You've Been Tricked! A User Study of the Effectiveness of Typosquatting Techniques

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

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

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

***Abstract*—The deceitful practice of *Typosquatting* involves deliberately registering Internet domain names containing typographical errors that primarily target popular domain names, in an effort to redirect users to unintended destinations or steal traffic for monetary gain. Typosquatting has existed for well over two decades and continues to be a credible threat to this day. While much of the prior work has examined various typosquatting techniques and how they change over time, none have considered how effective they are in deceiving users. In this paper, we attempt to fill in this gap by conducting a user study that exposes subjects to several uniform resource locators (URLs) in an attempt to determine the effectiveness of several typosquatting techniques that are prevalent in the wild. We also attempt to determine if the security education and awareness of cybercrimes such as typosquatting will affect the behavior of Internet users.**

Keywords. Domain Names, Typosquatting, Defenses.

I. Introduction

With nearly 300 million registered domain names (as of late 2015), the Domain Name System (DNS) has evolved to become a cornerstone for the operation of the Internet. While the majority of all domain names ultimately resolve to a web server that hosts meaningful content, there is an alarming amount of domain names that are deliberately registered with typographical variations that target popular domain names. *Typosquatting*, as this practice became known as, involves generating domain names in such a way as to exploit common typographical errors made by users that manually type URLs into web browsers in an attempt to steal traffic or redirect users to unintended destinations. These so-called "typosquatters" employ several techniques (*e.g.,* adding or deleting characters) when typosquatting domain names in order to sufficiently capture enough traffic for monetary or personal gain.

In this paper, we present the design and evaluation of a user study for gauging the effectiveness of several typosquatting techniques that are used in the wild. More specifically, we make the following contributions:

- We validate typosquatting techniques presented in prior studies by examining their prevalence using various carefully sampled domains from several data sources.
- We experimentally demonstrate how security education and awareness of cybercrimes, particularly typosquatting, will affect the behavior of Internet users.
- We highlight various correlations between attributes of participating subjects and their proneness to accepting typosquatted domains, and hint on leveraging cognitive traits of Internet users to strengthen the defense against typosquatted domains.
- We publicly release our data so others can verify and build upon our research findings and results.

II. Background and Related Work

Over the years, several studies have been conducted to understand models of typosquatting, including various features of typosquatted domain names [10]. In the following, we review the technical anatomy of these various models and features prevalent in typosquatted domain names.

A. Typo-Generation Models

One of the first and widely cited approaches in the area of typo domain name generation was introduced in 2006 by Wang *et al.* [12], where the following five typo-generation models were commonly used in the wild:

1) **Missing-dot typos:** the dot following "www" is removed, *e.g.,*, `wwwSouthwest.com`.
2) **Character-omission typos:** one character is omitted, *e.g.,*, `Diney.com` (a typo of the *Disney* brand).
3) **Character-permutation typos:** two consecutive characters are swapped, *e.g.,*, `NYTiems.com`.
4) **Character-substitution typos:** characters are replaced by their adjacent ones on a specific keyboard layout, *e.g.,*, `DidneyWorld.com` ("s" → "d").
5) **Character-duplication typos:** characters are mistakenly typed twice, *e.g.,*, `Googlle.com`.

Later studies, such as Banerjee *et al.* [3], looked at exhaustively generating typo domains using other methods:

6) **N-mod-inplace:** substitutes N characters in the original domain name with all possible alphabet letters.
7) **N-mod-inflate:** increases the length by N characters.
8) **N-mod-deflate:** removes N characters from the original domain name (or URL).

B. Features of Typosquatted Domains

Domain Name Length. Early observations showed that most typosquatted domain names had less than 10 characters [3]. However, it was later shown in [7] that no matter the length, typo domains within the Damerau-Levenshtein [4], [6] distance of one or adjacent-keyboard distance of one from popular domains were overwhelmingly typosquatted.

