Understanding the Effectiveness of Typosquatting Techniques

Jeffrey Spaulding
University of Central Florida
jeffrey.j.spaulding@gmail.com

DaeHun Nyang
Inha University
nyang@inha.ac.kr

Aziz Mohaisen
University of Central Florida
mohaisen@ucf.edu

ABSTRACT

The nefarious 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 stealing traffic for monetary gain. Typosquatting has existed for well over two decades and continues to be a credible threat to this day. As recently shown in the online magazine *Slate.com* [19], cybercriminals have attempted to distribute malware through `Netflix.om`, a typosquatted variant of the popular streaming site `Netflix.com` that uses the country code top-level domain (ccTLD) for Oman (`.om`).

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. Ultimately, we found that subjects tend to correctly identify typosquatting which adds characters to the domain names, while the most effective techniques to deceive users involves permutations and substitutions of characters. We also found that subjects generally performed better and faster at identifying typosquatted domain names after being thoroughly educated about them, and that certain attributes such as *Age* and *Education* affect their behavior when exposed to them.

1 INTRODUCTION

In today's connected world, billions of devices are able to access the Internet to allow their users to connect and exchange information. At the heart of this global information infrastructure is the Domain Name System (DNS), which allows people to use text-based domain names instead of numeric addresses to provide a distributed directory service. With nearly three hundred million registered domain names (as of late 2015), the Domain Name System 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 [23].

Typosquatters employ several techniques in the wild to register domain names that can sufficiently capture enough traffic for monetary gain. For example, a typosquatter may target a popular domain name and register an *identically similar* one in which only a single character is added or substituted. This is demonstrated by the famous case outlined in [21] where typosquatters targeted the immensely popular social-networking site *Facebook* to produce domain names such as `www.fagebook.com` and `www.facewbook.com`

Much of the research in identifying and understanding typosquatting generally falls under two approaches: 1) identifying typosquatted domain names from domain name registration information and 2) identifying typosquatted domain names through network traffic analysis. One of the first large-scale studies in the first approach was conducted in 2003 by Edelman [10], who located more than 8,800 registered domains that were minor typographical variations of popular domain names. A few years later, Wang *et al.* [24] introduced a system called the *Strider Typo-Patrol* for systematically identifying typo domains via "typo-generation models". Subsequent studies including Banerjee *et al.* [7] and Edelman [11] utilized the typo-generation models in [24] to produce a corpus of typographical variations of popular domain names, which were ultimately verified by domain registration look-ups or automated web crawlers. Recent studies such as Szurdi *et al.* [22] examined a wider scope of popular domain names while Agten *et al.* [6] examined the nature of typosquatting over time. Rather than validating potentially typosquatted domain names through registration records, the most recent study by Khan *et al.* [15] introduced a novel approach called the *conditional probability model*. This model essentially identifies typosquatted domains by investigating which domains have high proportions of visitors leaving for more popular domains with lexically-similar names soon after landing.

The prior work described above has proven to be valuable in understanding the landscape of typosquatting. First, it highlights the prevalence of the problem with direct empirical evidence on how domain names are being typosquatted. Second, it quantifies the various techniques utilized by these adversaries for generating typographical variations of the domains they target. Additionally, several of these studies have paved the way for introducing countermeasures to combat the problem. Defensive registration, for example, is perhaps the best countermeasure which involves domain name owners to register lexically-similar domain names to eliminate the opportunity for typosquatting altogether.

While these previous studies have showcased the prevalence of various typosquatting techniques, they fail to account for certain realities. For example, examining DNS queries can produce very useful data through DNS tracing–but it does have limitations. First,

HotWeb'17, San Jose / Silicon Valley, CA, USA

DOI: 10.1145/3132465.3132467

domain name queries are not necessarily the result of actual queries by users, but rather from browser pre-fetching or automated web crawling. In addition, the typical analysis of DNS traffic was based on a short period of collection time, from which the identification of typosquatted domain names is as only good as the users who explicitly query for them.

