Exploring the Characteristics and Security Risks of Emerging Emoji Domain Names

Mingxuan Liu[1,2](✉), Yiming Zhang[1](✉), Baojun Liu[1], and Haixin Duan[1,3](✉)

[1] Tsinghua University, Beijing, China
liumx18@mails.tsinghua.edu.cn, zhangyim17@tsinghua.org.cn
{lbj,duanhx}@tsinghua.edu.cn
[2] BNRist, Beijing, China
[3] Peng Cheng Lab, Shenzhen, China

Abstract. Emoji domains, such as `i❤️.ws` (`xn--i-7iq.ws`), are distinctive and attractive to registrants due to their eye-catching visuals. Despite its long history (over 20 years), little has been done to understand its development status and security issues. In this paper, we identify 54,403 emoji domains from 1,366 TLD zone files and a large-scale passive DNS dataset. And then, we correlate them with auxiliary data sources like domain WHOIS records. It allowed us to conduct by far the most systematic study to characterize the ecosystem, and retrieve multiple valuable insights. On one hand, the scale of emoji domains is constantly expanding in the wild, with dozens of ccTLD registries actively promoting registering domains with emoji characters and domain owners configuring emoji characters in sub-level domains. And emoji domains may act as promotional portals, as web requests are usually redirected to other websites. Besides, emoji domains are also leveraged to provide disposable email services, pornography or gambling pages, and even the distribution of malware. On the other hand, the concern is that the community still lacks best security practices in supporting and parsing emoji domains. Through empirical studies, we demonstrate that inconsistencies in rendering emoji characters can be exploited to launch visual phishing domain scams. Meanwhile, mainstream implementations may incorrectly parse or trans-code emoji domains, resulting in the security threat of traffic hijacking. Our study calls for standardization and best security practices for applications to handle emoji domains securely.

1 Introduction

Domain names are user-friendly alphanumeric names that make it easier for Internet users to navigate the online world. Conventionally, only a portion of ASCII characters (letters, digits, and hyphens) was allowed in domain names [42]. With the purpose to globalize the use of the Internet and make domain names more accessible, the IETF promotes the Internationalized Domain Name (IDN)

V. Atluri et al. (Eds.): ESORICS 2022, LNCS 13556, pp. 186–206, 2022.
https://doi.org/10.1007/978-3-031-17143-7_10

program, which allows non-native English speakers to adopt their native language or local script, i.e. Unicode characters, in domain names.

Emoji belongs to a special subset of Unicode characters. Today, it has been widely adopted on smartphones and social media, and plays a critical role in Internet communication. It also attracts the interests of domain registrants. With the advantages of being graspable and eye-catching, an emoji domain can be an effective tool for public marketing. Actually, many big companies have already been doing so. For example, Coca-Cola registered a whole bunch of domains containing smiley emojis like 😀.ws (xn--228h.ws) [51] (expired) for advertisement in 2015. Similarly, Budweiser registered 🍺🍺🍺.ws (xn--xj8haa.ws), and Mailchimp [16] registered 💌.ws (xn--rr8h.ws) for promotions. Besides, emoji domain also has been exploited for scam activities. In 2020, a collective defrauded over $200,000 through 👁👄👁.fm (xn--mp8hai.fm) in the guise of social justice [47]. And Weapon Depot utilizes the emoji domain, 🔫.ws (xn--bw8h.ws), to attract customers on some social media [12].

Despite the initiative of emoji domain having been proposed for about 20 years, little has been done to understand its ecosystem in the wild. In this paper, we report by far the first systematic study on emoji domains by answering a set of critical questions for understanding its development status and security risk, including: *What are the current scale and usage status? What are the characteristics of registrations? Are there any (new) security issues?* We made this study possible by a broad data collection, including 1,366 TLD zone files, a country-level passive DNS dataset and domain WHOIS records. Finally, *54,403* emoji domains are identified in total.

By analyzing the identified emoji domains, we discovered that discouragement from ICANN [27] has not hindered the development of emoji domains. In fact, the volume of emoji domains is constantly growing in the wild, increasing hundreds of folds compared to seven years ago. Although the registration of emoji domains under gTLDs has been restricted, registrants have turned to the registrars from ccTLDs (e.g., 😄.cctld (xn--i28h.cctld)), or embedding emoji characters (e.g., 👍.example.com (xn--yp8h.example.com)) in sub-level domains, which is prominent developing until now. Several ccTLDs registries even take emoji domain registration as a selling point for commercial promotion. As for registration intention, we find that high-profile emoji domains are created for promotion proposes, e.g., i❤️.ws (xn--i-7iq.ws) that received 7.96 million DNS requests is used for advertising emoji domain registration services, and ✉️.mail-temp.com (xn--4bi.email-temp.com) is designed for disposable temporary email service. However, pornographic sites and even malware distribution sites have also been witnessed leveraging emoji domains for user attraction.

Besides, we also reveal that the applications of emoji domains expose several security risks, especially in the trans-coding and rendering process. Through empirical study, three kinds of new security threats are uncovered. First, due to the inconsistent visuals of emoji rendering, we find that visual phishing attacks targeting emoji domains are feasible in the real world. Except for a few registrants who have noticed this risk and proactively registered visually similar domains for defense, the vast majority of phishing-vulnerable emoji domains are not yet protected. Second, mainstream implementations could not correctly parse emoji domains, resulting in text with emoji icons being unintentionally recognized as emoji domains. By inspecting one-day DNS queries from `B Root`, we uncover 6,372 "unintended" emoji domains as "parsing errors". Almost half of these domains are available for registration, leaving huge space for attackers to conduct traffic hijacking. Third, there is still a lack of best practices for handling special Unicode characters. Particularly, we find several mainstream browsers (e.g., Firefox and Safari) fail to trans-code `ZWJ` (Zero with Joiner) embedded emoji domains correctly, leading to the denial of service and hijacking threats.

In summary, our study shows that the development of emoji domain names is still at the early stage with a growing trend. And we recommend that security community should pay more attention to the ecosystem and propose best practice guidelines for harmonizing the usage and process of emoji domains.

2 Background

Domain Name Structure and Registration. A domain name is comprised of multiple layers and organized as a hierarchical structure. The boundary between hierarchy levels is separated with a dot, such as `esorics2022.compute.dtu.dk`. The top of the domain hierarchy is the DNS root. Below the root level are the Top-Level Domain (TLD, e.g., `dk`) and Second-Level Domain (SLD, e.g., `dtu.dk`).

TLDs are typically divided into three categories, including generic TLDs (gTLDs), country-code TLDs (ccTLDs), and sponsored TLDs (sTLDs). All TLDs are approved by Internet Corporation for Assigned Names and Numbers (ICANN), and operated by various registries. Of note, all registries operating gTLDs are *contracted* with ICANN [26], while ccTLDs are not necessarily required. For a registrant, domain names that are allowed to apply are SLDs (or apex domains). They are publicly offered and a domain name is registrable if it is not yet occupied. Domain owners are allowed to create *subdomains* under their apex domains, without asking permission from registrars.

