The Long "Taile" of Typosquatting Domain Names

Janos Szurdi[◇] Balazs Kocso[*] Gabor Cseh[*]
Jonathan Spring[◇] Mark Felegyhazi[*] Chris Kanich[†]

[◇]*Carnegie Mellon University* [*]*Budapest University of Technology and Economics*
[†]*University of Illinois at Chicago*

Abstract

Typosquatting is a speculative behavior that leverages Internet naming and governance practices to extract profit from users' misspellings and typing errors. Simple and inexpensive domain registration motivates speculators to register domain names in bulk to profit from display advertisements, to redirect traffic to third party pages, to deploy phishing sites, or to serve malware. While previous research has focused on typosquatting domains which target popular websites, speculators also appear to be typosquatting on the "long tail" of the popularity distribution: millions of registered domain names appear to be potential typos of other site names, and only 6.8% target the 10,000 most popular `.com` domains.

Investigating the entire distribution can give a more complete understanding of the typosquatting phenomenon. In this paper, we perform a comprehensive study of typosquatting domain registrations within the `.com` TLD. Our methodology helps us to significantly improve upon existing solutions in identifying typosquatting domains and their monetization strategies, especially for less popular targets. We find that about half of the possible typo domains identified by lexical analysis are truly typo domains. From our zone file analysis, we estimate that 20% of the total number of `.com` domain registrations are true typo domains and their number is increasing with the expansion of the `.com` domain space. This large number of typo registrations motivates us to review intervention attempts and implement efficient user-side mitigation tools to diminish the financial benefit of typosquatting to miscreants.

1 Introduction

Thousands of new domain names are registered daily that at first glance do not have completely legitimate uses: some contain random characters (possibly used by miscreants [23]), are a composite of two completely unrelated words (possibly used in spam [17]), contain keywords of highly-visible recent events (ex. `hillaryclingon.com` for political phishing in 2008 [28]) or are similar to other, typically well-known, domain names (ex. `twtter.com` [27, 32]). Domain purchasers use this final technique, often called "typosquatting," to capitalize on other domain names' popularity and user mistakes to drive traffic to their websites.

Many old and new domain names alike do not ever show up in search engines, spam traps, or malicious URL blacklists, yet still maintain a web server hosting some form of content. However, maintaining the domain registration, DNS, and web server expends resources, even if these domain registrations do not serve an obvious purpose. Investigating the purpose of domain registrations in the "long tail" of the popularity distribution can help us better understand these enterprises and their relationship to speculative and malicious online activities. In this paper, we specifically consider the hypothesis that typosquatting is a reason for many of these registrations, and scrutinize different methods for committing malice or monetizing this behavior.

In the Internet economy, monetizing on user intent has been a very profitable business strategy: search display advertising is effective because relevant ads can be shown based on user search queries. DNS is similar, as domain registrations provide ample opportunities for monetization through direct user navigation rather than search. Domain name front running, domain tasting and typosquatting domain names can all monetize this phenomenon. [1] [12] According to [22], domain tasting was nearly eliminated in the generic TLDs by the 2009 policy changes by ICANN. In addition, [12] reports that the

[1] *Domain name front running* is when registrars register domains that users have been looking for in order to monetize on their registration potential. *Domain tasting* is speculative behavior abusing the five-day grace period after domain registrations in some TLDs. This liberal registration policy gave refunds within a few days if the registrant wanted, however this policy resulted in short domain registrations en masse. ICANN has since changed policy, limiting the behavior [12, 22].

anecdotes about domain name front running by major registrars do not seem to hold. But typosquatting, the most prevalent speculative domain name registration behavior to date, continues apace.

Typosquatting wastes users' time and no doubt annoys them as well. As we show in Section 4.5, less than two percent of all domains we identify as "typo domains" redirect the user to the targeted domain, and the lion's share instead serve advertisements which previous research has shown to be profitable. [16, 26] These ad-filled pages give no clear indication to the user that they have typed the domain incorrectly; without a descriptive error, the user may abandon their task rather than double check their spelling. By monetizing these pages with advertisements, the typosquatter does a disservice both to the user and the victim web site. Protecting users from typosquatters can lessen the damage as well as disincentivize typosquatting by decreasing the squatters' profits.

If a typosquatter hosts a site that impersonates the legitimate brandholder it is certainly malicious and in some jurisdictions illegal. Such overt violations have been mitigated via legislation in the US and policy by ICANN [15, 21, 30]. For example, Facebook recently extracted a $2.8 million judgement against typosquatters impersonating their website; this successful litigation should serve as a strong deterrent against this form of malicious typosquatting against entities with the resources to litigate [18]. Several reports by commercial security teams have cited typosquatting domains' use in malicious campaigns for quiz scams [8], spam survey sites [37], in an SMS micro-payment scam [14], offering deceptive downloads or serving adult content [25], or in a bait-and-switch scam offering illegal music downloads [29]. However, until this paper, evidence regarding the extent of malicious typosquatting problems has not been available.

Typosquatting has been studied in depth in related work. In his first paper, Edelman points to the typosquatting phenomenon and discusses possible incentives for both squatters and defenders [15]. Wang *et al.* include a typopatrol service in their Strider security framework that focuses on generating typo domains for popular domains and protect visitors from offending content [35]. Moore and Edelman revisit the problem in [26] pursuing a more thorough study of the original thesis of Edelman. They explore various monetization methods and suggest intervention options. They pessimistically conclude that the best intervention options are hampered by misaligned incentives of the participants. Banerjee *et al.* [10] make another attempt to design a typosquatting categorization tool. Their method works well for a small set of sample domain names. These analyses have focused on active measurement of typosquatting sites which target the most popular domains – considering no more than 3,264 unique `.com` domain names. However, we find that no more than 4.9% of all lexicographically similar name registrations target these popular domains. While typos for the most popular domains likely account for a significant amount of typo traffic, it is unclear whether the long tail also supports a significant amount of typo traffic.

Here we present a systematic study of domain name registrations focusing on typosquatting perpetrated against the long tail of the popularity distribution. We design a set of algorithms that can effectively identify typosquatting domains and categorize the monetization method of its owner. We also design and implement tools to improve user experience by allowing them to reach their intended destination. Although various user tools exist in the wild, most are inaccurate and focus only on a limited set of targeted domains. Our typo identification algorithms combined with the user protection tools provide improved protection against being misled by typosquatting, even when it is perpetrated against less popular sites.

Section 2 provides background on typosquatting and the most common tricks used by typosquatters. Section 3 presents our data collection methodology and describes our typo categorization framework. Section 4 presents a characterization of the extent, purpose, trends, and malice involved in the perpetration of typosquatting. We present mitigation tools and intervention options in Section 5. Section 6 concludes.

2 Background

Popularity attracts speculation, and typosquatting is a showcase of this observation in the Internet ecosystem. Typosquatting maintains its popularity even in the face of the continuous effort to diminish its impact. In this section, we present a general overview of typosquatting and discuss efforts to protect legitimate domain owners from speculation.

2.1 Typo techniques and monetization

Typosquatters register domain names that are similar to those used by other websites in hope of attracting traffic due to user mistakes. The most frequent occurrences of mistyping are those that involve a one-character distance, also called the Damerau-Levenshtein (DL) distance one, from the correct spelling both in free text [13] and in case of domain names [10][2]. In this paper, we focus on typosquatting domains of Damerau-Levenshtein distance one (DL-1) that are generated using the most common operations: addition, deletion, substitution of one character, transposition of neighboring characters [13]. We extend this to include deletion of the period before the "www"

[2] Although some researchers have found that for longer original domains a small number of typosquatting domain names with larger DL distances exist [26].

