It's Not What It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains

Florian Quinkert[⋆], Tobias Lauinger[‡◇], William Robertson[‡], Engin Kirda[‡], and Thorsten Holz[⋆]
[⋆]Ruhr-University Bochum, *{firstname.lastname}@rub.de* [‡]Northeastern University [◇]University of Chicago

***Abstract*—International Domain Names (IDNs) may contain Unicode in addition to ASCII characters. This enables attackers to replace one or even more characters of a well-known domain with visually similar Unicode characters to create new, look-alike domains. These so-called *homograph domains* are attractive for malicious activities such as phishing or scams because they may appear legitimate to potential victims.**

In this paper, we propose two measurement setups to detect homograph domains and monitor their activity. Throughout eight months, we detected almost 3,000 homograph domains, targeting technology companies as well as financial institutions. To understand this phenomenon in more detail, we monitored the activity of these domains daily for more than five months and identified multiple instances of scamming and phishing, with some campaigns being active for several months. We also detected previously undiscovered domains used for a widespread scam in which attackers promise free shoes and other goods. In many cases, these domains were not detected by classical detection approaches such as VirusTotal or Google Safe Browsing, or reported only with a delay of several days or weeks compared to our approach. While we did observe defensive registrations of homograph domains by domain owners, we found that they were very limited in scope and did not cover all possible look-alike character replacements. To that end, we conclude our paper with recommendations for domain owners.

***Index Terms*—homograph domains, measurement study, domain registration purpose, phishing, defensive registrations**

I. Introduction

Domains are an important building block of today's Internet: They prevent users from having to remember plain IP addresses by providing easily memorable strings instead. More specifically, the Domain Name System (DNS) is used to translate domain names to IP addresses and vice versa. As such, it is crucial for surfing the Internet, exchanging e-mail messages, and similar online tasks. At the beginning of the Internet, allowed characters in domains included only ASCII characters such as Latin letters, numbers, and the dash sign. However, many people around the world use characters from additional alphabets, and hence they could not use domains in their native alphabets. To overcome this problem, the Internet Corporation for Assigned Names and Numbers (ICANN) introduced so-called *International Domain Names* (IDNs) which enable the usage of a variety of additional characters (i.e., Unicode characters) within a given domain name. From a technical point of view, this is implemented via a mechanism to represent IDNs in ASCII called *Punycode*. For example, the official website of the city Munich (*München* in German) can be encoded as *xn–mnchen-3ya.de* and the city of Krakow is spelled Kraków in Polish, which can be encoded as *xn–krakw-3ta.pl*.

An attacker can take advantage of Punycode as follows: some characters allowed in IDNs are visually very similar to (and sometimes even indistinguishable from) Latin characters. As a result, attackers can abuse IDNs by replacing characters in well-known domains with their visually undistinguishable counterparts to create new, look-alike domains that can be used in phishing attacks or other types of scams. For example, the Cyrillic letter *а* looks very similar to the letter *a* in the Latin alphabet used within ASCII, which allows an attacker to replace one of the two letters a in paypal.com to generate a domain that is hard to distinguish from the legitimate PayPal domain. This type of attack is a known threat, and in the literature, this attack is called *homograph attack*, while the used domains are called *homograph domains*. Recently, homograph domains attracted more attention and were covered in multiple blog postings [1]–[3], [6]. In contrast, the last scientific analysis of homograph domains is more than 12 years old [18]. Recently, Liu et al. gave an updated overview of IDN usage [22], but they did not focus on homograph domains.

In this paper, we perform a longitudinal analysis of homograph domains and study how they are used nowadays. To this end, we developed an analysis infrastructure to detect homograph domains in a systematic way, where the goal was to provide a current overview of the phenomenon with a focus on purpose of registration and used infrastructure. As input, our analysis setup takes a list of well-known domain names along with a list of Latin characters as well as numbers and their visually undistinguishable Unicode counterparts. We refer to these well-known domain names as *reference domains* and to the list of character pairs as *homograph pairs*. Based on this data set, we search in newly registered domains for IDNs which differ from our reference domains only by characters in homograph pairs. In this paper, we thus focus on substitutions and homograph domains that differ from reference domains only in homograph pairs because they represent domains that might be related to a homograph attack. We refer to such domains as *candidate domains*. To enrich our data set beyond information on domain names, we also built a measurement setup to track the activity of the candidate domains. More specifically, we collected activity information for each candidate domain daily, such as associated IP addresses and screenshots.

In an extensive measurement study with 10,000 reference domains and a daily feed of newly registered domains

covering eight months, we detected almost 3,000 candidate domains targeting a total of 819 distinct reference domains. We studied these candidate domains in detail by collecting activity information for them for more than five months. We found that especially technology companies (e.g., google.com, facebook.com, or apple.com) and financial institutions (e.g., binance.com or paypal.com) are targeted by homograph attacks. We found that the vast majority of candidate domains (2,393, 80%) consisted of a single character replacement, indicating that attackers typically take a low-effort approach by substituting only one character. With our analysis setup, we were able to identify multiple instances of scamming and phishing, with some campaigns being active for several months. For example, we uncovered scam schemes where attackers impersonate sports apparel and airline companies, among others: domains such as adidạs.com or deltå.com promised free shoes and other goods, but they were in fact only operated by a scamming group. Surprisingly, these domains were not detected by current industry tools such as VirusTotal (VT) or Google Safe Browsing (GSB): only one of the scam domains detected by us was reported by VT (with a delay of more than 1.5 months), and GSB flagged none. Overall, we discovered more than 200 scam domains that were never listed by current security solutions, suggesting that our approach can uncover domains sooner than the current industry standard. Noteworthy is the high number of candidate domains for the insurance company *Allstate* (70) and the bug bounty website *Hackerone* (59). A closer examination revealed that these domains were used in a defensive way, i.e., the company proactively registered these domains to prevent a homograph attack. In total, we detected 239 candidate domains (8%) that were likely used in a defensive way, a rather small number considering the threat potential of homograph attacks.

