Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse

Pieter Agten*, Wouter Joosen*, Frank Piessens* and Nick Nikiforakis†
* iMinds-DistriNet, KU Leuven,
{firstname}.{lastname}@cs.kuleuven.be
† Department of Computer Science, Stony Brook University,
nick@cs.stonybrook.edu

***Abstract*—Typosquatting is the act of purposefully registering a domain name that is a mistype of a popular domain name. It is a concept that has been known and studied for over 15 years, yet still thoroughly practiced up until this day. While previous typosquatting studies have always taken a snapshot of the typosquatting landscape or base their longitudinal results only on domain registration data, we present the first *content-based*, *longitudinal* study of typosquatting. We collected data about the typosquatting domains of the 500 most popular sites of the Internet every day, for a period of seven months, and we use this data to establish whether previously discovered typosquatting trends still hold today, and to provide new results and insights in the typosquatting landscape. In particular we reveal that, even though 95% of the popular domains we investigated are actively targeted by typosquatters, only few trademark owners protect themselves against this practice by proactively registering their own typosquatting domains. We take advantage of the longitudinal aspect of our study to show, among other results, that typosquatting domains change hands from typosquatters to legitimate owners and vice versa, and that typosquatters vary their monetization strategy by hosting different types of pages over time. Our study also reveals that a large fraction of typosquatting domains can be traced back to a small group of typosquatting page hosters and that certain top-level domains are much more prone to typosquatting than others.**

I. Introduction

Domain names and the underlying domain name resolution protocol are arguably one of the linchpin technologies that have allowed the modern web to expand to its current dimensions. Even though users increasingly rely on search engines to find interesting and relevant content, domain names are as important as ever. This is exemplified by ICANN's constant rollout of hundreds of new Top-Level Domains (TLDs) such as `.xxx`, `.guru`, and `.email`, which were created to allow institutes and individuals to obtain relevant domain names that are long unavailable in the overcrowded traditional TLDs.

The importance of domain names has not gone by unnoticed by unscrupulous individuals who wish to profit at the expense of others. In the nineties, some people started registering domain names including trademarks and brand names not belonging to them, with the hope of later selling them to their rightful owners at a higher price. This practice was named *domain squatting* and many variations of this type of attack have emerged over the years, with perhaps the most popular and exploited type being *typosquatting*.

In typosquatting, an attacker abuses the fact that real human users may mistype a URL while typing it in their browser's address bar or email client. As such, a typosquatter can register `vacebook.com` and capture the traffic of users who mistype `facebook.com` and would otherwise receive an error in their browsers. As a matter of fact, in May 2013, Facebook was awarded 2.8 million dollars in damages caused by typosquatting, as well as over 100 typosquatting domains that were registered and monetized by typosquatters [22].

The prevalence of typosquatting has attracted the attention of multiple researchers who attempted to map the typosquatting landscape, identify the domains that are targeted the most, and discover the preferred monetization strategies of typosquatters [2], [8], [16], [23], [25]. What these studies have in common, is that they either characterize a snapshot of typosquatting activity through a single crawling effort over a limited period of time, or limit the longitudinal aspect of their study to domain registration data only, without investigating changes in the domains' content.

In this paper, we present the first content-based typosquatting experiment that studies the typosquatting phenomenon longitudinally, i.e., *in time*. Instead of reporting on a single snapshot of the typosquatting landscape, we performed a seven-month-long experiment in which we visited the typosquatting domains targeting the 500 most popular sites of the Internet every day. Through the collection of more than 900 GB of typosquatting data, our study allows us not only to measure typosquatting at a large scale but also to investigate the *changes* of typosquatting domains over time, allowing us to answer questions that could not be answered with a single snapshot of typosquatting activity.

Among other results, we find that, even though 95% of the most popular domains on the Internet are targeted by typosquatters, most of them do *not* use defensive registrations as a means of protecting their identity and their clients. We also find that a large fraction of all possible typosquatting domains for short popular authoritative domains is already registered, and that typosquatters are hence increasingly targeting longer

NDSS '15, 8-11 February 2015, San Diego, CA, USA

http://dx.doi.org/10.14722/ndss.2015.23058

domains. Making use of the longitudinal aspect of our study, we discover that typosquatters are actively switching between monetization strategies for the domains that they own, and are also on the look-out for expiring registrations of popular domain names. We also show that 50% of all typosquatting domains can be traced back to just four typosquatting page hosters. Finally, on the policy-side, we observe that differences in domain price setting and the availability of out-of-court domain dispute resolution procedures between different TLDs, have a significant effect on the prevalence of typosquatting.

Our main contributions are:

- We report on the first content-based longitudinal study of typosquatting abuse, consisting of over 900 GB of data gathered over a period of seven months.
- We verify whether previously discovered typosquatting trends still hold today.
- We provide new results and insights in the typosquatting landscape, based on both the static and longitudinal aspects of our data.
- We show that the adoption of strict policies and easy dispute-resolution procedures from registries, can decrease typosquatting abuse.

The rest of this paper is structured as follows. In Section II, we provide background information on typosquatting in general and on the way our data gathering experiment was set up. Section III describes the main results found by our experiment. Section IV describes related work and finally Section V concludes.

II. Background

A. Typosquatting Models

The most frequently occurring domain name typos are those that have a Damerau-Levenshtein distance of one from a popular domain name [3], i.e., domain names resulting from a single character insertion, deletion, substitution or adjacent character permutation from a popular domain. When the inserted or substituted character is adjacent to the original character on a QWERTY keyboard, we say the typosquatting domain also has a "fat-finger distance" of one [16]. In 2006, Wang et al. categorized typosquatting domains into five different categories [25]. Based on those categories and assuming the authoritative domain `example.com` and intended URL `www.example.com`, we consider the following five typosquatting models for our study:

1) **Missing-dot typos**: The dot following "www" is forgotten, e.g., `wwwexample.com`
2) **Character-omission typos**: One character is omitted, e.g., `www.exmple.com`
3) **Character-permutation typos:** Consecutive characters are swapped, e.g., `www.examlpe.com`
4) **Character-substitution typos:** Characters are replaced by their adjacent ones, given a specific keyboard layout, e.g., `www.ezample.com` where "x" was replaced by the QWERTY-adjacent "z".
5) **Character-duplication typos:** Characters are mistakenly typed twice, e.g., `www.exaample.com`

While one can likely come up with more ways of including typos in a domain name, e.g., a wrong domain TLD, in this work, we limit ourselves to the typosquatting domains that can be generated following the aforementioned typo models. We also limit ourselves to domains resulting from a *single* application of *one* of these models, since those are more likely to be typed by a user than domains containing multiple typos.

B. Data Gathering

To gather the data required for our longitudinal study, we set up two automated crawlers, which where supplied with the Alexa top 500 domains of April 1, 2013 as input. The first crawler generates the typosquatting domains for each authoritative domain in the input, according to the aforementioned models. For each authoritative and generated domain, the crawler first determines whether the domain resolves to an IP address. If so, the crawler visits the web page hosted on the domain using PhantomJS[1], a headless JavaScript-enabled web browser. After loading the web page, the crawler waits for 10 seconds, allowing the page to load dynamic content or perform a redirect. Finally, the crawler saves the IP address, final URL, HTML body and a screenshot of the page to disk. The crawler was configured to process the entire list of domains daily, for a period of 7 months starting at April 1, 2013 and running until October 31, 2013. The crawl was duplicated onto a second machine to provide redundancy in case of system failure. In order to prevent excessive resource usage and to minimize the chance of our crawlers being blocked by typosquatting domains, duplicate typosquatting domains were filtered out and the rate at which the crawlers visit domains was set to the minimum value that still allows a crawl to finish within a small margin of 24 hours. In total, 28,179 potential typosquatting domains were generated, out of which 17,172 resolved to an IP address at least once during our study.

The second crawler was configured to perform a WHOIS lookup for every domain ever successfully resolved by the HTTP crawler. The WHOIS responses (if any) were parsed using Ruby Whois[2] and then saved to disk. The crawler was configured to process all domains once per week, over the same time period as the HTTP crawler. The slower one-week crawl interval was needed to respect the acceptable use policies of the queried WHOIS servers.