Since DNS is inherently complex and diverse, it is difficult to exactly quantify the relevance of typosquatting techniques at any point in time due to acquiring DNS zone files for several domains. For example, obtaining the top-level domain (TLD) zone file for `.com` requires approval from Verisign [5] which updates them daily (at the time of this writing, 120,517 new domain names were added in the `.com` TLD alone [4]). This problem is further exacerbated with the adoption of the new gTLD program, which provides even more opportunities for registering domain names and thus the potential for typosquatting [1]. Even worse, the registration of a domain name is generally a low-cost activity–to the point where typosquatted domain names are disposable resources: once the adversary is done with the malicious activity, the domain names are most likely blocked or de-registered (within the grace period of registration; often up to 7 days allowed by registries). In conclusion, while there is fair amount of work on the problem at hand, none of such work measured how users behave when exposed to a typosquatted domain name.

To ultimately address the issues described above, we chose to design and implement a user study which would provide several benefits: 1) it can be made scalable, 2) it can provide insight into how various (theoretical) typosquatting techniques compare to each other, 3) it can determine why certain people fall for typosquatting while others do not (*e.g.,* demographics), and 4) it can help discover the effectiveness of various techniques to mitigate the prevalence of typosquatting (*e.g.,* security education and its benefits)–which could hopefully be useful for developing strategic defenses.

Contributions. 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. **C1.** We validate typosquatting techniques presented in prior studies by examining their prevalence using various carefully sampled domains from several data sources. **C2.** We experimentally demonstrate how security education and awareness of cybercrimes, particularly typosquatting, will affect the behavior of Internet users. **C3.** 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. **C4.** We publicly release our data so others can verify and build upon our research findings and results.

2 BACKGROUND AND RELATED WORK

While typical Internet users may not have come across the term *typosquatting*, they may have inadvertently stumbled upon a typosquatted domain when they either typed the wrong URL into their web browser's address bar or clicked on an external hyperlink. The actual term *typosquatting* was coined in the late 1990's [12] to describe a new trend appearing alongside *cybersquatting*, a notorious practice where opportunistic individuals preemptively registered domain names in the hopes of selling them back to companies and trademark owners for a substantial profit. Over the years, several studies have been conducted to understand models of typosquatting, including various features of typosquatted domain names. In the following, we review the technical anatomy of these typosquatting models and provide an overview of the various features prevalent in typosquatted domain names.

2.1 Identifying Typosquatted Domains

Prior studies conducted on 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 studies 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 1 summarizes these several approaches which includes the 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, we describe the models that generated typo variations of an authoritative domain.

Typo-generation models. One of the first and widely cited approaches for typo domain name generation was introduced in 2006 by Wang *et al.* [24], where the following five typo-generation models were used in the wild:

(1) **Missing-dot:** this typo happens when the dot following "www" is forgotten, *e.g.,* `wwwSouthwest.com`.
(2) **Character-omission:** this typo happens when one character in the original domain name is omitted, *e.g.,* `Diney.com` (a typo of the Disney brand).
(3) **Character-permutation:** this typo happens when two consecutive characters are swapped in the original domain name, *e.g.,* `NYTiems.com`.
(4) **Character-substitution:** this typo happens when characters are replaced in the original domain name by their adjacent ones on a specific keyboard layout, *e.g.,* `DidneyWorld.com`, where "s" was replaced by the QWERTY-adjacent character "d".
(5) **Character-duplication:** this typo happens when characters are mistakenly typed twice (where they appear once in the original domain name), *e.g.,* `Googlle.com`.

While this previous study presented the first attempt to systematically understand the most prevalent typosquatting techniques 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.* [7] suggested the following for generating typosquatted domains:

(6) **1-mod-inplace:** the typosquatter substitutes a character in the original domain name with all possible alphabet letters.
(7) **1-mod-inflate:** the typosquatter increases the length of a domain name (or URL) by one character. Unlike in [24] characters are added based on distance (*e.g.,* using a keyboard layout), this work considers all characters as potential candidates.
(8) **1-mod-deflate:** similar to the approach in [24], this typo happens when a typosquatter removes one character from the original domain name (or URL).