In summary, we make the following contributions:

- We present the design and implementation of a measurement infrastructure to study homograph domains.
- We perform an extensive measurement study in which we identified about 3,000 candidate domains. Furthermore, we analyze these domains in detail and study this attack vector from several perspectives.

II. Related Work

There is a body of related work on various techniques that attackers use to generate domains similar to well-known ones, which is often referred to as *domain squatting*.

A. Homograph Domains

In 2006, Holgers et al. combined passive network monitoring and active DNS probing to measure the prevalence of homograph domains [18]. They found that popular websites often had multiple similar domain names registered. In most cases, only one character was substituted, and latin substitutions were used more often than IDN homographs. The registered domain names displayed advertisements, redirected to competitor websites, or spoofed content. The long time that has elapsed since this initial study calls for an updated measurement in order to reassess whether this form of attack has become more prevalent, and how the attacks have evolved since then. We provide more detail on the replaced character pairs and the domain registration purpose, and analyze novel aspects such as the underlying infrastructure.

In 2018, Liu et al. [22] conducted a study on the use of IDNs. As part of this study, they also performed a high-level analysis of homograph domains. In contrast to our work, their detection is based on the visual similarity of rendered images of domain names. This approach is highly dependent on the font used for rendering, and details are not provided in the paper. Our work uses a ten times larger set of reference domains and includes a much more detailed study of the websites hosted on homograph domains, as well as insights derived from their supporting infrastructure.

B. Other Types of Domain Squatting

To find phishing domains, Tian et al. [28] investigated multiple domain squatting techniques (e.g., homograph domains or *typosquatting*, where the attacker generates a domain that differs from a well-known domain name only in a typical typing error). The authors do not explain in detail how they identified homograph domains; their examples suggest a very broad definition, including, for instance, gougle.com for google.com. The focus of their work was not to analyze homograph domains in particular, but to detect phishing domains using different domain squatting techniques.

Agten et al. analyzed the use of typosquatting by monitoring the typosquatted versions of the 500 most popular domains during seven months [13]. Users might access typosquatted domains by their own initiative, such as by mistyping a well-known domain. To a certain extent, this may explain the higher frequency of typosquatted domains compared to homograph domains, as attackers need to advertise homograph domains through channels such as e-mail spam in order to reach their victims. Agten et al. concluded that 95% of the examined popular domains had typosquatted versions registered, that defensive registrations were uncommon, and that content on the typosquatted domains was changed regularly to earn money differently.

Kintis et al. introduced the concept of *combosquatting*, a domain squatting technique in which attackers add proper terms to a well-known domain name, e.g., paypal-secure-login.com [19]. The authors analyzed 468 billion DNS records and discovered a growing number of combosquatting domains every year as well as a variety of fraudulent use cases, such as phishing or social engineering. Nikiforakis et al. explored the use of flipped bits [26] and homophones [25] for the generation of squatted domains.

In 2017, Liu et al. presented a study on *domain shadowing* [23]. Instead of generating new domains, attackers gain access to legitimate domains and create subdomains for their malicious purposes. The authors underline that shadow domains are a rising threat and are often used for phishing. Recently, Lauinger et al. [21] and Miramirkhani et al. [24] analyzed what happens after domain names expire. Bilge et al. [15] and Antonakakis et al. [14] proposed two

systems to detect malicious domains in 2011, while Hao et al. [16] predicted malicious usage of a domain at the time of registration. Our approach complements such systems by covering homograph domains.

III. Approach

Our analysis setup consists of two separate steps. As the first step, it detects homograph domains (Figure 1). Subsequently, it collects additional metadata of these domains daily for learning more about their purpose (Figure 2).

A. Detecting Homograph Domains

To find homograph domains, we need a list of registered domains as input (step 1 in Figure 1). We use *Domain-lists.io* [8] to obtain an updated list of registered domains each day. In the first step, the module *Prefilter* (2) removes all non-IDNs from this list; we only consider homograph domains using a "special" character to replace a "normal" character. Furthermore, as the analysis pipeline is intended to run daily, we only process new domain registrations, unless it is the very first time the pipeline is running.

The module *Candidate Finder* takes the newly registered IDNs along with a list of reference domains, and a list of homograph pairs as input (3). The reference domains include well-known domain names often abused in malicious activities, such as *paypal.com*. For our analysis, we use the top 10,000 domains from the Majestic top 1 million list [4] downloaded on 2018-05-03 as the set of reference domains. The list of homograph pairs contains pairs of a "normal" Latin character or number and a potentially confusable, "special" Unicode character. We extracted these pairs from the *Recommended confusable mapping for IDN* [5], i.e., confusables for Latin letters, numbers, and the hyphen because these are the characters used in our reference domains. In total, we compiled a list of 836 confusable pairs, which we refer to as *homograph pairs*. Candidate Finder iterates over all combinations of newly registered IDNs and reference domains. For each pair, it first checks the lengths of the IDN and the reference domain and discards an IDN if the lengths are different, as our definition of a homograph domain only allows substitutions, but not additions or removals. Afterwards, Candidate Finder compares the IDN and the reference domain character by character. If a character pair *(a, b)* with *a* being the character in the reference domain and *b* being the character in the IDN is different, it checks whether (a, b) is in the list of homograph pairs. If (a, b) is not a known homograph pair, the IDN is excluded from further consideration since it is not a homograph domain according to our definition. If all character pairs are either identical or in the list of homograph pairs, Candidate Finder considers the IDN a possible homograph attack. We refer to these IDNs as *candidate domains* and store them in a database (4). Each candidate domain has at least one homograph pair because the input domains to Candidate Finder are IDNs, whereas all reference domains are non-IDNs.