Emoji Domain Names. An emoji domain refers to a domain name that contains at least one emoji character, regardless of the level at which the emoji is embedded. In the beginning stage, domain names were only allowed to be registered within letters, digits, and hyphens [42]. Most of the domain names came from a set of alphanumeric ASCII characters. To build a multilingual Internet, IETF instituted the Internationalized Domain Name (IDN) program in 2003. IDN program encourages Internet users around the world to adopt a domain that contains native scripts [7,14]. As a result, the scope of allowed characters in domain names has been extensively extended to *Unicode sets.*

At the time of writing (May 2022), 3,633 emoji code points are contained in the standard Unicode 14.0 [60]. Theoretically, registrants are permitted to apply

Table 1. Overview of datasets.

Data Source	# ED_sld	# ED_sub	# Emoji SLD	Unicode Domain
gTLD zone files	193	–	193	1,499,958
ccTLD zone files	1,732	–	1,732	3,246,266
Passive DNS	25,731	28,252	13,170	52,976,933
ALL	**26,151**	**28,252**	**13,581**	**55,887,203**

for domain names with embedded emoji characters, or add emoji characters to subdomains under the apex domain themselves.

Punycode Conversion. Although emoji domains are supported by DNS, they have to be converted to ASCII characters in order to maintain backward compatibility. IETF established technical standards to support domain names encoded with Unicode characters [13,32], named Internationalizing Domain Names in Applications (IDNA). IDNA is designed to convert a Unicode string (U-Label) into an ASCII-compatible encoding (ACE) string (A-Label), i.e., Punycode [7,13]. Punycode keeps all ASCII characters, and encodes the locations of non-ASCII characters, and re-encodes the non-ASCII characters with variable-length integers. As the algorithm design, a fixed prefix, "`xn--`", is added to the converted Punycode string after the above process. For example, xn–i-7iq.ws is the Punycode conversion of `i❤️.ws` .

Security Considerations. Due to the effect of attention-grabbing, emoji domain names have attracted a lot of attention from registrants worldwide, especially for marketing and advertising campaigns. However, DNS community has proposed several security concerns with the emoji domain applications, due to their potential impact on the stability and interoperability of the domain name system. Specifically, an advisory document has been proposed by ICANN, indicating that emoji domains may cause ambiguity and confusion [27].

Nonetheless, we believe it is still too early to claim the failure of the emoji domain initiative. Instead, we need to revisit the development of emoji domains, evaluate the real-world impact, and explore the practical security risks.

3 Data Sources of Emoji Domains

In this section, we first elaborate on how we collect large-scale datasets. Then, we describe technical details of how to identify emoji domains.

3.1 Collecting Large-Scale Datasets

We collect 1,366 TLD zone files and a country-level passive DNS dataset to exhaustively detect emoji domains in the wild. The details are presented in Table 1.

TLD Zone Files (gTLDs and ccTLDs). TLD zone files are maintained by registries, like Verisign. They contain active domains with their delegation information, and serve as an important data source in security research. ICANN provides a centralized zone data service (CZDS) [34] for interested parties to access zone files. It allows us to apply *1,254 gTLD zone files* in September 2021, which contain the up-to-date registered domains maintained by registries, including historical gTLDs (e.g., `.com`) and a range of new gTLDs (e.g., `.info`).

By contrast, ccTLDs do not (or no longer) provide publicly accessible zone files [25]. Several well-known public datasets utilized in previous studies, e.g., OpenIntel [52] and CAIDA-DZDB [4], also have quite limited coverage of ccTLD domains. To solve this issue, ViewDNS [63] continuously collects domains under ccTLDs by Internet crawlers with considerable domain coverage [8]. We purchased all ccTLD domain lists of ViewDNS in May 2021, and got *112 ccTLD zone files* in total, e.g., `.us`, and `.cn`, covering 35.44% of all (316) ccTLDs [25].

Passive DNS Dataset. In addition to registering domains with emoji icons directly, one can also place the icons on subdomain labels. However, TLD zone files have no information on subdomains configured by registrants. To this end, we leverage the Passive DNS dataset from DNS Pai Project [49] to extend our investigation scope to fully qualified domain names (FQDNs). The project was initiated by a world-leading security vendor, and has collected DNS requests from a large array of popular DNS resolvers since 2014. It handles around 240 billion DNS requests per day, and opens the collected DNS traffic data to the research community. In this study, we gain access to all records of historical domain names from Passive DNS spanning from 08/01/2015 to 7/27/2021.

Domain WHOIS Records. We also utilized WHOIS records to understand the registration trend of emoji domains. Specifically, the WHOIS dataset was collected with the help of our industry partner, 360 Netlab [50]. As several ccTLDs (e.g., `.to`) restrict crawlers from obtaining their WHOIS information, we finally got the WHOIS records of *8,638 (63.60%)* unique emoji SLDs as the best coverage effort. Then we used an open-source tool, `python-whois`[1], to parse the records. As this work only concerns the registrar/registry and the creation/expiration date of domains, our analysis would not be affected by the implementation of General Data Protection Regulation (GDPR) policies [40].

3.2 Identifying Emoji Domains

Definitions and Notations. In this study, we refer to any FQDNs embedded with at least one emoji character as an emoji domain, termed as ED. Depending on where the emoji characters are located in the domain structure, EDs could be further classified into two categories: ED_{sld}, whose SLD contains emoji characters directly, and ED_{sub}, whose emoji characters *only* appear in the subdomain labels. The two categories essentially denote the different sources of ED creation. ED_{sld} indicates the domain owner directly registering an apex domain

[1] https://pypi.org/project/python-whois/.

with emoji characters from registrars. ED_{sub} means that the domain owner only configured emoji characters into the subdomain in authoritative nameservers (Fig. 1).

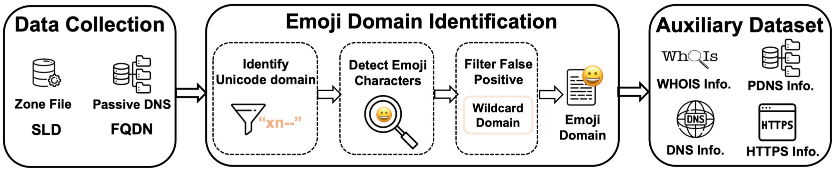


Fig. 1. The workflow of data collection and emoji domain identification.

Data-processing Workflow. Our emoji domain extraction workflow includes three steps.

(1) *Identify Unicode domains.* Emoji domain is only a subset of the Unicode domain. Given the rules of Punycode conversion, we are allowed to identify all Unicode domains by matching the fixed prefix "`xn--`", as described in Sect. 2.

(2) *Detect emoji characters.* Further, we convert the ASCII-compatible encoding string into Unicode string format, which is represented as a list of Unicode code points (e.g., U+1F600 for 😀). We also crawled Unicode code points for all emoji characters from the Unicode consortium [60]. Then the domains with at least one Unicode point inside the emoji range would be identified as emoji domains.

(3) *Filter false positives.* Through manual analysis, we find several emoji domains extracted from PDNS dataset are "false positives": *non-existent subdomains* caught by PDNS as their SLDs were enabled for wildcard resolution. To filter them, we replace the emoji characters with random strings and examine whether the newly generated domains could get the same resolution results.

Finally, we identified 54,403 unique emoji domains (*26,151* ED_{sld} and *28,252* ED_{sub}) from 55.89 million Unicode domains (Table 1). Among them, 4,947 emoji domains with SLDs are ranked within the Tranco Top 50k popular domain list. The list of the top 10,000 most queried emoji domains has been open-sourced[2].

