Domain Impersonation opportunities amidst TLS availability

Himanshu Goyal
Georgia Institute of Technology
Atlanta, Georgia, USA
hgoyal33@gatech.edu

Samina Shiraj Mulani
Georgia Institute of Technology
Atlanta, Georgia, USA
smulani6@gatech.edu

Abstract

This research work investigates the significant threat of domain impersonation attacks, wherein attackers create domains that resemble legitimate ones to deceive users into disclosing sensitive information. The effectiveness of existing security measures in detecting and preventing such attacks is analyzed. Specifically, the warning messages in the browsers and the Safe Browsing API, a prevalent browser-based security mechanism, are examined in identifying potentially harmful websites. Our Chrome-based browser study states that around 45% of look-alike domains are available to buy, and the browser application only shows warning messages for only 1%, providing an open area for attackers to launch malicious webpages. Our research also demonstrates the limitations of relying solely on these mechanisms, as attackers can easily obtain certificates for look-alike domains through free certificate authorities such as Let's Encrypt. We find that majority of the look-alike domains in our dataset have valid TLS certificates and that RPKI validity cannot be used to distinguish these sites. To address this issue, we propose a solution that combines tools such as DNSTwist and CT Monitors, which can assist domain owners in identifying potentially malicious look-alike domains. This solution can help mitigate the risks of domain impersonation attacks and improve the security of popular domains on the internet. The findings of this research work contribute to a more comprehensive understanding of the security risks associated with popular domains and inform the development of more effective security measures.

Keywords: Domain Typosquatting, Google Chrome, Safe Browsing, RPKI, TLS Certificates

1 Introduction

The use of *HTTPS* (Hypertext Transfer Protocol Secure) in website URLs [7] has become a widely recognized symbol of website security. When users see the HTTPS lock icon in their browser, they may feel confident that the website they are accessing is legitimate and secure. This is because HTTPS is designed to provide a secure and encrypted connection between the user's device and the website's server, helping to prevent unauthorized access, data theft, and other security threats.

However, it's important to note that the mere presence of the HTTPS lock icon does not guarantee the security of a website. Attackers can use various techniques to deceive users into thinking a website is secure, even when not. For example, they can use phishing URLs that mimic legitimate websites and display the HTTPS lock icon to trick users into entering their sensitive information. This is known as *"HTTPS phishing"* or *"HTTPS spoofing"* and it's a growing concern in the security community. The mimicked sites often employ *"typosquatting"* techniques to appear legitimate to unobservant eyes, wherein URLs of the phishing websites contain common misspellings, use a different top-level domain, or use homoglyphs.

In fact, recent research [6] [10] has also shown that phishing websites now use HTTPS encryption to make their fraudulent sites appear more legitimate and trustworthy to users. This demonstrates that HTTPS alone is not enough to protect users from all security threats, and that users need to be aware of the risks of phishing and other social engineering attacks.

Therefore, while HTTPS is an important security feature, users should not rely solely on the lock icon's presence to determine a website's legitimacy and security. Instead, they should look for other indicators of website security, such as the website's reputation, reviews, and domain name. Additionally, users should be cautious when clicking on links or entering sensitive information online and should always verify the legitimacy of a website before providing any personal information. By staying informed and vigilant, users can protect themselves from the growing threat of phishing and other online security threats.

The primary objective of our study is to determine the level of security of prominent domains on the Internet. To accomplish this, we have compiled a list of key queries that we intend to answer. First, we wish to determine how simple it is to impersonate well-known domains and what such attacks could entail. When attackers impersonate prominent domains, they can steal sensitive data from users who are unaware of the situation. This poses a significant risk to internet security. Because of this, it is essential to understand how attackers pose as prominent websites and how effective security measures are against such attacks.

Secondly, we would like to determine whether web browsers are an effective first line of defence against such attacks. Web

browsers are essential for ensuring the security of internet transactions and protecting users from harmful websites. In contrast, recent research [10] [5] indicates that web browsers may not be able to detect and prevent domain impersonation attempts. Therefore, we wish to determine what issues exist with the current browser-based security methods and where they could be enhanced. By providing answers to these questions, we aim to help people gain a better understanding of the security risks associated with well-known domains and improve their security measures.