Domain Name Popularity. While Banerjee *et al.* [3] initially suggested that typosquatting decreases significantly with declining domain name popularity, newer studies by Szurdi *et al.* [11] and Agten *et al.* [2] concluded that 95% of typo domains target the "long tail" of the popularity distribution.

Effect of the Top-Level Domain (TLD). Since `.com` is the dominant TLD of all registered domain names, most studies confirm that `.com` domain names have a high chance of being typosquatted–either by modifying the second-level domain (SLD) portions (*e.g.,* `googlle.com`) or creating a malicious counterpart in another separate TLD (*e.g.,* `Netflix.om`).

III. Study: Identifying Typo Domains

Our user study presented subjects with a list of actual domain name URLs, a subset of which were modified to represent possible typosquatted domains. For each URL, subjects were asked to select "Yes" if it appears to be typosquatted or "No" for an authoritative domain name.

Objectives. The primary objectives of the user study are to 1) gauge the effectiveness of various techniques of typosquatting on users and 2) study the benefits on how security education can improve users' awareness of typosquatted domain names. Secondary objectives include: 1) understanding correlations between user demographics and the outcomes of typosquatting (whether they fall for it or not) and 2) determining features of successful typosquatted domains. In particular, we hope to answer the following questions:

- Are users more susceptible to typosquatted domain names containing certain kinds of typos (*e.g.,* missing characters) than others (*e.g.,* substituted characters)?
- Does security education play a role in helping users correctly identify typosquatted domain names?
- Does a user's demographic (*e.g.,* age, education) affect how they perceive typosquatted domain names?
- Do users *more easily* identify typosquatted domain names that target popular domains?
- Are certain types of typosquatted domain names (*e.g.,* alphanumeric) more susceptible than others?
- Does the TLD (*e.g.,* `.com`, `.uk`) affect a user's identification of a typosquatted domain name?

To answer these questions and achieve our objectives, we rely on a systematic method for the selection of domains and subjects, as well as experimental design and evaluation criteria. In the following, we elaborate on each of those aspects.

Selection of Domain Names. A total of 200 domain names were chosen from the Alexa top 1 million websites (globally). To favor more popular domain names, the entire list was split into four unequal partitions with the first partition representing the top 1,000 domains. Subsequent partitions were increased in size by a factor of 10 and then 50 domain names were randomly sampled from each partition. Ultimately, 93 (46.5%) candidate domains were randomly chosen to be typosquatted using a random technique from §II-A. Since *Model 8* is similar to *Model 2*, we only considered *Models 1-7* in our study.

Selection of Subjects. The participants of the study primarily consisted of University students, staff and researchers. Despite the lack of choice in participants, we strived to include a good representation of demographics that would address the questions raised in our study's objectives. Additionally, we attempted to include diverse sample characteristics (with respect to subjects) so that we can understand other objectives of the study (*e.g.,* whether security education, familiarity, or educational background, help identify typosquatted domains).

Design of the User Study. To asses how prior knowledge and awareness of security concepts affect a user's behavior, the user study encompassed three phases which incrementally introduced subjects to all of the typosquatting techniques discussed in §II-A. We deployed the user study online, allowing participants access at anytime with their own devices.

For Phase One, participants were presented with a brief introduction to typosquatting followed by a series of questions to gather the following demographical data: *Name, E-mail, Gender, Age, Education,* and *Familiarity of Security Concepts* (on a scale 1-5). Each subsequent page of the online survey presented 10 candidate domain URLs with a "Yes" or "No" choice to indicate a typosquatted domain name. Phases Two and Three followed a similar template as Phase One, except the order of candidate domain names were randomly shuffled.

IV. Results and Discussion

A total of 34 participants completed all three phases of the survey over a one-week period, receiving their score (out of 200) for the number of correct responses after each phase.