Discussion. Although we try to make this study as comprehensive as possible, there are still limitations. First, our PDNS dataset may have geographical bias. However, given its huge DNS traffic volume and the longitudinal data collection period, we believe the dataset is still representative to reveal the ecosystem of emoji domains in the wild. Second, although we take the best effort to extend the observations on ccTLD domains by collecting zone files from ViewDNS [63], the coverage (112 out of 316 [25] ccTLDs) is still limited. The limitations indicate that our study may only reflect a lower bound in the real world.

[2] https://github.com/EmojiDomain/ESORICS22.

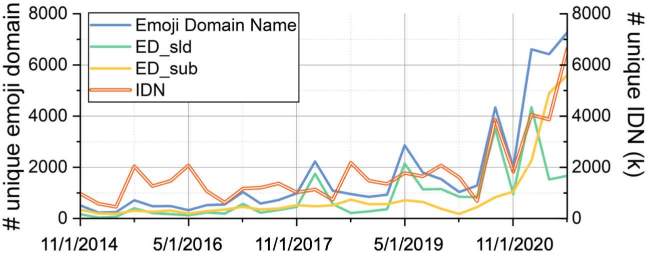


Fig. 2. Newly witnessed emoji domains and IDNs from passive DNS traffic.

4 Characteristics of Emoji Domain Ecosystem

In 2017, ICANN recommended discouraging registration activities for emoji domains. However, the impact of the recommendation has yet to be measured. This section reports our measurement results of the emoji domain ecosystem, including quantitative analysis of DNS statistics trends, registration distribution, usage strategies, as well as their web content and intention behind the registrations.

4.1 Growing Trend of DNS Statistics

The Passive DNS dataset is able to capture the DNS requests towards emoji domains among Internet users. The dataset could help to shed light on the first appearance and traffic volume of each emoji domain.

Figure 2 presents the trend of newly emerging emoji domains witnessed from passive DNS. Compared to 2014, the blue line indicates the volume of emoji domains witnessed in 2020 has increased hundreds of times and the entire scale is still increasing. The continuous growth trend of emoji domains is roughly similar to IDN (orange line). We also try to understand the reasons behind the four spikes in Fig. 2. By analyzing domain WHOIS records, we conclude these sharply emerged emoji domains are mainly caused by two reasons. First, the opening of support for emoji domains by several registries has sparked interest, like `.to` and `.ws`. Second, the update of the full list of emoji characters provides more options for the registration market.

The Passive DNS dataset could be also utilized to roughly estimate domain activities [30,39], including their popularity (query volume) and lifetime (intervals between the first and last occurrence).

Our results show that, the ecosystem is as yet in a "self-selling" phase, as a considerable percentage of the traffic and domains themselves are used for the purpose of promotion. Specifically, the DNS requests across the ecosystem were highly concentrated on several most popular ones (top 100 emoji domains hold 74.85% of DNS traffic). Further manual inspections confirm their activities for marketing emoji domains. For instance, the top popular emoji domain, `i❤️.ws`

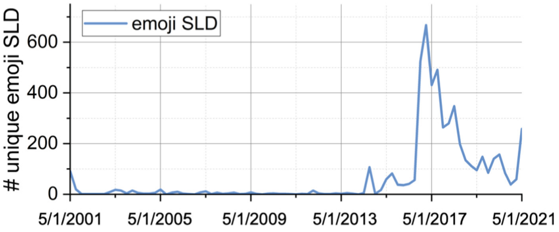


Fig. 3. Registrations of emoji domains.

Table 2. Statistics of collected emoji domains.

Category	**# TLD**	**# ED_sld**	**# ED_sub**	**Registered after 2017**
gTLD	178	1,894	22,552	0
ccTLD-registrable	16	23,600	384	96.15%
ccTLD-other	158	657	5,316	78.22%
All	352	26,151	28,252	75.43%

(`xn--i-7iq.ws`), with 7.96 million DNS requests, is hosting a promotional website for emoji domain registrations. And we find a large number of short-lived emoji domains, with 56.32% of which were active for only one day. A manual survey of 100 random-sampled 1-day domains showed that, 66 of them were "*FOR SALE*".

4.2 Registration Distribution and Usage Strategies

In total, we identify 54,403 emoji domains, 26,151 of which belong to ED_{sld}, i.e., apex domain registered with emoji characters. Associated with domain WHOIS records, we are able to learn the distribution of their registrars, creation dates and expiration dates.

Registration Activity (ED_{sld} **).** The earliest known registration event for emoji domain dates back to 2001 [41]. Benefiting from several ccTLDs supporting emoji domains, the registration volume started to increase rapidly around 2016, as shown in Fig. 3. Then the year 2017, when ICANN proposed the recommendation, was a turning point. By inspecting the sources of registrations, we find mainstream registrars stopped offering emoji domain registrations under gTLDs from 2017. However, several ccTLD registries actively promote the business of emoji domain registration [28, 41]. As a result, the registration activities have continued.

Registration Distribution (ED_{sld} **).** By clustering the registrar fields of domain WHOIS records, we find that 62 registrars have offered (perhaps no longer) the business of emoji domain registrations. Zooming into the distribution, a handful of popular registrars who dominate the global domain name market also play a major role in emoji domains. For example, Godaddy accounts for 26.43% of ED_{sld}.

To investigate the distribution at the registry level, we also categorize all emoji domains by their public suffix [44]. The result shows that the collected emoji domains come from 352 TLDs, including 178 gTLDs and 174 ccTLDs. We also conduct a manual survey of all the ccTLDs, and find 16 of them had explicitly announced their support for emoji domain registrations. These ccTLDs are then termed as ccTLD-registrable, and the remaining are termed as ccTLD-other. By checking the registration dates, we demonstrate ccTLDs have become the main source of emoji domains after 2017.

Emojis Embedding Location (ED_{sub}). While applying emoji domains from gTLD registries has been restricted, domain owners still have the freedom to adopt emoji characters under sub-level domains. In Fig. 2, the scale of newly observed ED_{sub} in passive DNS is rising rapidly. According to the statistics in Table 2, 79.8% (22,552 out of 28,252) of ED_{sub} belong to gTLDs.

We further investigate the usages of ED_{sub} that embed emoji characters under subdomains. Not only do we observe domain registrants themselves to leverage emoji characters for eye-catching, but we also find that third-party services create subdomains with embedded emoji characters. One example is the emoji-URL-shorten service. The service converts the input URL into a domain with a combination of emoji characters as the subdomain of e.mezw.com (1,667 observed in Passive DNS). For instance, `www.google.com` could be converted to `http://🐈🌻😃👸 😏🎹😤☁️🔓🔅e.mezw.com` . Another example is the cloud storage service provided by Amazon S3. The storage bucket would be accessed through an identifier as part of the subdomain under `s3.amazonaws.com`. This mechanism leads to the creation of emoji domains (4,021 observed), e.g., `🌲photo.s3.amazonaws.com` .

Conclusion. Although ICANN's guidelines of emoji domains have served a purpose, particularly for gTLDs, it has not discouraged the registration and use of emoji domains. Dozens of ccTLD registries still support and promote their commercial services for registering emoji domains. In addition, configuring emoji characters under subdomains is becoming a popular alternative, especially under gTLDs. Overall, the ecosystem of emoji domain is still thriving in the wild.