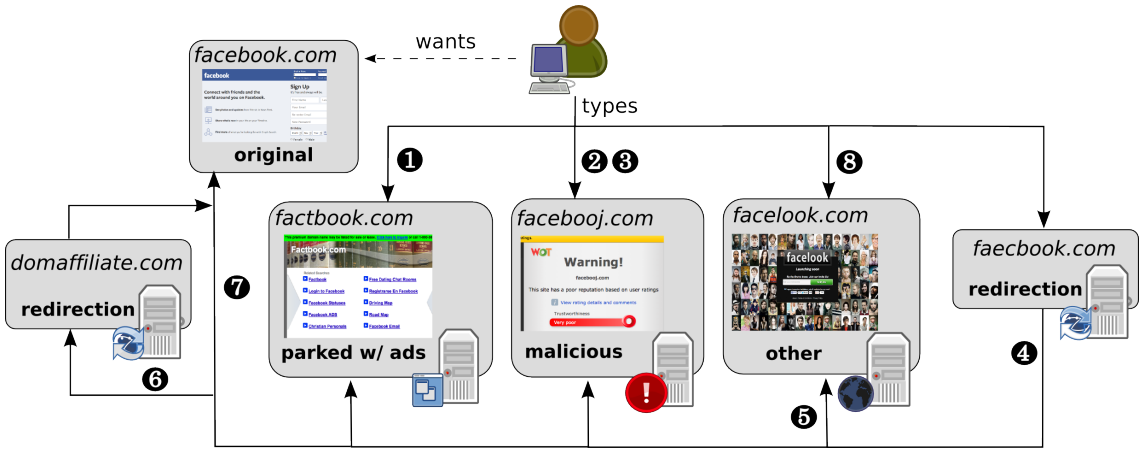


Figure 1: The typosquatting ecosystem with various monetization techniques.

commonly prepended to web server domain names [26]. We note that a special case of DL-1, called fat finger distance (FF distance), is considered when the mistyping occurs with letters that are adjacent on a US English keyboard. The rationale of this metric is that users are more likely to mistype letters in close proximity.

Typosquatters use various techniques to monetize their domain name registrations. The typosquatting domain can be *parked* and serve third-party advertisements to monetize the incoming traffic (❶ on Figure 1). The domain can also be set up to impersonate the intended domain for instance to host a phishing page [33] (❷), serve malware (❸), or perpetrate some other scam on the user [14, 37]. Many monetization techniques can also involve redirection to another domain (❹), the *landing domain*, that might employ the previously mentioned techniques. Speculators can also redirect visitors to *competitor domains* (❺) causing a direct loss to the owner of the original domain. Conversely, the typodomain owner can redirect traffic to the intended site, and monetize this traffic via *affiliate marketing* (❻). The original domain owner can also perform *defensive registrations* of typos for their main domain name and set up the redirections themselves (❼). Finally, in some cases, the typo domain owner can serve content that is unrelated to the original domain (❽).

2.2 Intervention attempts

Typosquatting exists within a legal and moral gray area; consequently, intervention has traditionally been weak to reduce the effect of typosquatting. ICANN provides the Uniform Domain-Name Dispute-Resolution Policy (UDRP) to mediate domain registration disputes for a relatively small filing fee. Unfortunately, cheap domain registration allows for mass typo-domain registrations and this gives a significant advantage to speculators. Against mass registrations of typo-domains UDRP mitigation becomes infeasible. Companies have initiated legal procedures in cases where cybersquatting and trademark infringement was applicable (see for example [32] on a recent court order against `twtter.com` and `wikapedia.com`, and a more recent court order against typosquatters of `facebook.com` [31]). The Anti-cybersquatting Consumer Protection Act (ACPA) (15 USC §1125(d)) offers legal protection to push such cases to court.

Policy intervention is more effective when targeting the registration process either at a national scale for specific TLDs or on a registrar level [24]. One can also mount an effective defense by targeting the monetization infrastructure [23, 24]. Unfortunately, the agility of domain speculators in registering new domains and the difficulty of determining their ill intent makes this a difficult prospect.

There have been some efforts to provide technical tools to mitigate typosquatting, notably the Microsoft Strider Typopatrol system which protects trademarks and childrens' sites [35]. At the user level, the OpenDNS has a typo correction feature which corrects major TLD misspellings [27] and the Mozilla URLFixer Firefox plugin [6] can suggest corrections to typed URLs. A common property of these solutions is that they only cover a relatively small set of typos, typically those that target the most popular domain names. As we show in Section 5.3, our mitigation solution is based on an extensive set of investigated domain names and hence provides significantly better coverage to detect typosquatting. Moreover, our extended set of detection features allows for more accurate detection of typosquatting than solutions in previous work.

3 Methodology

This section presents our data collection and domain categorization framework in detail as illustrated it in Figure 2.

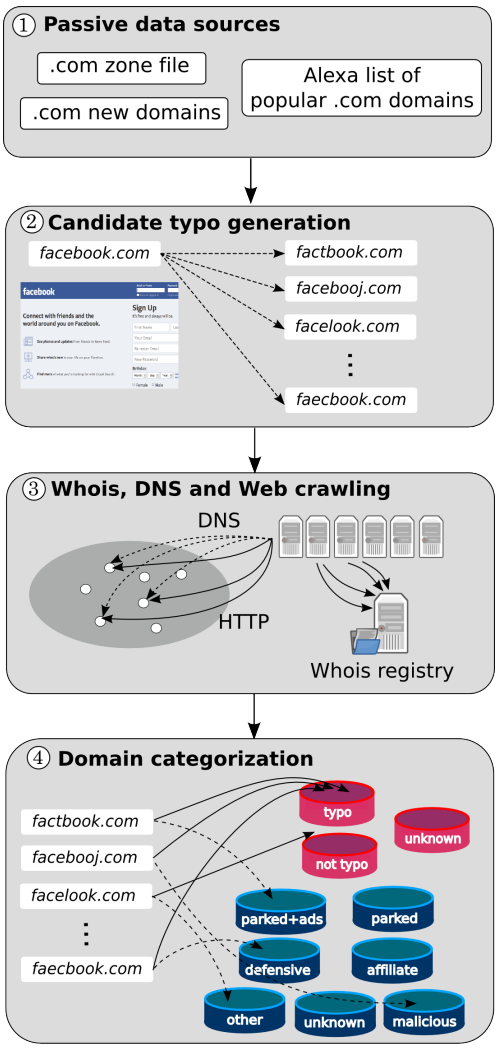


Figure 2: The data collection and typo categorization framework. The framework uses (①) large domain lists (zone file, Alexa popular domains list), (②) derives candidate typos based on lexical features and registration data in the zone file, (③) acquires additional information using active crawlers (Whois, DNS, Web), and finally (④) decides about typo domains and assigns them into typosquatting categories.

Terminology. Throughout this paper, we will refer to domains available for direct registration under a public suffix as *registered domains*, for instance *example.com* or *example.co.uk*. Generated typo domains, or *gtypos*, are domain names which are lexically similar (e.g. at DL-1) to some set of target domains. Candidate typo domains, or *ctypos*, are the subset of registered domains within the *gtypo* set which have been registered. Below we describe both how we select the target set and how we generate the gtypo set.

3.1 Data sources and scope

`.com` zone file. We leverage a variety of data sources to infer the prevalence of typosquatting in domain registrations. Our primary source is the `.com` zone file, which contains records of every domain registered within that TLD. As a popular generic domain name, the `.com` zone file contains millions of registered domain names `.com` and is available to researchers making it an ideal candidate for a representative investigation of typosquatting. Our comprehensive study is based on the March 15, 2013 version of the zone file provided by Verisign Inc containing approximately 106 million domain names. For trend analysis we collected the daily newly added and deleted domains from the zone file from October 01, 2012 to February 20, 2014.

Alexa list. The Alexa list of the top 1 million sites from March 15, 2013 serves as a benchmark for popularity [1], out of which 523,960 domains belong to the `.com` TLD, with 488,113 unique registered domains five characters long or more. For our study, we split the Alexa list into three categories: *Alexa top* containing domains ranked higher than 10,000, *Alexa mid* containing domains ranked 10,000-250,000, and *Alexa tail* containing the remaining `.com` domains ranked below 250,000. While Alexa cautions that rankings below 100,000 are not statistically significant, we are not concerned with exact comparative ranking or traffic counts for these domains but consider the Alexa list rather as a rough indicator of popularity. We also collected the Alexa top 1 million for the October 01, 2012 to February 20, 2014 period for trend analysis.

Domain blacklists. To shed light on the malicious use of typo domains, we check the typo domains from the `.com` zone file against twelve different domain name blacklists. The black lists come from abuse.ch's list of Zeus and SpyEye servers, malwaredomainlist.com, malwaredomains.com, malwarepatrol.com, Google Safe Browsing, and a commonly used commercial list. We also derive lists of malicious domains from recorded requests to DNS-based black lists (DNSBL). This method does not capture the complete list, but rather only includes domains actively marked as malicious and looked up by users during the collection time frame.

3.2 Generating candidate typos

We generated a list of all possible typo domains using the most common typo operations: addition (*add*), deletion (*del*), substitution of one character (*sub*), transposition of neighboring characters (*tra*), and supplement this set with a "." deletion operation specific to "www." domain names (e.g. a user typed (*wwwexample.com*). We define this list as the "generated typo" or *gtypo* list. The subset of the gtypo list which was registered within the `.com` TLD

includes approximately 4.7 million domains, which we refer to as "candidate typos" or *ctypos*.