Another essential aspect of our research is determining how many impersonated popular domains use Transport Layer Security (TLS) certificates and identifying any patterns or trends in the Resource Public Key Infrastructure (RPKI) data pertaining to TLS certificates for these websites. TLS certificates are a vital component of internet security because they ensure the privacy and accuracy of data transmitted over the internet. Many prominent domains may not have TLS certificates or may have certificates that are invalid or have expired. Therefore, we wish to determine how well TLS certificates are utilised on prominent websites and what may be preventing their greater adoption. The Certificate Transparency (CT) system [8], supported by Chrome and Safari, is leveraged for collecting information regarding the certificates. The CT system consists of a distributed, independent, append-only ledger of certificate logs secured using a Merkle Tree, making them cryptographically verifiable by Monitors.

We would also like to examine RPKI data to determine if there are any commonalities or patterns among these sites' TLS certificates. RPKI is a framework that improves security for the Border Gateway Protocol (BGP) by allowing cryptographic verification of the origin Autonomous System (AS) in BGP announcements. By examining RPKI records, we expect to identify any issues with the way TLS certificates are issued and revoked. By answering these questions, we hope to assist people gain a better understanding of the current state of internet security and develop more effective security measures. Our ultimate goal is to enhance the security of well-known domains on the Internet and ensure that people can trust the websites they visit.

2 Datasets Description

- Tranco List - Tranco List [9] provides a ranking of the top million popular domains by averaging data from multiple providers (like Alexa Internet Top 1 Million, Cisco Umbrella Popularity List, and so on). We use the Tranco List obtained on 2^{nd} April for our study.
- RIPEstat Application Programming Interface (API) - RIPEstat is an online tool developed by the Regional Internet Registry for Europe, the Middle East and parts of Central Asia (RIPE NCC) that provides a wide range of information and analysis on internet resources such as IP addresses, autonomous system numbers (ASNs), domain names and network-related data. We leverage the RIPEstat Data API [2], which is the public data interface for RIPEstat.
- CT logs - We use two public APIs, namely crt.sh and SSLMate, to access cryptographically verified TLS certificate data recorded in the public CT ledger. The first data source, crt.sh, is often unstable (resulting in Gateway errors), and limits entries if the domains are large. Thus, the data from this source is supplemented by SSLMate, which is up-to-date but does not show information about expired certificates.

2.1 Domain Twist

We use a tool called DNSTwist [1] to generate look-alike domains (working described in Section 3). The following depict the variations the tool detects -

- Typo squatting: banrkofamerica.com
- Hyphenation: bankofamerica-signin.com
- Homographs: bànkofamerica.com
- Omission: bankofamrica.com
- Repetition: bankoffamerica.com
- More variations: vowel-swap, subdomain, replacement etc.

3 Methodology

In order to assess the vulnerabilities of domain impersonification in the TLS ecosystem, our methodology is broken down into two parts. The first examines the behavior of look-alike domains in a popular web browser like Google Chrome. The second part looks at the data available from RIPEstat and Certificate Transparency logs to look for patterns with respect to RPKI and TLS. The entire workflow is highlighted in Figure 1.

3.1 Browser Study

1. **Tranco List:** We randomly sampled 150 domains from the top 1500 list, which included organizations from various sectors such as government, education, healthcare, and online shopping. We then used these domains as input to the domain permutation engine, which we describe below.
2. **Domain Permutation:** We employed DNSTwist [1] to perform domain permutations on the target 150 domains. DNSTwist takes the original domain as input and generates permutations with a minimum edit distance with respect to the given domain. It also considers domain alteration types and inserts or removes characters accordingly. As part of our study, we aimed to generate approximately 30 close permutations of each target domain, covering the domain manipulations outlined in Section 2.1. We then passed the generated domains to our local Domain Name Server (DNS)

to obtain the corresponding Name Server's IP address. We used the collected information to understand the current landscape of domain impersonation from the perspective of Google Chrome Browser, TLS certificates, RPKI status, and the ease of buying a similar-looking domain to the original. We further contrasted our browser findings with Google's Safe Browsing API.