Table 1: Summary of typo domain identification approaches. † Notice that 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 [24]	Alexa Top 10,000	(1) Missing-Dot	10,000	51% (5,094)
	Alexa Top 30	(1-5)	3,136	71%(2,233)
	MillerSmiles† Top 30	(1-5)	3,780	42%(1,596)
	Top 50 Children's Sites	(1-5)	7,094	38%(2,685)
Keats 2007 [14]	Top 2,771 (*Various Sources*)	(1-5)	1,920,256	7% (127,381)
McAfee Labs 2008 [11]	Top 2,000 (*Unknown Source*)	*Unknown*	*Unknown*	80,000
Banerjee 2008 [7]	Top 900 (*Various Sources*)	(6-8)	~3 million	35%
Moore 2010 [18]	Alexa Top 3,264	(1-5)	1,910,738	~49%(938,000)
Szurdi 2014 [22]	Alexa Top 1 million	(1-5)	~4.7 million	~20%
Agten 2015 [6]	Alexa Top 500	(1-5)	28,179	61% (17,172)

Certain aspects of the techniques proposed in [7] can be viewed as generalization of the techniques proposed in [24]. 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 [9, 16] 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 the Hamming distance).

2.2 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 (§2.2.1), popularity of domain names (§2.2.2), popularity of top-level domain (TLD) (§2.2.3), and domain landing behavior (§2.2.4).

2.2.1 Domain Name Length. While investigating if domain name length affects the chances of being typosquatted, Banerjee *et al.* [7] 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 [18], the authors show that no matter the length of the popular domain, typo domains within the Damerau-Levenshtein distance of one or an 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.* [6], 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. Our study, on the other hand, confirms that users are equally likely to fall for typosquatted domains regardless of their length.

2.2.2 Domain Name Popularity. Another feature of domain 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 (*i.e.,* the number of visits by unsuspecting users). The results of Banerjee *et al.* [7] 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.* [22], 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.* [6] also confirms this trend, suggesting a shift in trends and behaviors of typosquatters.

2.2.3 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.* [7] 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.* [6] 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 (*i.e.,* cheaper domain names are more likely to be typosquatted).

Additionally, the TLD portion of a domain name may also be a target for exploitation. As mentioned previously, typosquatters have targeted popular `.com` domain names by registering a similar domain name in the country code TLD (ccTLD) for Oman (*e.g.,* `Netflix.om`). Furthermore, `.com` domains may have a malicious `.org` counterpart unbeknownst to the original registrant of the `.com` domain. A noteworthy example of this was mentioned in [8], where unsuspecting viewers inadvertently typed `www.whitehouse.com` instead of `www.whitehouse.gov` and were exposed to questionable content in lieu of the official United States White House website. Banerjee *et al.* [7] 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.

2.2.4 Probability Models for Domain Landing. The 2015 study by Khan *et al.* [15] 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 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.

3 STUDY: IDENTIFYING TYPO DOMAINS

To gauge the effectiveness of typosquatting techniques discussed in §2.1 that are prevalent in the wild, we conducted a user study. Subjects were presented with a list of actual domain name URLs where a subset of them were modified using all of the techniques shown in §2.1 to represent possible typosquatted domain names. The subjects were asked to indicate that for each given domain name URL, select "Yes" if it appears to be a typosquatted domain name or "No" if it is an authoritative domain name.

3.1 Objectives

The primary objectives of the user study are to 1) gauge the effectiveness of various techniques of typosquatting domain names on users and 2) study the benefits on how security education can improve users' awareness of typosquatting. Secondary objectives are: 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:

Q1. Are users more susceptible to typosquatted domain names containing certain kinds of typos (*e.g.,* missing characters) than others (*e.g.,* substituted characters)?
Q2. Does security education play a role in helping users correctly identify typosquatted domain names?
Q3. Does a user's demographic (*e.g.,* age, education) affect how they perceive typosquatted domain names?
Q4. Do users *more easily* identify typosquatted domain names that target popular domains?
Q5. Are certain types of typosquatted domain names (*e.g.,* alphanumeric) more susceptible than others?
Q6. Does the TLD (*e.g.,* `.com`, `.uk`) affect a user's identification of a typosquatted domain name?

To answer these questions and achieve the objectives stated above, 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.

3.2 Selection of Domain Names

The list that was presented to each subject comprised of 200 domain names that were chosen from the Alexa top 1 million websites (globally). Rather than sample domain names randomly, the entire collection of 1 million domain names were split into four unequal partitions with the first partition representing the top 1,000 domain names. Subsequent partitions were increased in size by a factor of 10 and then 50 domain names were randomly sampled from each partition to bring us a total of 200 candidate domain names. This method allowed us to favor domain names that appeared towards the popular end of the spectrum.