3.3 Typosquatting definitions

To define the scope of our work, we provide a concise definition of typosquatting.

Definition 1 *A candidate typo domain is called a* typosquatting domain *if (i) it was registered to benefit from traffic intended for a target domain (ii) that is the property of a different entity.*

It is important that both conditions have to be met simultaneously. Typosquatting domain names are registered with the parasitic intent to reap the mistyped traffic of popular domains belonging to someone else. This includes parked domains serving ads, phishing domains, known malicious domains, typo domains redirecting to unrelated content and affiliate marketing. Arguably, these conditions cannot always be checked with confidence, for example ownership information could be disguised[3].

According to our definition, parked domains that do not serve ads are excluded from our definition of typosquatting, because they are not making any visible profit from parking. We still consider them as typos until it becomes clear if they are performing typosquatting on the target or serving unrelated content. Candidate typo domains that are defensively registered by the original domain owner are also excluded from typosquatting, because the owner of the typo domain and the original domain are the same. Although defensive typo registrations cannot be considered as typosquatting, they are born as an unwanted consequence of typosquatting.

We define *true typo* domains as follows.

Definition 2 *We call the union of typosquatting domains, parked domains not serving ads and defensive registrations the* true typo *domain set.*

Finally, all candidate typos that are at DL-1 from an original domain yet have unrelated content are considered as incidental registrations, although they can surely benefit from the lexical proximity[4].

3.4 Active crawling

We developed a set of active crawlers to collect additional information about the ctypo domains.

[3]For example, the name servers `*.aexp.com` of `americanexpressl.com` belong to American Express Inc., but that is the only indicator of ownership. This can only be marked using manual inspection.

[4]Here we face another uncertainty presented by scam pages that generate legitimately looking random content. We observed several such cases for suspiciously looking webshops. We make a conservative assessment and categorize them as other (O) in spite of their questionable content

Whois crawler. First, we collect registration data from the WHOIS global database. We restrict our crawler to the thin whois information as provided by Verisign Inc. for the `.com` domains. From the thin whois record, we use the registrar and registration date information.

DNS crawler. We collect DNS data to explore the background infrastructure serving these domains. Our crawler queries separately for A, AAAA, NS, MX, TXT, CNAME, and SOA records for each domain. The crawler then tests for random strings under the registered domain to infer whether wildcarding is present. Wildcarding is the practice when a name server resolves any subdomain under the domain belonging to its authority in the DNS hierarchy.

Web crawler. We use a web crawler to obtain the rendered DOM of each page, along with any automatic redirections that take place during the page load. This crawler uses the PhantomJS WebKit automation framework to provide high volume, full fidelity web crawling with javascript execution, cookie storage, and page rendering capabilities [20]. The crawler follows JavaScript redirections even when they may be obfuscated or contained in child iframes; it then reports the method of redirection and the destination for intermediate and final redirections. We also collect rendered screenshots of a subset of pages for manual inspection.

3.5 Clustering and categorization

Clustering. We group domains together according to various attributes obtained from available datasets and active analysis. Our goal with this clustering is twofold: to identify typo domains that might have been registered for the same purpose and to point to infrastructure elements that host a large number of typo domains. First, we identify domain sets that are at DL-1 distance from each other, forming a cluster of *typo neighbors*.

Understanding the infrastructure support and the content of the typo domains is required to make an informed decision about their real purpose. To characterize the infrastructure support for typosquatting, we cluster the candidate typo domains based on their registration and hosting information. In particular, we identify the major registrars and name servers (NSs) that host candidate typo domains.

Domain features. We derive a feature set including lexical, infrastructure and content features of the candidate typos as shown in Table 5 in Appendix A. We selected the features after carefully considering related work, collecting 40+ features in various attribute categories, and focusing only on relevant ones. To assess the efficiency of the selected feature set, we perform a systematic evaluation based on manual sampling in Section 4.1

and we use the results of this evaluation as a benchmark.[5]

Among the chosen features, domain length is a key indicator for typosquatting behavior as longer ctypo domains are more likely to indeed typosquat on the original domain they are close to [26]. Intuitively, the Alexa rank of the original domain indicates that more popular domains are more likely a target of typosquatting. Based on the zone file, we are able to observe the ratio of ctypo domains versus all domain names on a given NS and we deem hosting a large of proportion of potential typo domains suspicious for an NS. Similarly, if the registered domain of the NS contains keywords indicating parking behavior, then ctypo domains hosted on this NS are more likely to belong to typosquatting domains. NXDOMAIN wildcarding is used by major parking service providers to serve ads for web requests regardless of the subdomain. It has been shown that NXDOMAIN wildcarding is a precursor of suspicious behavior and quite often indicates parked typosquatting domains [7, 36]. Thus, we also consider it an indicator for typosquatting when the page content matches some collected parking keywords[6]. Finally, several redirections usually imply suspicious behavior, and we deem them important if the redirection targets a registered domain different from the typo domain and the target domain. The features we selected resulted in a significant improvement over existing methods in identifying typosquatting domains across the whole range of `.com` domains. We leave a more complex feature set selection and parameter calibration using machine learning techniques as future work.

Categorization. Using these features, we attribute typosquatting to candidate typo (ctypo) domains by assigning the tag *typosquatting (T)*, *not typosquatting (NT)* or *unknown (U)*. Unknown is typically used when the domain returns an HTTP or DNS error which prevents successfully downloading the page. We also tag the usage type of the typosquatting domains according to the monetization categories presented in Figure 1. We also present the novel approach of categorizing domains based on their monetization strategy. Hence, we tag ctypo domains which do not redirect the user to the target site as *parked (P)* without ads (not on Figure 1), *parked serving ads (PA)* (❶ on Figure 1), employing a *phishing (PH)* scam (❷), or serving *malware (M)* (❸). When redirection is used, then the ctypo domain can be tagged as *defensive (D)* registration (❼), defensive registration using *affiliate (A)* marketing (❻) in addition to the previously mentioned categories. When a ctypo domain redirects to another domain, then we tag it as *other (O)* (❹, ❺) no matter if it is a competitor or a completely unrelated site[7]. Finally, we mark all uncategorized domains as *unknown (U)*, a set that typically contains unreachable domains.

3.6 Checking Maliciousness

To analyze how the typo domains are used, 12 black lists are checked for an indication that the domains are malicious. To check a black list, we look for anything that was on that list during the first quarter of 2013. A "match" is a second-level domain match, since this is the relevant typo label.

To perform a check, a superset of all the domains for Q1 2013 per list was made, and the typo and Alexa domains were compared against that superset. For Google Safe Browsing, due to Google's technical constraints, the each set of domains was checked using the provided python client against data for May 1, 2011 to July 31, 2013. The results are presented in subsection 4.6.

4 Analysis

In this section, our goal is to characterize the current state of typosquatting. For this purpose, we use the `.com` zone file as the most popular and versatile TLD for domain registrations.

4.1 Typosquatting distribution

Experts believe that most newly registered domains are speculative or malicious. Paul Vixie posits that "most new domain names are malicious" [34]. The subset of registered typo domains from the generated typo domains is widely accepted as true typo domains ([26, 35]), and [26] has shown that this assertion mostly holds for the top 3,264 `.com` domains in the Alexa ranking.

We believe, however, that this assertion does not necessarily hold if we extend our scope to less popular domains. In order to investigate this possibility, we first perform a manual sampling from various sets of the `.com` zone file to systematically control the accuracy of typosquatting identification and also to provide a credible ground truth for investigation. We conduct a manual inspection of four thousand domain names because the typosquatting definitions in the academic literature [26, 35] are very crude. Moreover, we present our mitigation tool analysis in Section 5, and in so doing also discuss the limitations of existing defense tools that typically only focus on correcting typos for a limited set of popular domain names.

[5] Manually generated datasets are widely used as indicators for malicious behavior; for example, the PhishTank phishing list is a major component of SURBL, the leading domain blacklist. [2].

[6] Here, we improve on the techniques used by [7] and [19] to find parking services and parked domains

[7] Determining domain competitors is beyond the scope of this work; we summarized redirections to third-party domains independently of the typosquatter's intent. While these redirections might simply be to other parked sites, any redirection away from the original site is a traffic loss for the original domain owner.

We first take a sample of 1000 ctypo domains randomly with uniform distribution from the Alexa top domain list to match the sampling methodology of [26]. We then complete this with three additional samples of 1000 ctypo domains each derived from the `.com` zone and the Alexa domain list. Our four sample sets are thus the following: ctypos of the the Alexa top/mid/tail domains (recall their description from Section 3.1) and ctypos of a random sample taken over the whole `.com` zone file. With these multiple sets, our goal is to check whether the conclusions from prior work regarding the frequency of typosquatting hold for less popular domains.