A. Evaluation Criteria

For our evaluation, we primarily observed two performance metrics among users: 1) correct responses and 2) the amount of time to complete each phase of the study. For a given domain name, a correct response is defined as to whether the user answered "No" if the given domain name was an authoritative domain name (*unaltered*) or "Yes" if the given domain name was indeed typosquatted (altered according to §II-A). We examined the total completion time for each user and calculated the average amount of time spent (in seconds) to answer each question of the 200-question survey.

As will be shown, users generally performed better (*i.e., correctly identified typosquatted domain names*) with each subsequent phase of the survey. However, if we drill down and examine the users' demographical data, we can see that variables such as their *Age* and *Education* affect not only how they perceive potentially typosquatted domain names–but also how long they spend analyzing them. These interesting findings are discussed further in §IV-C.

B. Participant Scores and Completion Time

With each phase, the average number of correct responses improved and the average response time decreased slightly. For Phase 1, scores ranged from 78 to 186 correct responses with a mean, standard deviation and variance of 142.2, 23.6 and 557.1, respectively. In Phase 2, the minimum score increased to give us a range from 110 to 188 (Mean=147.1, s.d.=18.6,

variance=345.7). Phase 3's minimum score increased slightly to range of 117 to 183 (Mean=149.9, s.d.=15, variance=225).

Those results are interesting in several ways. First, the more subjects were educated about the security problem at hand, the faster they became at identifying typosquatted domains. Second, subjects also became better at identifying those domains.

C. Demographics

Age. The ages of the participants ranged from 22 to 39 (Mean=25, s.d.=4.1, variance=16.5). Interestingly, younger participants generally scored higher than older participants across all phases of the study. However, surprisingly, while the younger participants scored higher, they also spent more time per question on average compared to their older counterparts.

In the fields of Psychology and Optometry, it is generally understood that older adults often take longer to read than young adults. As Paterson *et al.* [9] points out, this age-related difference is due to optical changes and changes in neural transmission that occur with increasing age. Conversely, the results of our study show a trend where older participants spent less time per survey question on average than their younger counterparts. This trend of spending more time on each question could explain why the younger participants scored better, as the older participants appeared less patient and tended to perform worse at identifying typosquatted domain names.

Education Level. The majority of participants were students with University staff making up the rest of the test subjects. Of the participants, there was only 1 High School Graduate and 1 participant who reported some College Education. For the rest, 17 participants (50%) had a Bachelors degree, 13 participants (38.2%) had a Masters degree, and 2 held Doctorate degrees (5.88%). Participants holding higher degrees of education actually *scored worse* than participants with less education.

Familiarity of Security Concepts. On a scale of 1 to 5 in the familiarity of security concepts, only 1 participant chose a value of "2". 15 participants (44.1%) chose the middle value of "3", 14 participants (41.1%) chose the higher value of "4", and the remaining 4 participants (11.8%) stated they were very familiar with security concepts by choosing "5". Naturally, the self-identification of one's familiarity with security coincides with how well they performed as scores generally increased.

D. Domain Name Features

Domain Ranking. As mentioned previously, the Alex top 1M domain names were split into four unequal partitions to favor popular domain names during sampling. As expected, participants were more successful in correctly identifying typosquatted domain names that targeted popular domains.

Typosquatting Model. As shown in Fig. 1, participants were very likely to distinguish typosquatting that used *Model 1* (Missing-dot), *Model 5* (Character-duplication) and *Model 7* (1-mod-inflate). The models that caused most participants to incorrectly identify typosquatted domain names were *Model 2* (Character-omission) and *Model 6* (1-mod-inplace). Essentially, users tend to correctly identify typosquatting which adds

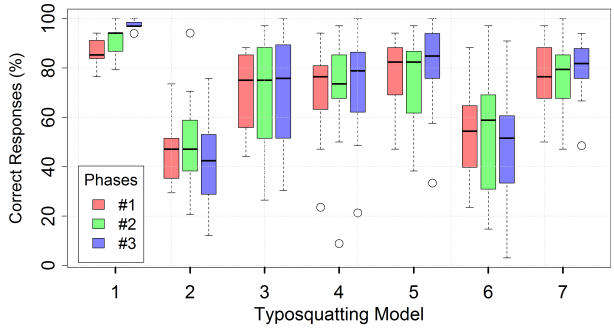

Fig. 1. Correct Responses Per Typosquatting Model listed in §II-A.

characters (*e.g.,* duplicate or random) while *the most effective typosquatting involves permutations and substitutions*.