4.3 Infrastructure Analysis

We perform an infrastructure analysis to understand the motivation for registering emoji domains, including their DNS resolution status and web content. Besides, we also evaluate the adoption of security practices on their websites.

DNS Resolution Analysis. Until December 2021, 43,184 (79.4%) emoji domains are still active and resolvable, i.e., could fetch IP addresses through

Table 3. Security deployment of Emoji domains and general popular domains.

Deployment Rate	DNSSEC	HTTPS	HSTS
Regular Domains	1.85% ([6], 2017)	75.51% ([38], 2014)	6.9% ([37], 2017)
Emoji Domains	0.00%	44.61%	4.27%

DNS resolution. Since active emoji domains have configured `NS` records, we collect 2,687 nameservers in total. By comparing with NS records of popular domain parking services [64], 423 emoji domains are found in parking status, suggesting that their owners are seeking to gain profit through traffic monetization. In addition, 2,430 (12.0%) emoji domains also have enabled "`MX`" records, indicating the adoption of email-related services. As an example, `✉️.mail-temp.com` (`xn--4bi.email-temp.com`) is utilized for temporary disposable email services.

Types of Web Content and Intention. We further analyze the web content of all emoji domains (54,403 including ED_{sld} and ED_{sub}) to understand their usage. We perform automatic web crawls (including HTTP and HTTPS) towards all active emoji domains. As a result, 34.21% of emoji domains may act as promotional portals, as their original requests would be redirected to other websites. Totally, we find 9,265 landing domains which are redirected from emoji domains, with 33.39% belonging to the top 10k domains from Tranco List [48]. Our manual inspections show that the redirection targets include social application and registration websites of registries.

Due to the lack of ground truth, it is difficult to automate an accurate content classification of all emoji websites. Therefore, we randomly select 500 emoji domains and manually render their web contents in a controlled browser (Chrome) to inspect their categories. The results show that, 46.4% of them provide meaningless content, such as the default configuration page of the web server (e.g., Nginx) or plain responses with the HTTP status code (`404`, `503`, etc.). 43.8% of the web pages we inspected display contact information of registrants, indicating that these domains are for sale. Besides, 11 domains are employed for personal homepages, and 24 domains for parking advertisements to make profits. Notably, we find 12 emoji domains being utilized for porn or gambling businesses, e.g., `we❤️models.to` (`xn--wemodels-gf7e.to`). In particular, with the help of intelligence information from VirusTotal and Qihoo 360, 2 emoji domains have been categorized as *malicious* as they are associated with malware distribution.

Adoption of Security Practices. We also investigate the deployment of DNSSEC and HTTPS-related security policies for emoji domains, as shown in Table 3. As a whole, the adoption status is significantly less desirable than regular domains. First, by fetching DNSKEY records and HTTPS content, we find no emoji domains have deployed DNSSEC, and the HTTPS adoption rate (44.61%) is also lower than regular domains (75.5% [38] in 2014). Further, we find that 2,322 (4.27%) emoji domains enable HTTP Strict Transport Security (HSTS) by setting the `max-age` HTTPS header. But the deployment rate is also lower than that of regular domains (6.9% [37] in 2017). In addition, the proportion of invalid certificates on emoji domain websites (7.47%, including 1,153 expired certificates and 496 self-signed certificates) is also higher (4.6% [10] of regular domains in 2017). As for the reasons behind such a poor security deployment status, we speculate that, on the one hand, it may be due to the lack of attention to domain security by emoji domain owners. While on the other hand, the inadequate emoji compatibility of security implementations [61] would also matter. For example, OpenSSL is a popular open-source toolkit for implementing TLS, while one critical python library it relies on, `idna`[3], does not support emoji domain processing (unable to trans-coding its Punycode).

Conclusion. The majority of emoji domains could be successfully resolved, with most of them hosting websites, and some even providing email services. Besides, by analyzing the web content, we find emoji domains are now mainly used for promotion, with 34% of them redirecting to other websites. In addition, security implementations of emoji domains are inferior to that of normal domains and need to be improved.

5 Security Threats of Emoji Domain Applications

Until now, little has been done to understand the security risks of emoji domains in real-world applications. In this section, we report an empirical study to explore the threats of visual phishing, parsing and trans-coding errors, aiming to provide guidelines for the correct and safe handling of emoji characters in the future.

5.1 Visual Phishing Threat of Emoji Domains

Threat Model. The eye-catching visual rendering effect of emoji boosts its popularity in domain names. However, the enrichment of rendering without standards from the Unicode community introduces new security risks. In practice, rendering results of the same emoji vary from platform to platform, and even from application to application. As a result, the visual boundaries between different emoji characters may be obscured. Two emoji characters may be rendered quite similarly, even closer on another platform/application than one emoji itself.

Table 4 presents two real-world examples of such visual ambiguity: "`xn--i-7iq.ws`" on Apple renders quite similarly to "`xn--i-n3p.ws`" on Google, and is even more visually equivalent than "`xn--i-7iq.ws`" itself on Windows. As the unique resource identifier in DNS, it raises the security threats of visual phishing. Although previous studies have analyzed the visual phishing attacks of IDN [1,9,22,30,39,46,56,57], the threat has not been well investigated with

[3] https://pypi.org/project/idna/.

Table 4. Examples of phishing emoji domain names.

	Apple (iOS 15.4)	Windows (Win 10)	Samsung (Galaxy M30s)	Google (Pixel 5)	Facebook (Website)	Twitter (Website)	JoyPixels (Website)
U-Label	i❤️.ws	i ❤ .ws	i🧡 .ws	i♥ .ws	i 💖.ws	i 🤎.ws	i💘 .ws
A-Label	xn--i-7iq.ws	xn--i-7iq.ws	xn--i-9h5s.ws	xn--i-n3p.ws	xn--i-u92s.ws	xn--i-744s.ws	xn--i-y92s.ws
U-Label	🍀.yshi.org	🍀 .yshi.org	☘.yshi.org	🌿 .yshi.org	✳ .yshi.org	❇ .yshi.org	🌳 .yshi.org
A-Label	xn--9h8h.yshi.org	xn--9h8h.yshi.org	xn--83h.yshi.org	xn--8h8h.yshi.org	xn--cdi.yshi.org	xn--wdi.yshi.org	xn--wh8h.yshi.org

** Website means that this is rendered from a web service in Chrome browser.*

emoji domains. Below, we provide a quantitative analysis to evaluate the feasibility of emoji domain phishing.

Terminology. In this work, we denote the rendered image of emoji x on platform a as E_{xa}. By calculating image similarities of arbitrary two images, we define one potential "visual phishing attack" against emoji x exists, when:

$$\exists y \neq x, \exists a, b, s.t, Similarity\left(E_{xa}, E_{yb}\right) > \underset{c \neq d}{MAX}\left(Similarity\left(E_{xc}, E_{xd}\right)\right)$$

That is, the similarity between the rendering of emoji y on platform b and emoji x on platform a is quite high, even exceeding the maximum of the internal similarities among x's own rendering results on different platforms/applications.