One irregularity that occurred during our data gathering period is that the crawl rate of our crawlers had to be increased during the week of August 21, to accommodate for a planned power interruption of our crawling machines on August 27 and 28. The possible effects of this irregularity are discussed in Section III-B.

C. Analysis

Our crawlers collected over 900 GB worth of data during the data gathering period, consisting of 3,389,137 web pages and 424,278 distinct WHOIS records. In order to analyze this data, we classified the collected pages into the categories listed in Table I. The third column of this table indicates for each category whether we consider it a legitimate, malicious or undetermined use of a domain name. Although most of

[1] http://phantomjs.org/
[2] http://ruby-whois.org/

the malicious categories are actually in a legal gray area, we mark them as malicious because these practices are set up to *deceptively* extract profit from users' mistypings.

As a first step towards classifying the collected pages into the selected categories, we tried to automatically divide the pages into clusters. Since some types of typosquatting pages (such as generic ad parking pages) are likely to target many different authoritative domains, we decided to cluster the pages based on visual appearance rather than domain-specific properties such as the page's domain name or the corresponding authoritative domain. To measure the visual appearance of a page, we used the concatenation of a perceptual hash of the page's screenshot and a locality-sensitive hash of its HTML body. The perceptual hash of an image is a fixed-length bit vector that represents a fingerprint of that image. Unlike cryptographic hashes, which give a drastically different output for small changes in the input, perceptual hash functions are designed to produce similar outputs for similar inputs. A locality-sensitive hash is based on the same principle, but works on textual data instead of multimedia files. These hashes allow us to programmatically compare the screenshots and HTML bodies of the collected pages, in order to group them into clusters of similar pages. The concrete perceptual and locality-sensitive hash algorithms we used are the aHash[3] and the Nilsimsa[4] algorithm respectively, which were selected based on an evaluation of their performance on a small subset of the dataset that was clustered manually. Using these hash functions and a custom clustering algorithm based on fastcluster [17], the 3,389,137 collected web pages were automatically grouped into 8,102 clusters.

After this initial clustering, we spent approximately two and a half man-months performing an extensive manual analysis of the data. The goals of this manual analysis were twofold: i) to improve the automatic visual clustering results, and ii) to categorize the clusters into the categories of Table I. In order to assign the appropriate category to each page, clusters of visually similar pages were, in many cases, split up into smaller clusters. For instance, the cluster of pages looking like `amazon.com` was split up into sub-clusters for authoritative pages, affiliate abuse pages and ad parking pages. To facilitate the manual analysis, we developed a custom web application that presents all of the collected data, i.e., the screenshots, WHOIS records, IP addresses and final URLs of the visited web pages, in a structured way. Ultimately, 40% of all clusters were classified, representing 95% of all collected pages. Because of this second-phase manual analysis, we have high confidence in the quality and accuracy of our data. Our dataset has been made available for download at https://distrinet.cs.kuleuven.be/software/typos15/.

D. Types of Abuse

In this section we briefly describe the non-self-explanatory categories listed in Table I.

1) Affiliate abuse: We consider a page to be performing affiliate abuse when it redirects its visitors to a legitimate website, taking advantage of an affiliate program offered by that legitimate site. Affiliate programs are arrangements in which a website owner (the advertiser) pays a commission to a third party (the affiliate) for sending traffic to her website. For instance, `amazon.com` pays a commission for every purchase made by visitors coming from websites participating in their affiliate program. To identify what traffic comes from which affiliate, each affiliate is assigned a unique identifier that she should specify in the URLs toward which she forwards her visitors. Sometimes an intermediate company sits in between the advertiser and the affiliate to handle the technical issues of organizing an affiliate program and to make it easier for advertisers and affiliates to find each other. Most affiliate programs have strict conditions that limit the ways in which affiliates are allowed to bring traffic to the advertiser's website. Banners and hyperlinks to the advertiser's site are allowed, but automatic forwarding typically is not. For instance, Amazon explicitly disallows automatic forwarding.

An example of affiliate abuse by typosquatting that is active at the time of this writing, is the following: users who mistype `match.com` as `ma5ch.com` ("`t`" substituted by the QWERTY-adjacent "`5`") are eventually brought back to the `match.com` domain, but the typosquatting page appends an affiliate identifier to the URL when it redirects the user's browser from the typosquatting domain to the authoritative one. As such, the owners of the authoritative domain will now have to pay an affiliate commission to the typosquatter, for a visit that should have been theirs in the first place.

Identifying affiliate abuse is not always easy. In particular, a naive analysis cannot differentiate affiliate abuse from so-called *defensive* registrations, which are typosquatting domains proactively registered by an authoritative domain owner, to prevent abuse from typosquatters. Both types of domains forward their visitors to the authoritative domain, typically adding an identifying parameter to the forwarding URL. In the case of affiliate abuse, this parameter is used to identify the affiliate, while in the case of defensive registrations it is typically used to identify the forwarding domain (for traffic analysis purposes). Because of this similarity, we did not attempt to automatically identify affiliate abuse, relying on our second-phase *manual* analysis for this instead. For this manual analysis, we took into account several factors to identify affiliate abuse, including (1) the names of parameters added to the forwarding URL, (2) the values of these parameters, (3) whether the authoritative site advertises the fact that it has an affiliate program, and (4) the WHOIS records of the authoritative and typosquatting domains. We did not attempt to distinguish between typosquatting pages abusing an affiliate program of their own authoritative domain or an affiliate program of other unrelated domains, e.g., a typosquatting domain of `target.com` redirecting the user to the affiliate program of `amazon.com`.

2) Scam: A scam page is a page that tries to trick users into performing an action that is undesirable for the user and profitable for the attacker. Two popular types of scams are "surveys" and malicious advertisements (malvertising). In surveys, users are asked to perform a series of steps in return for some reward, for example a $100 coupon for a big box store. The steps almost always involve users entering their email address, name, phone number and potentially physical address. These details can then be used to subscribe users to

[3]https://github.com/JohannesBuchner/imagehash

[4]http://ixazon.dynip.com/~cmeclax/nilsimsa.html

TABLE I. THE COLLECTED TYPOSQUATTING PAGES WERE CLASSIFIED INTO CATEGORIES, BASED ON THE MOST LIKELY INTENT OF THE PAGE. THE THIRD COLUMN INDICATES THE CATEGORY TYPE: L STANDS FOR LEGITIMATE, M STANDS FOR MALICIOUS AND U STANDS FOR UNDETERMINED.

Category	Description	T
Authoritative	Pages redirecting to or displaying the authoritative domain without any abuse	L
Coinciding	Pages containing legitimate content that happen to reside on a typosquatting variant of an authoritative domain	L
Protected	Pages notifying the user that she made a typo and/or link to the authoritative domain	L
Ad parking	Pages that have no content other than showing advertisements	M
Adult content	Pages showing adult/pornographic content	M
Affiliate abuse	Pages taking advantage of an affiliate program offered by another domain (see Section II-D1)	M
For sale	Pages that have no content other than being advertised as for sale	M
Hit stealing	Pages redirecting to a legitimate domain without abusing an affiliate program	M
Scam	Pages persuading the user to enter personal information or to download malware (see Section II-D2)	M
No content	Pages that have no content (e.g., blank pages or pages under construction)	U
Server error	Pages displaying an error, which was caused by a server-side problem	U
Crawl error	Pages for which the crawler failed or that explicitly block the crawler's IP address	U
Other	Unclassified pages and pages that do not fall into any of the above categories	U

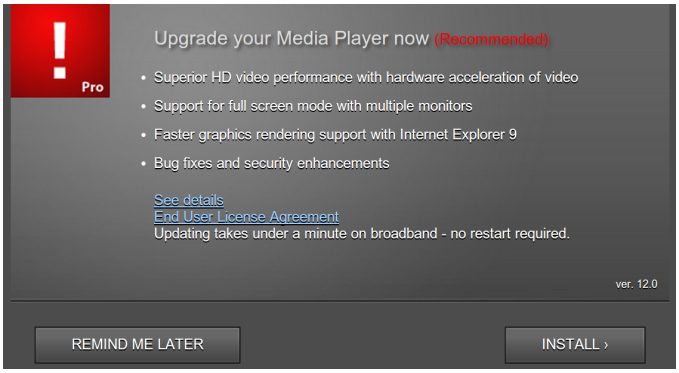

Fig. 1. A scam page trying to trick users mistyping `youtube.com` into downloading malware.

spam lists, expensive mobile services, and even potentially sold to larger data aggregators.

