
# Certified Phishing: Taking a Look at Public Key Certificates of Phishing Websites

**Vincent Drury and Ulrike Meyer,** ***Department of Computer Science, RWTH Aachen University***

https://www.usenix.org/conference/soups2019/presentation/drury

**This paper is included in the Proceedings of the Fifteenth Symposium on Usable Privacy and Security.**

**August 12–13, 2019 • Santa Clara, CA, USA**

ISBN 978-1-939133-05-2

**Open access to the Proceedings of the Fifteenth Symposium on Usable Privacy and Security is sponsored by USENIX.**

# Certified Phishing: Taking a Look at Public Key Certificates of Phishing Websites

Vincent Drury
*Department of Computer Science*
*RWTH Aachen University*
*drury@itsec.rwth-aachen.de*

Ulrike Meyer
*Department of Computer Science*
*RWTH Aachen University*
*meyer@itsec.rwth-aachen.de*

## Abstract

The share of phishing websites using HTTPS has been constantly increasing over the last years. As a consequence, the simple user advice to check whether a website is HTTPS-protected is no longer effective against phishing. At the same time, the use of certificates in the context of phishing raises the question if the information contained in them could be used to detect phishing websites. In this paper we take a first step towards answering this question. To this end, we analyze almost 10 000 valid certificates queried from phishing websites and compare them to almost 40 000 certificates collected from benign sites. Our analysis shows that it is generally impossible to differentiate between benign sites and phishing sites based on the content of their certificates alone. However, we present empirical evidence that current phishing websites for popular targets do typically not replicate the issuer and subject information.

## 1 Introduction

Phishing is still an important and direct risk to many Internet users. The Anti-Phishing Working Group (APWG) recorded more than 50 000 unique phishing websites in September 2018, a number that has been relatively stable for the last year [19]. These websites follow the general trend of the Web [17] in that they are steadily adopting the usage of HTTPS: 49.4 % of the phishing sites were using SSL/TLS in the third quarter of 2018, up from less than 5 % in 2016. This rapid development, in conjunction with the availability of easy-to-obtain certificates, has already led to changes in browser design (e.g., Google Chrome aims to remove some positive security indicators [6]) and will render the general advice to "look for the lock icon" to detect phishing less and less effective.

There are mainly two lines of work aiming at protecting against phishing from two different directions: technical solutions and user educational approaches. These two lines complement one another rather than competing against each other as the former addresses the technical component while the latter addresses the social engineering aspect of phishing. The technical approaches include reactive measures, like blacklists, as well as preventive measures, like heuristic and machine learning-assisted detection approaches. Educational endeavors on the other hand focus on users, and try to improve their phishing detection and prevention abilities as soon as technical solutions fail. Previous educational efforts (e.g., [3, 22, 33]) mainly focus on noticing the absence of HTTPS usage or detecting suspicious URLs as indicators for phishing. With the rise of HTTPS-hosted phishing sites, the usage of HTTPS is no longer a strong indication for a benign site. However, certificates used by phishing sites are now a new potential source of information that might be useful in the context of automatic or user-based phishing detection.

Whether or not certificate information can be used in this context depends on the answers to the following open issues: First, it depends on whether there are differences between certificates of benign websites and phishing websites in the first place. Second, it depends on whether these differences are robust, i.e., whether it is safe to assume that they persist even if adversaries try to actively reduce these differences. Third, even if there are robust differences, it remains an open question, whether these differences can effectively be used to discriminate between phishing and benign certificates as part of a technical solution and/or whether these differences can be exposed to a user in a way such that they increase the user's ability to detect a phishing website.

In this paper, we focus on addressing the first of these issues and briefly touch the second one. The third of the above issues is not addressed in this paper.


*USENIX Symposium on Usable Privacy and Security (SOUPS)* 2019.
August 11–13, 2019, Santa Clara, CA, USA.

More specifically, we empirically investigate the following two research questions:

1. Are there general differences between the certificates of phishing websites compared to those of benign websites and if so which ones?

2. Are there differences between the certificate of a phishing website and the certificate of the corresponding targeted benign website and if so which ones?

To answer these questions, we collect and compare 9 479 certificates from 31 264 phishing websites and 39 478 certificates from 50 000 benign websites.

We find no obvious differences between phishing and benign websites in general. However, we find that the phishing certificates of the 15 most popular phishing target's websites do currently differ from their benign counterparts in particular with respect to the issuer and subject information provided in the certificate.

We also identify the threat of hosting services, that make it easy for attackers to present a valid certificate that looks very similar to the target's certificate, in particular if the target is the hoster itself. Even users that have knowledge of URLs, may fall victim to such an attacker as the fake website is hosted on a legitimate domain.

The rest of this paper is structured as follows: The next section introduces some preliminary terms and concepts. Section 3 presents related work on the topics of phishing education. Section 4 details the certificate collection process, as well as the results of our analysis. Section 5 takes a look at the representation of certificate information in browsers. Finally, we conclude with a summary and future work.

## 2 Preliminaries

In this section we present some basic concepts and terms that will be used throughout this paper.