Feasibility of Visual Phishing Attacks. Here, we introduce our methodology to quantitatively assess the feasible space for visual phishing attacks on emoji domains. First, we extensively collect the rendering results of thousands of emoji characters on mainstream applications (Google, Facebook, Twitter, JoyPixels) and operating systems (Apple, Windows, Samsung) [11], yielding a dataset of 12,169 images of 1,816 emoji characters (excluding the GIF images). Specifically, image E_{xa} is the rendered result of emoji x ($1 \leq x \leq 1816$) on platform a ($1 \leq a \leq 7$), which is a 72×72 matrix with each element (pixel) ranging from 0 to 255. Then, we test five classical image similarity metrics to evaluate the visual similarity, including Peak Signal-To-Noise (PSNR) [23,65], Feature Similarity Indexing Method (FSIM) [67], Information theoretic-based Statistic Similarity Measure (ISSM) [3], Signal to Reconstruction Error ratio (SRE) [36], and Spectral Angle Mapper (SAM) [66].

As there is no ground-truth dataset for this task, we started by manually labeling an emoji icon similarity dataset by two researchers, with 125 randomly selected emoji image pairs. Image pairs with inconsistent labels will be double-checked. Following, we input similarity results of each pair using five metrics separately for similarity classification via a random forest (RF) model [2], with a 16:9 training-test ratio. As shown in Table 5, **FSIM** performs the best, which could achieve an accuracy of 80%, and has been chosen as our final method. We also open source the labeled emoji similarity dataset[4] to facilitate future work.

[4] https://github.com/EmojiDomain/ESORICS22/.

Table 5. Evaluation of each image similarity metric.

	PSNR	FSIM	ISSM	SRE	SAM
Accuracy	66.67%	80.00%	71.11%	73.33%	68.89%
Precision	66.67%	86.20%	76.92%	82.75%	71.43%
Recall	75.00%	83.33%	74.07%	77.42%	76.92%
F1 score	70.59%	84.75%	75.47%	80.00%	74.07%

Finally, based on the similar results of **FSIM** among 148 million emoji image pairs, we find that **1,332** (73.35%) emoji characters could be threatened by the above visual phishing attacks.

Visual Phishing in the Real World. We also try to answer the question of whether visual phishing attacks already happening in the real world, and the registration space of phishing domains from the perspective of adversaries. In total, 1,112 pairs of the collected emoji domains satisfy the similarity requirement of visual phishing attacks. Through manual inspections, we do observe some suspected examples, i.e., `i😀.ws` (`xn--i-jv3s.ws`) and `i😃.ws` (`xn--i-pv3s.ws`) both promoting the service of emoji domain registration. However, we could not further verify whether they were actually exploited for phishing. Besides, we also speculate that some of the similar domain pairs are caused by defensive registering, i.e., registrants pre-register domains similar to their own to prevent others from registering for phishing. For example, the website owner of `i❤️.ws (xn--i-7iq.ws)` also has registered another 4 emoji domains with similar "heart" characters, e.g., `i♥️.ws (xn--i-n3p.ws)`. However, most of the vulnerable emoji domains have not been protected yet. Taking the top 100 popular domains with the most queries as examples, we find that 78 of them are phishable, and 67 of them even have more than 10 potential phishing domains. By requesting the registration API of Godaddy [18], we find that 23.38% of visually similar emoji domains are available for registration, leaving considerable space for adversaries.

5.2 Parsing Error of Emoji Domains

Threat Model. To optimize usability, mainstream online social media and chatting platforms would automatically parse URLs in the text and render them into clickable links. However, this automatic process is not always reliable, and the unanticipated parsing may lead to "unintended URLs". Beliz et al. [29] explored the "unintended URLs" caused by *typos*, where users forget the space after a full stop and the next sentence happens to begin with a "TLD" word (e.g., .to and .online). Attackers could exploit such parsing errors by registering the domains in "unintended URLs" and hijacking the traffic.

The introduction of emoji expands the character space of domains, which raises new challenges for URL parsing. By empirical analysis and manual inspections of open-source projects, we find one common approach to parsing URLs is regular expression matching. For instance, Android 11 (with SDK version 30)[5] has a predefined character set for URL recognition and also includes 168 emoji characters. However, simply expanding the character set may introduce URL parsing errors.

We present two cases below, where the emoji characters are incorrectly recognized as part of the URLs. Attackers can hijack the traffic towards www.google.com and `i❤️.ws` (`xn--i-7iq.ws`) by registering `com⭐.to` (`xn--com-x19a.to`) and `😄i❤️.ws` (`xn--i-7iq2158q.ws`).

> Case-I *Check www.google.com⭐.To hurry up.*
>
> Case-II *You can register your own emoji domains 😄i❤️.ws.*

Parsing Errors in the Real World. Through manual testing, we confirmed that such parsing errors are prevalent on the Android platform even in multiple Android systems (e.g., version 5–8 with SDK version 21–27) and applications (e.g., Short Message Service), indicating the developers are not yet aware of such vulnerabilities.

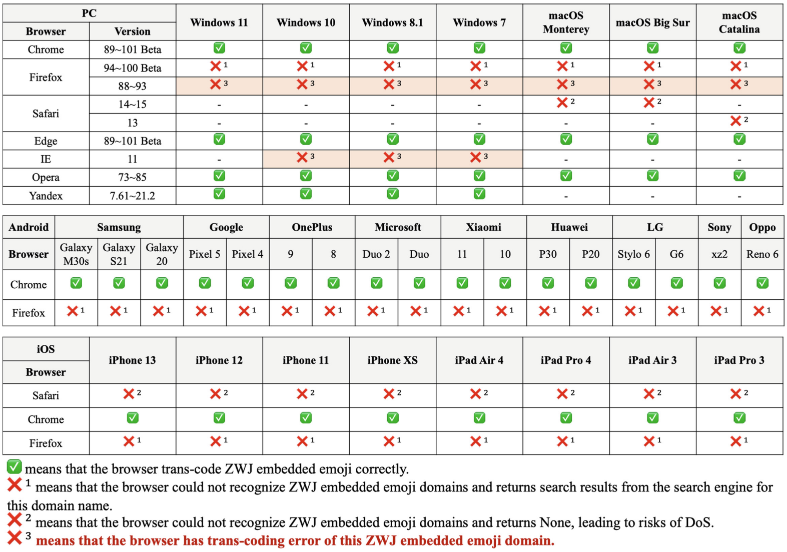
Moreover, to evaluate the impact of this security threat, we apply for one day of DNS request data (April 13, 2021) from `B Root` [62]. As most "unintended domains" would be not resolvable, Root traffic could provide a holistic observation of parsing errors. Based on the two cases above, the structural features of wrongly parsed emoji domains could be summarized as follows: the emoji character appears on the right-most and before a TLD (Case-I) or left-most (Case-II) side of a valid domain, and is then incorrectly parsed as a new (most likely NXDomain) domain name. Therefore, we first filter out the emoji domains from the DNS requests in Root traffic, and divide the domains into left and right sub-strings by the emoji character. When one origin domain is NXDomain, while its left sub-string is a valid domain, it would be tagged as Case-I; when its right sub-string is a valid domain, it would be tagged as Case-II. Finally, a total of 6,372 emoji domains are reported as "parsing error", including 1,591 Case-I (e.g., `youtube.com👍.dlink` with A-Label of `youtube.xn--com-3113b.dlink`) and 4,781 Case-II (e.g., `☝️meet.google.com` with A-Label of `xn--meet-uk3b.google.com`). Based on the Godaddy registration API [18], we find that 43.13% of emoji domains with "parsing error" are available for registration. To conclude, we speculate this security threat does have a real-world impact and needs to be taken seriously by individual applications.