In malvertising, the scam page is trying to convince the user to willingly download and execute a malicious program. Fig. 1 shows the ad we got when purposefully mistyping `youtube.com` as `outube.com`. If the user downloads and installs the purported software update, she will be infected with malware (11/51 virus engines at virustotal.com identified the downloaded executable as malicious).

III. RESULTS

A. Malicious vs. Defensive Registrations

Our data indicates that typosquatting is still very prevalent for the list of authoritative domains we considered. Out of these 500 domains, 477 have at least one malicious typosquatting domain. We considered a domain to be malicious when it is classified as such for at least 7 days during the data gathering period. These numbers indicates that on the attack side, typosquatters have no trouble registering and exploiting typosquatting domains, despite long-standing anticybersquatting legislation [1].

On the defense side, trademark owners can protect themselves against typosquatting by proactively making defensive typosquatting domain registrations whenever they register an authoritative domain. Many registrars provide a service to automatically register a wide range of possible cybersquatting domain names when a trademark owner wants to register a domain. Nevertheless, our data shows that only 156 of the authoritative domains in our list have defensive domain registrations, meaning that 344 domains (representing 68.8% of the 500 most popular sites of the Internet) have no defensive registrations whatsoever. Thus, anyone who makes a typo for these domains and does not receive an error, is sure to land on a malicious typosquatting page.

The top 3 of authoritative domains with the most *defensive* registrations consists of `huffingtonpost.com` with 57 defensive domains, `americanexpress.com` with 42 domains and `bloomberg.com` with 39 domains. The top 3 of authoritative domains with the most *malicious* typosquatting domains are `adultfriendfinder.com` with 132 typosquatting domains, `constantcontact.com` with 103 typosquatting domains and `odnoklassniki.ru` with 97 such domains. Alarmingly, out of the three banks in our top 500 list (`bankofamerica.com`, `hdfcbank.com` and `icicibank.com`), only `bankofamerica.com` has defensive registrations. This means that if a user enters a typo for the domain of one of the two other banks, she could easily land on a phishing page, thinking she entered the proper domain name of her bank. Although we did not encounter any phishing pages for these banks during our study, our data shows `hdfcbank.com` had 42 active malicious typosquatting domains, `icicibank.com` had 43, and `bankofamerica.com` had 46. Any of these domains could start hosting phishing pages at any time or redirect users to the websites of competing financial institutes.

It is surprising to see that, in a time where companies are estimated to spend 7% of their information technology budgets on security, and global cyber crime costs are estimated between $300 billion and $1 trillion [15], many companies do not bother to make any defensive registrations at all for their domains. In particular, one would expect the financial sector to take a leading role in protecting their reputation and their customers. It seems these companies are either not aware of the problem, or simply do not care about it. The fact that large Internet companies such as Microsoft [21] and Facebook [22] are successfully contending with cybersquatters through defen-

sive domain registrations and legal actions, demonstrates that the problem is real and should not be ignored.

B. Daily Typosquatting Domain Count

Taking the Alexa top 500 domains as input, our HTTP crawler generated 28,179 potential typosquatting domains, based on the models discussed in Section II-A. From those *potential* typosquatting domains, 17,172 domains (61%) at least once during our data gathering period resolved to an IP address that was hosting a web page. From those *active* typosquatting domains, 13,526 domains (79%) were hosting malicious content for at least one day. Fig. 2 illustrates the daily growth in the number of discovered domains. The crossed-out line shows the total number of malicious typosquatting domains discovered up until a certain date. Since this is a cumulative measure, this line is strictly increasing. The figure shows an increase from about 11,000 domains at the start of the period to 13,526 domains at the end, indicating that typosquatters are continuously acquiring new typosquatting domains.

The solid line in the same figure shows the daily number of *active* typosquatting domains serving malicious content. This line does not follow the monotonic increase of the crossed-out line, indicating that typosquatters are not only registering new domains, but are also getting rid of domains at approximately the same rate. On average, there are 10,510 active malicious typosquatting domains per day, which amounts to 21 such domains per authoritative domain. Apart from three significant dips, which are discussed below, the number of active malicious typosquatting domains remains fairly constant over time.

As can be seen from the figure, the first dip in the number of active malicious typosquatting domains took place from June 4 to 6. Our data indicates this dip can be attributed to a single large cluster of ad parking pages that contained significantly fewer domains during these days. This cluster, which is the single largest cluster of our data set, has an average of 2,136 domains per day, but contained only 1,049 domains per day during these three days. Furthermore, 97% of the domains that dropped out during this period resolved to a single subnet in the days before, namely `208.73.210.0/23`, an IP address block in an autonomous system owned by Oversee.net. These results indicate that a single large ad parking hoster, related to Oversee.net, for some reason blocked our crawlers or was having network connectivity problems.

The second dip took place from August 21 to 25 and is preceded by a more subtle reduction in the number of active typosquatting domains in the weeks before. Our data shows that 1,675 domains that were previously hosting ads started serving blank pages or showed an error during this period (see also Fig. 3). More than 90% of those domains were parked with Sedo.[5] The dip coincides with the period during which our crawl rate was increased (see Section II-C), which leads us to believe that some of Sedo's ad parking servers blocked our crawlers due to the increased traffic.

On October 18 and 20, the number of active malicious typosquatting domains again drops by a significant amount. Our data indicates this dip is again due to Sedo servers temporarily returning blank pages or showing errors, although we could not find any specific reason for these servers to block our crawlers during these days.

[5]http://www.sedo.com/

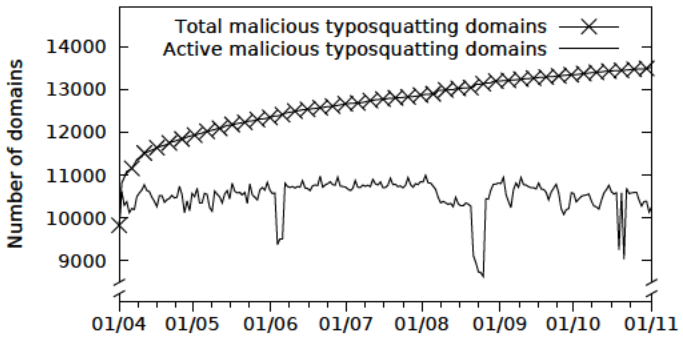

Fig. 2. The daily domain count shows an average of 10,500 active malicious typosquatting domains per day and a steady increase in the total number of discovered malicious typosquatting domains over time.

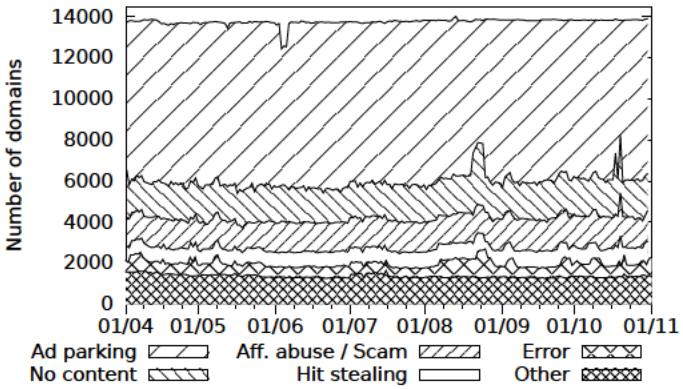

Fig. 3. Apart from the three dips discussed in Section III-B, the category distribution remains relatively constant over time.

C. Category Distribution

Fig. 3 shows how the collected typosquatting pages are distributed over the categories listed in Table I. To improve readability, some related or very small categories have been grouped together and the legitimate pages have been excluded from the graph. We can see that, apart from the dips discussed above, the distribution stays relatively constant over time. Table II shows the overall distribution of the visited typosquatting pages over all identified categories. Clearly, the ad parking category is by far the largest, confirming the results of Moore and Edelman [16]. The second largest category consists of the pages showing no content, including completely

TABLE II. THE DISTRIBUTION OF TYPOSQUATTING PAGES OVER ALL IDENTIFIED CATEGORIES. THE AUTHORITATIVE DOMAIN VISITS ARE EXCLUDED FROM THESE NUMBERS.