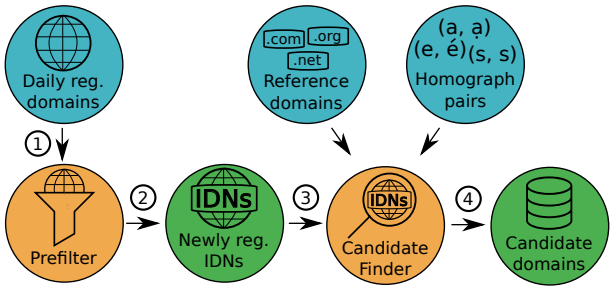


Fig. 1: Overview of candidate domain detection. *Prefilter* takes a list of domains as input ① and filters all newly registered IDNs ②. *Candidate Finder* takes the newly registered IDNs along with a list of reference domains and a list of homograph pairs as input ③, searches for candidate domains and stores them in a database ④.

B. Tracking the Activity of Homograph Domains

Our second analysis pipeline is intended to understand the purpose of the candidate domains. Figure 2 describes its workflow. A module called *Activity Checker* takes the previously found candidate domains as input and verifies if they are already known as malicious by VT [10] and GSB [9].

Furthermore, we collect WHOIS information for each candidate domain. Some fields of the WHOIS information, such as the registrant name, are often obfuscated, either due to privacy concerns or to cover the tracks in case of malicious activity. Yet, WHOIS data still provides valuable insights, such as the registration timestamp, or the name of the sponsoring domain registrar. Attackers may need multiple domains and register them in a short time period, potentially using the same company. WHOIS information can help uncover such patterns. Since the WHOIS data format differs among top-level domains and parsing WHOIS information is a tedious task, we use WhoisXmlApi [11], which we query to receive parsed WHOIS information for a domain.

Additionally, attackers may be identified because they use the same infrastructure across multiple domains. For example, they might use the same web server IP address multiple times. Therefore, we collect DNS information (resource records A, AAAA, TXT, CNAME, NS) as well as the autonomous system numbers.

Attackers may also use certificates to make malicious domains appear more trustworthy. For example, almost half of all phishing domains use certificates [20]. Thus, we collected a set of Certificate Transparency Logs (CTL) between 2018-02-02 and 2018-11-17 via the Certstream API [7] to understand whether any certificates were generated for the candidate domains.

The content hosted on a candidate domain can provide insights into the purpose of the domain. Therefore, we built a distributed crawler to collect screenshots and page source code of candidate domains.

C. Classifying Candidate Domains

To categorize how candidate domains are used, we assign to each domain one or more of the following labels:

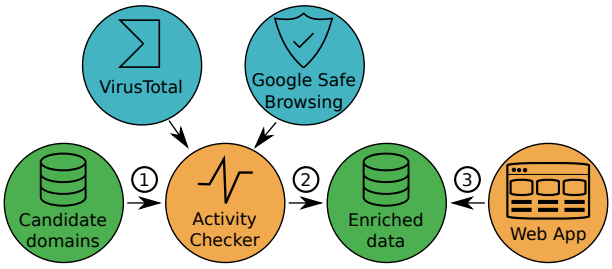


Fig. 2: Overview of daily activity tracking. *Activity Checker* takes all candidate domains as input ①, uses VT and GSB to verify their maliciousness and collects daily status information such as associated IP addresses and screenshots. The enriched data is stored in a database ②, which is accessible by a *Web App* ③.

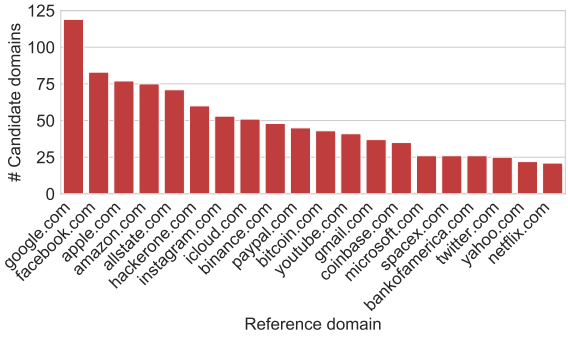


Fig. 3: Number of candidate domains for the 20 most targeted reference domains, which mostly belong to technology companies and financial institutions.

a) Scam: The candidate domain displays a website that tricks visitors into buying products or downloading malware.

b) Phishing: The candidate domain poses as a well-known brand and aims at convincing a victim to enter personal information, such as e-mail addresses or passwords.

c) Parked: The domain displays advertisements, often provided by a commercial domain parking vendor.

d) Referrer Fraud: The candidate domain forwards to another domain while adding an affiliate referrer in order to earn a commission.

e) False Positive: The candidate domain is benign, and the name is authentic. For example, the city of Krakow is spelled Kraków in Polish and has both domains registered.

f) Defensive: The reference domain's owner registered the domain to prevent a homograph attack.

g) Educational: The domain was registered to inform about homograph attacks, often with a note on the website.

h) Registered: The candidate domain is registered but we cannot make a statement about its purpose, e.g., because it is not reachable or displays a blank page.