5.3 Trans-coding Issue of Emoji Domains

Benefiting from the existing disclosure of IDN vulnerabilities, applications would trans-code domains with non-ASCII characters into A-Labels (strings starting with "xn–") to mitigate phishing threats. Hu *et al.* [24] found that mainstream

[5] https://developer.android.com/studio.

Table 6. Trans-coding test results of ZWJ embedded emoji domains.



PC		Windows 11	Windows 10	Windows 8.1	Windows 7	macOS Monterey	macOS Big Sur	macOS Catalina
Browser	**Version**							
Chrome	89~101 Beta	✅	✅	✅	✅	✅	✅	✅
Firefox	94~100 Beta	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]
	88~93	❌[3]	❌[3]	❌[3]	❌[3]	❌[3]	❌[3]	❌[3]
Safari	14~15	-	-	-	-	❌[2]	❌[2]	-
	13	-	-	-	-	-	-	❌[2]
Edge	89~101 Beta	✅	✅	✅	✅	✅	✅	✅
IE	11	-	❌[3]	❌[3]	❌[3]	-	-	-
Opera	73~85	✅	✅	✅	✅	✅	✅	✅
Yandex	7.61~21.2	✅	✅	✅	✅	-	-	-

Android	**Samsung**			**Google**		**OnePlus**		**Microsoft**		**Xiaomi**		**Huawei**		**LG**		**Sony**	**Oppo**
Browser	Galaxy M30s	Galaxy S21	Galaxy 20	Pixel 5	Pixel 4	9	8	Duo 2	Duo	11	10	P30	P20	Stylo 6	G6	xz2	Reno 6
Chrome	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
Firefox	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]

iOS	**iPhone 13**	**iPhone 12**	**iPhone 11**	**iPhone XS**	**iPad Air 4**	**iPad Pro 4**	**iPad Air 3**	**iPad Pro 3**
Browser								
Safari	❌[2]	❌[2]	❌[2]	❌[2]	❌[2]	❌[2]	❌[2]	❌[2]
Chrome	✅	✅	✅	✅	✅	✅	✅	✅
Firefox	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]	❌[1]

✅ means that the browser trans-code ZWJ embedded emoji correctly.
❌[1] means that the browser could not recognize ZWJ embedded emoji domains and returns search results from the search engine for this domain name.
❌[2] means that the browser could not recognize ZWJ embedded emoji domains and returns None, leading to risks of DoS.
❌[3] **means that the browser has trans-coding error of this ZWJ embedded emoji domain.**

browsers would selectively trans-code IDNs in the address bar of browsers. In this study, we conducted a similar investigation on how emoji domains are displayed by browsers.

However, the trans-coding process itself is also a complex task and could be error-prone, especially when dealing with special functional characters for emoji rendering. The most representative special character is Zero With Joiner (ZWJ, "U+200D" and "U+200C"). It is **invisible**, but can change the rendering results of its contiguous emoji (e.g., 🏳️+ZWJ+🌈 would be rendered as 🏳️‍🌈). Unfortunately, the community has not developed a uniform (and strict) standard for how to handle ZWJ in domain names [31], e.g., IDNA 2003 recommended removing ZWJ in trans-coding while IDNA 2008 considered keeping it as a valid character. Such insidious characters would introduce serious ambiguity when used as unique identifiers, thus discouraging being used in domain names [55].

We witnessed 1,026 emoji domains with ZWJ being used in the wild based on our collected dataset, and further confirmed they do trigger ambiguity during trans-coding. The test was performed on LambdaTest [35], a cloud-based framework that supports remote testing on different versions of browsers across multiple operating systems. By configuring the versions of browsers, operating systems and domains to be tested, we can remotely control LambdaTest to load the domain in the address bar of the specified browser and get the result in video form. A total of 7 browser vendors on 7 PC operating systems and 3 browser vendors on 10 mobile brands were tested. According to the results shown in Table 6, we find that trans-coding of emoji domains is primarily implemented by the browser vendors themselves, independent of the platforms they are running on. Specifically, Chrome is correctly implemented in all versions on all platforms, keeping the `ZWJ` and trans-coding it (e.g., trans-codes `💇‍♂️.ws` to `xn--g5hz810o.ws` correctly). However, other browser implementations are not satisfactory. Most seriously, lower versions (88–93) of Firefox and IE stand in the "drop `ZWJ`" branch, e.g., `💇‍♂️.ws` would be improperly trans-coded as `xn--1ug66vqx45b.ws`, which is totally a different domain and could lead to security risks of traffic hijacking. There are also flaws where browsers could not recognize `ZWJ` embedded emoji domains and then use them as keywords to fetch results from search engines (e.g., higher versions of Firefox), or return the navigation page directly (e.g., Safari), causing the denial of access failures. Considering the prevalence of special characters used in the emoji ecosystem (e.g., `ZWJ` can be combined with at least 202 sets of emoji characters for special effects), we need to explore the best security practice for emoji domain trans-coding and propose consistent standards to mitigate the above risks.

6 Discussion

Recommendations. In this study, we provide a landscape of how emoji domains are parsed in mainstream platforms and applications. Most of them are "compatible", but from the perspective of adversaries, they are not prepared to deal with potential security risks. Given the rapidly growing trend of emoji domains in the wild, we believe it is necessary to take action for mitigation. Here, we provide three recommendations based on our observations:

- **Unicode community: provide guidance for emoji domain processing.** Our study reveals that the Unicode standards still have ambiguous and unspecified fields on emoji domain processing, which should be specified and regulated in the near future. For example, we need best security practices on how to securely parse special emoji characters during the trans-coding, such as `ZWJ`. Furthermore, despite the fact that it could be difficult to uniform the rendering of emojis across all platforms, the Unicode community should propose guidance to prevent visual phishing attacks.
- **Domain registry and registrar: adopt proactive anti-phishing defenses.** As mentioned above, a dozen of ccTLD registries are supporting and promoting the registration of emoji domains. Considering potential security concerns, the related registries and registries should take proactive approaches. In particular, a previous study proposes a series of anti-phishing defenses for IDN domain registrations [15], including enumerating potential phishing domains in advance based on emoji similarity, and encouraging users to register them proactively.
- **Application: elaborate emoji-compatible implementation.** Applications should balance both usability and security of emoji domains, particularly in the parsing and trans-coding processes. The threat models and test cases presented in this paper could be considered as references for secure testing of applications.

Generality of Proposed Security Risks. The essence of this work is the security pitfalls when special Unicode characters are adopted as unique identifiers. Therefore, the security risk is generally applicable in multiple scenarios beyond emoji domains, e.g., the Windows registry [58] and file paths with IDNs [54]. We believe that the first exploration perspectives in this work, such as trans-coding, parsing and rendering of special Unicode characters, are also applicable to other areas. We leave the exploration of broader scenarios as our future work.