### 2.1 Website-based phishing attacks

In this paper, we look at the following website-based type of phishing attack, that we will generally refer to as phishing: The **attacker** clones the website of a **target** and sends a link to a user of the target, the **victim**. In particular, we do not restrict the transmission channel to email, other methods might also be possible. The victim then clicks on the link and opens the attacker's (fake) website and interacts with the phishing website as if it was the website of the target. This interaction will typically include entering the victim's username and password information into the fake website, thus enabling the attacker to impersonate the user to the target in the future. We are not concerned about specifics of the attack (e.g., circumvention of two-factor authentication, website cloning techniques, etc.), as long as a fake website is involved.

| Field | |
|---|---|
| Subject: | `Common Name` (`CN`) |
| | `Organization` (`O`) |
| | `Organizational Unit` (`OU`) |
| | `Locality` (`L`) |
| | `Country of Residence` (`C`) |
| | `Business Category` |
| Issuer: | `Common Name` (`CN`) |
| | `Organization` (`O`) |
| | `Country of Residence` (`C`) |
| | `Valid From` |
| | `Valid To` |
| Extensions: | `Subject Alternative Name` (`SAN`) |
| | `Certificate Policies` |

Table 1: Certificate fields and shortnames used in this paper.

### 2.2 HTTPS and public key certificates

HTTPS allows web servers to authenticate themselves to users based on X.509 certificates using TLS [31]. Such certificates are issued by Certification Authorities (CAs) and bind the public key of a web server to the identity of the web server. Table 1 illustrates how the identity of the web server and its issuing CA are represented in an X.509 certificate and what other fields included in a certificate are of interest in the context of this paper. Note that the web server's domain name is included in the certificate either in the subject `CN` field or as an entry in the `SAN` extension field.

The CAs used on the web today are ordered in a hierarchy, where CAs on higher levels issue certificates for CAs on lower levels, and the CAs at the lowest level issue certificates for the individual web servers. The certificates of the CAs at the highest level, the root CAs' certificates, are shipped in web browsers and are thus readily available on the client side. When a client connects to the web server with HTTPS, the web server presents a chain of certificates to the client and the client can validate the certificates in the chain, starting with checking that the last certificate in the chain was issued by one of the pre-established root CAs and thus obtaining a public key to check the next certificate in the chain. While certificates certainly help in validating public keys, the mere fact that a website is able to present a valid chain of certificates is not a guarantee that the website itself is trustworthy as CAs may follow different policies while issuing certificates. Thus, it is possible that a request for a certificate, e.g., for an intentionally misleading domain name, is indeed signed by a CA if the policy used by the CA to validate the identity of the requester is rather lax.

### 2.3 Types of Validation

There are several levels of vetting a CA can perform before signing a certificate for a subject, that can also influence the

amount of information included in the certificate. These validation types represent different levels of trust or effort by the CAs and are briefly introduced in the following. We use the CA/Browser Forum's (CAB) guidelines as reference for the different validation levels [4].

According to these guidelines, all CAs have to ensure certain qualities regardless of the type of validation, that include basic employee vetting as well as logging and auditing requirements. The CAs also have to ensure that all information that is included in a certificate was verified, taking reasonable steps to ensure correctness. In the context of phishing, it is worth mentioning that CAs are required to maintain a database of "high-risk" names, that are at risk for phishing or other fraudulent usage. This database has to be checked for each certificate that is issued, and if a high-risk name is found, additional scrutiny on the part of the CA is expected to make sure that the certificate is issued to a valid entity. There are, however, no specific requirements on how "high-risk" names are to be handled in the CAB documents.

#### Domain Validation

Domain Validation (DV) is the most basic form of validation. Here, the CA only checks that the Certificate Signing Request is valid and that the subject has control over the domain in question (indicated in the `CN` or `SAN` field of the certificate). This might include a challenge, e.g., setting a specific DNS entry or uploading a file with some predefined content. Since no further review is required to validate control over a domain, this process can be automated, e.g., as is the case with the CA "Let's Encrypt" [23].

#### Organization Validation

Certificates where the CA has asserted the validity of the subject's organization identity are called Organization Validated (OV) certificates. In the CAB documents [4], this requires more rigorous validation of the subject, beyond simple control of the domain. Verification of the organization identity means, that the issuing CA has to verify name and address of the organization entity, e.g., via consulting the government agency in the jurisdiction of the organization, or a site visit. Additionally, the CA has to verify the authenticity of the certificate applicant, e.g., via a reliable method of communication.

As a result of the organization validation, the CA is able to add organization information (i.e., the `O`, `OU`, `C`, `L` fields) to the certificate. They might also include the CAB policy ID for OV certificates (2.23.140.1.2.2) in the `Certificate Policies` field of the certificate, and must then include the subject field `O` as well as location information (i.e., country and state or province).

#### Extended Validation

The most thorough validation level is called Extended Validation (EV) and is used to validate the legal entity that controls a website [5]. Preventing phishing is explicitly mentioned as a secondary purpose, a consequence of the more reliable information included in the certificate. The main difference to OV certificates is, that the process for issuing EV certificates is defined in much more detail and adds some additional requirements. In theory, a CA could issue a non-EV certificate using the EV validation processes.