To ascertain why certain typosquatting techniques are more effective than others, we can look at studies in Cognitive Science. For example, Grainger and Whitney [5] highlight the fact that a text composed of words whose inner letters have been re-arranged can be can be `raed wtih qutie anazimg esae`! This so-called "jumbled word effect" is due to the special way in which the human brain encodes the positions of letters in printed words.

Domain Name Type. Given our sample size of 200 domain names, 167 (83.5%) were made up of all alphabetic characters with the remaining 33 (16.5%) containing alphanumeric characters. Our results show that participants were more likely to identify a domain name that contained all alphabetic characters as opposed to alphanumeric characters.

Top-Level domain. According to the Internet Assigned Numbers Authority (IANA), the organization who delegates administrative responsibility of TLDs (Top-Level Domains), there are different "groups" of TLDs which include [1]:

- infrastructure top-level domain (ARPA)
- generic top-level domains (gTLD)
- restricted generic top-level domains (grTLD)
- sponsored top-level domains (sTLD)
- country code top-level domains (ccTLD)
- test top-level domains (tTLD)

According to the list above, the "original" TLDs such as `.com` and `.net` that were created early in the development of the Internet are now grouped under the "generic" category. For our purposes, we will consider the following TLDs as "historic", since they are widely-known to the average Internet user: `.com`, `.org`, `.net`, `.int`, `.edu`, `.gov`, and `.mil`. Out of our sample size of 200 domain names, 119 (59.5%) fall into the "historic" TLD group, 75 (37.5%) fall into the "country-code" TLD group, and 6 (3%) fall into the "generic" TLD group. Our participants performed better when presented with a domain name with a "historic" TLD than domain names from either the "generic" or "country-code" groups.

Typosquatting difficulty. Table I lists the top 10 domain names which caused the most incorrect responses (averaged over all phases) for the participants of the study, which coincides with our earlier statement that the most effective

TABLE I
TOP 10 INCORRECTLY IDENTIFIED DOMAINS (unmodified domains shown in gray).

Typo Domain	Authoritative Domain	Typosquatting Model	Rank	Phase 1 Correct	Phase 2 Correct	Phase 3 Correct	Average Correct
---	onlainfilm.ucoz.ua	---	6,345	24%	9%	9%	14%
ngbus.com	tgbus.com	6 (1-mod-inplace)	998	24%	15%	3%	14%
afg.com	avg.com	4 (keyboard-sub)	366	24%	9%	21%	18%
---	umblr.com	---	506	18%	26%	15%	20%
egadget.com	engadget.com	2 (char-omission)	403	29%	21%	12%	21%
vc.cn	ivc.cn	2 (char-omission)	1,778	29%	24%	18%	24%
zasgames.com	oasgames.com	6 (1-mod-inplace)	7,942	32%	29%	15%	26%
hispress.com	hespress.com	6 (1-mod-inplace)	536	41%	26%	24%	31%
---	05tz2e9.com	---	5,988	32%	29%	36%	33%
rudupoint.com	urdupoint.com	3 (char-permutation)	443	44%	26%	30%	34%

techniques involve permutations and substitutions.