3. **Chrome Browser Study:** Our objective was to understand what Google Chrome shows for all the permuted domains. We wanted to know how many domains served HTTP or HTTPS connections, whether Chrome displayed warnings for these closely related domains, and how many domains were available for sale. To answer these questions, we used a web crawler software, Selenium, to automate the process of visiting each of the domains in Chrome. We extracted information from the Hypertext Markup Language (HTML) text and the TLS certificates returned by Chrome. We searched for keywords such as "domain sale" and "free domain parked" in the HTML text and relied on security statistics received after visiting the domain using Selenium's Chrome API to determine whether the domain served HTTP or HTTPS. We used the HTML text to understand the warnings displayed by Chrome since the warning is mostly standard. To validate our findings, we took screenshots of the homepage of each domain we visited and manually reviewed them to remove any false positives or true negatives.
4. **Google's Safe Browsing:** We used Google's Safe Browsing API to understand whether the warnings displayed by Chrome in the user application were based on the API's information.

3.2 RPKI and TLS Certificate Study

The list of permuted domains and their corresponding Internet Protocol (IP) addresses obtained from the first part are queried under the RIPEstat API to obtain data about the AS number, RPKI status and Regional Internet Registry (RIR) registration. The domain names are used to look up Certificate Transparency logs to extract relevant information such as validity of certificate, country distribution and Certificate Authority distribution. When searching for the relevant certificate for a domain, those which have an exact match with the domain in the common name field are picked. With the data from RIPEstat and CT logs, trends in the data are observed by grouping and plotting the different columns.

4 Results

4.1 Google Chrome

The statistics for the domain resolution pertaining to all generated twisted domains are displayed in Figure 2. Out of all generated domains, around 83.5% of the domains resulted in successful webpage loading; however, the remaining 16.5%

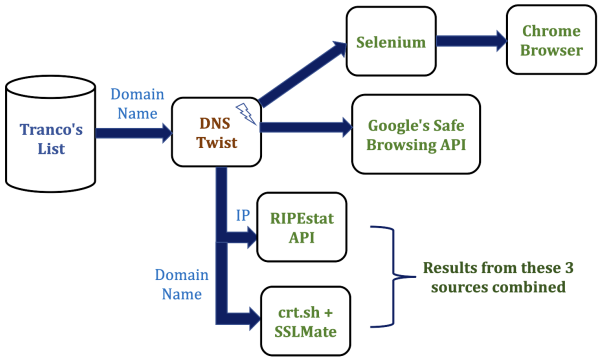

Figure 1. Measurement Workflow

domains were either unreachable or the local DNS server failed to resolve it. Throughout the measurement process, we encountered two common errors, namely HTTP 403 and HTTP 404 errors, which signify that the webpage is not accessible. Additionally, unresolved domain errors were encountered, whereby the domain in question does not exist or the Regional Internet Registry (RIR) does not support registering domains that contain certain special Unicode characters. As we had generated look-alike domains that contained homographs, this resulted in the presence of non-allowed characters. Figure 3 provides an overall view of Chrome's measurement results. It is noteworthy that approximately 45% of the generated look-alike domain names are available for purchase by regular users. This highlights the ease with which a phishing attack could be carried out against well-known websites, and we posit that this may also be true for less popular websites. We further explain the infrastructure behind available-to-(not)buy domains in the following section. The findings revealed that approximately 3000 domains were successfully resolved, indicating the ease with which attackers can launch phishing attacks using domain impersonation techniques. These results underscore the need for robust security measures to detect and prevent such attacks, as they can lead to severe consequences such as financial losses and reputational damage.