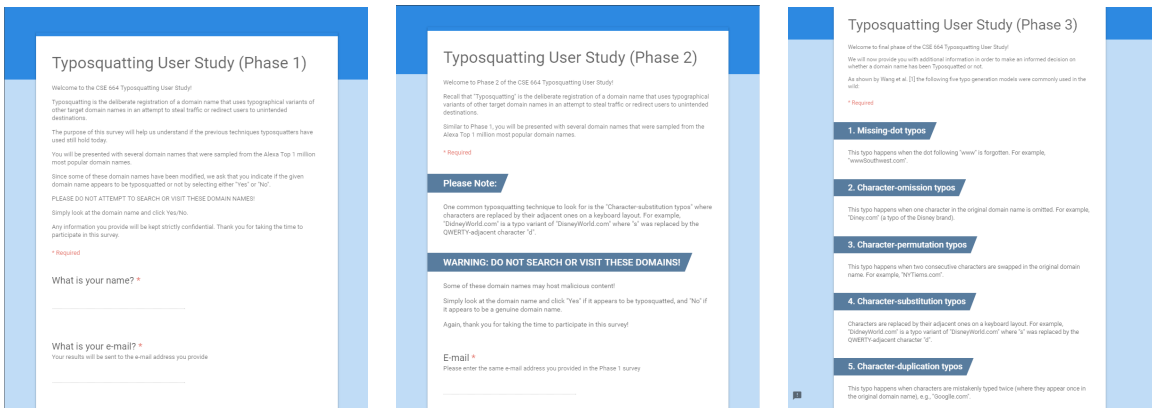
Next, we simply iterated through the candidate domains and randomly decided if it should be typosquatted or not. If chosen to be typosquatted, a candidate domain is then modified using one of the techniques from §2.1, which was also randomly chosen. Ultimately, 93 (46.5%) candidate domains were randomly chosen to be typosquatted, with Table 2 showing a breakdown of how many candidate domains were modified using a particular typosquatting model. Since *Model 8* (1-mod-deflate) is similar to *Model 2* (Character-omission), we only considered *Models 1-7* in our study.

Table 2: Number of candidate domains per model

Typosquatting Model	# of Domains
1 (missing-dot)	15
2 (character-omission)	11
3 (character-permutation)	12
4 (keyboard-substitution)	11
5 (character-duplication)	19
6 (1-mod-inplace)	12
7 (1-mod-inflate)	13
Total Domains Modified:	93