Category		Category	
Ad parking	50.55%	**Scam**	2.40%
No content	11.62%	**Adult**	1.90%
Affiliate abuse	6.89%	**For sale**	1.56%
Authoritative	5.37%	**Crawl error**	0.99%
Hit stealing	4.90%	**Other**	0.30%
Coinciding	4.88%	**Protected**	0.29%
Server error	3.03%	***Unclassified***	*5.20%*

blank or black pages and pages that label themselves as under construction. By manual analysis, we discovered that some of the domains showing blank and black pages did show content when visited through a foreign proxy server, indicating that some typosquatters were using IP blacklists or geolocation information to block our crawlers.

The affiliate abuse category is ranked third. During our manual analysis, we discovered that many of the pages in this category are using third-party services to hide the referring URL when abusing an affiliate program. For instance the typosquatting domain `hostgatkr.com` forwards its visitors to `tracking.warmmedia.com` (with a parameter indicating where to redirect to further on), which then redirects the visitor to a URL within the `hostgator.com` domain, specifying an affiliate id parameter. Because of the intermediate redirection, the exploited site `hostgator.com` sees visitors coming from `warmmedia.com` instead of `hostgatkr.com`. This makes it much harder for HostGator to discover that they are paying a commission for traffic that should have been theirs in the first place. Other commonly used redirection services we discovered are `trafficinterface.com` and `world-redirect.com`.

Hit stealing is ranked fifth in Table II but is the third largest *malicious* category. We discovered two distinct types of domains in this category. The first type consists of domains owned by a competitor of the authoritative site that is being typosquatted. These domains typically just forward their visitors to the competitor's site, effectively stealing the traffic of the authoritative domain. While we saw this behavior mostly with adult sites, some non-adult sites are stealing hits from their competitors as well. For instance, the Russian search site `tochki.ru` has typosquatting sites registered for `google.com.ua`, `google.ru`, `rambler.ru` and `yandex.ru`. The other type of hit stealing domains are those owned by Internet marketing companies trying to draw traffic to the sites of their customers. These domains typically forward their visitors to unrelated pages, often changing the destination domain at regular time intervals or even on every visit. For instance, over a dozen Czech typosquatting domains of `google.cz` and `seznam.cz` were forwarding their visitors to a different legitimate Czech domain on every visit. Similarities in the way the forwarding is implemented on these different typosquatting domains and in the parameters specified in the destination URL, lead us to believe that all of these domains are owned by a single Internet marketing company that is using them to draw traffic to its customers. This is of course a questionable way of increasing the traffic to their customers' websites, since the typosquatting domains used for this purpose are typically unrelated to the various landing pages, making it unlikely that a redirected visitor will stay on the landing domain. As such, while the landing page will likely receive hundreds or thousands of extra visitors, it is doubtful that these visitors will be of value. One could argue that these companies are providing "typosquatting-as-a-service" rather than proper Internet marketing services.

D. Typosquatting Models

To investigate to what degree typosquatters and legitimate domain owners are aware of the different models that can be used to generate typosquatting domains, we calculated the number of registered domains for each of the five models described in Section II-A. Our data indicates there are 6,399 malicious and 376 defensive registrations for domains generated by the character substitution model, making it the most popular model in absolute numbers. This is to be expected however, because the character substitution model generates much more typosquatting domains for a given authoritative domain than any other model. Conversely, the missing dot model only generates a single possible typosquatting domain per authoritative domain. To accurately compare the popularity of the different typosquatting models, we should hence look at the *active fraction* of all possible typosquatting domains, per typosquatting model. That is, out of all typosquatting domains generated according to some model (starting from our list of authoritative domains), we look at how many of them are active (i.e., resolve to a web page). We call this relative measure the *typosquatting saturation* per typosquatting model. Fig. 4 shows the typosquatting saturation for each of the five models we consider, differentiating between defensive and malicious registrations. We can see that, by this measure, the character substitution model is actually the *least* popular model, i.e., out of all possible typosquatting domains generated by this model, less than 40% are in use. The figure shows that the missing dot model is the most popular model, for defensive as well as for malicious registrations. Our data indicates no significant change in the popularity of the different models over time.

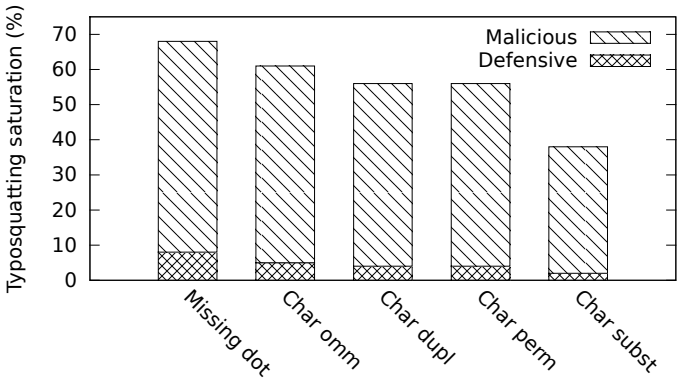

Fig. 4. The typosquatting saturation per typosquatting model. Although the substitution model has the most domains in absolute numbers, it is the least popular taking into account the number of domains generated by each model.

Note that if we would rank the five models based on their corresponding typosquatting saturation, we get the same ranking whether we consider malicious or defensive registrations. This suggests that attackers and defenders have a similar perception of what typosquatting domains are worthwhile to register.

E. Influence of Domain Name Length

Since four of the five typosquatting models we consider can be applied for each character in a domain name, the number of possible typosquatting domains for a given authoritative domain increases linearly with the length of the authoritative domain. Previous work published by Banerjee et al. [2] indicated that the *active* number of typosquatting domains for a given authoritative domain does not follow this relation: shorter domains were targeted much more frequently than longer domains. Our data shows that this is no longer the

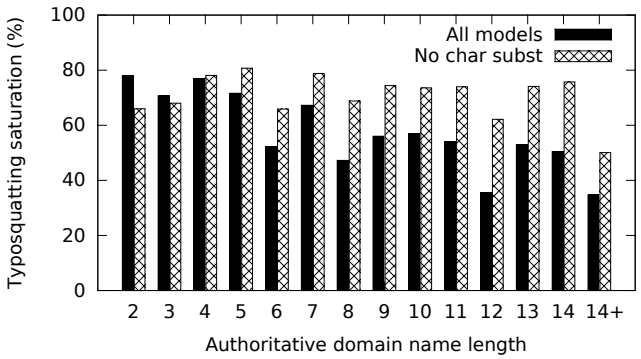

Fig. 5. Domains with short names suffer more from typosquatting, but this effect is not nearly as outspoken as in the '08 results of Banerjee et al. [2].

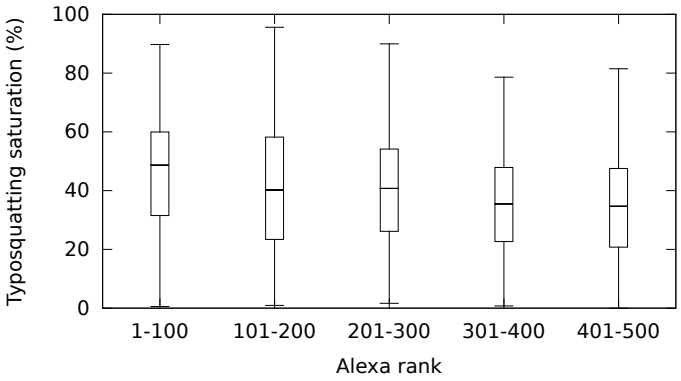

Fig. 6. Our data shows no significant correlation between Alexa rank and saturation.

case, as illustrated by Fig. 5. The solid black bars in this figure show the average typosquatting saturation per authoritative domain name length when taking all five typosquatting models into account. That is, the Y-axis shows the percentage of all possible typosquatting domains of authoritative domains of a certain length that are active (i.e., resolve to an IP address hosting a website). As mentioned in Section III-D, the typosquatting saturation is an averaging over time and in this case takes into account the fact that longer authoritative domains have more *possible* typosquatting domains. Although the solid bars show a decrease in typosquatting saturation as the authoritative domain length increases, this decrease is not nearly as outspoken as in the results of Banerjee et al. in 2008 [2], where the typosquatting saturation quickly drops to under 20% for domains longer than 5 characters.