Typosquatting domains are notoriously difficult to identify. In several cases, only a careful investigation shows the potentially speculative behavior. We performed manual verification to establish a ground truth for identifying typosquatting domains. Clearly, manual classification is not perfect, but it allowed us to go in depth at domains that were ambiguous. In manual classification, we go beyond simple rules, like identifying simple one-hop defensive redirections and consider the environment, like the owner of name servers (`ns*.aexp.com` indeed belongs to American Express Inc) or potential relation between brands (Oldnavy is a subsidiary of GAP and thus `oldnavy.com` redirects to `oldnavy.gap.com`). We could further establish a ground truth based on crowdsourcing typosquatting identification. This would remove the bias introduced by the mindset of the authors, yet it could introduce significant inaccuracies due to the lack of experience and understanding of typosquatting by the crowd.

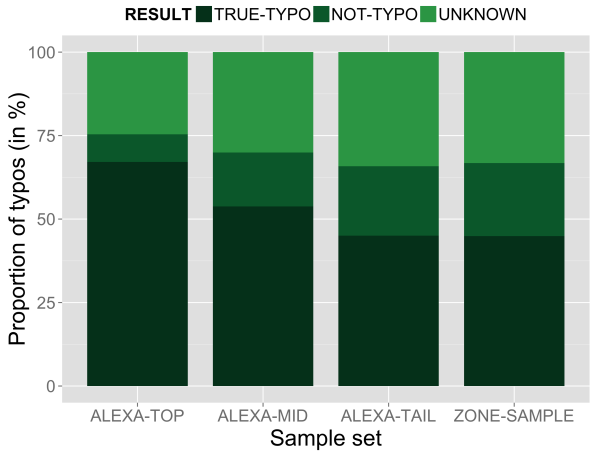


Figure 3: The prevalence of true typo domains in the four sample sets drawn popular and less popular `.com` domain names. The domain sets are ctypo samples of the Alexa top/mid/tail domains and the domains in the `.com` zone file. The number of true typo domains decreases with the Alexa rank of original domains, yet their ratio in the whole population remains high.

According to our manual inspection, a majority of the ctypo domains registered against the Alexa top domains are true typo domains (as shown in Figure 3). This result confirms the finding of [26]. We note here that there is a significant number of ctypo domains for which we cannot reliably decide if they are typo domains or not (U). This is mostly due to the fact that domains return "not accessible" responses for DNS or HTTP queries. The number of true typo domains steadily decreases when we perform the same experiment for the Alexa mid and tail domains, yet it remains high (around 50% within the set of all ctypo domains). While this indicates that thousands of domains are indeed typosquatting on less popular domains, to present defenses we need to develop a more reliable strategy to predict whether a domain is involved in typosquatting.

4.2 Accuracy of identification

We developed an automatic categorization tool based on the domain features presented in Section 3.5 called *Yet Another Typosquatting Tool (YATT)*. YATT has three modes. In the *passive mode, YATT-P* uses the information readily available from static files, such as lexical features, zone information and Alexa information. In the *DNS mode, YATT-PD* includes Whois and DNS features collected from the active crawler infrastructure, and finally in the *content mode, YATT-PDC* content features obtained via crawling are added to the categorization. The complexity of the algorithms increases from YATT-P to YATT-PDC. We expect that YATT-PDC will show the best performance in categorizing typo domains, but the other variants can still provide useful information if one wants to avoid the tedious work of collecting content features.

As presented before, we fine-tuned the parameters of YATT, but further improvement might be possible with additional features and a more complex feature selection process. At the moment, this optimization is left as future work.

In addition to YATT, we tested notable typosquatting identification methods from related work. First, we consider the method in [26], which showed that most ctypo domains of DL-1 are indeed true typos. Their primary feature is the domain length so we repeat their experiment for DL-1 and we name their method *AllTypo*. Then, we implemented the most important features of the *SUT-net* algorithm in [10] and compared it to various modes of YATT.

In Figure 4, we compare the accuracy of the typo identification methods in related work and the three modes of YATT to the established benchmark of manual evaluation. We perform this accuracy evaluation on the four ctypo domain samples described in Section 4.1. In Figure 4, we see that all five algorithms mark ctypo domains

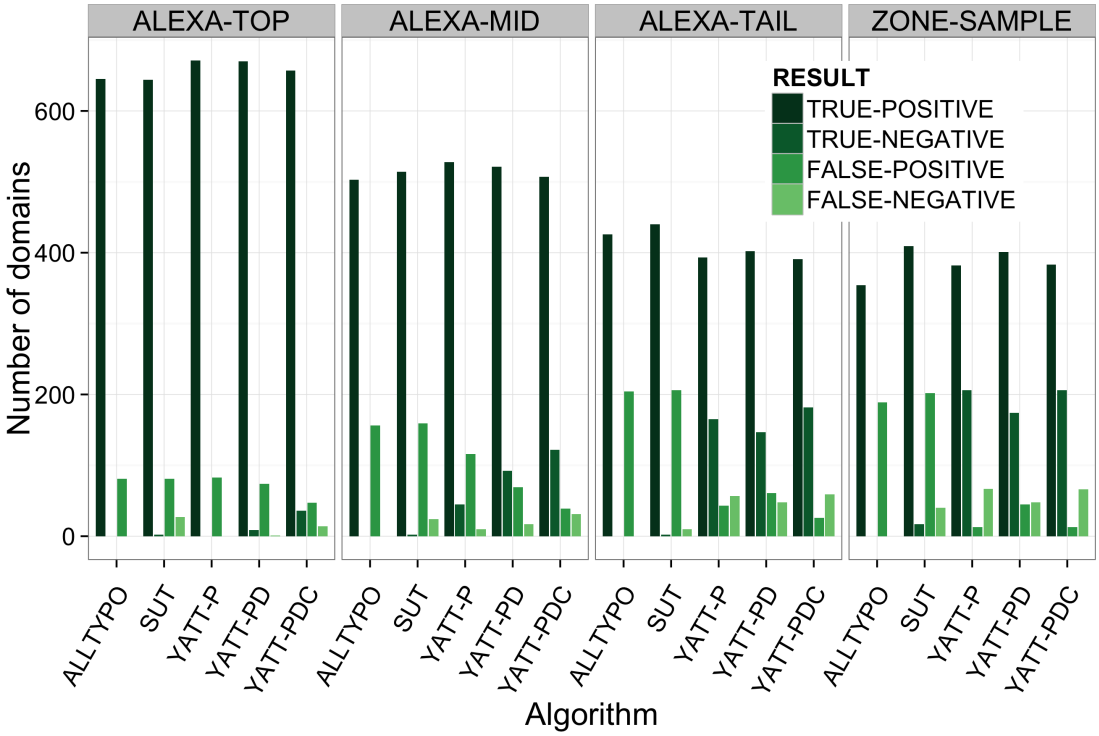


Figure 4: Accuracy of four typosquatting prediction tools. We tested (a) AllTypos, (b) SUT-net-based content features, (c) YATT-P, (d) YATT-PD, and (e) YATT-PDC for the four ctypo domain sample sets of (1/2/3) the Alexa top/mid/tail domains and (4) the domains in the `.com` zone file.

as positives in the Alexa top dataset. This assertive categorization results in a good true positive (TP) rate, a reasonably small number of false positives (FP) and with almost no false negatives (FN). Only the full YATT-PDC can identify a small set of true negatives (TN) in the population. In the Alexa mid, the agressive typo identification of AllTypo and SUT results in a high FP number whereas YATT keeps the FPs low while correctly identifying TNs (with YATT-PDC being the most accurate as expected). For the Alexa tail and zone datasets, the number of true typos further decreases and both AllTypo and SUT overwhelmingly categorize these domains as typos resulting in a very large false positive rate. All versions of YATT keep the FPs low and correctly categorize TNs at the expense of a small number of FNs. It is clear that perfect categorization is difficult to do, but YATT does not sacrifice much precision as the number of non-typo domains get introduced.

Next, we study the accuracy of the YATT-PDC to identify parked domains and other typosquatting indicators based on our manual sampling in Table 1. Note that related work on typosquatting identification usually focuses on typo identification and leaves the categorization aside. Only the active mode of the algorithm can perform this categorization, because it requires content features. YATT-PDC uses regular expression-based matching for the identification of parking domains. It matches these domains with about 85% precision, the error stemming from the incompleteness of the set of regular expressions we use. YATT-PDC still finds the majority of the parking sites and lists a significantly larger number of parking sites than methods in related work [7, 19]. For the defenisve domain registrations, YATT-PDC fares worse. It only finds 60-85% of the defensive registrations. This is due to the complexity of defensive registration patterns that can mostly be caught by a human eye. Finally, for affiliate registrations, YATT-PDC performs quite well, correctly categorizing almost all domains. We also checked the existence of malicious and phishing domains in our sample dataset, but we could not find any in such a small sample. Our results from more rigorously checking for maliciousness in typo domains is described in subsection 4.6, however maliciousness was not used to classify typo domains as typos.