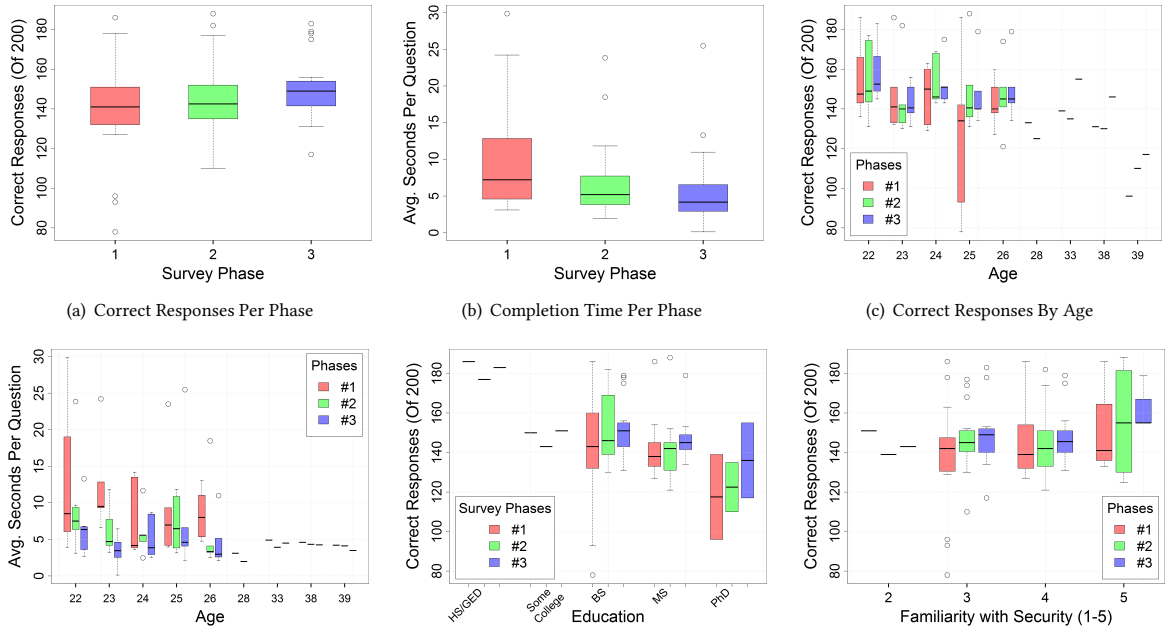
3.3 Selection of Subjects

The participants of the study primarily consisted of University students, followed by University staff and researchers. Despite the lack of choice in participants, we strove to include a good representation of demographics that would address the questions raised in our study's objectives (*i.e., Does a user's demographic affect how they perceive typosquatted domain names?*). Additionally, we attempted to



(a) Phase 1. (b) Phase 2. (c) Phase 3.

Figure 1: Screenshots of each phase of the user study. Notice that each phase introduces the participants to increasing levels of knowledge about typosquatting.



(a) Correct Responses Per Phase (b) Completion Time Per Phase (c) Correct Responses By Age

(d) Completion Time By Age (e) Correct Responses By Education (f) Correct Responses By Familiarity

Figure 2: Plots of the user study results, highlighting various correlations: identification of the typosquatted domains is positively correlated with security education and familiarity, and negatively correlated with response time, age, and education.

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 domain names).

3.4 Design of the User Study

To assess 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 §2.1. To deploy the user study, we created an online survey form so each participant could complete the survey anytime they wish using their own computers and devices.

For Phase One, participants were given a URL link to the online survey website and were presented with a brief introduction to typosquatting (omitting any of the previously-discussed techniques from §2.1). Following the introduction section, a series of questions were included to gather the following demographical data: *Name, E-mail, Gender, Age, Education,* and *Familiarity of Security Concepts* (on a scale 1-5). Before participants could proceed to the next page of the online survey, they must enter the current time (this was

necessary in order to calculate the amount of time each subject completed the survey). Each subsequent page of the online survey presented 10 candidate domain URLs with a "Yes" or "No" choice to indicate a typosquatted domain.

Phases Two and Three followed a similar template as Phase One, except they only asked for the e-mail provided in Phase One (to uniquely identify subjects) and the current time. Additionally, the same corpus of candidate domain name URLs used in Phase One are shown *except the order in which they appear are randomly shuffled*. To provide subjects with more knowledge about typosquatting techniques, the introduction sections of Phases Two and Three include typo-generation models mentioned in §2.1.

4 RESULTS AND DISCUSSION

In our user study, we recruited 34 participants who completed all three phases of the survey over a one week period. After completing each phase of the survey, each participant received an automated e-mail containing the total number of correct responses (out of 200) which forms their score.

4.1 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 §2.1). 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 in the next section, 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 our demographical results outlined in §4.3.

4.2 Participant Scores and Completion Time

As illustrated in Figure 2(a), the average number of correct responses improved with each subsequent phase. For Phase One, the 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 Two, the minimum score increased to give us a range from 110 to 188 (Mean=147.1, s.d.=18.6, variance=345.7). Phase Three's minimum score increased slightly to range of 117 to 183 (Mean=149.9, s.d.=15, variance=225). As depicted in Figure 2(b), the average number of seconds per response decreases slightly with each phase. For Phase One, the average elapsed time per response ranged from 3.1 to 29.8 seconds (Mean=9.6, s.d.=6.7, variance=44.5). In Phase Two, the minimum average time per response decreased to give us a range from 1.9 to 23.8 seconds (Mean=6.7, s.d.=4.7, variance=21.8). Phase Three's minimum score also decreased slightly to range of 0.1 to 25.5 (Mean=5.5, s.d.=4.6, variance=20.8).

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 (*i.e.,* better identification rate).

4.3 Demographics

Age. The ages of the participants ranged from 22 to 39 (Mean=25, s.d.=4.1, variance=16.5). As we can see in Figure 2(c), younger participants generally scored higher than older participants across all phases of the study. Interestingly, Figure 2(d) shows that even though 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.* [20] 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 in the survey 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.