Ethical Consideration. The major ethical considerations for this study include data collection and security threat disclosure. First, the datasets we collected are publicly available and used for research purposes only. No personally sensitive information is involved in the data collection. Second, we propose three security threats and evaluate their feasibility in the real world. Our results demonstrate that ambiguous understanding and mishandling of emojis are prevalent in the wild. As a result, it is possible that these attacks will be initiated by adversaries. However, we consider that our study gains more benefits than exposing threats, which makes the security community aware of the unique security threats introduced by emoji characters.

7 Related Works

IDN Domain Security. The initiative of IDN has been proposed for a long time, and attracted the security community to study its ecosystem and implications. Since registrants are free to choose characters from the Unicode consortium, an adversary can carefully craft an IDN domain that looks quite similar to a popular domain by replacing ASCII characters with Unicode ones. Such an attack is named IDN homograph. Security accidents show that homographic IDN has been utilized by cyber-criminals [21,22]. In 2018, a reexamination study was conducted to detect registered homographic IDNs and estimate the scale of available ones [39]. After that, the methodology of homograph attack detection was optimized by a series of studies [56,59]. To mitigate this risk, mainstream browsers introduced defense policies. However, almost all implementations have weaknesses in their rules, leaving opportunities for attackers and re-allow homograph IDNs [24].

Domain Abuse. Continuous expansion of domain space led to the security risk of domain squatting [20,46,59]. Besides homograph IDNs, deceptive domains could be constructed by typos [1,43,57], flipping a bit [46], using a hyphen to connect related keywords [30], the sound similarity [45], or even the long-length of domain name [9]. Previous studies demonstrated that newly released TLDs may be exploited to create look-alike domain names of popular brands [5,17,19,20,33]. And recent work shows that domain impersonation attacks also have a negative impact on the issuance of TLS certificates [53]. To the best of our knowledge, there is no prior work has attempted to explore the security implication of emoji domains for the DNS ecosystem.

8 Conclusion

This work is the first to propose a systematic study of emoji domains based on a comprehensive dataset, including 1,366 TLD zone files, and long-period country-level passive DNS datasets. We identify 54,403 emoji domains by matching characters with the emoji code point lists. We show that the scale of emoji domains is constantly growing in the wild. The proliferation of emoji domain registrations under ccTLDs and configuring emoji icons in subdomains have enabled the entire ecosystem to remain developing after 2017 and up to now. About half of emoji domain names are associated with meaningful web content, with most for promotion and redirection, or even pornographic sites. However, it still lacks best security practices in supporting and parsing emoji domains, which exposes serious security risks, including phishing threats, parsing errors and trans-coding issues. Overall, we believe that the development of emoji domain names is at an early stage. And different communities should pay more attention to the security issues, and take efforts to find the best practice for processing emoji domains.

Acknowledgments. We thank all the anonymous reviewers for their valuable comments to improve this paper, and our industry partner, 360 Netlab, for their help and support. We acknowledge the University of Southern California for providing B Root data for security analysis. The authors were supported by the National Natural Science Foundation of China (U1836213, U19B2034, 62102218), the Alibaba Innovative Research (AIR) programme, and a gift from China Telecommunications Corporation. Baojun Liu was also partially supported by the Shuimu Tsinghua Scholar Program.

References

1. Agten, P., Joosen, W., Piessens, F., Nikiforakis, N.: Seven months' worth of mistakes: a longitudinal study of typosquatting abuse. In: NDSS. Internet Society (2015)
2. Alasalmi, T., Suutala, J., Röning, J., Koskimäki, H.: Better classifier calibration for small datasets. ACM Trans. Knowl. Discov. Data (2020)
3. Aljanabi, M.A., Hussain, Z.M., Shnain, N.A.A., Lu, S.F.: Design of a hybrid measure for image similarity: a statistical, algebraic, and information-theoretic approach. Eur. J. Remote Sens. (2019)
4. CAIDA and Ian Foster: Caida-dns zone database (dzdb) (2020). https://dzdb.caida.org/
5. Chen, Q.A., Thomas, M., Osterweil, E., Cao, Y., You, J., Mao, Z.M.: Client-side name collision vulnerability in the new gtld era: a systematic study. In: SIGSAC 2017. ACM (2017)
6. Chung, T., et al.: A longitudinal, end-to-end view of the DNSSEC ecosystem. In: USENIX Security 2017. USENIX Association (2017)
7. Costello, A.: Punycode: a bootstring encoding of Unicode for internationalized domain names in applications (IDNA). Technical report, RFC 3492, March 2003
8. Du, K., Yang, H., Li, Z., Duan, H., Zhang, K.: The ever-changing labyrinth: a large-scale analysis of wildcard DNS powered blackhat SEO. In: USENIX Security 2016. USENIX Association (2016)

9. Du, K., et al.: TL;DR hazard: a comprehensive study of levelsquatting scams. In: Chen, S., Choo, K.-K.R., Fu, X., Lou, W., Mohaisen, A. (eds.) SecureComm 2019. LNICST, vol. 305, pp. 3–25. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37231-6_1
10. Durumeric, Z., Kasten, J., Bailey, M., Halderman, J.A.: Analysis of the https certificate ecosystem. In: IMC 2013. ACM (2013)
11. Emoji Community: Full emoji list, v14.0 (2022). https://unicode.org/emoji/charts/full-emoji-list.html
12. Emoji Domain Registration: Emoji domain case study: Weapon depot and 🔫.ws (2017). https://emoji-domains.medium.com/emoji-domain-case-study-ws-bd070f31090f
13. Faltstrom, P.: The Unicode code points and internationalized domain names for applications (idna). Technical report, RFC 5892, August 2010
14. Faltstrom, P., Hoffman, P., Costello, A.: Internationalizing domain names in applications (IDNA). Technical report, RFC 3490, March 2003
15. Fu, A.Y., Deng, X., Wenyin, L., Little, G.: The methodology and an application to fight against Unicode attacks. In: Usable privacy and security (2006)
16. Gandi News: Can you use emojis in your domain name? yes! you can! (2020). https://news.gandi.net/en/2020/07/can-you-use-emojis-in-your-domain-name-yes-you-can/
17. Pouryousef, S., Dar, M.D., Ahmad, S., Gill, P., Nithyanand, R.: Extortion or expansion? An investigation into the costs and consequences of ICANN's gTLD experiments. In: Sperotto, A., Dainotti, A., Stiller, B. (eds.) PAM 2020. LNCS, vol. 12048, pp. 141–157. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-44081-7_9
18. Godaddy: Search and buy domains in bulk. https://sg.godaddy.com/domains/bulk-domain-search . Accessed January 2022
19. Halvorson, T., Der, M.F., Foster, I., Savage, S., Saul, L.K., Voelker, G.M.: From .academy to .zone: an analysis of the new TLD land rush. In: IMC 2015. ACM (2015)
20. Halvorson, T., Levchenko, K., Savage, S., Voelker, G.M.: XXXtortion? Inferring registration intent in the. XXX TLD. In: WWW 2014 (2014)
21. Hannay, P., Baatard, G.: The 2011 IDN homograph attack mitigation survey. Edith Cowan University Publications (2012)
22. Holgers, T., Watson, D.E., Gribble, S.D.: Cutting through the confusion: a measurement study of homograph attacks. In: USENIX ATC 2006 (2006)
23. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369. IEEE (2010)
24. Hu, H., Jan, S.T., Wang, Y., Wang, G.: Assessing browser-level defense against IDN-based phishing. In: USENIX Security 2021 (2021)
25. Internet Assigned Numbers Authority (IANA): Resources for country code managers. https://www.icann.org/resources/pages/cctlds-21-2012-02-25-en. Accessed 25 Feb 2012
26. Internet Corporation for Assigned Names and Numbers : ICANN gtld registries and registrars required to implement new interim registration data policy by 20 May 2019 (2019). https://www.icann.org/en/announcements/details/icann-gtld-registries-and-registrars-required-to-implement-new-interim-registration-data-policy-by-20-may-2019-17-5-2019-en
27. Internet Corporation for Assigned Names and Numbers: SSAC advisory on the use of emoji in domain names (2017). https://www.icann.org/en/system/files/files/sac-095-en.pdf