YATT results in an accurate prediction of true typo domains and domain categories for the whole range of the domain population and hence its results can be used as a basis for intervention attempts and tools. Using YATT, we compile a typosquatting blacklist and use it in a set of mitigation tools (see Section 5).

	PARKED			DEFENSIVE			AFFILIATE		
	False Positive	True Positive	False Negative	False Positive	True Positive	False Negative	False Positive	True Positive	False Negative
Alexa top	3	402	76	0	39	15	0	27	1
Alexa mid	3	358	50	0	18	3	0	15	0
Alexa tail	1	295	59	0	9	3	0	0	0
Zone	0	265	43	1	7	4	0	0	0

Table 1: The accuracy of YATT to identify parked, defensive and affiliate registrations across the sample datasets.

4.3 Presence of typosquatting registrations

Having designed an accurate typosquatting identification tool, we now study the existence of typosquatting in current domains registrations. We first obtained 4.7 million ctypos targeting the `.com` domains in the Alexa top 1m domain list and existing in the `.com` zone file using the methodology described in Section 3. Recall, that we split the original domains according to their Alexa rank into the Alexa top/mid/tail categories.

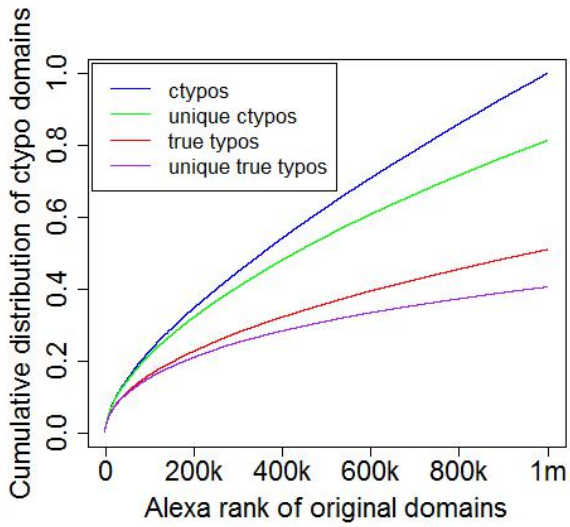


Figure 5: The cumulative distribution of true typo domains in ctypos and unique ctypos as a function of the Alexa rank of the original domains.

The first and foremost question is the extent of typosquatting targeting the Alexa domain set. We use YATT to determine typosquatting behavior and partition ctypo domains into the categories described in Section 3.5. In Figure 5, we plot the cumulative distribution of ctypo domains as a function of the originals' Alexa rank, and we also plot the cumulative distribution of true typo domains. We see that the number of true typos steadily increases as the Alexa rank decreases, although at a slower pace than the number of ctypos. In addition, we also plot the cumulative distribution of unique ctypos and true typo domains.

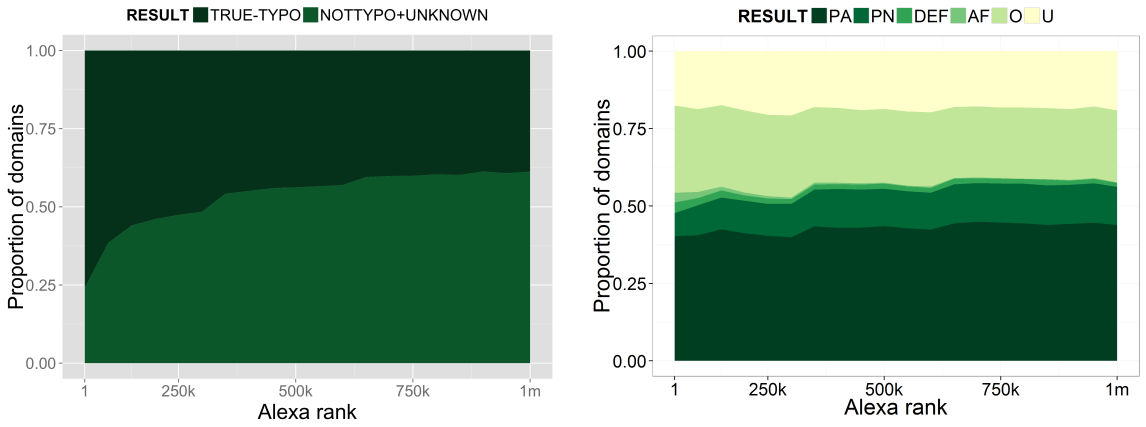
We then show the fraction of true typos in the population of ctypos in Figure 6(a). We calibrated YATT to make a decision about each ctypo and thus it conservatively categorizes the majority of unknown domains as not typos. For Alexa top sites, the fraction of true typos is higher, but for lower Alexa ranks the number of not-typo and unknown domains increases. This is consistent with our benchmarking results in Figure 3. Finally, in Figure 6(b), we present the typosquatting categories as a function of the original domains' Alexa rank. We observe that the bulk of the true typo registrations profits from parked domains with advertisements. The number of defensive and affiliate registrations is higher for the Alexa top sites, but then then the affiliate registrations disappear as we head to the Alexa tail while the defensive registrations persist. Finally, there is a significant number of non-typo domains incidentally close to the domains in the Alexa domain list.

Projecting our results to the total number of `.com` domains in the zone file, we estimate that about 53% of them are candidate typo domains and hence 20% of the total domain set are true typo domains. Based on our results, we estimate that about 21.2m domains are true typo domains in the `.com` zone file.

4.4 Trend analysis

We analyzed trends in typo domain registrations for a period of approximately one year (from 2012-10-01 to 2013-10-15). We considered domains from four datasets: domains from the `.com` zone file, ctypos from the `.com` zone file, ctypos targeting the whole Alexa list and ctypos targeting the Alexa top list.

For the purposes of our analysis, we use visibility into the `.com` zone file as a proxy for domain registration. Because the actual registration and registration lapse events are not visible to us, we use presence in the zone file as a proxy for registration events. We define a registration



(a) (b)

Figure 6: The existence of typosquatting domains targeting the Alexa domain set. The fraction of (a) true typo domains and (b) various typo categories in the true typo population.

event as one where a domain was not in a daily zone dump, and was present in the subsequent day's zone file, and vice-versa for a registration lapse, or deregistration.

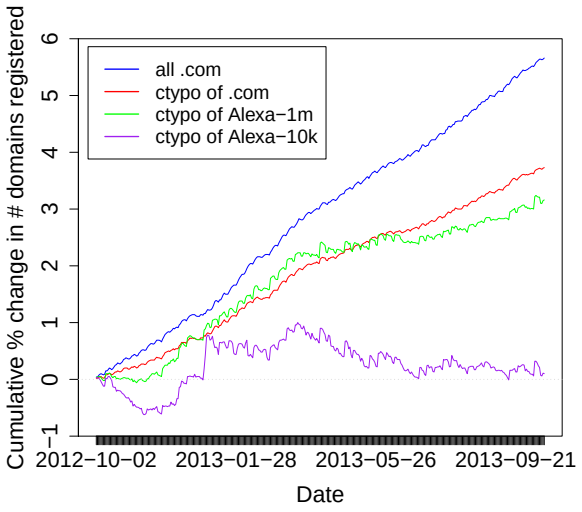


Figure 7: Cumulative change in the total number of domains registered over time.

We looked at the change in domain registrations over time. Figure 7 plots the cumulative changes in the number of domains registered in the above mentioned domain sets. While the overall registration rate is steady, the difference between the rate of Alexa-10k targeted and Alexa-1m targeted typos suggests that, through enforcement or typosquatter preference, the overall increase in registrations targeting popular domain typos is far smaller even though

	Stable	Mean uptime	Reregs
Alexa-1m ctypo	72.3%	458 days	49.5%
Alexa-10k ctypo	71.0%	454 days	49.5%
Alexa-1m	93.3%	501 days	67.1%
Alexa-10k	99.0%	506 days	86.8%
Random sample	70.4%	440 days	28.5%

Table 2: Speculation trend analysis between 2012-10-01 and 2014-02-20. Alexa list and zone file used was from 2012-10-01. The "stable" column indicates what proportion were registered throughout the dataset. "Reregs" indicates how many domains experienced at least one lapse in visibility at the zone file, indicating that the domain was decommissioned and then reactivated. "Random sample" is a selection of 2 million random domain names from the `.com` zone file of 2012-10-01.

many DL-1 typos of popular domains are still available. It is also interesting to note that the spike centered on January 1 2013 is due to four organizations (sedoparking, 1and1.com, dsredirection, and graceperioddomain.com) registering a large number of domains: these four account for 87% of all domains registered at that time.

