status, registrant, IP, etc.)

C. Patterns among malicious domains

Domains which have been confirmed as malicious by the analysts and signaled to the domain owner concur to the set of "Previously detected malicious domains", which was initiated by the domain owner itself. Patterns are looked for in this set, using clustering and machine learning techniques, resulting in regular expressions (e.g., *brandpayment.com, *shop.com, *2018*, etc.). If a domain matches a malicious pattern, it should be assigned a higher test frequency.

A possible implementation of this pattern detection consists in grouping previously detected malicious domains into clusters accounting for the Levenshtein distance [3] among them. Finally, common strings can be selected from clusters. This solution allows to easily identify patterns not only at the beginning of the domain names, but also in the middle and at the end, and also for domains of variable length.

D. Domain testing

According to the time-stamp of the DB data for a specific domain and the testing frequency corresponding to its status, a new testing for the domain is triggered. The testing include a set of tests to check:

- The domain registration status, i.e., if the domain has been registered or not, and eventually the registrant identifier and the registration time-stamp.[1]
- Whether the domain has an IP address associated or not.
- Whether the domain has an MX server associated or not.
- Whether the domain has a web-page associated or not. In the case in which a domain corresponds to a web-page, a further opportunity is to test the page content (for logos, content, etc.). This opportunity is technically very complex and requires to mix text and image comparison, as well as to set thresholds for similarity, which are very difficult to select. As such, we leave this opportunity to future developments.

[1]Changes in the domain registration status may range from owner change, i.e., selling the domain, to a simple change in the registered email address, being eventually irrelevant.

IV. Experimental Results

The tests have been executed on a virtual machine whose characteristics are detailed in Table I

TABLE I
CHARACTERISTICS OF THE TESTING ENVIRONMENT.

OS	Ubuntu 14.04.5 LTS
Processor	2x Intel Xeron CPU E5-2650 0 @8 2.00GHz
Memory	4137MB
IP Address	172.16.0.29

A. Dictionary Generation

In order to get an idea of the size of the generated dictionary of tested domains, starting from the set of .com domains in the top 500 Alexa ranking (i.e., 340 domains), the *keyboard typos* variations generated a total of 66,880 domains. Table II reports how the generated domains are distributed among the different variations included among the keyboard typos, as well as the share of each of them resulting in a resolving address. In

particular, in order to give a relative view, the table reports (i) the domains generated by each variation, (ii) the percentage it represents with respect to the total amount of domains generated through keyboard typos, (iii) the number of domains resolving for the variation in object, (iv) the percentage it represents with respect to the number of domains generated by the variation in object, and (v) the percentage it represents with respect to the total amount of domains resolving among the ones generated by keyboard typos.

TABLE II
RESULTS FOR THE KEYBOARD TYPOS VARIATIONS. THE SECOND COLUMN FROM THE RIGHT SPECIFIES THE PERCENTAGE OF DOMAINS RESOLVING TO AN IP ADDRESS WITH RESPECT ALL THE DOMAINS GENERATED BY THE SPECIFIC VARIATION, WHILE THE MOST RIGHT COLUMN SPECIFIES THE PERCENTAGE OF DOMAINS RESOLVING FOR THE SPECIFIC VARIATION WITH RESPECT TO THE TOTAL AMOUNT OF RESOLVING DOMAINS.

	Num of gen domain names	**Perc. of total**	**Num of solved addresses**	**Perc. resolved**	**Perc. of total resolved**
Addition	8,440	12.8%	2,817	33.4%	11.8%
Bitsquatting	11,941	18.1%	4,201	35.2%	17.6%
Hyphenation	2,228	3.4%	403	18.1%	1.7%
Insertion	20,743	31.4%	5,439	26.2%	22.8%
Omission	2,499	3.7%	1,820	72.8%	7.6%
Repetition	1,566	2.4%	913	58.3%	3.8%
Replacement	12,393	18.7%	5,384	43.4%	22.5%
Subdomain	2,228	3.4%	808	36.3%	3.4%
Transposition	2,152	3.3%	1,230	57.2%	5.2%
Bitsquatting	1,950	2.9%	826	42.4%	3.5%
Total	66,880	100%	24,234	36.0%	100%

B. Related keywords

This architectural block has been implemented through a crawler for web-pages, a crawler for social networks and language dictionaries for common words (which will probably not be of interest, as representing common words and generating noise). As a first step, the official web-page(s) of the target brand will be crawled for keywords, as well as the social network official account(s). Then, the two resulting word sets are merged keeping only the common ones, and finally the resulting set is filtered removing common words from the dictionary of the official language(s) of the target brand.

This solution has been tested for the Telefónica brand. For this test case, the web crawling generated about 4000 keywords, while the social media crawler generated about 7000 keywords. After merging and filtering, the resulting dictionary is composed by 37 keywords, among which more than a half (54.1%) were directly correlated relevant keywords (e.g., Movistar, TGS - Telefónica Global Service, LTE, Wayra - Telefónica's startup accelerator, Luca - Telefónica's BigData unit for corporate services, etc.), and the rest equally shared between correlated words (21.6%, e.g., Rafa Nadal - official ambassador of Telefónica in the World, Blau - Competence, i.e., Telefonica.de, Huawei, etc.), and uncorrelated words (24.3%, e.g., eeuu, sourcing, ceo, etc.).

C. Twitter trending topics

The related keywords are looked for among the Twitter trending topics. Twitter trending topics may be filtered by amount (e.g., top 10 trending topics), by frequency check (e.g., during last hour), and by geographical region (e.g.,World, Europe, Spain, Madrid, etc.).