Furthermore, since the number of *possible* typosquatting domains following the character substitution model rises very quickly as the domain length increases, and the previous section has shown that this is the least popular typosquatting model, we can automatically expect the typosquatting saturation to drop with increasing domain length. To remove this bias, the crossed-out bars in Fig. 5 discard the character substitution model. Here we can see even more clearly that there is no correlation between typosquatting saturation and authoritative domain length for the domains we investigated.

These results indicate that typosquatters have started targeting longer authoritative domains in the past six years. The most likely reason for this is that most short typosquatting domains were already in use: the figure illustrates that the average typosquatting saturation for domain names up to 8 characters is over 75%. A large fraction of the possible typosquatting domains of relatively short, popular websites is hence already registered.

F. Influence of Alexa Rank

Since typosquatters are trying to get as many page hits as possible, more popular authoritative domains are presumably targeted more often than less popular domains. Fig. 6 shows a box-and-whisker plot indicating the typosquatting saturation per Alexa rank. The ends of the whiskers show the minimum and maximum typosquatting saturation, the box shows the upper and lower quartiles and the line within the box shows the median typosquatting saturation. The plot indicates that the saturation within each rank bin varies widely and shows no significant correlation between rank and saturation. This contradicts the '08 results of Banerjee et al. [2], which indicated that the percentage of active typosquatting domains for a given authoritative domain reduces significantly with decreasing popularity, reaching only about 20% for the domain ranking 500 in their list of authoritative domains. Our results hence indicate typosquatters have started focusing on lower ranking domains in the past six years, in addition to the top ranking domains. These results are consistent with the recent findings of Szurdi et al. [23], who investigated the typosquatting activity for all `.com` domains in the Alexa top 1 million. Their study indicates that, although there is a positive correlation between typosquatting saturation and authoritative domain popularity, the typosquatting saturation is still at 40% near the Alexa 1 million rank.

To further investigate the influence of authoritative domain popularity on typosquatting activity, we investigated whether there is a correlation between the *change* in Alexa rank of an authoritative domain over our data gathering period, and the *change* in number of active malicious typosquatting domains of that authoritative domain. We did this by comparing the authoritative domain ranks and the number of active malicious typosquatting domains during the first week our data gathering period with the same figures during the last week of data gathering period, but we found no significant correlation.

G. Typosquatting Domain Volatility

One of the main objectives of our longitudinal study is to evaluate the volatility of the field of typosquatting. In particular, we would like to see (1) whether domains are changing hands from typosquatters to legitimate owners and (2) whether typosquatters vary the type of content they host on their domains. For this, we look at the number of category transitions per domain. To avoid overestimating the number of transitions, we assume that the content hosted on a domain does not change when we see a transition from a legitimate or malicious category to an undetermined category. That is, suppose a domain is in a legitimate or malicious category C at date x and is in an undetermined category at date $x+1$, then we still consider the domain to be in category C at date $x+1$. We used this methodology to generate the graph in Fig. 7, which hence gives a *lower bound* on the number of category transitions per domain over time.

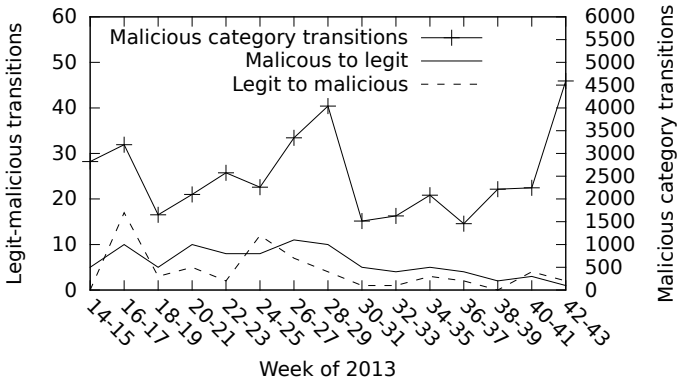

Fig. 7. The number of category transitions per domain. The two bottom lines use the left Y-scale, the top line uses the right Y-scale. Domains are moving from typosquatters to legitimate owners and vice versa, and are often changing category while being malicious.

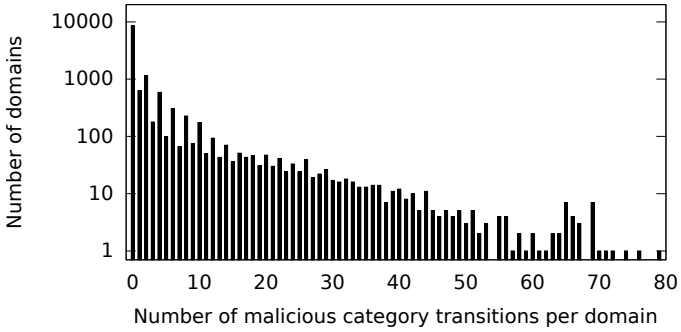

Fig. 8. Frequency distribution of the number of malicious category transitions per domain. Note that the Y-axis has a logarithmic scale.

The two bottom lines of this figure refer to the left-hand Y-scale and show the total number of malicious to legitimate or legitimate to malicious transitions. During our study, we saw an average of about 3 malicious-to-legitimate, and about 2 legitimate-to-malicious transitions per week. These numbers indicate legitimate owners are taking over domains from typosquatters and vice versa, albeit not in great numbers. With the exception of one domain, all domains that moved from malicious to legitimate stayed legitimate until the end of the data gathering period, and likewise for the legitimate to malicious transitions. The one exception is `os.com`, a domain being leased by `lexidot.com` and which transitioned from the "Coinciding" category to the "For sale" category and back, twice during the data gathering period. Some examples of malicious to legitimate transitions are `tumblr.com` taking over `timblr.com`, `umblr.com` and six other typosquatting domains. Some examples of transitions in the other direction are `livedoor.com` losing `livedooor.com` and `bleacherreport.com` losing `bleachereport.com`. For some of these legitimate to malicious transitions, the WHOIS records clearly indicate a change of hands, while for others the records are incomplete or simply do not change, which could indicate that those domains were already owned by typosquatters but were not being used for malicious purposes initially.

The upper line of Fig. 7 uses the right-hand Y-scale and shows the number of malicious category transitions, e.g., a transition from the "Ad Parking" category to the "Affiliate abuse" category. With an average of 1,239 transitions per week, we can see that these kinds of transitions occur much more frequently than legit-malicious transitions. If we average the total number of category transitions over the number of discovered typosquatting domains, we get a value of 2.84 transitions per domain over the seven month data gathering period. This means that, on average, a typosquatting domain serves pages from the same category for about 75 days in a row. However, in practice most domains did not change category at all during the data gathering period, while some changed very often. The bar chart in Fig. 8 shows the frequency distribution of the number of malicious category transitions per domain. The data indicates that 8,521 domains, representing 65% of all malicious typosquatting domains, stayed in the same category for the entire duration of the study, and 95% of all malicious typosquatting domains made less than 20 category changes. Nevertheless, there are some domains that switch malicious categories up to 79 times.

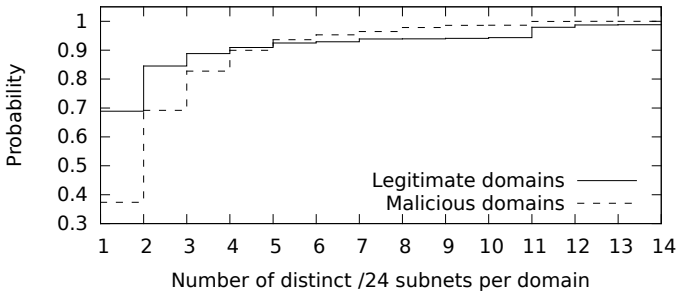

Fig. 9. Cumulative distribution function of the number of /24 subnets per domain. Malicious domains are more likely to resolve to multiple subnets over time than legitimate domains.

