Typosquatting
A More Comprehensive View.

Josh Engelsma & Andrew Kalafut
Grand Valley State University, School of Computing and Information Systems

Contact Information:
School of Computing and Information Systems
Grand Valley State University
1 Campus Drive, Allendale MI, 49401, United States

Email: engelsjo@mail.gvsu.edu

Abstract

Typosquatting is the practice of registering domains very similar to legitimate domains with hopes that people will land on the aforementioned domain via a typo. Typo domains are used maliciously to park ads and phish information from users. Defensive typo registrations are made by the owners of genuine web sites to prevent exploitation of users trying to access the legitimate site. Our research aims at quantifying the extent to which typosquatting occurs and to determine the actual cost of defending one's site. While previous research has focused predominantly on quantifying typosquatting within the .com top level domain or amongst the more popular .com sites, our research looks more broadly across most general top level domains - approximately 400, (including .net, .org, .biz, .mobi, and .name). Our research also looks across a larger spectrum of .com domains. Examining typosquatting across all top level domains gives us a superior view of the true cost of typosquatting.

Main Objectives

1. Revisit previous work on the extent to which typosquatting occurs on popular sites (eg. Google, Facebook, Twitter etc.)
2. Push research further by examining the extent of typosquatting on less popular sites and across less popular top level domains than .com (approximately 400 plus top level domains such as .fishing, .dating, .wedding, .pictures, .lawyer etc.)
3. Classify whether a typo is malicious (phishing, malware, ad parking) or defensive (gooogle.com is defensively registered for google.com)
4. Attempt to quantify the true cost of adequately defensively registering one's popular website.
5. Warn / Aid the public by exposing the risks of not investing in defensively registering one's site.
6. Aid in fighting against dishonesty and fraud.

Methodology

In performing our research, we first obtained hundreds of zone files for the numerous top level domains. Each zone file contains information about all of the domains that exist underneath the respective top level domain (eg. the .com zone file contains information for facebook.com, google.com, instagram.com, and all other domains the end with .com).

In addition to the collection of zone files, we obtained a list of the top one million most popular websites. This list is referred to as the Alexa List. We use the Alexa List to provide us with a baseline of legitimate domain names.

Using the python programming language, scripts were written to individually navigate a random sampling of domains from within the Alexa List top one million domains. For each of the domains that we sampled, we first generated all possible distance one typos (typos that vary from the original domain by one character). Next, for each of the generated typos (g-typos) for a particular domain, we determined if it was actually a registered domain by looking throughout the zone files and by doing a name server lookup. If the typo domain was registered, we assigned the typo domain to the original legitimate domain as a candidate typo (c-typo).

With a list in place of all the randomly selected domains from the Alexa List with their respective c-typos, we moved towards finding information for both the legitimate domain and its c-typos - that would help us classify whether or not the candidate was malicious.

The information we chose to target for each domain includes: a hash of the page HTML, a perceptual hash (similarity based hash) of the screenshot image of the page, number of redirects, creation date, whether or not the domain was privacy protected, whether or not the domain was hosted using one of 15 popular parking servers, and the name server of the domain. Using these features we plan to categorize the various c-typos for each domain as malicious (phishing, ads, malware), defensive (redirection to intended site), or coincidental.

In order to collect the textual hash of the domain's HTML, the perceptual hash of the webpage screenshot, and the number of redirects for a domain, we programmatically load up every single domain using the PhantomJS webcrawler. Phantom provides the ability to download the HTML of a webpage to a file, save a screenshot of the page to a png, and determine the number of redirects. All other features were discovered using the Unix WHOIS protocol and by parsing information from the zone files.