Our next analysis focuses on the amount of speculation present within the market for typosquatting domains between 2012-10-01 and 2014-02-20. Table 2 shows the percentage of stable domains, the average uptime, and the percent of domains experiencing at least one rereg-istration event during our measurement time period. As might be expected, random domains are purchased and left to lapse very often, with less than one third being reregistered after being abandoned. Domains which are a typo of a popular domain, however, experience almost

twice as much interest, although they are not active for significantly more time. This trend suggests that the information asymmetry of the typosquatting marketplace is such that new speculators register old typos at a much higher rate than random domains.

4.5 Typosquatting redirections

Now, we scrutinize the affiliate redirections via third-parties. This third-parties can be legitimate brand protection companies, but more frequently they are typosquatting affiliates collecting type-in traffic from a large number of typo domains.

Domain redirections that lead back to the targeted original domains without intermediate domains are considered defensive registrations, as explained in Section 2.1. If the redirection leads back to the target domain via a third-party, then we call it an affiliate defensive registration. In Figure 8 the first graph shows that in the cumulative distribution of third party landing pages, eleven domains (less then 0.1 percent of all of these landing pages) get redirections from more than 50 percent of ctypos redirecting to a third party domain. The second graph in Figure 8 shows defensive affiliate domains, where the landing pages is the original domain, but the traffic goes through an intermediate affiliate domain. 18 such intermediate domains (1.3 percent of all domains) are responsible for more than 80 percent of defensive affiliate marketing. Even though this set has a very small overlap with the non-defensive affiliate domains, a small fraction of affiliate domains are controlling 80 percent of the affiliate market.

Finally, if the redirection leads to a third-party domain, that is away from the original target, then this is considered truly speculative. The third graph in Figure 8 shows redirections to third-party pages with only one redirection. Here the domains are more widely distributed: there is only one big landing domain hugedomains.com which receives traffic from more than 21 percent of this type of redirection. The last graph shows the cumulative distribution of all affiliate domains participating in third-party redirections with a non-defensive purpose. That means that these affiliate domains lead away the users from the targeted original sites. 41 of these non-defensive affiliate domains (0.4 percent of all such domains) control the traffic originating from more than 80 percent of candidate typo domains. This means that, here too, a relatively small set of domains control the majority of such traffic going to a few landing pages.

4.6 Maliciousness of Typo Domains

In order to test the assertion that typo domains are more malicious than other domains, the candidate typo (ctypo) and true typo (ttypo) domains extracted from the `.com` were checked against a variety of available black lists. These results are compared against the same test on the Alexa domains. By using 12 available black lists from various sources fluctuations due to the idiosyncrasies of any individual list can be controlled.

	# Malware Hits	% of List Marked Malware	# Phish Hits	% of List Marked Phish
Alexa	9990	1.907%	27	0.005153%
ctypos	17485	0.3716%	272	0.005781%
ttypos	3720	0.1585%	125	0.005329%

Table 3: Google Safe Browsing results for domains in Alexa, *ttypos*, and *ctypos*.

The Alexa top 488,133 `.com` domains (all the `.com` domains in the top 1m) are more likely to appear on black lists than the typos of them, either *ctypos* or *ttypos*. This result is consistent across all 12 black lists investigated. In each case, the Alexa domains are more likely to host malicious activity. The percentage of .com domains from the Alexa list on each black list is always higher than the percentage of *ttypo* domains on the same list.

Google's Safe Browsing list requires a different checking method, due to their storage method. The list also distinguishes between a match due to malicious content or attempts at phishing. However, the results show a similar trend. The Alexa domains are more likely to be purveyors of malicious software. Table 3 shows the results for Google Safe Browsing checking for any listing from May 1, 2011 – July 31 2013.

There are several possible causes for this pattern, and several of them would be uninteresting. A possibility is that there is a pocket of malicious activity using typos, but that most of it is benign. The first place to look for this would be the name servers hosting predominantly typo domains. There are 10 name servers for which most of the domains they host are typos of other domains—for these name servers, between 20-80% of their domains are typos.

The typo domains hosted on these 10 name servers seem to be even less likely to appear on a black list. The average percentage of these name servers' domains on any of the black lists is 0.051%, and the maximum percentage of typo domains hosted by one of these name servers on any one list is 0.27%. Both of these numbers are below those both for typos generally as well as the results for the Alexa domains.

5 Intervention options

Just as defining typosquatting remains one of the grey areas of domain name security, developing effective in-

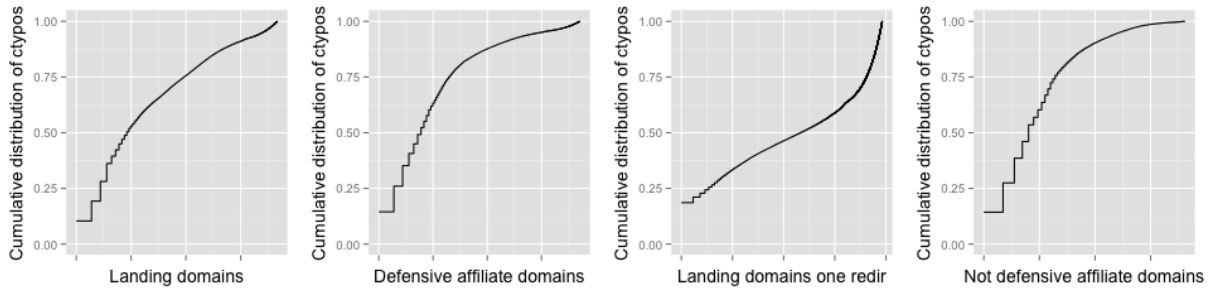


Figure 8: The leftmost figure shows the cumulative distribution of landing pages targeted from ctypo domains. The second figure shows the cumulative distribution of intermediate domains in case of defensive redirections. The third figure is when the length of the domain redirection chain is one. Finally, the rightmost figure shows the cumulative distribution of intermediate domains in case of redirections targeting a third party.

tervention techniques is similarly difficult. So far, most intervention attempts remain ineffective. In the following, we present viable typosquatting mitigation options and present a set of practical tools to prevent typosquatting from negatively effecting users.

5.1 Policy intervention

Much of the effort to crack down on typosquatting focuses on policy options. Two major tools exist for policy intervention. The first is the UDRP arbitration framework provided by ICANN [21]. Unfortunately, only a small fraction of typosquatting domains enters the UDRP procedure [26], although domains are claimed by their trademark holders very often.

The Anti-cybersquatting Consumer Protection Act (ACPA) (15 USC §1125(d)) offers an alternative to the UDRP through legal action. The act "was designed to thwart cybersquatters who register Internet domain names containing trademarks with no intention of creating a legitimate web site, but instead plan to sell the domain name to the trademark owner or a third party." While originally aimed at preventing cybersquatting, in May 2013 Facebook successfully litigated a case including typosquatting domains, earning a US $2.8 million judgement [18]. As with any legal action, the enforcement of this act is costly and only major trademark holders have exercised their legal rights [25, 31, 32]. Additionally, the bad faith of typosquatting registrations is difficult to prove and hence the legal action might not always be efficient [30]. Unfortunately, even vigilant companies seem overwhelmed by the number of typosquatting domains targeting their brands, motivating them to litigate; even so, many of their domains are still controlled by typosquatters.