By manual inspection, we saw some typosquatting domains redirect to a different landing page on a regular basis, sometimes even on every visit. We posit that the domains operating in this fashion can be subdivided into two types. The first type of domains shows different categories of landing pages, often switching between ad parking, affiliate abuse and scamming pages. Typosquatters most likely use this scheme to diversify their monetization strategy, possibly even switching to the most profitable strategy dynamically. An example of a typosquatting domain of this type is `yuotube.com`, a domain that, at the time of writing, switches between pages classified under the ad parking, affiliate abuse and scam categories on every visit. The other type of domains shows many different pages over time, but all within the general hit stealing category. These domains are owned by the rogue Internet marketing company for drawing traffic to the sites of its customers, as discussed in Section III-C.

H. IP Address Statistics

In the previous section we found some typosquatting domains to be very volatile, i.e., changing typosquatting categories many times during the data gathering period. To see whether the IP address these domains resolve to also changes regularly, we investigated the number of distinct IP addresses and subnets associated with each typosquatting domain over time. Fig. 9 shows the cumulative density function of the number of distinct /24 subnets per domain, for both the legitimate domains and the malicious domains. We considered a domain

TABLE III. IP SUBNETS AND CORRESPONDING AUTONOMOUS SYSTEMS (AS) SERVING THE MOST MALICIOUS TYPOSQUATTING PAGES AND DOMAINS. TOGETHER THESE NETWORKS ACCOUNT FOR 36% OF ALL MALICIOUS PAGES AND 50% OF ALL MALICIOUS DOMAINS VISITED.

Network	AS owner	Nb visits	Nb domains
208.73.210.0/23	Oversee.net	259,781	2,405
199.59.243.96/28	Bodis	209,388	1,741
82.98.86.160/27	Sedo	187,174	1,388
69.43.161.128/25	Castle Access	140,098	1,216

to be legitimate when it is classified as such for at least 90% of the data gathering period, and we again considered a domain to be malicious when it is classified as malicious for at least 7 days. We only take into account the 24 most significant bits of an IP address to group together addresses that are close to each other, for instance belonging to servers in the same data center. Changing the subnet mask by a couple of bits has no significant influence on the graph. We can see from the graph that malicious domains are much more likely to resolve to more than one subnet over time than legitimate domains. For up to 90% of all domains, malicious domains resolve to more distinct subnets over time than legitimate domains.

Conversely, we can also look at the number of malicious typosquatting domains per IP subnet. The 13,526 malicious typosquatting domains discovered during our study resolve to only 2,377 distinct IP addresses, which means that on average each of them hosts 5.7 typosquatting domains. In practice, however, a much smaller range of IP addresses hosts a very large fraction of the typosquatting pages: the four subnets shown in Table III together host 36% of all malicious typosquatting pages and 50% of all malicious typosquatting domains visited during the study. The table also shows the owner of the autonomous system corresponding to each network. Except for Castle Access, which operates enterprise data centers, each of these companies is in the domain parking business.

I. Registrar statistics

Based on our collected WHOIS records, we investigated the most commonly used registrars for both legitimate and malicious domain registrations. Ranking the registrars by their number of malicious typosquatting domain registrations shows, unsurprisingly, that the overall most popular registrars also have the most malicious domain registrations. Unfortunately, we do not know the total number of domain registrations for each of the encountered registrars, nor have we investigated enough domains to make statistically significant claims about whether or not particular registrars are used more often than others for malicious registrations.

Nonetheless, our data does show that some registrars have a significantly higher ratio of legitimate to malicious registrations than others. These are mostly reputable registrars that explicitly market themselves as brand protection or corporate domain portfolio management companies, such as Safenames and Corporation Service Company (CSC). However, we also found the company Network Solutions to have a high ratio of legitimate to malicious domains in our study, even though we could not find any evidence that this company has a better

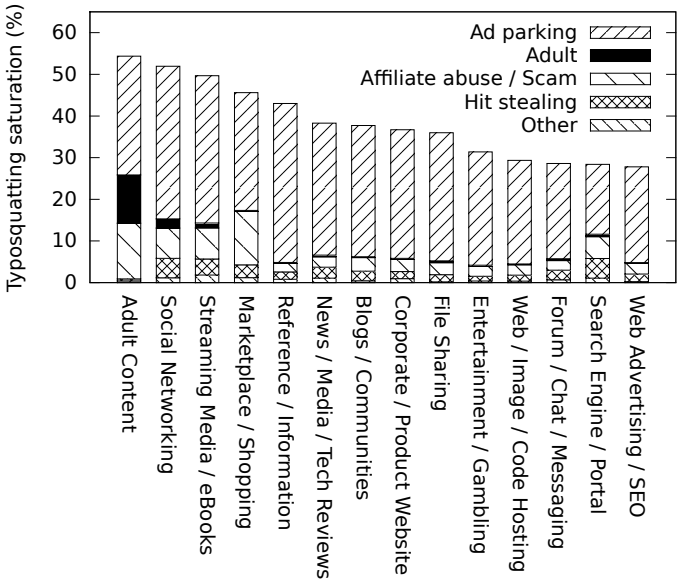

Fig. 10. The typosquatting saturation per authoritative domain category. Adult typosquatting domains are clearly targeting adult authoritative sites more than other authoritative sites.

than average reputation[6], nor that it positions itself as a brand protection specialist. We suspect many legitimate domains are registered through this company because of its historic role, first as registry for the .com, .org and .net zones and later as the sole domain registrar for these TLDs, until '98. Our data indeed indicates that their legitimate domains are mostly high-profile domains with a length of 2 to 5 characters, registered in the 1990's.

We posit that the reason for these results is that the total set of typosquatting domains for any given authoritative site is spread among many typosquatters, a claim which is supported by our WHOIS data. Since each typosquatter can opt for a different registrar, there is a dilution of malicious activity over the most popular registrars. On the other hand, when a legitimate company uses a brand protection service, all defensive domain registrations are likely to be done through the same registrar. As such, the legitimate registrations will tend to cluster, while the malicious registrations will tend to get diffused.

J. Influence of Authoritative Domain Type

To see whether typosquatters indiscriminately target all popular domains, or rather target certain kinds of authoritative domains more often than others, we classified our list of authoritative domains into content-based categories. This classification was done manually, with the assistance of Trend Micro's Site Safety Center[7] and McAfee's Threat Intelligence[8] domain categorization services. Fig. 10 shows the typosquatting saturation for these authoritative categories and also illustrates the relative contribution of the most important malicious typosquatting categories to the saturation. Since our data indicates no significant change of the typosquatting

[6]http://en.wikipedia.org/wiki/Network_Solutions#Controversies

[7]http://global.sitesafety.trendmicro.com/

[8]http://www.mcafee.com/threat-intelligence/domain/

saturation per authoritative domain category in time, we show an averaging over time. In order for the results to be representative, authoritative categories containing less than 10 domains have been excluded.

We can see from the figure that adult sites are targeted the most and that adult typosquatting domains are clearly targeting adult authoritative domains more often than other types of authoritative domains. The social networking category comes second and also experiences some adult typosquatting, caused mainly by the typosquatting domains of `odnoklassniki.ru` and `renren.com`. There are also substantial differences in the contribution of the affiliate abuse or scam typosquatting domains between the different authoritative categories. It would be wrong to attribute these differences to the fact that some authoritative categories (e.g. marketplace and shopping) are more likely to offer an affiliate program than others, because a typosquatting site in the affiliate abuse category does not necessarily exploit an affiliate program of *its own* authoritative site. In fact, we saw many cases of typosquatting domains exploiting an affiliate program of an unrelated domain during our manual classification. Hence, we cannot pinpoint a specific reason for this difference in the contribution of the affiliate abuse or scam typosquatting domains between the different authoritative categories based on our data.

K. Influence of TLD

The Internet Corporation for Assigned Names and Numbers (ICANN) is responsible for coordinating the global domain name system, but has delegated the responsibility of managing top-level domains to other commercial or non-profit organizations, known as registry operators. This is true for generic TLDs (`.com`, `.org`, `.net`, etc.) as well as for country-code TLDs (`.de`, `.jp`, etc.). These registry operators may also fulfill the role of registrar, or may delegate this responsibility to other companies. A registrar must be officially accredited by ICANN for it to directly do business with a registry operator, while non-ICANN-accredited registrars can only be resellers for other registrars. For the generic TLDs, all accredited registrars have adopted the Uniform Domain-Name Dispute-Resolution Policy (UDRP) [14]. This is a policy to be agreed between a registrar and a registrant for a domain name registration, giving an official domain dispute panel the right to take down or transfer ownership of the domain in case a third party complainant can prove the registration violates her rights. The goal of this policy is to reduce the prevalence of cybersquatting by saving the complainant from having to start an expensive court procedure in order to take down the domain.