We rely on a semi-automated approach to label our data. Some screenshots are straightforward to identify by comparing them with reference screenshots. For example, the websites generated by domain parking vendors differ only minimally. Therefore, in a first step, we compare each screenshot with a set of reference screenshots and, if there is a match, label the candidate domain accordingly for that particular day. Second, when we were unable to obtain sufficient metadata, such as when the website was unreachable, we label it as *registered*. Third, a human annotator labels each remaining candidate domain manually. Our system automatically carries over previous labels if the screenshot and IP address remain the same on subsequent days, thereby reducing the workload.

IV. Analysis

In this section, we first present an overview of our results. Afterwards, we provide details on the registration purpose of the candidate domains and analyze their used infrastructure. We conclude with a case study on scam domains.

A. Overview

We collected candidate domains from 2018-03-30 through 2018-11-17 and analyzed 2,864,449 unique IDNs provided by *Domainlists.io* – 2,237,638 IDNs had already been registered when our measurement began, and 626,811 new registrations were added during our measurement period. In total, we found 2,984 candidate domains for 819 unique reference domains. We tracked the activity of candidate domains daily from 2018-06-10 through 2018-11-17. Overall, we collected 440,818 days of activity information for the 2,984 candidate domains, including associated IP addresses and screenshots.

1) Reference domains: While we find that the candidate domains target a total of 819 distinct reference domains, the majority of reference domains has only very few candidate domains. Only 99 reference domains have more than five candidate domains. Therefore, we focus on the reference domains with most candidate domains. Figure 3 shows the number of candidate domains for the 20 most targeted reference domains. Together, the top 20 reference domains account for 984 candidate domains (33%). These top reference domains mostly belong to large technology companies such as *Google*, *Facebook* or *Apple*, and financial institutions like *PayPal* or *Bank of America*. Noteworthy is the high number of candidate domains for the insurance company *Allstate* and the bug bounty website *Hackerone*. We will show in Section IV-B3 that these candidate domains were most likely registered for defensive purposes. Compared to the 2006 results of Holgers et al., the number of homograph domains has increased significantly. For example, Holgers et al. reported four homograph domains for google.com whereas we identify 120. At a high level, our results are in line with those reported by Liu et al. based on 2017 data. The authors found 55 homograph domains for amazon.com (we identify 75 candidate domains) and 98 facebook.com homograph domains (83 in our data), for instance.

2) Replaced characters: Unlike typosquatted domains, homograph domains are not accessed by chance, and attackers need to advertise them via e-mail or instant messenger, for instance. In doing so, they need to ensure that the victims do not notice the difference between their fake

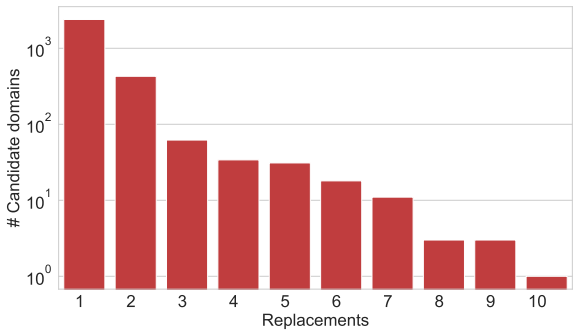


Fig. 4: Histogram of character replacement frequency in candidate domains. 80% contain only a single replaced character.

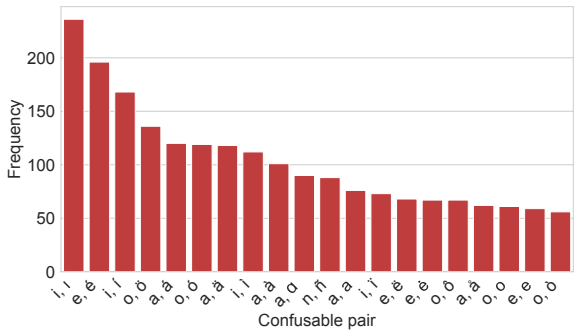


Fig. 5: Histogram of candidate domains using the 20 most frequent homograph character pairs. Character pairs with very similar visual representation are seeing surprisingly little adoption compared to pairs with a clearer visual difference.

homograph domain and the authentic reference domain. Figure 4 displays how many characters are replaced in the candidate domains. The vast majority of candidate domains (2,393, 80%) has only a single character replacement, with an observed maximum of ten. Presumably, a lower number of replaced characters has a lower likelihood of a candidate domain being detected as malicious by a potential victim. Candidate domains with more than three replaced characters consist completely of unicode characters. Holgers et al. only considered one, two or three replacements in their 2016 study. Our data show that candidate domains with a higher number of replacements exist, even though to a minor degree.

In total, our list of homograph pairs contains 836 pairs. Only 189 of them were observed in registered candidate domains. These 189 homograph pairs account for 3,972 replacements (1.3 replacements per candidate domain). However, the 20 most frequently used homograph pairs already account for 2,073 replacements (53%). Therefore, Figure 5 shows how many candidate domains use each of these top 20 homograph pairs. Surprisingly, homograph pairs that are difficult to tell apart visually, such as “a” and “а” (small cyrillic a), saw much less use than pairs with clearer visual distinction, such as “a” and “á.” Thus, a difference between the original character and the replacement is visible in most cases. Additionally, none of the replacement characters in the top 20 homograph pairs have a lower component, such as a dot below the actual character. We conducted a study on how mail programs display IDNs and determined that a homgoraph domain using a character with a lower part interrupts the underline below a link, which eases identification of homograph domains and may explain their infrequent use.