The documents include, among others, detailed requirements for certificate requests. For EV certificates, the certificate applicant has to name several contact people, who have to fulfill certain roles (certificate requester, certificate approver, certificate signer, applicant representative), all of which have to be authenticated by the CA. The CA also has to verify the organization's legal, physical and operational existence, verify the authority of all roles of requesters and ensure reliable means of communication in addition to verifying domain control. The guidelines also introduce additional constraints to EV certificates, including the prohibition of wildcard certificates. CAs will also have to look out for high risk certificates, that include websites with the risk of fraud (e.g., websites with an Internationalized Domain Name (IDN) [21] that looks similar to an existing business). An EV certificate must include several fields:

- The subject organization name.
- The subject business category (e.g., private organization or government entity).
- The subject jurisdiction of incorporation or registration.
- The subject registration number (identifying the subject in the registration agency at the jurisdiction of the subject).
- An EV policy identifier that confirms the CA's compliance to the CAB EV documents. This can be specific to each CA.

All steps of the issuance process have to be documented and reviewed before granting the certificate request, all discrepancies have to be resolved. In particular, no single person must be able to grant an EV certificate, corresponding control procedures have to be enforced. The CA's employees have to be trained and their trustworthiness ensured via background checks (e.g., employment history, professional references, education, criminal history).

A client, for example a browser, checking the validity of an EV certificate, has to check for the corresponding policies in the certificate and confirm, that the issuing CA is valid and known to adhere to the EV guidelines.

## 3 Related Work

In our work, we investigate at a large scale whether and if so how certificates of benign and phishing websites differ. As such, it is related to several fields of study. In the following, we will look at previous work in phishing user studies, educational and technical approaches to prevent phishing, as well as browser evaluations regarding the presentation of certificate information and validation level.

Phishing is an attack that directly targets users, such that several **user studies** have set out to understand how and why it works. Phishing has been shown to be effective, even if users are primed to look for it, and even if they have technical knowledge [2, 32]. In 2006, Dhamija et al. published the results of a user study to find out, why users are susceptible to phishing [9]. They find, that users generally focus on the body of a website to decide if it is legitimate, ignoring more robust indicators like the URL. More recently, these results are confirmed by Alsharnouby et al., who track eye movement of users and find that they do not spend much time looking at security indicators. Less than 15% of the time is spent looking at browser UI, only about 6% is spent focusing on "areas of interest" like the URL bar or lock icon. The authors do however find, that browser indicators can be very helpful: detection correlates to focus on browser UI [2]. Downs et al. look at detection strategies and their effectiveness, especially for phishing emails [13]. They find, that knowledge about cues and past experience is not enough to reliably detect phishing.

Consequently, to get users to behave more securely, researchers have designed and evaluated several **educational approaches**. For example, Kumaraguru et al. conducted a large-scale (>500 participants) study that shows, that phishing education using PhishGuru, an embedded training method, can be effective and even have long-term benefits [22]. To create an engaging and immersive experience, researchers have also created and evaluated learning games to teach phishing detection. Sheng et al., with Anti-Phishing Phil in 2007, identified the problem that the browser UI is largely ignored in favor of the website body and try to teach users to understand and focus on URLs [33]. Arachchilage et al. design and evaluate a mobile game to prevent phishing [3]. Similarly to Phishing Phil, the game focuses on URLs. They show, that participants were motivated and improved their test scores after playing the game. These games mainly focus on URLs as indicators for phishing while, to the best of our knowledge, certificates have not been evaluated for user education so far.

User education as an approach has been shown to be somewhat successful, but no "perfect" results have been achieved. This leads to a different research direction, that focuses on automated **technical approaches** to phishing prevention to support and complement user efforts. A widely represented approach are blacklists, that maintain lists of known phishing websites and prevent users from opening them. These lists can be successful to prevent the spreading of known attacks but leave a window of opportunity to attackers until the malicious website is added to the list and distributed to users [30, 36]. Therefore, other techniques were developed that include more proactive approaches. These are generally better at finding unknown phishing, but can have false positives and are not as widely used (e.g., integrated into browsers like Google Safe Browsing [18]). Here, Dou et al. compiled a list of approaches and their effectiveness [12]. Recently, machine-learning-based approaches to classify websites as phishing or benign based on features extracted from certificates have been proposed (e.g., [11, 24, 35]). Specifically, Dong et al. use and compare several machine learning algorithms to classify phishing websites [11]. They extract several features including information on the validity period and relation between subject and issuer fields. The best approach achieves a precision of more than 95%. Other machine learning approaches that focus on certificates, like the one by Torroledo et al., include features like the existence of several subject fields and validation levels [35]. Mensah et al. try to classify phishing and benign websites using features extracted from certificates and handshake information but conclude, that it is not possible to discriminate the two using only this type of information [24]. We come to a similar conclusion in that there are no general differences between certificates of benign and phishing websites. However, we go one step further by directly comparing the certificate of a phishing website to its target's certificate.

Even though these tools perform quite well (especially when compared to humans), there seems to be no solution employing these techniques widely, possibly due to the still rather high false positive rates. Unfortunately, the positive results do not translate well to user education: Not only do some features require complex computations to evaluate, but the classification process itself is also not applicable to users. In this paper, we extend the domain knowledge required to create effective classifiers by evaluating and arguing about certificate information as potential features.