While this policy is being applied for all generic TLDs, not all country-code TLDs follow the same approach. Some country-code TLD registry operators that do not implement the UDRP have decided to implement their own similar out-of-court domain dispute resolution, but others only intervene in disputed registrations after an official court ruling. This can potentially have an influence on the prevalence of typosquatting in those TLDs. In particular, we can expect that TLDs providing a swift, out-of-court procedure for domain disputes have fewer typosquatting domains than those that do not.

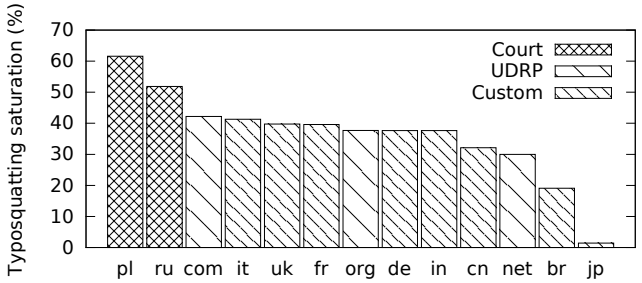

Fig. 11. The typosquatting saturation per TLD, for those TLDs with at least 5 domains in our top 500 list. The different patterns indicate the type of dispute arbitration provided.

A second way in which local domain authorities have an influence on the prevalence of typosquatting, is in their price setting. Obviously typosquatting can only be profitable if the revenue from a domain name is greater than its cost. Hence, we can expect more expensive TLDs to attract less typosquatters than cheaper TLDs.

A final policy component that can have an effect on typosquatting, is the restrictions placed by registry operators on registrants that can apply for a domain name. For instance, some country-code registry operators require registrants to be a citizen of the corresponding country or to have a local correspondence address. This makes it more difficult (and hence more expensive) or even impossible for foreign parties to legally register a domain name under those TLDs, hence potentially lowering their attractiveness for typosquatters.

Fig. 11 shows the typosquatting saturation per TLD, for the TLDs with at least 5 domains in our top 500 list. The typosquatting saturation per TLD does not change significantly over time, hence we show an averaging over time. The different bar patterns in the figure indicate the type of dispute arbitration provided by the corresponding TLD authority. "Court" means the authority provides no out-of-court arbitration whatsoever, "UDRP" means the authority has adopted the UDRP and "Custom" means the authority implements some custom form of out-of-court arbitration. We can see that the two TLDs without an out-of-court arbitration option have the highest typosquatting saturation.

On the other side of the scale, the Brazilian TLD and especially the Japanese TLD have a significantly lower typosquatting saturation. Since we did not find any evidence to indicate that domain names under these TLDs are particularly difficult to obtain, we believe the scarcity of typosquatting domains for these TLDs can be attributed to the high cost of acquiring such domains. We calculated the yearly price of a domain under these TLDs by averaging the yearly price advertised by the global top 3 domain registrars[9] GoDaddy, eNom and Network Solutions, and found that a `.com.br` domain costs $43 and a `.jp` domain costs $107, while domains under all other TLDs shown in Fig. 11 sell for less than $15. Clearly, these higher domain prices are effective at deterring typosquatters. Based on the steep drop in typosquatting saturation between the `.br` and `.jp` domains, we can estimate the annual revenue of a typosquatting domain of a popular website to be somewhere between $43 and $107. However, since there are still plenty

[9]According to http://www.webhosting.info/registrars/top-registrars/global/

of domains available under cheap TLDs, there is also no real incentive for typosquatters to target the more expensive TLDs, hence the annual revenue might actually be higher.

IV. Related work

To the best of our knowledge, our work is the first *content-based* study that examines the problem of typosquatting *in time*. Prior work on typosquatting almost always operated by taking a snapshot of the typosquatting problem and characterizing that snapshot, or by limiting the longitudinal aspect of the study to domain registration records only. Our work is complementary to this type of prior work because it adds to the understanding of typosquatting abuse and reveals trends that otherwise remain hidden. In this section we review prior work on typosquatting as well as other types of domain squatting attacks.

Cybersquatting: Cybersquatting refers to the registration of domains that include trademarks belonging to other persons and companies. Cybersquatting was popular in the early nineties, when long-existing brick-and-mortar companies did not yet operate websites. Various opportunists registered their trademarks as domain names before them, so that they would sell the domains back to their rightful owners for profit [11].

Today, this type of domain squatting is not as popular, since companies usually register all appropriate domains well before the company and its trademarks become popular. There are, however, still cases where opportunists try to speculate the name of future products and services and register them, before the company marketing the product or service gets a chance to.[10] This phenomenon has been studied by Coull et al. [5] together with other domain registration abuses, such as *domain-name front running*.

Typosquatting: Cybersquatting evolved into *typosquatting*, i.e., the act of registering domains that are mistypes of popular authoritative domains, with the intention of capturing the traffic of users who mistype URLs in their browser address bar. This practice can be traced back to over 15 years, since the 1999 Anticybersquatting Consumer Protection Act (ACPA) already mentioned URLs that are "sufficiently similar to a trademark of a person or entity" [1].

Edelman, in 2003, reported on thousands of mistyped and cybersquatting domains that served sexually-explicit content, which were likely registered by the same individual [8]. Wang et al. [25] later described a system for automatically discovering and analyzing typosquatting by simulating typing errors. The researchers also brought attention to the fact that the majority of the discovered typosquatting domains were pointing to companies specializing in domain parking, which were used to automatically serve ads related to the mistyped domain name. In this paper, we used Wang et al.'s models to generate typosquatting domains for our seven-month long study.

Banerjee et al. [2] showed that typosquatting extends beyond the models of Wang et al., including the abuse of domain suffixes, such as typosquatting `.org` domains, for the equivalent `.com` authoritative ones. Even though the authors have collected typosquatting data over a two-month period, they do not perform a longitudinal study and thus do not include the parameter of time in their reported results.

Halvorson et al. [12] recently showed that over three quarters of the domains registered with the `.xxx` TLD are defensive registrations where companies and individuals pre-emptively register their trademarks with the `.xxx` extension, e.g., `obama.xxx` and `president.xxx`, out of fear that someone could, in the future, use these domains to harm their image.

Moore and Edelman performed a similar study for discovering typosquatting domains in 2010 [16] and estimated that approximately one million typosquatting domains targeted the top 3,264 `.com` sites. The authors also bring attention to the fact that large advertising networks willingly cooperate with typosquatters by showing ads on the mistyped domains. As such, these networks can be held equally responsible for the damage against the authoritative domains that are being attacked. Next to the use of ads as a monetization strategy, there have also been documented cases of typosquatting domains used to serve malware [9]. Nikiforakis et al. [20] showed that typosquatting can also occur outside the confines of a browser's address bar, in remote JavaScript inclusions. There, developers mistype the domains of remote code providers and thus make their sites susceptible to malicious script injections served through the appropriate mistyped domains.

Szurdi et al. [23] recently investigated typosquatting registrations targeting `.com` domains that are in the "long tail" of the popularity distribution. The authors present a tool named *Yet Another Typosquatting Tool* (YATT), which uses passive domain features and (optionally) active domain features such as DNS, WHOIS and content information, to identify and categorize typosquatting domains into categories similar to ours. The authors use this tool to classify 4.7 million potential typosquatting `.com` domains derived from the Alexa top 1 million. Their results indicate that, although the typosquatting saturation decreases for less popular authoritative domains, it is still at 40% near the Alexa 1 million rank and an estimated 20% of all `.com` domains appear to be typosquatting domains. The study also includes a longitudinal component, tracking the domain registrations between October 2012 and October 2013 based on daily dumps of the `.com` zone file. In contrast to our study, the content of these domains is not taken into account for the longitudinal analysis. That is, when a new domain appears in the zone file, it is compared against the list of all potential typosquatting domains generated from the Alexa 1 million, to determine whether or not it is an active typosquatting domain. The authors find, amongst other results, that typosquatting domains of popular authoritative domains appear to change hands much more frequently than other domains.