3) Certificate Transparency Logs: We searched for all 2,984 candidate domains in our CTL dataset to understand the adoption of certificates for homograph domains. In total, we discovered certificates for 498 candidate domains (17%). At first glance, this seems to be a low number but taking into consideration that our CTL dataset covers only nine months in 2018, and 1,841 candidate domains were registered before 2018, it is likely that certificates have also been generated for many candidate domains before. There is a surprisingly long time of more than 50 days between domain registration and certificate generation for 328 candidate domains. However, it is possible that these domains already had certificates generated earlier, which were renewed during our measurement. In the case of 124 candidate domains, there are less than ten days between domain registration and certificate generation. A short time between domain registration and certificate generation increases the likelihood of a candidate domain being used.

B. Registration Purpose

Prior work by Holgers et al. [18] focused on detection of homograph domains, and Liu et al. [22] presented only a brief summary of how registered homograph domains were being used. Therefore, we focus our analysis on elucidating possible motivating factors behind the registration of homograph domains. To that end, we monitored the activity of candidate domains from 2018-06-10 through 2018-11-17 with the Activity Checker described in Section III-B. In total, we collected 440,818 days of activity and labeled them according to Section III-C. A single candidate domain can exhibit different behavior on different days. Thus, it may be annotated with multiple labels. E.g., a candidate domain may be parked on one day and host a phishing website on another day. Figure 6 shows a histogram with each activity label and the number of candidate domains exhibiting at least one day of the respective activity.

1) Registered: The most prevalent label is *registered* because we labeled a domain as *registered* if it was not reachable or did not contain actual content and displayed, for example, a blank page. Almost every candidate domain had at least one such day (2,895 candidate domains). Especially candidate domains older than two years were not reachable.

2) Parked: About one-third of the candidate domains were *parked* for at least one day during our analysis period (1,220 candidate domains). It is a common strategy by both malicious actors and legitimate users and partially even registrars to park and monetize a domain if it is not used for other purposes. In contrast to Agten et al. [13], we do not consider parked domains as malicious because we do not find parking a domain after registration as malicious.

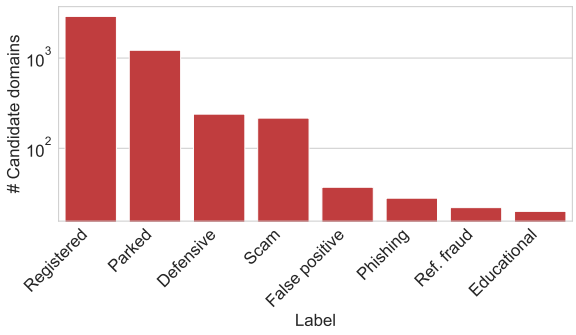
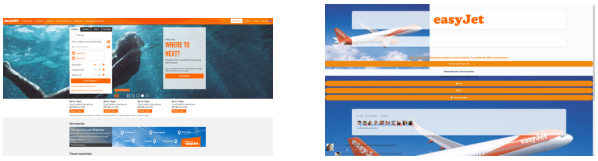


Fig. 6: Number of candidate domains per domain use label. Labels were assigned daily, and one candidate domain can have multiple different labels assigned when it changed behavior. Nearly all domains were unreachable (*registered*) on at least one day, and about one third were parked for at least one day. Malicious activity exceeds defensive registrations.

3) Defensive: We labeled 239 candidate domains (8%) as *defensive*. Typically, an e-mail address in the WHOIS information or the name servers of the candidate domain belonged to the company of the reference domain. The defensive registrations were made for 23 distinct reference domains, with the insurance company *allstate.com* (70 candidate domains) and the bug bounty website *hackerone.com* (59) being the most prolific. Overall, the number of defensive registrations is low, especially when taking into account that only 23 out of 10,000 reference domains have any defensive registration. Furthermore, even when defensive registrations exist for a reference domain, only a small number of all possible variants are actually registered, given that very high numbers of such variants exist. Considering only the first two letters of *allstate.com*, for instance, there are over 100 possibilities to replace one of the two letters (43 for *a*, 65 for *l*), and 2,795 possibilities when replacing both letters. Also replacing the remaining letters of the domain further increases the number of variants, making it impractical to achieve comprehensive coverage. We did not observe any pattern in defensive registrations (such as always registering homograph domains with a dot below the replaced letter).

4) Scam: In total, we labeled 216 candidate domains (7%) as *scam*. They target 113 reference domains, i.e., the number of candidate domains per reference domain is considerably lower than for defensive registrations. We divide the candidate domains into two groups: first, candidate domains that display content related to their reference domain, e.g., candidate domains using *adidas.com* as reference domain and promising free sports shoes when a user completes a survey. Second, candidate domains that display content unrelated to their reference domains, e.g., dubious advice on earning "thousands of dollars a week" with bitcoin trading on a *facebook.com* homograph domain.

Candidate domains in the first group tend to be registered solely for one purpose and display only one page. Surprisingly, a minority of 45 candidate domains belonged to this group. Most of these candidate domains were registered during our analysis period. Their activity ranged from a few days up to multiple months. Even though a user might access such a website because the candidate domain looks legitimate, the design of the websites can differ (see Figure 7). One reason for the different design might be to evade detection mechanisms that compare screenshots. Another possible reason is that users access even a non-legitimate looking website, so that more effort is not necessary.



(a) (b)

Fig. 7: The authentic EasyJet website (left), and a scam website emulating its visual language (right).

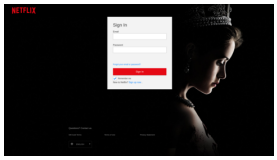
The majority of 171 candidate domains belong to the second group. These usually display different content and often switch almost every day, e.g., they show a raffle on one day, pornography the next day and try to infect a victim with malware on another day. These domains were registered before we started our analysis, thus we hypothesize that attackers may initially display content related to the reference domain, and later attempt to monetize the domain with less relevant, but rotating content.