Lastly, taking a look at **browser evaluation**, Biddle et al. set out to understand users' perceptions of the trustworthiness of a website when looking at certificates. They start with the assumption that users do not really understand certificates, and that the browser UI does therefore not help them make informed decisions. As such, they create an alternative UI for different validation levels and evaluate it in a user study. They find, that users' understanding of the original UI greatly varies, and that users do on average understand the level of trust a certificate provides better when using the proposed UI. Similarly, Sobey et al. also propose an alternative indicator for EV certificates [34]. They find using eye-tracking technology, that users did not notice the original EV indicators at all. However, the UI of Firefox has changed since then, making this information much more accessible (see Section 5). In this paper we analyze whether the certificate information relevant in the context of this paper is available to users in the browser

UIs and which steps users have to perform to get to this information.

## 4 Certificate Collection

This section describes the process and results of our certificate collection efforts in detail. The main goals are to answer research questions (1) and (2) as described in Section 1. We therefore look at general differences between the certificates of phishing and benign websites, as well as differences between the certificates of popular targets and their corresponding phishing websites. To achieve our goals, we collect certificate information from benign and phishing websites, extract features, and compare phishing and benign certificates.

### 4.1 Data Collection

For our analysis we retrieved 39 478 benign and 9 479 phishing certificates. In the following, we first describe how we collected benign and phishing domains and then describe, how we retrieved certificates from these domains.

#### 4.1.1 Data sources and preprocessing

In order to collect popular benign domains, we used the Alexa Top million list [1] and crawled the top 50 000 entries. Unfortunately, the Alexa data set does not include subdomains and for some domains, querying the domain without subdomains leads to a result that is different from querying the domain with its subdomains. A prominent example for this is PayPal, the most popular target for phishing campaigns in our data set. In this case, querying "`paypal.com`" leads to a certificate that differs from the one returned when querying "`www.paypal.com`". In order to mirror the experience of users more closely, we therefore apply a preprocessing step and query all benign websites using curl [7] to follow auto-redirects. We then use the resulting domain names for all further steps.

The phishing data set was obtained from Phishtank [29], a website that collects phishing websites collaboratively. Users can submit potential phishing websites and verify others, resulting in a peer-reviewed data set of phishing websites. However, this data set is not completely free of false positives: We did encounter some false positives when looking at specific certificates. We assume that this is due to one of the following reasons:

- The websites has been cleared and phishing content removed, but is still shown as "online and valid" by Phishtank.
- The websites were falsely flagged and the verification of users was wrong.

Either way, these cases seem to be rare in comparison to the data set of true phishing websites (we found less than ten cases in our detailed analysis in Section 4.2.2). We queried the Phishtank database for online and valid (i.e., verified by other users) phishing websites once daily over a period of 54 days (one day was missed due to technical problems). In this time, we collected 31 264 unique Phishtank entries.

#### 4.1.2 Certificate Collection

We use the following process to retrieve certificates from benign and phishing websites: First, we obtain the data sets for phishing and benign websites, using or converting to JSON representations of the data. For phishing websites, since we do not want to download certificates that have already been considered on a previous day, we merge the new data sets with a list of previously visited websites. Thus, we reduce the queried websites from several thousands to several hundred new phishing domains per day. This is not necessary for the larger benign data set, as these domains can be queried all at once.

After acquiring the websites to be queried, we start the crawling process using OpenSSL [28]. OpenSSL is an open source toolkit for the TLS and SSL protocols. We use the *s_client* component of OpenSSL to query websites and get certificate information [27]. The version of the program is "OpenSSL 1.1.1a FIPS 20 Nov 2018", as root certificates we utilize the Mozilla CA Certificate Store, which is, among others, also used by the Firefox browser [25]. We use *s_client* to connect to the specified domains on port 443 and retrieve a certificate, if possible. The certificates and additional information about the connection are saved on success for further analysis.

All in all, we were able to obtain 25 777 certificates from the 31 264 phishing domains. From these, we removed 11 712 duplicate certificates with respect to domain names in order to avoid polluting our data set with several entries for a single phishing campaign. To be precise, we create a database that only contains unique domain names and for each domain name exactly one certificate. This results in 14 065 certificates, but introduces a bias in our dataset, which now includes phishing campaigns using different subdomains, but disregards campaigns using different URL paths. Note that not all of the remaining certificates are unique. It is still possible that several different domain names are included in the same certificate. Next, we also decided only to look at certificates that are valid (as recognized by OpenSSL), since browsing to websites with invalid certificates generates a visible error in all major browsers to warn users. An overview of the validity status of phishing and benign certificates can be seen in Figure 1. *Name mismatch* errors (the domain name of the website does not match the subject `CN` or `SAN` of the certificate) were the most common, followed by *expired* certificates (validity period is in the past) and *self-signed* certificates. Overall,

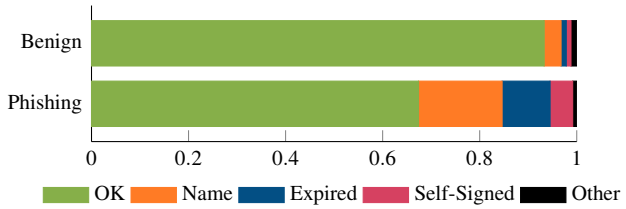


Figure 1: The validty status of certificates from phishing and benign websites.

| Field | |
|---|---|
| Subject: | `CN` |
| | `Organization` |
| CA: | Issuer `CN` |
| | Root `CN` |
| Validity: | Validity Period |
| | isValid |
| Extensions: | `SAN` |
| | Extended Validation |

Table 2: Features selected for further analysis.

phishing websites are more likely to present an invalid certificate than benign websites. Our final data set of valid phishing certificates contains 9 479 entries.