5.2 Infrastructure support

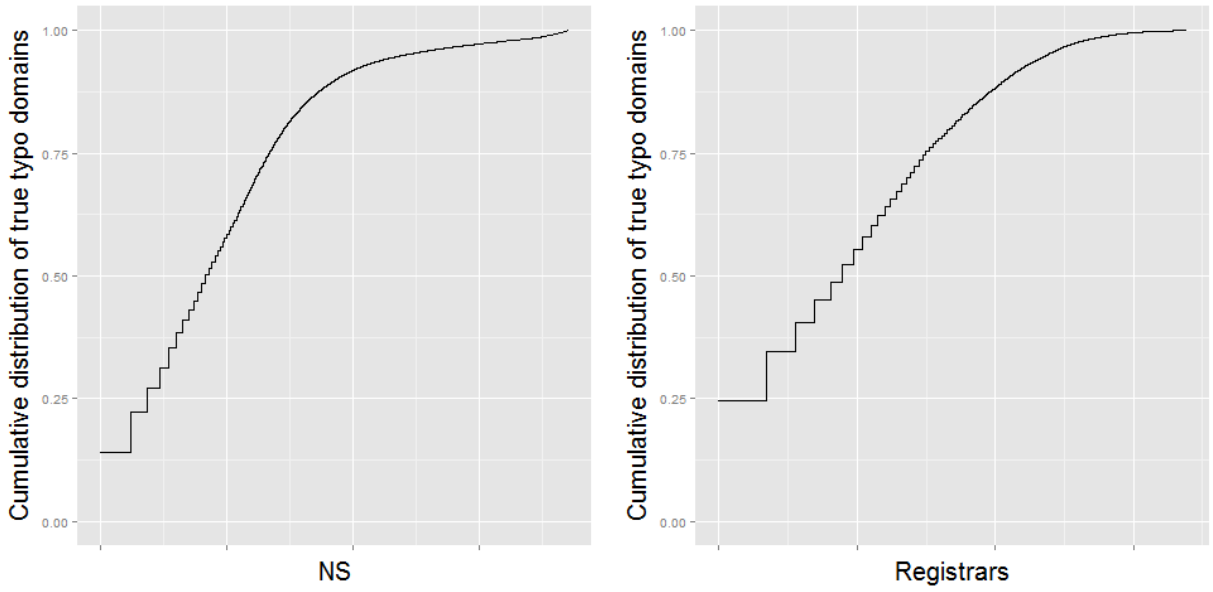
Another option for intervention is to motivate registrars and hosting providers to scrutinize domain name registrations when they happen (with a mandatory light-weight UDRP procedure for example). Let us now look at the potential of registration intervention at the infrastructure side. Figure 9 shows the distribution of typosquatting domains (a) as a function of the registrars and (b) as a function of the supporting NSs (setting the x axis to a log scale to improve visibility). We observe that most true typo domains cluster at major registrars and are hosted at a few NSs. In particular, 12 NSs and 5 major registrars are responsible for hosting 50% of the true typo domains. Forcing these major registrars to enforce prudent registration practices with respect to typosquatting may be a viable policy option.

NS	True typos	All domains	Typo ratio
`a0f.net`	5221	6332	82%
`citizenhawk.net`	8819	12004	73%
`easily.net`	18281	36890	50%
`domainingdepot.com`	51854	132864	39%
`next.org`	9426	30252	31%
`domainmanager.com`	23493	90929	26%

Table 4: Worst offender NSs in true typo hosting with at least 5000 true typo domains. All NSs in the top list have higher than 25% of true typo / all domain ratio.

Based on the `.com` zone file, we are also able to collect the ratio of true typo domains to the total number of domains. Table 4 presents the top offenders with at least 5000 true typo domains hosted. Interestingly, there are only 65 NSs with such a high number of true typo domains. We see that the worst offenders almost exclusively host true typo domains, and none of them belong to the major hosting companies[8]. Further investigating

[8] An interesting case might be `citizenhawk.net`, a brand protec-



(a) (b)

Figure 9: Intervention potential at domain registrars and hosting companies. We present the distribution of typosquatting domains (a) as a funtion of the registrars and (b) as a function of the supporting NSs (while setting the x axis to a log scale for better visibility)

these typo domains we found two interesting results. First, out of the 6 name servers with the highest true typo ratio, 5 have domains that are privately registered and only `citizenhawk.net` is not, showing that the others are aware that their monetization strategy is questionable. Second, we found that on the average 24.5 percent of the domains hosted by these NSs is in the top Alexa, which is 2.5 time higher number than for the rest of the name servers. This indicates that these name servers are more effectively targeting popular typo domains than major hosting services who are not focusing on typosquatting. These hosting companies with an unusually high number of true typo domains could be regulated to effectively decrease the effect of speculative typosquatting.

Infrastructure intervention is promising if it can be enforced globally by ICANN on the supporting providers. Unfortunately, it is unlikely that such a global action will emerge as this is counterproductive for the domain registrars, and thus miscreants can always shift their businesses to negligent or accomplice providers who are financially motivated to assist their businesses. Registrar- and hosting-level intervention remains ineffective against spammers [23, 24] and it is unlikely that it will be effective against typosquatting. Registrars and hosting companies do not suffer from typosquatting, thus there is little economic incentive for them expend resources to defend against it.

tion company who probably registered a large number of domain names for protecting their customers.

5.3 Mitigation tools

The last option to counter typosquatting is the application of technical tools to reduce the impact of typosquatting. There exist mitigation tools to this end, but most tools suffer from either trivial errors or from small coverage of typosquatting domains.

Related work. Wang et al. developed *Strider Typopatrol*, a tool to automatically discover typo domains of popular domains [35]. They focus on a small subset of the Alexa top domain list [1], phishing targets, and childrens' websites. OpenDNS [27] provides typosquatting correction in their DNS services, but only for major TLDs. A similar tool called *URLFixer* [6] was introduced in the Adblock Plus advertisement blocking tool. The URLFixer tool includes misspellings of top Alexa domains, but fails to correct less popular domain names and includes some short domain names leading to false corrections. Chen et al. [11] develop a browser plugin to check typo domains based on a user-customized local repository. Banerjee et al. [9, 10] propose *SUT*, a method to identify typosquatting domains mostly based on HTML properties. Finally, the autocomplete feature of most major browsers can also decrease the instance of typos, albeit only for previously visited sites.

Initial tests show that most existing solutions are lim-

ited in scope (the most popular domains or most frequent typos), in features (only TLD correction or HTML features) or in the information used (search typing or local browser history) and consequently these tools are missing a large set of typosquatting domains.

The YATT framework. We developed a typosquatting categorization tool, YATT, that uses an extended domain feature set to provide accurate typosquatting identification. Based on the output provided by YATT, we implemented three typosquatting detection and protection services. The first service is a typosquatting blacklist (YATT-BL) compiled from the output of one of the versions of the YATT tool. As a DNS based blacklist, this access method is quick and lightweight. The tool works similarly to major domain blacklists such as URIBL [5], SURBL [4] or the Spamhaus DBL [3] and it can be used to filter out typo domains from live traffic. The DNS server uses RPZ [34] to efficiently distribute the typo list.

Second, we implemented a Firefox browser plugin and a corresponding typo protection server to protect users from typosquatting domains. Our plugin contacts the typo protection server each time a user types in a domain and raises a warning if the domain typed by the user is found on the typosquatting domain list. The user is provided with the option of accepting the automatic correction or rewriting it to her needs. The typo protection server uses YATT-BL DNS blacklist described above.

Third, we are in the process of implementing a YATT DNS server for organizations that want to avoid typosquatting yet do not want to expose their DNS traffic to a third party server. Using this tool, a company could periodically download an updated typosquatting blacklist and query it locally.

6 Conclusion

Typosquatting has caused annoyances for Internet users for a long time. Since users lack effective countermeasures, speculators keep registering domain names to target domains and exploit the traffic arriving from mistyping those domain names. Existing studies of typosquatting focused on popular domain names and thus have only shown the tip of the iceberg. Similar to traditional cybercrimes like spamming or financial credential fraud, typosquatting has minimal transparency, allowing what may be an unprofitable activity to continue because new entrants see its effects and attempt to become profitable typosquatters themselves. Investigating such speculative, "gray area" behavior longitudinally can give us insights which might generalize to traditional cybercrime and cybercriminals.

In this paper, we performed a thorough study for an extensive set of potential target domains. We found that 95% of typo domains are targeting less popular domains. We designed an accurate typo categorization framework and find that typosquatting using parked ads and similar monetization techniques not only exists for popular domains, but a whole range of domain names in the Alexa domain list. We showed that a large number of incidental domain registrations exist with close lexical distance to the target domains. Our conservative estimates indicate that as much as 21.2 million `.com` domain registrations are confirmed true typo domains, which accounts for about 20% of all `.com` domain registrations. Additionally, we found that the typosquatting phenomenon is only continuing to thrive and expand.

The difficulty of categorizing typosquatting domains partially explains the inefficiency of existing mitigation techniques. Much like typosquatting itself, mitigation is a gray area: one cannot easily classify a new registration as an example of typosquatting based on the name alone. As such, typo domains rarely appear on blacklists. To counter this problem, we designed several defense tools that rely on a broad range of features. We provide a typosquatting blacklist and a corresponding browser plugin to prevent mistyping at the user side. While typosquatting will likely continue to exist, these analyses and tools may improve user experience and further decrease the profit available to typosquatters.