4.1.1 Available-to-buy domains: Out of all the generated twisted domains, around 45% of domains were available to purchase. Figure 4 illustrates the infrastructure behind the domains that are available to buy among the resolved twisted domains. Of the available domains, approximately 72% are served on HTTP webpages, while the remaining domains are served on HTTPS webpages. When accessing these domains using a browser application, warnings are shown for only approximately 1% of the domains. This suggests that it is relatively easy for an individual to purchase a similar-looking domain without being warned by the Chrome browser, thereby increasing the surface area for launching attacks.

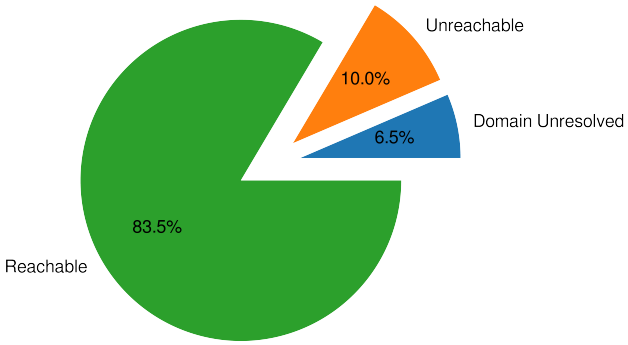

Figure 2. Twisted Domain Resolution Status

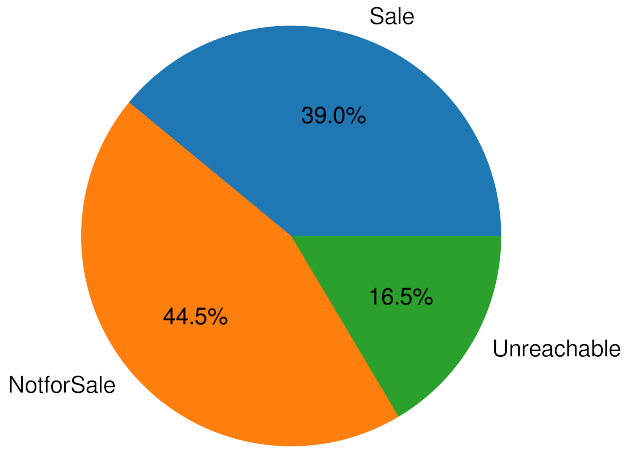

Figure 3. Overview of Chrome Browser Study

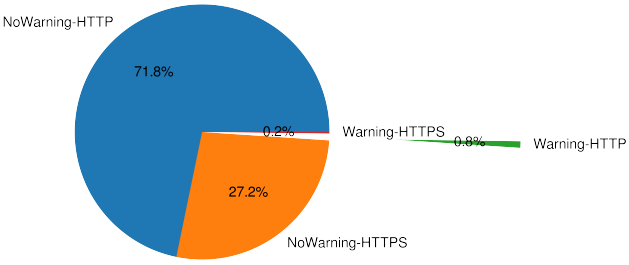

Figure 4. Available-to-buy domains

4.1.2 Not available-to-buy domains: Some domains are not available to purchase because the parent domain owner owns closely related domains or someone else has already purchased them for other user services. Even with the availability of TLS, a significant proportion (around 39%) of these closely related domains still operate on HTTP connections as shown in Figure 5. During automated crawling using Selenium, malware downloads were detected without explicit permission, highlighting the potential security risks associated with these domains. Around 6% of all the domains showed warnings on the Chrome browser user application. To ensure accuracy, the results were manually verified to eliminate any false positives or negatives.