For benign websites, we remove 698 certificates with duplicate domain names and 2 842 invalid certificates and end up with a data set containing 39 478 benign certificates.

#### 4.1.3 Analysis and feature extraction

The analysis starts in a second pass, after all certificates are downloaded. Here, we scan all certificates, extract features of interest (see Table 2) and save them in a database. The features are divided into three groups:

- **Subject Information**: This group contains the subject `Organization` as well as validity and EV information. These are usually easily available to users and directly correspond to the websites a user might expect to be on.

- **Issuance Information**: This group contains the issuer and root `CN`, as well as the validity period. These are features that go beyond subject information, but are still easily available to users (see Section 5). We use the `CN` of the issuer rather than the `O` information of the issuer, as it is usually more detailed in our dataset.

- **URL information**: This group includes the subject `CN` and `SAN`, as well as the domain name of the website in question. We include this information to determine if looking at the certificate can be more effective than looking at the URL of a phishing website.

We disregard other fields commonly found in certificates for several reasons: Some fields are very similar (or the same) for all certificates issued by the same issuer (e.g., signature algorithm, policies). Thus, they only differ for certificates issued by different issuers, i.e. a field we already consider. Other fields consist only of long strings of numbers, that would be impractical to deal with in the context of user education and are unlikely to be usable in the context of automated phishing detection as well (e.g., public key, serial number). Lastly, some fields simply do not offer much variation at all (e.g., key usage, basic constraints).

### 4.2 Results

In the following, we first analyze the differences between our benign and phishing certificate collections w.r.t. to the information described in the last section and thus address research question 1. We then take a closer look at the differences between certificates of phishing websites and the certificates of their targets and thus continue with research question 2.

#### 4.2.1 General information in phishing and benign certificates

To address the first research question, we look at the distribution of features for benign and phishing certificates in general and try to find out how well they separate benign and phishing websites.

As described in Section 2, some CAs offer different levels of validation. It stands to reason that the more complex types of validation, i.e., organization and extended validation, make it harder for attackers to present a corresponding certificate. Still, we found that 1 444 certificates, about 15 % of all phishing certificates, include an `Organization` in their subject fields. We use the subject `O` field to decide if a certificate is OV, assuming that CAs follow best practices and do not include unverified information in the certificates they issue. For benign certificates, 13 852 or about 35 % of websites have an `Organization` in their subject fields. We assume that this difference is particularly pronounced for the higher ranks in the Alexa list, as these companies, with high user counts, have more incentives to buy organization validation or extended validation certificates. Taking this distribution into account, organization validation is not a deciding factor for differentiating phishing and benign websites, and would lead to many false positives if it were to be used as such.

EV certificates on the other hand are a more interesting matter. To decide if a certificate is EV, we use the subject `business category` field (OID: 2.5.4.15). This field is a requirement for EV certificates [5], checking for it is therefore an over approximation if we assume compliance to the CAB documents. We found that this approximation works quite well for otherwise valid certificates (we did not find any false classifications, even after working with and randomly

sampling our data set several times). Using this method, we identified only 39 phishing websites, that is about 0.4 %, with a valid EV certificate. These consist of compromised servers, as well as websites that are abused to host malicious content (e.g., `dropbox.com`, `jsfiddle.net`, `medium.com`), but also include several false positives (e.g., `paypal-notice.com`). We assume that such domains are less useful when phishing for user credentials, as they prominently display a different company in the URL bar of several popular browsers, but are still used for scams and other types of deception. As such, it seems that extended validation is less likely to be available to phishing websites, even though possible (social engineering) attacks were demonstrated before (e.g., [20]). Still, it is much harder to correctly fake an existing organization, including business registration details, as required for extended validation. On the other hand, only about 7 % (2746) of the benign websites use an extended validation certificate. Even among the top ten ranks, none protect their landing page with an extended validation certificate. This shows that even though an EV certificate (if it is valid and has the correct organization displayed) can be a good indicator that a website is legitimate, it does not provide a robust method to detect phishing websites. We found that some OV and EV certificates are used for phishing in connection with services that allow users to host content on their platforms. This includes Tumblr, Dropbox, Heroku and Medium. The interesting part of this phenomenon is, that these organizations have at least organization validated certificates. As such, a user that expects to be on bankingsite.com might open the certificate, look at the `Organization` and realize they are in fact on somehostingsite.com, which might awake suspicion.

The most popular issuers for benign and phishing websites are shown in Table 3. Again, we do not find any distinct features for phishing: the 10 most popular issuers, making up for 8598($\approx$ 90.7 %) of all phishing certificates, are also popular among benign websites (26046 certificates $\approx$ 66 %). As such, issuer information alone is not enough to separate benign from phishing domains. More detailed numbers for popular issuers for benign and phishing websites can be found in Tables 6 and 7 in Appendix A.