During our activity measurements, only three candidate domains were detected by GSB, and those were domains registered *before* the beginning of our measurement period. We discovered more than 200 scam domains that were never listed by GSB, suggesting that our approach can uncover domains sooner than the current industry standard.

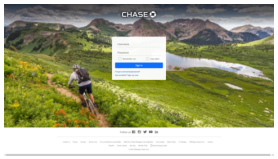
5) False positive: We labeled 37 candidate domains as *false positive*. False positives mostly occurred when a person or institution had a name with a special character and registered both the version with the corresponding Latin character and the IDN. In some cases, it was difficult to decide whether a candidate domain was a false positive or a defensive registration. For example, when a company initially spells its name with a Unicode character but usually uses the version without Unicode character, it is not clear if the homograph domain was a defensive registration or a false positive. When we had only one candidate domain for a reference domain, and we knew that the original spelling was with a Unicode character, we labeled the candidate domain as *false positive*. Furthermore, we labeled a candidate domain as *defensive* when we had reasons to believe that it was part of a defensive registration, e.g., because multiple candidate domains were registered for a reference domain.

6) Phishing: We labeled 28 candidate domains as *phishing*. Figure 8 shows examples for Netflix and PayPal phishing websites that we encountered during our analysis period. In general, the phishing websites were very close to the content of the reference domains.

The 28 phishing candidate domains belong to 17 dis-



(a)



(b)

Fig. 8: Phishing sites hosted on candidate domains for Netflix and Chase.

TABLE I: Summary of properties used for clustering.

Feature	# Elements
Reference domain	819
Registration time	1,723
Registrar	158
Contact e-mail	1,271
Nameserver	1,636
IP address	3,793
Autonomous system numbers (ASNs)	259

tinct reference domains, with *instagram.com*, *netflix.com*, *chase.com* and *paypal.com* being the reference domains with more than one candidate domain. The targeted reference domains belong to big technology companies and financial institutions, which is in line with common phishing targets.

Out of the 28 phishing domains, 17 were active for more than ten days and 12 of them for more than 30 days, indicating that they were able to run undetected for a long time. In contrast, Hao et al. reported that 60% of spam domains they found were active for only one day [17]. While we know that the phishing websites were available for a long time, we do not know when attackers started advertising them to victims. The remaining 11 candidate domains were active for less than ten days and most likely detected rather soon. These candidate domains included three domains targeting the American bank *Chase*, suggesting that the bank may have a good detection.

One phishing domain was already blacklisted by GSB before the beginning of our measurement, one domain was blacklisted on the same day as we detected it, and six domains were detected by Google one to ten days after we labeled them as phishing (these domains often ceased operation once detected). The remaining twenty phishing domains were not detected by Google at all. Overall, these results indicate that our approach can detect malicious homograph domains faster and in greater quantities than GSB.

7) Referrer fraud: Overall, we labeled 22 candidate domains as *referrer fraud*, targeting 18 distinct reference domains. Three candidate domains each targeted *Snapchat* and the cryptocurrency exchange *Binance*. Additionally, we found multiple cryptocurrency related reference domains targeted, such as *coinmarketcap.com*, *localbitcoins.com* or *coindesk.com*. Usually, attackers forward requests made to the candidate domain to another, potentially unrelated domain and earn a commission, e.g. if the user trades bitcoins.

8) Educational: Twenty candidate domains displayed an explanation of homograph domains to educate their visitors. To the best of our knowledge, prior studies on homograph domains did not report any such domain.

C. Infrastructure Analysis

1) Motivation and properties: In the following, we analyze the infrastructure of the candidate domains in more detail. In addition to registering domains, attackers need to rent servers and configure supporting infrastructure such as DNS. Often, it is possible to use patterns in the registration process, such as registration of multiple similar domains in a short time period, to link together candidate domains. Furthermore, attackers reuse at least parts of their infrastructure because it is tedious and costly to start from scratch for every campaign. We use these patterns and cluster the candidate domains into groups being for example registered in the same hour, or using the same IP addresses. Furthermore, we show that domains labeled as malicious can be connected to other candidate domains that we did not label as malicious initially. For that purpose, we create a graph with each of the 2,984 candidate domains being a node and two nodes having an edge if the two candidate domains:

- target the same *reference domain*,
- were registered in the *same hour* of the day,
- were registered using the *same registrar*,
- share the *same contact e-mail address* in their WHOIS information,
- share a *common name server* at least once a day,
- share a *common IP address* at least once a day, or
- share a *common Autonomous system number (ASN)*.

We add at most one edge between two candidate domain nodes and call the number of shared properties the edge weight. Additionally, we add a tag to each node if we labeled the corresponding candidate domain as *scam*, *phishing*, *referrer fraud*, *defensive* or *educational* for at least one day during our measurement time period. We omit the labels *registered* and *parked* because they are less relevant when analyzing attackers' infrastructure.

When we first created the graph, we observed many candidate domains sharing the same IP addresses. A closer examination revealed that these IP addresses belonged to parking services. We decided to remove these IP addresses from consideration for clustering purposes because of the relatively high likelihood that independent parties could park their domains using the same service. In total, we removed 250 IP addresses. Furthermore, we removed the 15 corresponding autonomous system numbers because they are closely related to the IP addresses.

Table I summarizes the properties along with the number of elements we collected. We have 819 distinct reference domains and 1,723 distinct hours during which at least one candidate domain was registered.