28. Johnson, P.: Emoji domains are the future (2018). https://gizmodo.com/emoji-domains-are-the-future-maybe-1823319626
29. Kaleli, B., Kondracki, B., Egele, M., Nikiforakis, N., Stringhini, G.: To err. is human: characterizing the threat of unintended URLs in social media. In: NDSS 2021 (2021)
30. Kintis, P., et al.: Hiding in plain sight: a longitudinal study of combosquatting abuse. In: SIGSAC 2017 (2017)
31. Klensin, J.: Internationalized domain names for applications (IDNA): background, explanation, and rationale. Technical report., RFC 5894, August 2010
32. Klensin, J.: Internationalized domain names for applications (IDNA): definitions and document framework. Technical report, RFC 5890, August 2010
33. Korczynski, M., et al.: Cybercrime after the sunrise: a statistical analysis of DNS abuse in new gTLDs. In: AsiaCCS 2018 (2018)
34. Lab, V.: The centralized zone data service. https://czds.icann.org/home. Accessed January 2021
35. Lambdatest: Lambdatest: Cross browser testing cloud (2022). https://app.lambdatest.com/. Access April 2022
36. Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K.: Super-resolution of sentinel-2 images: learning a globally applicable deep neural network. ISPRS J. Photogramm. Remote Sens. (2018)
37. Li, X., Wu, C., Ji, S., Gu, Q., Beyah, R.: HSTS measurement and an enhanced stripping attack against HTTPS. In: Lin, X., Ghorbani, A., Ren, K., Zhu, S., Zhang, A. (eds.) SecureComm 2017. LNICST, vol. 238, pp. 489–509. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78813-5_25
38. Liang, J., Jiang, J., Duan, H., Li, K., Wan, T., Wu, J.: When https meets CDN: a case of authentication in delegated service. In: IEEE S&P. IEEE (2014)
39. Liu, B., et al.: A reexamination of internationalized domain names: the good, the bad and the ugly. In: DSN (2018)
40. Lu, C., et al.: From WHOIS to WHOWAS: a large-scale measurement study of domain registration privacy under the GDPR. In: NDSS 2021 (2021)
41. Michael Cyger: The definitive guide to emoji domain names (2017). https://www.dnacademy.com/emoji-domains
42. Mockapetris, P.V.: RFC 1034: domain names-concepts and facilities (1987)
43. Moore, T., Edelman, B.: Measuring the perpetrators and funders of typosquatting. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 175–191. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14577-3_15
44. Mozilla Foundation.: Public suffix list. https://publicsuffix.org/. Accessed December 2021
45. Nikiforakis, N., Balduzzi, M., Desmet, L., Piessens, F., Joosen, W.: Soundsquatting: uncovering the use of homophones in domain squatting. In: Chow, S.S.M., Camenisch, J., Hui, L.C.K., Yiu, S.M. (eds.) ISC 2014. LNCS, vol. 8783, pp. 291–308. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13257-0_17
46. Nikiforakis, N., Van Acker, S., Meert, W., Desmet, L., Piessens, F., Joosen, W.: Bitsquatting: exploiting bit-flips for fun, or profit? In: WWW 2013 (2013)
47. Organisation of 👁👄👁: what it really is? https://xn-mp8hai.fm/statement. Accessed October 2021
48. Pochat, V.L., van Goethem, T., Tajalizadehkhoob, S., Korczynski, M., Joosen, W.: Tranco: a research-oriented top sites ranking hardened against manipulation. In: NDSS 2019. The Internet Society (2019)
49. Qihoo 360 Netlab: DNSPai: Passive DNS. https://passivedns.cn/. Accessed October 2021

50. Qihoo 360 Netlab: Qihoo 360 Netlab. https://netlab.360.com/. Accessed October 2021
51. Rainey, C.: Coca-cola bought a whole bunch of emoji web addresses. https://www.grubstreet.com/2015/02/coke-emoji-websites.html. Accessed 20 Feb 2015
52. van Rijswijk-Deij, R., Jonker, M., Sperotto, A., Pras, A.: A high-performance, scalable infrastructure for large-scale active DNS measurements. IEEE J. Sel. Areas Commun. **34**(6), 1877–1888 (2016)
53. Roberts, R., Goldschlag, Y., Walter, R., Chung, T., Mislove, A., Levin, D.: You are who you appear to be: a longitudinal study of domain impersonation in TLS certificates. In: SIGSAC 2019. ACM (2019)
54. Smith, B.: Chinese characters in windows registry: are they safe? https://internet-access-guide.com/chinese-characters-in-windows-registry/. Accessed 23 Feb 2021
55. Suignard, M.: Proposed update Unicode technical report# 36 (2014). https://unicode.org/reports/tr36/
56. Suzuki, H., Chiba, D., Yoneya, Y., Mori, T., Goto, S.: Shamfinder: an automated framework for detecting IDN homographs. In: IMC 2019. ACM (2019)
57. Szurdi, J., Kocso, B., Cseh, G., Spring, J., Felegyhazi, M., Kanich, C.: The long taile of typosquatting domain names. In: USENIX Security 2014 (2014)
58. The MITRE Corporation.: Masquerading: Right-to-left override. https://attack.mitre.org/techniques/T1036/002/. Accessed 14 Oct 2021
59. Tian, K., Jan, S.T., Hu, H., Yao, D., Wang, G.: Needle in a haystack: Tracking down elite phishing domains in the wild. In: IMC 2018. ACM (2018)
60. Unicode Community: Emoji list (2020). https://www.unicode.org/Public/emoji/13.1/emoji-test.txt
61. Universal Acceptance: Reviewing programming languages and frameworks for compliance with universal acceptance good practice, May 2019
62. USC/B-Root Operations with USC/LANDER project: Day in the life of the internet (ditl) april, 2015 dataset, impact id: web1191, for dataset: Usc-lander/ditl_b_root_message_question-20210413. https://ant.isi.edu/datasets/dns/index.html. Accessed January 2022
63. ViewDNS: Viewdns info. https://viewdns.info/. Accessed October 2020
64. Vissers, T., Joosen, W., Nikiforakis, N.: Parking sensors: analyzing and detecting parked domains. In: NDSS 2015. Internet Society (2015)
65. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
66. Yuhas, R.H., Goetz, A.F., Boardman, J.W.: Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In: Proceedings of the Summaries 3rd Annual JPL Airborne Geoscience Workshop (1992)
67. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. (2011)