Similar to the issuers, we also find slight differences in other certificate details. The validity period for benign websites is on average longer than that of phishing websites (about 252 days for phishing and about 412 days for benign websites). We assume this is mainly due to the distribution of issuers: phishing websites more often use issuers with short validity periods like "Let's Encrypt" (90 days on average for both phishing and benign) and "cPanel"(average validity period of about 93 days for phishing, about 98 days for benign).

All in all, we do not find simple indicators for whether a certificate originates from a benign website or a phishing website. Attackers that set up their own websites have restrictions similar to benign administrators, resulting in similar choices for issuers and in similar certificates. Even though we found that phishing certificates often do not include an organization in the respective field, we found that this is also the case for many benign websites. The similarity in certificates is even more prominent if a benign website is used (compromised or not) to host an attacker's content.

| Issuer `CN` | Phishing | Benign |
|---|---|---|
| Let's Encrypt Authority X3 | 34.4 % | 17.4 % |
| cPanel, Inc. Certification Authority | 22.2 % | 1.6 % |
| RapidSSL TLS RSA CA G1 | 9.1 % | 0.2 % |
| COMODO RSA Domain Validation Secure Server CA | 5.3 % | 10.2 % |
| COMODO ECC Domain Validation Secure Server CA 2 | 5.2 % | 18.2 % |
| CloudFlare Inc ECC CA-2 | 5.0 % | 6.5 % |
| DigiCert SHA2 Secure Server CA | 3.4 % | 4.4 % |
| Go Daddy Secure Certificate Authority - G2 | 2.9 % | 4.4 % |
| Google Internet Authority G3 | 2.0 % | 0.5 % |
| RapidSSL RSA CA 2018 | 1.4 % | 2.6 % |

Table 3: Percentages of benign and phishing certificates issued by the 10 most popular issuers of phishing certificates.

### 4.2.2 Popular target websites

Next, we try to answer research question 2, i.e., the question whether the certificates of phishing websites differ significantly when comparing them to their target's certificate. For this, we look at the 15 most popular target websites of phishing attacks and their certificates (covering 2771 of 3275 valid phishing websites with a target label in the Phishtank database), and try to find out if and how well the phishing attacks are able to mimic their targets' certificates. We start by determining the login pages for the targets and noting their certificate information. Then, we look at the phishing data set and compare the target certificates with the phishing websites imitating these targets. The full results can be found in Table 4. Note, that all entries greater than one indicate unique domain names, that might still host several phishing websites on different URL paths.

First, we look at target organizations, and find that only few phishing websites are able to fake this information. To determine organization similarity we use the Python `difflib.SequenceMatcher` class [10], and manually verify all matches with a ratio of more than 0.3. We found no evidence of any phishing website obtaining a certificate with a spoofed organization name (even beyond the targets listed in Table 4). All entries in the table with a similar organization are hosted on the target's own infrastructure. For example, Microsoft offers several cloud services (e.g., Azure, SharePoint

| Target | Domain name | Number of phishing websites | Similar Organization | Same Issuer | Similar Issuer | Target in URL DN | Target matches wildcard |
|---|---|---|---|---|---|---|---|
| PayPal | www.paypal.com | 1169 | 0 | 1 | 24 | 84 | 12 |
| Facebook | www.facebook.com | 571 | 0 | 4 | 221 | 32 | 31 |
| Microsoft | login.live.com | 297 | 47* | 0 | 58 | 10 | 9 |
| ABSA Bank | www.absa.co.za | 214 | 0 | 0 | 0 | 5 | 0 |
| RuneScape | secure.runescape.com | 87 | 0 | 0 | 1 | 74 | 0 |
| eBay | signin.ebay.com | 67 | 0 | 1 | 0 | 5 | 0 |
| MyEtherWallet | www.myetherwallet.com | 62 | 0 | 1 | 2 | 15 | 0 |
| Blockchain | www.blockchain.com | 46 | 0 | 1 | 1 | 0 | 0 |
| Allegro | allegro.pl | 44 | 0 | 0 | 0 | 35 | 0 |
| Apple | appleid.apple.com | 42 | 0 | 0 | 2 | 8 | 3 |
| Steam | store.steampowered.com | 39 | 0 | 0 | 0 | 6 | 0 |
| Dropbox | www.dropbox.com | 37 | 1* | 1* | 0 | 2 | 1 |
| Binance | www.binance.com | 34 | 0 | 0 | 0 | 3 | 1 |
| Google | accounts.google.com | 33 | 1*[a] | 1* | 0 | 1 | 0 |
| ASB Bank Limited | online.asb.co.nz | 29 | 0 | 0 | 0 | 4 | 0 |

Table 4: Certificate and URL similarities for popular phishing targets. False Positives we found were removed. Entries marked with an asterisk are hosted on the target's own infrastructure.

[a] No text input, refers to different website

and OneDrive), that allow users to host content on domains owned by Microsoft. These domains are protected by Microsoft's own certificates and therefore match the target's `Organization`. We will encounter and argue more about this type of attack later on in this section.