Finally, Vissers et al. [24] very recently performed a study which explored the ecosystem of domain parking services. While these services have previously been investigated in the context of cybersquatting, this study explores the domain parking ecosystem in its own right. The authors discover, among other results, that between 1.63% and 9% of the investigated parking domains are typosquatting domains. The study also shows that some of the investigated parking service providers have a significantly higher ratio of typosquatting domains than others, with only one provider serving no typosquatting domains at all.

[10] Parked domain with ads - `www.iphone9.com`

Homograph attacks: In domain-homograph attacks, attackers take advantage of the perceived visual similarity between two or more letters, in order to trick the user into believing that she is interacting with a specific authoritative website while she is interacting with a malicious one. This confusion can be abused to convince users to willingly submit their credentials and other sensitive information. The main difference between these attacks and the previously mentioned domain squatting attacks, is that the homograph domains are usually spread-out through spam emails and social networks, instead of relying on user mistakes, since their construction cannot usually be achieved through typical typing mistakes.

Gabrilovich and Gontmakher showed that characters from non-Latin character-sets that look like Latin characters can be substituted to confuse the user of the nature of a given domain [10]. For instance, an attacker could register `paypal.com` using the Cyrillic letter р (lower case "r", Unicode U+0440), which looks almost identical to the Latin letter "p". Today, this type of attack is significantly harder since browser vendors revert to the *punycode* format of URLs [4], whenever they think that a domain is maliciously crafted.

Dhamija et al. [6], investigate why phishing works, and mention "visually deceptive text", i.e., domains that substitute characters with look-alikes within the same character-set, such as `bankofvvest.com` (two "v"s instead of a "w"). Holgers et al. [13] performed a large-scale study of homograph attacks by gathering popular domains and searching for their homographed versions. Their results showed that, even though the monetization strategies of homograph domains are very similar to traditional domain squatting, the practice of registering homograph domains is significantly less popular than typosquatting.

Other types of domain squatting: In 2011, Dinaburg proposed a new type of domain squatting, *bitsquatting* [7]. Dinaburg hypothesized that random bit-flips that occur in hardware memory that is failing or operating outside normal temperatures, can be abused for domain squatting purposes, assuming that the flips would occur in the bits holding the textual representation of a domain name. Domain squatters can register domains that have a one-bit difference from popular authoritative domains and capture the traffic resulting from such erroneous bit-flips. Dinaburg registered 30 *bitsquatting* domains, such as `mic2osoft.com`, and recorded more than 50,000 requests over an eight-month period. Later research showed that domain squatters have already adopted this technique [18]. Recently, Nikiforakis et al. also discovered that domain squatters, next to typos, are also abusing the sound similarity of words to construct malicious squatting domains and attract users to them [19].

V. Conclusion

Typosquatting has been known and studied for over 15 years, yet we have shown that it remains a popular form of domain abuse up until this day. By performing the first content-based longitudinal study of typosquatting abuse, we have investigated to what extent previously discovered typosquatting trends still hold today, and have also discovered many new typosquatting facts. Our main results include that (1) few trademark owners protect themselves against typosquatting by defensively registering typosquatting domains for their own domains, (2) over 75% of all possible typosquatting domains for short popular authoritative domains are already registered, and that typosquatters are increasingly targeting longer domains, (3) typosquatters are varying their monetization strategy over time, (4) some companies choose not to renew their defensive registrations of typosquatting domains, leading these domains back into the eager hands of typosquatters, (5) up to 50% of all typosquatting domains can be traced back to just four typosquatting page hosters and (6) certain top-level domains are much less prone to typosquatting than others, due to their price setting and local registration and arbitration policies. We hope this paper and the corresponding dataset, which has been made publicly available online, can serve as a new reference for the state of the typosquatting landscape.

Acknowledgments

We thank Christian Kreibich and our anonymous reviewers for their valuable comments and suggestions that have improved the quality of this paper. This research was performed with the financial support of the Prevention against Crime Programme of the European Union (B-CCENTRE), and the Research Fund KU Leuven. We acknowledge the support of EURid, the European Registry of Internet Domain Names. Pieter Agten holds a PhD fellowship of the Research Foundation - Flanders (FWO).

References

[1] "Anticybersquatting Consumer Protection Act (ACPA)," http://www.patents.com/acpa.htm, November 1999.

[2] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan, "Cyber-fraud is one typo away," in *Proceedings of the 27th Conference on Computer Communications, IEEE INFOCOM*, 2008.

[3] A. Banerjee, M. S. Rahman, and M. Faloutsos, "Sut: Quantifying and mitigating url typosquatting," *Comput. Netw.*, vol. 55, no. 13, pp. 3001–3014, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2011.06.005

[4] A. Costello, "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)," http://www.ietf.org/rfc/rfc3492.txt.

[5] S. E. Coull, A. M. White, T.-f. Yen, F. Monrose, and M. K. Reiter, "Understanding domain registration abuses," in *Proceedings of the 25th International Information Security Conference (IFIP SEC)*, 2010.

[6] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ser. CHI '06. ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1124772.1124861

[7] A. Dinaburg, "Bitsquatting: DNS Hijacking without Exploitation," in *Proceedings of BlackHat Security*, July 2011.

[8] B. Edelman, "Large-scale registration of domains with typographical errors," September 2003.

[9] F-Secure, "W32/Googkle," http://www.f-secure.com/v-descs/googkle.shtml.

[10] E. Gabrilovich and A. Gontmakher, "The homograph attack," *Communucations of the ACM*, vol. 45, no. 2, p. 128, Feb. 2002. [Online]. Available: http://doi.acm.org/10.1145/503124.503156

[11] J. Golinveaux, "What's in a domain name: Is cybersquatting trademark dilution?" in *University of San Francisco Law Review 33 U.S.F. L. Rev. (1998-1999)*, 1998.

[12] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker, "XXXtortion?: Inferring Registration Intent in the .XXX TLD," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, 2014, pp. 901–912.

[13] T. Holgers, D. E. Watson, and S. D. Gribble, "Cutting through the confusion: a measurement study of homograph attacks," in *Proceedings of the 2006 USENIX Annual Technical Conference*, 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267359.1267383

[14] ICANN, "Domain name dispute resolution policies," https://web.archive.org/web/20141028140919/https://www.icann.org/resources/pages/dndr-2012-02-25-en, accessed: 2014-12-01.

[15] J. Lewis and S. Baker, "The economic impact of cybercrime and cyber espionage," *Center for Strategic and International Studies, Washington, DC*, 2013.

[16] T. Moore and B. Edelman, "Measuring the perpetrators and funders of typosquatting," in *Financial Cryptography and Data Security*, vol. 6052, 2010, pp. 175–191.

[17] D. Müllner, "fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python," *Journal of Statistical Software*, vol. 53, no. 9, pp. 1–18, 5 2013. [Online]. Available: http://www.jstatsoft.org/v53/i09

[18] N. Nikiforakis, S. V. Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting: Exploiting bit-flips for fun, or profit?" in *Proceedings of the 22nd International World Wide Web Conference (WWW)*, 2013.

[19] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Proceedings of the 17th Information Security Conference (ISC)*, 2014.

[20] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "You Are What You Include: Large-scale Evaluation of Remote JavaScript Inclusions," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2012.

[21] Recap, "Microsoft corporation v. shah et al," http://archive.recapthelaw.org/wawd/166997/, accessed: 2014-07-29.

[22] C. Roth, M. Dunham, and J. Watson, "Cybersquatting; typosquatting Facebooks $2.8 million in damages and domain names," http://www.lexology.com/library/detail.aspx?g=7088bf09-8a9e-4449-a179-d90bdfad3310.

[23] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich, "The long "taile" of typosquatting domain names," in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, 2014, pp. 191–206. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/szurdi

[24] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking Sensors: Analyzing and Detecting Parked Domains," in *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS '15)*, Feb 2015. [Online]. Available: http://dx.doi.org/10.14722/ndss.2015.230053

[25] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels, "Strider typo-patrol: discovery and analysis of systematic typo-squatting," in *Proceedings of the 2nd conference on Steps to Reducing Unwanted Traffic on the Internet - Volume 2*, ser. SRUTI'06. Berkeley, CA, USA: USENIX Association, 2006, pp. 5–5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251296.1251301