We ran a test for the Telefónica brand using the following settings: top 10 trending topic per hour, accounting for World, Spain and Madrid. The matching trending topics are tested for domain resolution using the following set of TLDs: .com, .es, .uk, .net, .co, .edu, .org, .int,.gov, .mil. Globally, the 22% of the tested resulting domain (i.e., trending topic + TLD) resolved. The corresponding distribution among different TLDs is reported in Table III.

TABLE III
PERCENTAGE OF RESOLUTION OF THE TRENDING TOPICS MATCHING RELATED KEYWORDS, WITH DIFFERENT TLDS.

TLD	.com	.es	.uk	.net	.co
Perc. resolved	64.5%	37.2%	4.7%	42.7%	31.6%
TLD	.edu	.org	.int	.gov	.mil
Perc. resolved	1.3%	44.5%	0%	0.4%	0%

D. Pattern detection

The pattern detection functionality has been tested over a set of about 3,500 domains signaled by the analysts to Telefónica customers. We used a Levenshtein distance of 0.15 and generated 873 clusters. Analyzing the clusters, two kind of them were highlighted: (i) the ones corresponding to a searched keyword, and (ii) the one not corresponding to any searched keyword. By analyzing the clusters in the first set, we are able to establish on which position in the domain name the keywords are more likely to appear in suspicious domains. On the other hand, the second set of clusters allows to identify further keywords. In our testing we identified here keywords such as *services*, *shop*, *website*, *billing*, etc.

E. Domain testing

The solution has been tested over a set of about 36M domains. All the domains have been checked correctly, without service interruption by the third party services, resulting in an average time per domain check of about 0.618s. This testing speed allows to test an amount of about 140K domains per day on a single machine. Using more parallel threads, or multiple replicas of the architecture on different machines and adjusting the checking frequency of the different priority levels allow to scale to any amount of tested domains.

The proposed solution has been tested on a part of Telefónica's customers. On two different weeks, i.e., 24-30/07/2017 and 5-11/10/2017, the solution results has been compared to the current practice in Typosquatting detection. Results of such comparison is reported in Table IV. In particular, the proposed solution results in an increase of about 23% of the detected domains (i.e., potentially malicious, to be submitted to analysts for manual analysis), and in a

much lower false positive rate (i.e., signaling domains not related to the brand). It is interesting to notice as the sets of domains signaled by the current tool and the proposed one are mostly disjointed (i.e., only about 5% of the signaled domains are common among the two solutions, while the rest of the domains are signaled by only one of the two solutions). As such, the best possible solution would be to use both tools in an integrated manner.

TABLE IV
COMPARISON OF THE PROPOSED SOLUTION AND THE CURRENT PRACTICE. ABSOLUTE VALUES ARE FOR WEEK AVERAGE.

	Current practice	Proposed solution
True positives	17 domains	21 dom. (+23.5%)
False positives	97%	6 (22.2%)
New detections	32 (94.1%)	40 (95.2%)

V. RELATED SOLUTIONS

Concerning the trending topic analysis for the registration of the corresponding domain (trending topic + TLD), other solutions are currently available. In particular, **Domain Check - Trending Domain Names** (http://domaincheckplugin.com/trending) is a tool that checks trending topics on Twitter, Facebook and Google, "on real time". It checks only for the .com TLD, and only for the global trending topics (as geographical target). At the same time, **Aquí Hay Dominios** (https://www.aquihaydominios.com/dominios-libres-trending-topic) is a service that checks for the resolution of domains made by a trending topic + a TLD (only .com and .es), for the top 80 Spanish trending topics, every 30 minutes. None of these solutions provide automatic signaling related to a specific brand nor to a set of specified keywords.

Similar approaches to ours have been developed, for instance in [4], where an automatic analysis system is proposed, but using a much smaller set of variations for domain generation, not accounting for social networks nor brand related keywords, and using a much simpler automatic analysis, where domains are only checked through a browser, making the analysis slower and less relevant for the analysts.

VI. FUTURE WORK

Being able to access statistics from Domain Name Servers (DNSs) about searched domains may be of great interest as it may help different components of the proposed system. (i) For what concerns priority among domains, higher priority could be granted to domains which have actually been searched by users, or on the basis of the number of searches. (ii) Keywords may be searched among domains which have actually been searched by users and matches may be integrated into the dictionary. (iii) Finally, statistics on searched domains may give a better understanding on the importance of the different techniques for domain generation and allow to optimize resources. Furthermore, it may highlight eventual domains to be monitored much before they are registered.

In the case of resolving domains, the analysis will benefit from taking into account also the content of the resulting web-page. Analysis possibilities include statistics on the keywords resulting from the target web-page and from the original one of reference, or image similarity. Both analysis may return similarity indicators with respect to the original web-page, to be provided to analysts as support and as risk indicators.

Finally, when distributing the domains to be checked among different processes and machines, a proper distribution algorithm may be implemented, as different domains result in different analysis time and a simple distribution on numeric basis is not efficient.

REFERENCES

[1] Kelly M. Slavitt, Protecting Your Intellectual Property from Domain Name Typosquatters, FindLaw, 2004, https://corporate.findlaw.com/intellectual-property/protecting-your-intellectual-property-from-domain-name.html

[2] J. M. Rao and D. H. Reiley, The Economics of Spam, The Journal of Economic Perspectives, pp. 87–110, 2012.

[3] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10(8):707–710, 1966.

[4] Wang, Y. M., Beck, D., Wang, J., Verbowski, C., & Daniels, B., Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. SRUTI, 6, 31-36, 2006.