Basic Process Flow

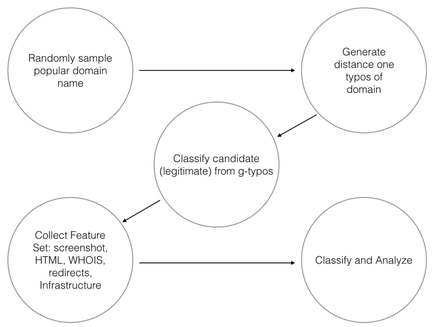

Figure 1: System Overview

Results

While research for this project is ongoing, we present the following initial results and analyzation to our data collection.

For this first chart, we divide the candidate typos up into two main categories - candidates found within the same top level domain and candidates found across different top level domains. On average we find that candidates are more likely to be found within the same TLD (16 per domain) than across different TLDs (4 per domain). We also see that the number of candidates can greatly deviate from the average as *youtube.com* has 456 candidates with the same TLD and *amazon.es* has 301 candidates across TLDs.

	Average	Max
Within TLD	16	456
Across TLD	4	301

Table 1: Candidated within/across TLDs

By looking at the CDF below, we see that around 45 percent of our legitimate domains have only 1 or less candidate domains. We can conclude from this that while the average number of candidate domains per legitimate domain is relatively high (20 per domain) there are a significant number of domains that have a very small number of candidate domains.

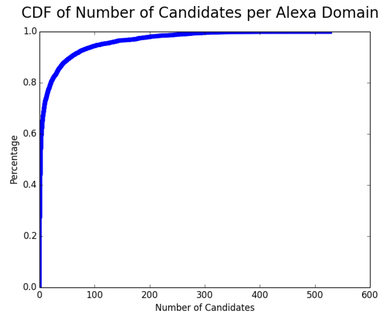

Figure 2: Percentage of Domains With x Candidates or Less

The chart below tells us the average number of redirects for candidates (indicating maliciousness for large numbers) and the percentage of candidate domains hosted with popular parking name servers (indicating a parking page). The legitimate domains we sampled averaged 0.54 redirects - lower than the averages of our candidates which come in at .78 and .62 redirects per domain. Surprisingly, less than 1 percent of candidates within TLDs and across TLDs were hosted on one of the more popular parking Name Servers.

Candidate-Type	Average Redirects	Percentage Parking NS
Within TLD	0.783232678894	0.712711945589
Across TLD	0.622566690699	0.0558386263698

Table 2: Average Redirects and Parking NS for Candidates

Finally, the table below provides us with a view on the textual matching of candidates to their legitimate Alexa Domains. The threshold we chose to assign was 50 percent or more similar. Our results show that roughly 1000 domains have 0 percent of candidates at or above a threshold match of 50 percent similarity and that roughly 1000 domains have 100 percent of candidates at or above a threshold match of 50 percent similarity (of a sample size of 4000 Alexa List domains).

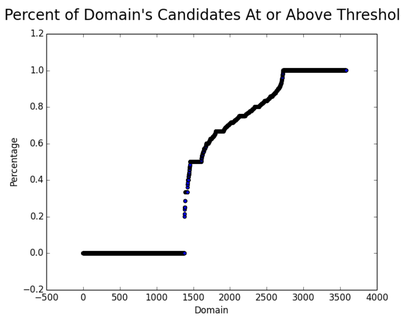

Figure 3: Candidate Textual Similarity

Conclusions

- While previous research focused mostly on a small set of popular domains, our research shows that typosquatting also occurs on less popular sites - abeit at a much smaller level. We show that popular and non-popular sites have multiple candidate domains with similar textual hashes (indicating similar content), and a larger number numbers of redirects (indicating malicious or phony behavior).

Forthcoming Research

Future research will involve extracting more useful information from our data set. We would like to further analyze the relationships of the candidate domains we are finding into either parking pages, malicious pages, or coincidentally related pages. In particular, rather than looking at the data across all of the domains, we would like to analyze similarities and difference amongst domains that are very popular and domains that are less popular. Doing this will enable us to more accurately assign a cost for the average user looking to host a site for purposes such as a local e-commerce store.