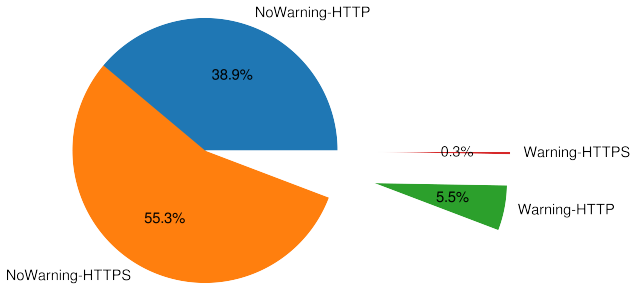

Figure 5. Not-available-to-buy domains

4.2 Google's Safe Browsing API

The study produced a surprising result regarding the effectiveness of Chrome's warning system and the Safe Browsing API in detecting potentially malicious domains. Out of the approximately 3000 domains that were successfully resolved, Chrome displayed warnings for only around 70 domains. In contrast, the Safe Browsing API only showed warnings for 6-7 domains, indicating a significant gap in the detection capabilities of the two systems. This led to the inference that the open-source Safe Browsing API may not expose updated information, thus limiting its ability to detect and prevent domain impersonation attacks. Consequently, we think that Google uses a multi-modal approach instead of relying solely on Safe Browsing API to display warning messages in the browser application.

4.3 RPKI status and ASN distribution

From the generated list of twisted domains, 57% had an RPKI status of valid, 42% had a status of unknown and 4 of them were of invalid length. When looking at the subset of domains which had unexpired TLS certificates, 65% had a status of valid and 34.7% had a status of unknown. The relatively high percentage of domains being covered by RPKI successfully indicates that RPKI is not a good indicator to filter out the look-alike domains.

Table 1 outlines the popular ASNs encountered in the list of domains obtained from DNSTwist. Popular providers like Google, Amazon, and Cloudflare are no surprises on the list.

Table 1. Popular ASNs encountered

ASN	Count	Owner
16509	483	Amazon
6461	226	Zayo Group
206834	172	Team Internet (Germany)
13335	144	Cloudflare
396982	118	Google
133618	108	Trellian Pty. Limited (Australia)
14618	108	Amazon

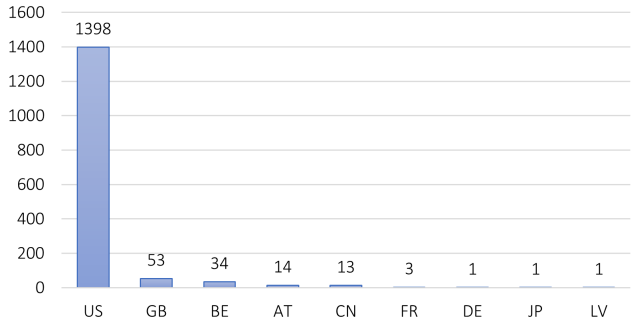

Figure 6. Country Distribution of TLS certificates

It is interesting to see marketing services like Team Internet and Trellian Pty. Limited (which offers services related to targeted domain redirect traffic) on this list. Domain redirect traffic for advertisements [4] typically involves a group that owns domains for misspelt URLs of popular websites. When a user mistakenly types this URL, their profile is forwarded to advertisement agents which bid in real time to show their ad to the specific user.

4.4 TLS Certificates

From DNSTwist's list of permuted domain names, 1519, or 53% of the domains were found to have TLS certificate information.

Figure 6 shows the country wise distribution of the certificates. Nearly 92% are located in the US. This finding also agrees with the RIR registration information of the list of IP addresses for the domains, wherein 68% of the prefixes were delegated by American Registry for Internet Numbers (ARIN).

When looking at the Certificate Authority (CA) issuing these certificates (refer Figure 7), it is found that the majority (64%) are issued by Let's Encrypt and 16% are issued by DigiCert. The popularity of Let's Encrypt is explained by the fact that it allows TLS certificates to be issued for free in an open, automated manner with the help of the Automated Certificate Management Environment (ACME) protocol for domain validation. DigiCert's share is not surprising as it is ranked