2) Results: When looking for clusters with the edge weight of seven (i.e., all properties were the same), we found 20 clusters with 140 nodes. The majority of clusters are small (eight clusters with two nodes each and eight clusters with three nodes each). Eight clusters contained nodes labeled

as *defensive* (four clusters with three nodes each and the clusters with six, ten, 14 and 70 nodes) and one cluster contained three nodes labeled as *referer fraud*. These clusters did not include any unlabeled nodes. The high number of defensive registrations is expected because a company does not need to hide defensive registrations and can use the same infrastructure for all of them.

Requiring clustered nodes to share all seven properties limits the utility of the clustering, and appears to be overly restrictive. For example, two nodes cannot be clustered together when they are registered on different days, or target different reference domains. Hence, we repeat our previous experiment in a less strict configuration by searching for components with edge weights of at least five, i.e., they can have different values in up to two properties.

In total, we found 235 clusters with 1,006 nodes, i.e., more than a third of the candidate domains are clustered. Again, the majority of clusters contain a small number of nodes (134 clusters with two nodes each and 36 clusters with three nodes each). Eighteen clusters contain nodes labeled as *defensive* with three clusters having one candidate domain not yet labeled as *defensive*. Furthermore, we identified 31 clusters with nodes labeled as *scam*. In 13 clusters, all nodes were labeled as *scam*, while in 18 clusters the nodes were only partially labeled as *scam*. At the time of writing, the 55 unlabeled nodes in these clusters have not yet been observed in malicious activity, but their similarity to the *scam* nodes suggests that they are worth monitoring in the future.

We found four clusters with nodes labeled as *phishing*, out of which two clusters were entirely labeled as *phishing*. The other two clusters target *Snapchat* and *Apple*, respectively, and contain two unlabeled nodes each. In both cases, the unlabeled candidate domains fit the cluster (two more Snapchat domains and two iCloud domains in case of Apple), and we suspect that we may have identified four phishing domains before they were put into use. Another four clusters contained only nodes labeled as *referer fraud*. One cluster contained five nodes labeled as *educational*, and another three clusters contained two such nodes. This suggests that security researchers register multiple example domains in their effort to educate the public.

D. Case Study

In the following, we illustrate the role played by homograph domains in attacks by revisiting three categories of candidate domains labelled as *scam* in Section IV-B4. These domains impersonate sports apparel and airline companies, among others. A selection of domains is shown in Table II. In all three cases, attackers use a domain of a well-known brand and replace characters with similar looking Unicode characters. The design of the corresponding websites recalls that of the authentic domains, e.g., they use a similar color scheme, yet there may still be a clear difference to the original website, as shown in Figure 7 for EasyJet. The websites display a set of questions that victims must answer in exchange for free flights, shoes or vouchers (that are never delivered). Some of these scams have already been mentioned in blog posts [12], [27], but our analysis pipeline also uncovered new, previously unknown scam domains such as the ones listed in Table II. Some of these domains were registered before 2018-03-30, the start of our measurement period, which illustrates that scam domains may be surprisingly long-lived. Typically, domains begin their malicious activity within a few days of being registered. Our perhaps most concerning finding is that detection of these scam domains by current industry tools is very poor. VT reported only one of these scam domains with a delay of more than 1.5 months. GSB flagged none of them.

TABLE II: Selection of scam domains identified by our pipeline, with first day and duration of malicious activity, showing scarce detection by VT and GSB. The websites promise free shoes, flights, etc.

Domain	Added	First Active	Days	VT	GSB
adidą̇s.com	2018-10-20	2018-10-20*	4	✗	✗
adidaṡ.com	2018-08-09	2018-08-14	54	✗	✗
nikė.com	2018-03-30	2018-06-10	157	✗	✗
ą̇irą̇sia.com	2018-10-16	2018-10-16*	13	✗	✗
airasıa.com	2018-09-07	2018-09-08	67	✗	✗
easyjȇt.com	2018-07-29	2018-08-14	1	✗	✗
deltå.com	2018-06-21	2018-06-21*	35	2018-08-07	✗
chıck-fil-a.com	2018-08-11	2018-08-14	92	✗	✗
pepsı.com	2018-03-30	2018-09-06	17	✗	✗
tęsco̧.com	2018-10-01	2018-10-01*	15	✗	✗
tesčo.com	2018-03-30	2018-06-10	3	✗	✗

* the domain was possibly already previously active

V. Discussion

During our analysis, we encountered only a low number of defensive registrations. This suggests that some domain owners consider malicious homograph domain registrations a security risk, whereas other domain owners do not, or they may be unaware of this type of attack. Due to the huge number of visually similar character substitutions, it is typically not feasible to defensively register all possible homograph domains. Instead, we propose a two-prong defensive strategy. Domain owners could identify the most commonly used character substitutions (e.g., from Figure 5) and preemptively register those homograph domains. Furthermore, domain owners should continually monitor whether any of the remaining homograph domains are registered by miscreants, using, for example, the pipeline that we have proposed in Section III-A. Domain owners could then use existing conflict resolution mechanisms to request ownership transfer of homograph domains in case of trademark infringement, and ensure that malicious activity is reported to blacklists.

VI. Limitations

Our work has some limitations due to its approach that we discuss in the following. First, the system can detect homograph domains only for a given list of reference domains. In our study, we used the top 10,000 domains from the Majestic Top 1 million list to represent the largest websites, but our system could also work with different lists of reference domains, such as all domains owned by a company that wishes to detect homograph attacks against their brand.