The first-most incorrectly identified site could be attributed to the fact that most participants were based in the United States and therefore unfamiliar with the `.ua` TLD, which is the ccTLD for Ukraine. However, it should be pointed out that the fourth-most incorrectly identified domain name, `umblr.com`, turned out to be an edge case that is in fact an actual typosquatted domain. Most participants most likely thought it was a typosquatted variant of `tumblr.com`, a popular microblogging and social networking website. During the study design phase, our algorithm selected the domain `umblr.com` but did not actually modify it with a typosquatting model and subsequently marked it as *not* typosquatted. According to WHOIS, the domain `umblr.com` is actually owned by "Tumblr, Inc." which makes this a perfect example of a defensive registration against potential typosquatters.

V. CONCLUSIONS & RECOMMENDATIONS

This study has allowed us to gain valuable insight into the effectiveness of various typosquatting techniques and how security education affects the behavior of users. Our results confirm that participants generally performed better and faster at identifying typosquatted domain names *after* being educated about typosquatting models between each phase of the study.

Recommendations. Based on the results of our user study, we offer some recommendations for strengthening the defenses against typosquatting. As demonstrated by the improved scores with each subsequent phase of the study, we can confidently say that thoroughly educating users about all known typosquatting techniques will surely help them fend off against malicious domain names. As more organizations and businesses adopt security training programs in this day and age, it would be most beneficial to incorporate a training module that specifically explores typosquatting in more detail (perhaps alongside the commonly-taught *Phishing* attacks).

The results of the study, pertaining to the particular features of the domain names that were used, can most certainly aid in the design of a defense system that uses heuristic analysis. While typical defense systems use blacklists (*e.g., Google Safe Browsing*), a heuristic-based defense system that dynamically analyzes URLs can incorporate the findings in this study to help "rank" potentially malicious domain names. For example, domain names from a gTLD or ones with alphanumeric characters can be "flagged" for closer inspection since users are more likely to fall victim to their typosquatted variants.

Future work. Looking forward, we will continue to pursue research that will utilize the findings in this study to build suitable defenses against typosquatting, taking into account vital aspects such as a user's behavior and cognitive ability. For example, we can incorporate *user profiles* into a typosquatting defense system that considers frequently-visited domain names (thereby automatically populating blacklists with typosquatted variants). We will also explore gathering additional data by conducting the user study with a custom-developed mobile application, which provides participants even further freedom to complete the survey anytime using their own devices.

VI. ACKNOWLEDGMENT

This work was supported by the Global Research Lab (GRL) Program of the National Research Foundation (NRF) funded by Ministry of Science, ICT (Information and Communication Technologies) and Future Planning (NRF-2016K1A1A2912757).

REFERENCES

[1] —. Root Zone Database. http://bit.ly/1TBSeck, 2015.
[2] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse. In *NDSS*, 2015.
[3] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. Cyber-Fraud is One Typo Away. In *IEEE INFOCOM*, 2008.
[4] F. J. Damerau. A technique for computer detection and correction of spelling errors. *CACM*, 1964.
[5] J. Grainger and C. Whitney. Does the huamn mnid raed wrods as a wlohe? *Trends in cognitive sciences*, 8(2):58–59, 2004.
[6] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
[7] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In *FC*, 2010.
[8] L. H. Newman. Be Careful. Mistyping a Website URL Could Expose You to Malware. http://slate.me/1Pey1m3, 2016.
[9] K. B. Paterson, V. A. McGowan, and T. R. Jordan. Effects of adult aging on reading filtered text: Evidence from eye movements. *PeerJ*, 1:e63, 2013.
[10] J. Spaulding, S. Upadhyaya, and A. Mohaisen. The Landscape of Domain Name Typosquatting: Techniques and Countermeasures. http://arxiv.org/pdf/1603.02767.pdf, 2016.
[11] J. Szurdi, B. Kocso, G. Cseh, M. Felegyhazi, and C. Kanich. The Long Taile of Typosquatting Domain Names. In *USENIX Security*, 2014.
[12] Y. Wang, D. Beck, and J. Wang. Strider typo-patrol: discovery and analysis of systematic typo-squatting. *USENIX SRUTI*, 2006.