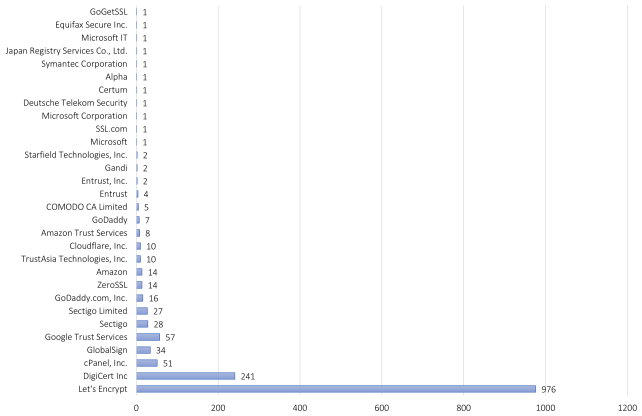

Figure 7. CA distribution

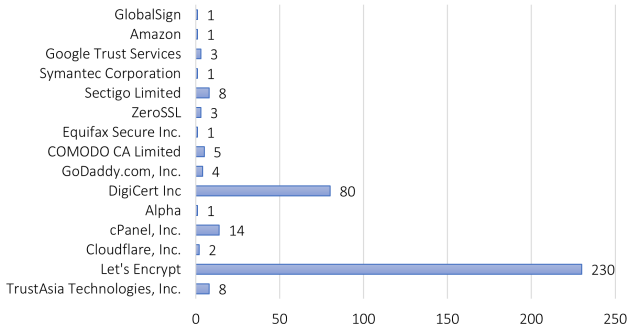

Figure 8. CA distribution for expired certificates

second in terms of market share amongst certificate authorities [3]. Interestingly, we did not find any certificates issued by IdenTrust, which occupies 53.6% of the market share.

Of all the certificates, only 22% were expired or revoked. Figure 8 shows the distribution of the expired certificates with respect to the certificate authorities. This means that most look-alike domains still have valid certificates (all issued in 2022 or later).

5 Discussion

Out of all the twisted domains corresponding to the randomly sampled 150 domains from the top 1500 domains in the Tranco list, approximately 45% of look-alike domains are available to host a website. This finding is concerning, as it indicates that adversaries could potentially exploit these domains to launch phishing attacks or other forms of malicious content that could harm unsuspecting users. Additionally, there is a disparity between the warnings shown by the Chrome browser application and the Safe Browsing API. There are two possible explanations for this disparity. Firstly, it is possible that the Safe Browsing API is outdated and does not provide recent results. Secondly, instead of relying solely on the Safe Browsing API, the Chrome browser application

uses a multi-modal approach to warn users about potentially malicious domains.

We saw over half the domains in our dataset have TLS certificate information which tells us that it is relatively common for impersonated domains to have TLS certificates. This is made easy with the help of certificate authorities like Let's Encrypt which offer free automated services to generate certificates for domains. This is validated by the high presence (64%) of certificates issued by Let's Encrypt in the dataset used. The fact that most of the certificates are still valid reaffirms the real concern of HTTPS phishing or spoofing.

From the perspective of RPKI, we do not see any correlation between impersonated domains and lack of RPKI configuration, as most of the domains were valid with respect to RPKI. Thus, RPKI cannot be used an indicator to spot these look-alike domains.

The distribution of ASNs reveals the trend of impersonated domains being used for advertisements.

6 Limitations

It is important to note that there are several limitations to our research work that need to be taken into consideration. Firstly, the usage of a small dataset may limit the generalizability of our findings to a larger population of popular domains. We used 150 domains from the Tranco list (dated 2^{nd} April) of a million domains, which may not represent the diversity of the internet as a whole. Therefore, caution must be exercised when interpreting our results.

Secondly, we were restricted to using Firefox and Chrome for our analysis, which may limit the applicability of our findings to other web browsers. Other browsers may have different security mechanisms and protocols that may affect the effectiveness of our proposed solutions.

Thirdly, we used two datasets (crt.sh and SSLMate) for collecting certificate information. This may have led to some inconsistencies as we did not have access to Censys, which is another popular source of certificate information. Therefore, our findings may not fully reflect the current state of TLS certificate adoption and revocation.