Second, our approach can detect a homograph domain only if the Latin character in the reference domain and its counterpart in the homograph domain is a known homograph pair. To address this issue, we based our work on a well-known list of typical confusable characters [5]. Furthermore, we study IDNs under the angle of homograph domains, that is, as domains that need to be advertised by attackers and that look (nearly) indistinguishable from the authentic domain. In some regions, these domains could also be considered typosquatting domains because the characters they use can be found on keyboards in the local language.

Third, the value of the collected screenshots depends on whether we accessed the candidate domain at the right time, i.e., when malicious content was hosted. It is possible that cloaking or crawler detection techniques prevented us from seeing a candidate domain in the same way a victim would.

VII. Conclusion

In this paper, we presented an analysis approach to systematically detect homograph domains based on a list of homograph pairs, and a second analysis approach to examine their activity in detail. For a period of eight months, we used the first analysis setup to perform a long-term measurement study to detect homograph domains and found 2,984 such domains targeting 819 reference domains. While the number of candidate domains is low compared to other domain squatting techniques, homograph domains are difficult to tell apart from their original counterpart making them a dangerous threat. We used the second analysis setup to monitor the activity of candidate domains and found multiple instances of scams and phishing. Financial institutions like PayPal or Chase, and technology brands like Instagram or Netflix were often targeted. Our system was able to identify malicious homograph domains sooner and to a far greater extent than current industry standards.

Acknowledgments

The research leading to these results has received funding from the European Union's Marie Sklodowska-Curie grant agreement 690972 (PROTASIS). The paper reflects only the authors' view and the Agency and the Commission are not responsible for any use that may be made of the information it contains.

This was partially-supported by the National Science Foundation under grant CNS-1703454.

Additionally, this work was supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0112. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government.

References

[1] Hackers, not users, lose money in attempted cryptocurrency exchange heist. https://www.bleepingcomputer.com/news/security/hackers-not-users-lose-money-in-attempted-cryptocurrency-exchange-heist/. Accessed: 2019/01/03.

[2] How a Bitcoin phishing gang made 50 million with the help of Google AdWords. https://www.tripwire.com/state-of-security/featured/bitcoin-phishing-million-google-adwords/.

[3] Look-alike domains and visual confusion. https://krebsonsecurity.com/2018/03/look-alike-domains-and-visual-confusion/.

[4] Majestic million. https://majestic.com/reports/majestic-million. Accessed: 2018/05/03.

[5] Recommended confusable mapping for IDN. http://www.unicode.org/Public/security/8.0.0/confusables.txt. Accessed: 2019/01/03.

[6] Who is behind those fake Whatsapp campaigns supposedly giving for free Nike shoes and alike? http://blog.emiliocasbas.net/2018/02/who-is-behind-those-fake-whatsapp.html. Accessed: 2019/01/03.

[7] Certstream. https://certstream.calidog.io/, 2017. Accessed: 2019/01/03.

[8] Domainlists.io. https://domainlists.io, 2018. Accessed: 2019/01/03.

[9] Google Safe Browsing. https://safebrowsing.google.com/, 2018. Accessed: 2019/01/03.

[10] VirusTotal. https://www.virustotal.com, 2018. Accessed: 2019/01/03.

[11] WhoisXmlApi.com. http://whoisxmlapi.com, 2018. Accessed: 2019/01/03.

[12] R. Abel. Adidas phishing campaign promises free shoes, offers $50 subscription instead. https://www.scmagazine.com/home/security-news/cybercrime/adidas-phishing-campaign-promises-free-shoes-offers-50-subscription-instead/. Accessed: 2019/01/03.

[13] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *NDSS*, 2015.

[14] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting malware domains at the upper DNS hierarchy. In *USENIX Security*, 2011.

[15] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. Exposure: Finding malicious domains using passive DNS analysis. In *NDSS*, 2011.

[16] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster. Predator: Proactive recognition and elimination of domain abuse at time-of-registration. In *CCS*, 2016.

[17] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck. Understanding the domain registration behavior of spammers. In *IMC*, 2013.

[18] T. Holgers, D. E. Watson, and S. D. Gribble. Cutting through the confusion: A measurement study of homograph attacks. In *USENIX ATC*, 2006.

[19] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. In *CCS*, 2017.

[20] B. Krebs. Half of all phishing sites now have the padlock. https://krebsonsecurity.com/2018/11/half-of-all-phishing-sites-now-have-the-padlock/, 2018. Accessed: 2019/01/03.

[21] T. Lauinger, A. Chaabane, A. S. Buyukkayhan, K. Onarlioglu, and W. Robertson. Game of registrars: An empirical analysis of post-expiration domain name takeovers. In *USENIX Security*, 2017.

[22] B. Liu, C. Lu, Z. Li, Y. Liu, H. Duan, S. Hao, and Z. Zhang. A reexamination of internationalized domain names: The good, the bad and the ugly. In *DSN*, 2018.

[23] D. Liu, Z. Li, K. Du, H. Wang, B. Liu, and H. Duan. Don't let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains. In *CCS*, 2017.

[24] N. Miramirkhani, T. Barron, M. Ferdman, and N. Nikiforakis. Panning for gold.com: Understanding the dynamics of domain dropcatching. In *WWW*, 2018.

[25] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen. Soundsquatting: Uncovering the use of homophones in domain squatting. In *International Conference on Information Security*, 2014.

[26] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen. Bitsquatting: Exploiting bit-flips for fun, or profit? In *WWW*, 2013.

[27] M. Schiffman. Free airline tickets: The latest internationalized domain name-based homograph scam. https://www.farsightsecurity.com/2018/08/13/mschiffm-freeticketsscam/. Accessed: 2019/01/03.

[28] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang. Needle in a haystack: Tracking down elite phishing domains in the wild. In *IMC*, 2018.