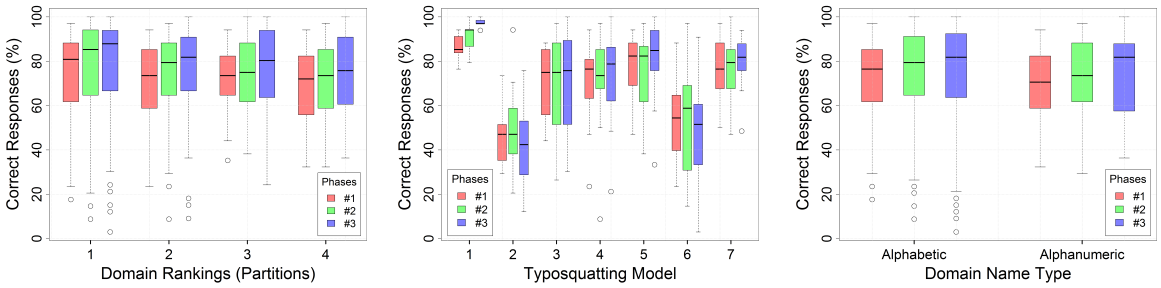
Education level. The majority of participants were University students with University staff making up the rest of the test subjects. Of the participants, there was only 1 High School Graduate and 1 participant who reported some College Education. For the rest, 17 participants (50%) had a Bachelors degree, 13 participants (38.2%) had a Masters degree, and 2 held Doctorate degrees (5.88%). As shown in Figure 2(e), 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". As we can see in Figure 2(f), the self-identification of one's familiarity with security concepts coincides with how well they performed as scores generally increased.

4.4 Domain Name Features

Domain ranking. As mentioned previously, the entire collection of 1 million domain names were split into four unequal partitions with *Partition 1* representing the top 1,000 domain names. Subsequent partitions were increased in size by a factor of 10, so *Partition 2*, *Partition 3*, and *Partition 4* represented the ranges: [1,001-10,000], [10,001-100,000], and [100,001-1,000,000], respectively. As expected, Figure 3(a) demonstrates that participants were more successful in correctly identifying typosquatted domain names that targeted popular domains.

Typosquatting model. Since *Model 8* (1-mod-deflate) is similar to *Model 2* (Character-omission), we only considered *Models 1* through *7* in our study. As shown in Figure 3(b), participants were very likely to distinguish a typosquatted domain name 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



(a) Responses By Domain Ranking (b) Responses By Typo Model (c) Responses By Domain Type

Figure 3: Plots of the user study results. Notice the great variability in the identification of typosquatted domains across the different models, and the slight increase in chances of accepting typosquatted domains that are alphanumerical.

correctly identify typosquatting which adds 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 [13] 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. Figure 3(c) illustrates 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, there are different "groups" of TLDs which include [3]: 1) infrastructure top-level domain (ARPA), 2) generic top-level domains (gTLD), 3) restricted generic top-level domains (grTLD), 4) sponsored top-level domains (sTLD), 5) country code top-level domains (ccTLD), 6) test top-level domains (tTLD).

The initial set of TLDs that were created in the early development of the Internet (*e.g.,* `.com` and `.net`) are now grouped under the "generic" category mentioned above. For our purposes; however, we labeled the following TLDs under the "historic" group, since they are widely-known to the average Internet user: `.com`, `.org`, `.net`, `.int`, `.edu`, `.gov`, and `.mil`. As a result, our sample size of 200 domain names were only categorized into three groups where 119 (59.5%) fell into the "historic" TLD group, 75 (37.5%) fell into the "country-code" TLD group, and 6 (3%) were in the "generic" TLD group. As illustrated by Figure 4, 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 3 lists the top 10 domains which caused the most incorrect responses (averaged over all phases) for the participants, which coincides with our earlier statement that the most effective 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

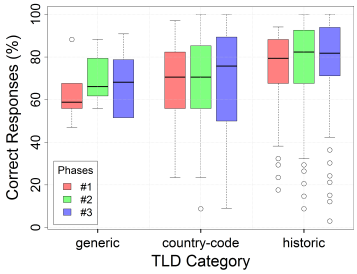


Figure 4: Correct responses by TLD type.

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 records, the domain name `umblr.com` is actually owned by "Tumblr, Inc." which makes this a perfect example of a defensive registration against potential typosquatters.

Table 4 lists the top 10 domain names that were correctly identified (averaged over all phases) by the participants of the study. Again, this coincides with that fact that users easily spot typosquatted domain names which adds characters–especially if they target popular domain names such as `google.*` or `blogger.com`.