Lastly, we did not factor in domain reputation when analyzing the security of popular domains. This includes factors such as the presence of domains in spam filters or blocklists, which may affect the likelihood of a domain being used for malicious purposes. Therefore, our findings may not fully capture the security risks associated with popular domains on the internet.

Despite these limitations, we believe that our research work provides valuable insights into the security of popular domains on the internet and identifies potential areas for improvement. Future studies may want to address these limitations and expand on our findings to further enhance our understanding of internet security.

7 Conclusion

Our research addresses the security risks associated with domain impersonation attacks and evaluates the effectiveness of current security measures in detecting and preventing such attacks. Our findings indicate that relying solely on browser-based security mechanisms, such as Google's Warning messages and Safe Browsing API, has limitations in identifying potentially malicious websites. In the current state, we believe that Google may use additional sources beyond the Safe Browsing API to detect potentially harmful websites in their browser applications. Additionally, it is worth noting that attackers can easily obtain certificates for domains with a similar appearance from free certificate authorities like Let's Encrypt.

To assist domain proprietors in identifying potentially malicious domains with a similar appearance, we believe that a potential solution could combine publicly available information using DNSTwist and CT Monitors. With this approach, domain owners can receive alerts when a new certificate is detected for websites that resemble their domain. This solution can mitigate the risks associated with domain impersonation attacks and enhance internet security for popular domains.

In conclusion, our research contributes to a better understanding of the security risks associated with popular internet domains and proposes solutions to improve security measures. By addressing the limitations of existing security mechanisms and understanding the current context pertaining to domain impersonation, we believe our work could help in making a more secure and trustworthy Internet for all users.

Acknowledgments

The authors would like to express their gratitude to the entire SII class students for generously providing improvement and suggestions pertaining to our research study. The questions asked and insights obtained from the class discussions were essential for our study, and we are grateful for their support and cooperation. We would also like to thank Prof. Cecilia Testart, and Amanda Hsu for their helpful feedback. Their expertise, guidance, and resources were invaluable for this project.

References

[1] 2023. DNSTwist. https://github.com/elceef/dnstwist. Accessed: May 1, 2023.

[2] 2023. RIPESTAT data API. https://stat.ripe.net/docs/02.data-api/ Accessed: May 1, 2023.

[3] 2023. Usage statistics of SSL certificate authorities for websites. https://w3techs.com/technologies/overview/ssl_certificate

[4] Kinga Gawron. 2022. The Ultimate Guide to Domain Redirect Advertising in 2022. https://zeropark.com/blog/ultimate-guide-domain-redirect-traffic/

[5] Hang Hu, Steve T.K. Jan, Yang Wang, and Gang Wang. 2021. Assessing Browser-level Defense against IDN-based Phishing. In *30th USENIX*

Security Symposium (USENIX Security 21). USENIX Association, 3739–3756. https://www.usenix.org/conference/usenixsecurity21/presentation/hu-hang

[6] Doowon Kim, Haehyun Cho, Yonghwi Kwon, Adam Doupé, Sooel Son, Gail-Joon Ahn, and Tudor Dumitras. 2021. Security Analysis on Practices of Certificate Authorities in the HTTPS Phishing Ecosystem *(ASIA CCS '21)*. Association for Computing Machinery, New York, NY, USA, 407–420. https://doi.org/10.1145/3433210.3453100

[7] Platon Kotzias, Abbas Razaghpanah, Johanna Amann, Kenneth G. Paterson, Narseo Vallina-Rodriguez, and Juan Caballero. 2018. Coming of Age: A Longitudinal Study of TLS Deployment. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 415–428. https://doi.org/10.1145/3278532.3278568

[8] Ben Laurie. 2014. Certificate Transparency. *Commun. ACM* 57, 10 (sep 2014), 40–46. https://doi.org/10.1145/2659897

[9] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society. https://doi.org/10.14722/ndss.2019.23386

[10] Richard Roberts, Yaelle Goldschlag, Rachel Walter, Taejoong Chung, Alan Mislove, and Dave Levin. 2019. You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2489–2504. https://doi.org/10.1145/3319535.3363188