7 Acknowledgements

We thank the anonymous reviewers and the PC at large for their helpful feedback. We also thank Nicolas Christin for his support. This work was made possible in part by the National Science Foundation grant NSF CNS-1351058. Mark Felegyhazi was supported by the Bolyai Janos Research Fellowship Nr: BO/00273/12. Janos Szurdi has been supported by the Ann and Martin McGuinn Graduate Fellowship.

References

[1] Alexa top sites. http://www.alexa.com/topsites.

[2] PhishTank. http://www.phishtank.com.

[3] Spamhaus DBL. http://www.spamhaus.org/dbl/.

[4] Surbl domain blacklist. http://www.surbl.org/lists.

[5] Uribl domain blacklist. http://www.uribl.com/about.shtml.

[6] Urlfixer for mozilla firefox. http://urlfixer.org/.

[7] ALMISHARI, M., AND YANG, X. Ads-portal domains: Identification and measurements. *ACM Transactions on the Web (TWEB) 4*, 2 (2010), 4.

[8] AVAST. Misspelling goes criminal with typosquatting. https://blog.avast.com/2012/03/23/misspelling-goes-criminal-with-typosquatting/, Mar 23 2012.

[9] BANERJEE, A., BARMAN, D., FALOUTSOS, M., AND BHUYAN, L. N. Cyber-fraud is one typo away. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE* (2008), IEEE, pp. 1939–1947.

[10] BANERJEE, A., RAHMAN, M. S., AND FALOUTSOS, M. SUT: Quantifying and mitigating url typosquatting. *Computer Networks 55*, 13 (2011), 3001–3014.

[11] CHEN, G., JOHNSON, M. F., MARUPALLY, P. R., SINGIREDDY, N. K., YIN, X., AND PARUCHURI, V. Combating typo-squatting for safer browsing. In *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on* (2009), IEEE, pp. 31–36.

[12] COULL, S. E., WHITE, A. M., YEN, T.-F., MONROSE, F., AND REITER, M. K. Understanding domain registration abuses. *Computers & security* (2012).

[13] DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Communications of the ACM 7*, 3 (1964), 171–176.

[14] DANCHEV, D. Legitimate software typosquatted in SMS micro-payment scam. blog, http://ddanchev.blogspot.com/2009/07/legitimate-software-typosquatted-in-sms.html, Jul 7 2009.

[15] EDELMAN, B. Large-scale registration of domains with typographical errors. http://cyber.law.harvard.edu/people/edelman/typo-domains/, Sep 2003.

[16] EDELMAN, B. Estimating visitors and advertising costs of typo domains. http://www.benedelman.org/typosquatting/pop.html, 2010.

[17] FELEGYHAZI, M., KREIBICH, C., AND PAXSON, V. On the potential of proactive domain blacklisting. In *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more* (2010), USENIX Association, pp. 6–6.

[18] GROVE, J. V. Facebook wins millions in case against typo squatters. http://www.cnet.com/news/facebook-wins-millions-in-case-against-typo-squatters/, 2013.

[19] HALVORSON, T., SZURDI, J., MAIER, G., FELEGYHAZI, M., KREIBICH, C., WEAVER, N., LEVCHENKO, K., AND PAXSON, V. The BIZ top-level domain: ten years later. In *Passive and Active Measurement* (2012), Springer, pp. 221–230.

[20] HIDAYAT, A. Phantomjs. http://phantomjs.org/, 2013.

[21] ICANN. Uniform domain name dispute resolution policy (UDRP). http://www.icann.org/en/help/dndr/udrp, 1999.

[22] ICANN. The end of domain tasting - status report on AGP measures. http://www.icann.org/en/resources/registries/agp/agp-status-report-12aug09-en.htm, Aug 12 2009.

[23] LEVCHENKO, K., PITSILLIDIS, A., CHACHRA, N., ENRIGHT, B., FELEGYHAZI, M., GRIER, C., HALVORSON, T., KANICH, C., KREIBICH, C., LIU, H., ET AL. Click trajectories: End-to-end analysis of the spam value chain. In *IEEE Symposium on Security and Privacy, 2011* (2011), IEEE, pp. 431–446.

[24] LIU, H., LEVCHENKO, K., FELEGYHAZI, M., KREIBICH, C., MAIER, G., VOELKER, G. M., AND SAVAGE, S. On the effects of registrar-level intervention. In *Proceedings of 4th USENIX LEET* (2011).

[25] MICROSOFT TECHNET. The trouble with typosquatting. http://blogs.technet.com/b/microsoft_on_the_issues/archive/2010/04/15/the-trouble-with-typosquatting.aspx, Apr 15 2010.

[26] MOORE, T., AND EDELMAN, B. Measuring the perpetrators and funders of typosquatting. In *Financial Cryptography and Data Security* (2010), Springer, pp. 175–191.

[27] OPENDNS. There's no "i" in twtter: How to outsmart typosquatting. http://blog.opendns.com/2011/09/02/there%E2%80%99s-no-%E2%80%9Ci%E2%80%9D-in-twtter-how-to-outsmart-typosquatting/, Sep 2 2011.

[28] SOGHOIAN, C., FRIEDRICHS, O., AND JAKOBSSON, M. The threat of political phishing. In *The International Symposium on Human Aspects of Information Security & Assurance* (2008), Citeseer.

[29] SOPHOS, NAKED SECURITY. Typosquatting - what happens when you mistype a website name? http://nakedsecurity.sophos.com/typosquatting/, Dec 14 2011.

[30] SUNDERLAND, S. D. Domain name speculation: Are we playing whac-a-mole. *Berkeley Tech. LJ 25* (2010), 465.

[31] TECHCRUNCH.COM. U.S. Court Rules For Facebook In Its Case Against Typosquatters On 105 Domains; $2.8M In Damages. http://techcrunch.com/2013/05/01/u-s-court-rules-for-facebook-in-its-case-against-typosquatters-on-105-domains-2-8m-in-damages/, May 1 2013.

[32] THE NEXT WEB. Typosquatting sites 'wikapedia' and 'twtter' have been fined $300,000 by UK watchdog. http://thenextweb.com/insider/2012/02/16/typosquatting-sites-wikapedia-and-twtter-have-been-fined-300000-by-uk-watchdog/, Feb 16 2012.

[33] THE REGISTER. Typosquatters set up booby-trapped High Street names. http://www.channelregister.co.uk/2011/12/13/typosquatting_scams_target_xmas_shoppers/, Dec 13 2011.

[34] VIXIE, P. Taking back the DNS. http://www.isc.org/community/blog/201007/taking-back-dns-0, Jul 29 2010.

[35] WANG, Y.-M., BECK, D., WANG, J., VERBOWSKI, C., AND DANIELS, B. Strider typo-patrol: discovery and analysis of systematic typo-squatting. In *Proc. 2nd Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI)* (2006).

[36] WEAVER, N., KREIBICH, C., AND PAXSON, V. Redirecting DNS for ads and profit. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI), San Francisco, CA, USA (August 2011)* (2011).

[37] WEBSENSE SECURITY LABS. The rise of a typosquatting army. http://community.websense.com/blogs/securitylabs/archive/2012/01/22/The-rise-of-a-typosquatting-army.aspx, Jan 22 2012.

Appendices

A Features used for domain categorization

Feature description	Priority	Comment
Lexical attributes		
domain length	M	[26]
highest-ranked neighbor's operation	M	diff. from the most popular original domain
is any neighbor at fat finger distance one?	M	FF typos are more likely to be true typos [26]
nr. of neighbors	L	
nr. of neighbors with *op*	L	where *op*={*add*,*del*,*sub*,*tra*,*www*}
Popularity (Alexa) attribute		
Alexa rank of original domain	H	
Zone file attributes		
total nr of ctypo-s on NS	M	
ctypo/alldomain ratio on NS	H	
total nr. of domains on the NS in the zone	L	
parked keywords in NS domain	H	
Whois attributes		
total nr of ctypo-s at registrar	M	
registration date	L	
DNS attributes		
NXDOMAIN wildcarding	H	
TXT google auth	L	Google ads affiliate auth
total nr of ctypo-s on IP address	M	[10]
Content attributes		
Parked	H	by RE keywords
Serving ads	M	by RE keywords
Total redirection length	M	# of redirections [10]
Domain redirection length	H	# of redirections between registered domains
DERPContent size	M	[10]
Affiliate marketing	M	[26]

Table 5: Domain and infrastructures features to categorize candidate typo domains. The column Priority indicates the relative importance in idenfitying typosquatting behavior.