5 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 our results, 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

Table 3: 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%

Table 4: Top 10 correctly identified domains (unmodified domains shown in gray).

Typo Domain	Authoritative Domain	Typosquatting Model	Rank	Phase 1 Correct	Phase 2 Correct	Phase 3 Correct	Avg. Correct
googlje.dz	google.dz	7 (1-mod-inflate)	370	97%	100%	94%	97%
---	wayfair.com	---	568	94%	97%	100%	97%
blogger.comm	blogger.com	5 (char-duplication)	72	91%	97%	100%	96%
---	office.com	---	63	94%	94%	100%	96%
---	audible.com	---	840	94%	100%	94%	96%
wwweromode.net	eromode.net	1 (missing-dot)	697,652	94%	94%	97%	95%
---	popsugar.com	---	837	91%	94%	100%	95%
syosetu.comm	syosetu.com	5 (char-duplication)	930	94%	97%	94%	95%
wwwiklan-oke.com	iklan-oke.com	1 (missing-dot)	688,829	91%	97%	94%	94%
financial-spread-bettin.gcom	financial-spread-betting.com	3 (char-permutation)	729,388	85%	97%	100%	94%

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 our findings to help "rank" potentially malicious domains. 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. We will continue to pursue research that will utilize the findings in this study to build 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 they wish using their own devices, as well as explore typosquatting in new gTLDs [17].

Acknowledgment. This work was supported by NSF under grant CNS-1643207 and the Global Research Lab (GRL) Program of the National Research Foundation (NRF) funded by Ministry of Science, ICT and Future Planning (NRF-2016K1A1A2912757).

REFERENCES

[1] —. FTC Warns That Rapid Expansion of Internet Domain Name System Could Leave Consumers More Vulnerable. http://1.usa.gov/1LVjyQ5, 2011.
[2] —. History Behind .COM. http://bit.ly/1Uhdba0, 2015.
[3] —. Root Zone Database. http://bit.ly/1TBSeck, 2015.
[4] —. Daily DNS Changes and Web Hosting Activity. http://bit.ly/2i9Lqs0v, 2017.
[5] —. TLD Zone File Access Program. http://bit.ly/1pxjRrv, 2017.
[6] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse. In *NDSS*, 2015.
[7] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. Cyber-Fraud is One Typo Away. In *IEEE INFOCOM*, 2008.
[8] C. G. Clark. The Truth in Domain Names Act of 2003 and a Preventative Measure to Combat Typosquatting. *Cornell Law Review*, 2004.
[9] F. J. Damerau. A technique for computer detection and correction of spelling errors. *CACM*, 1964.
[10] B. Edelman. Large-Scale Registration of Domains with Typographical Errors. http://bit.ly/1IEGvql, 2003.
[11] B. Edelman. Typosquatting: Unintended Adventures in Browsing. *McAfee Security Journal*, 2008.
[12] 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.
[13] J. Grainger and C. Whitney. Does the huamn mnid raed wrods as a wlohe? *Trends in cognitive sciences*, 8(2):58–59, 2004.
[14] S. Keats. What's In A Name: The State of Typo-Squatting 2007. http://bit.ly/1mqonpI, 2007.
[15] 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.
[16] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
[17] A. Mohaisen and K. Ren. Leakage of. onion at the dns root: Measurements, causes, and countermeasures. *IEEE/ACM Transactions on Networking*, 2017.
[18] T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquatting. In *FC*, 2010.
[19] L. H. Newman. Be Careful. Mistyping a Website URL Could Expose You to Malware. http://slate.me/1Pey1m3, 2016.
[20] 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.
[21] C. Roth, M. Dunham, and J. Watson. Cybersquatting; typosquatting Facebook's $2.8 million in damages and domain names. http://bit.ly/1SnahSF, 2013.
[22] J. Szurdi, B. Kocso, G. Cseh, M. Felegyhazi, and C. Kanich. The Long Taile of Typosquatting Domain Names. In *USENIX Security*, 2014.
[23] M. Thomas and A. Mohaisen. Kindred domains: detecting and clustering botnet domains using DNS traffic. In *WWW, Companion Volume*, 2014.
[24] Y. Wang, D. Beck, and J. Wang. Strider typo-patrol: discovery and analysis of systematic typo-squatting. *USENIX SRUTI*, 2006.