Next, we look at issuers and their similarities to the target websites. The column "similar issuer" lists the number of phishing websites with similar issuers, meaning the same CA organization (e.g., DigiCert High Assurance is similar to DigiCert Extended Validation). We find, that many popular targets have few or no exact matches for the issuing CAs of phishing websites. Disregarding false positives and misclassifications again, only seven targets' issuers were replicated by phishing websites, and these cases are very rare (only one case for six targets, four for Facebook). Still, issuers seem to be a less precise metric than organizations as described above. This is also supported by the fact that there are many phishing websites with a similar issuer. It is also notable that among the 15 most popular targets we analyzed in detail, 9 are using EV certificates for their login pages. These require a thorough investigation of the entity requesting the certificate (see Section 2.3), making it less likely that organization information is spoofed. Looking at the details for similar and identical issuers reveals an interesting finding: Most of these entries come from websites that host user content, protecting it with their own certificate. In many such cases, users might still be able to recognize that they are not on the website they expect if they look at organization information. However, this is not the case if an attacker targets the service they are hosting their website on. We found this to be the case for Microsoft, as well as Google and Dropbox. To prevent such attacks, user content could be protected with a different certificate from the one used to login. Logins might be preferably protected with EV certificates.

Lastly, we look at URL similarities. We label a URL as similar to its target if it contains the organization or original domain name. We found that there are often far more similar URLs than either organizations or issuers. As shown in previous user studies (e.g., in [33]), complex phishing URLs can be difficult to detect even for users that were previously educated on the subject of phishing URLs. Interestingly, we find that attackers seem to be able to add the target name to the domain name in many cases (see Table 4). Therefore, even though many browsers offer a reduction in complexity by only showing the domain name, this part can still lead to users mistaking a phishing site for a benign site. As an aside, our database did not include a single valid certificate for a URL consisting of an IP-address, making this type of URL obfuscation less relevant than before (e.g., [16, 26]).

## 4.3 Discussion of collection results

We found that, unsurprisingly, there are no straightforward features extractable from certificates that instantly separate certificates of phishing and benign websites. We therefore answer research question 1 in the negative, concluding that there are no general differences between the certificates of phishing and benign websites. On the other hand, we found

that phishing websites currently do not seem to recreate the information included in the certificates of popular targets. So, as for research question 2, we find that currently there are some differences between the certificates of a phishing website and the certificate of its target. We particularly find that to date, the subject `O` and issuer `CN` seem not to be actively replicated.

On the other hand, it remains an open research question, whether it would be possible to expose the differences we observed to users in a way such that it would help them to detect phishing websites. In addition, it is unclear, how these differences could be used in the context of automated phishing detection.

Furthermore, while some information is currently not replicated, it is an open question how robust these findings are, i.e., how difficult it would be for an attacker to replicate the information on its target's certificate. Since organization validated certificates do not require the same level of vetting that EV certificates do, it is possible that attackers might get a fraudulent OV certificate without the risk of compromising their operations. It is also possible that a CA is compromised or misses a spoofed or fake organization in a certificate request. Replicating the issuer of a website is generally less complicated, as it does not require the attacker to spoof any information. As such, we conclude that it is not a robust feature to consider when analyzing a website.

In our analysis of popular target websites, we found that phishing websites with certificates that are similar to their target's certificate are often using hosting services and are not self-hosted. If the user content on such hosting services is protected by a wildcard certificate that includes information about the hosting service, it might still be possible to recognize this type of attack looking at the certificate. However, this is not the case if the service provider itself is the target.

Another potential problem with certificates as source of information for any future phishing detection tool is, that the tool might have problems with false positives if websites change their certificates. This includes a change from one issuing CA to another, or a change in validation level, both of which are possible scenarios for an organization.

A further potential problem of using certificates to detect phishing is, that automated tools (or users) might not be able to retrieve the certificate for a given website if they are affected by TLS interception [8, 14]. Here, the tool (or user) would only be able to retrieve the certificates of its interception middleware regardless of the website that is visited, rendering the detection of malicious websites with the help of certificate information entirely impossible.

For the sake of completeness, we also include some considerations that might have influenced our collection process. Firstly, our queries are performed from our country of residence, which might have influenced the results. This is more likely the case for larger websites that use content distribution and serve different content to users from different countries. Both phishing and benign websites were likely influenced by this, as attackers can use larger services to host their websites (see Section 4.2.2). This bias, however, is hard to remove and still represents a large amount of users that would have been served similar results.

Secondly, it is possible that attackers notice the crawling efforts and blacklist our client at some point (e.g., [26]). In this case, we would no longer be able to capture some of the attackers' methods, which might include more sophisticated techniques. We currently do not have any indications of this being the case.

In the next section, we will look at how certificate information is presented in popular browsers.

## 5 Browser Evaluation

In this section, we look at the presentation of certificates in major current browsers, taking into account the results from the previous section. There we looked at organization and issuer information included in certificates, as well as different levels of validation. In the following, we will compare the UI of several browsers with respect to their certificate presentation.

### 5.1 Browser UIs

We look at five browsers that cover a wide range of users: Google Chrome (Desktop[1] and Mobile[2]), Mozilla Firefox (Desktop[3]), Microsoft Edge (Desktop[4]), and Safari (Mac[5]).

First, we look at the browser's URL bars. We find that none of the browsers make a distinction between OV and DV certificates. However, all but Chrome for Android have special UIs for EV certificates. The indication for EV certificates ranges from additional information displayed next to the lock icon to more prominent highlighting of the lock and URL.

Next we count how many clicks are needed to get to subject `Organization` and issuer information. This varies greatly between the browsers. A comparison for non-EV certificates is given in Table 5. All browsers will open a smaller window after clicking on the lock (e.g., Figure 2), that includes general information about the current page.

The next part is where the browsers start to differ more prominently. All of the browsers offer a certificate viewer, that contains an overview of the certificate of the current website (e.g., Figure 3). We find that the subject `O` information is first available in the certificate viewer for all browsers we tested. Reaching this information takes different amounts of user input for the different browsers. While Edge and Chrome Desktop make this menu available after only one additional

[1] Chrome Desktop Version: 71.0.3578.98 (64-bit)
[2] Chrome Mobile Version: 74.0.3729.136
[3] Firefox Version: 64.0 (Build ID: 20181212110248)
[4] Edge Version: 42.17134.1.0
[5] Safari Version: 12.0.2 (13606.3.4.1.4)

| Browser | Subject O | Issuer CN or O |
|---|---|---|
| Chrome Desktop | 2 | 2 |
| Chrome Mobile | 3 | 2 |
| Edge | 2 | 2 |
| Firefox | 4 | 2 |
| Safari | 3 | 2 |

Table 5: Number of clicks required to get to subject and issuer information.

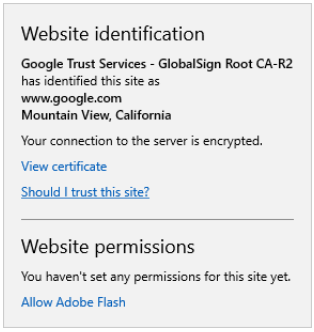


Figure 2: Pop-up after clicking on the lock symbol for an Organization validated website in Edge.

click, Firefox takes two more, for a total of four clicks, to open the certificate viewer. We also note that not all viewers are equally detailed, some are missing fields. For example, neither Edge nor Chrome Mobile includes information on extensions like certificate policies or basic constraints.

Note that the certificate viewers for Firefox, Chrome (Desktop and Mobile), and Safari offer an additional feature: The domain names are not translated from punycode, even if they are shown as IDN in the URL bar. This helps in preventing homograph attacks (e.g., [15]). We did not find websites that were translated to IDN in Edge's URL bar in the first place, making this less relevant in the case of Edge.

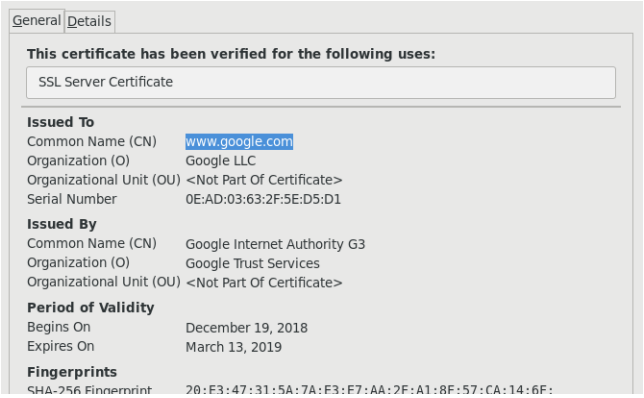


Figure 3: "General" tab of the certificate viewer in Firefox.

## 5.2 Discussion of Browser Evaluation

In Section 4.2.2 we looked at possible ways to recognize phishing websites, looking at issuer, subject and URL information. In this section, we discuss the certificate information presented by the different browsers in consideration of these findings.

We first make a distinction between EV and non-EV certificates, since the URL bar in most browsers is notably different for websites with EV certificates and those with non-EV certificates. In this case, some browsers (Edge, Chrome Desktop and Firefox) also show the subject `Organization` next to the URL, making the information readily available.

However, things are different for non-EV certificates. Here, no browser shows additional information by default without any user input. Only Edge displays some information after one click (issuer information and location if available), and all information discussed in this paper after two clicks. For Chrome Desktop, it takes users two clicks to get an overview of the certificate information, including `Organization` and issuer `CN`. Chrome Mobile requires an additional click to get to the certificate viewer, as does Safari. This is even more pronounced for Firefox: even though users will be able to verify the issuer `Organization` after two clicks, they will have to click through an additional window, four clicks in total, to get any information on the subject `Organization`.

Furthermore, some browsers did not include all fields of the certificate in their certificate viewer, though all of them contained the information discussed in this paper.

We also saw how hosting services can be abused and could offer a serious threat to unsuspecting users. Here, the browsers do include information about the current domain name, which might help mitigate the risk of hosting services.

All in all, we found that all of the fields discussed in this paper are available in all browsers we analyzed, yet this certificate information is available to users after different amounts of steps.

# 6 Conclusion and Future Work

Our analysis shows, that it is hard to differentiate phishing from benign websites using only information included in the certificate of a visited website, as certificates used by phishing websites include information that is very similar to that of benign websites, especially if both use certificates issued by the same issuer. This is plausible, considering the fact that phishers are often able to misuse the certificates of compromised servers, and that they will make decisions similar to the ones taken by administrators of benign websites when setting up their own servers.

We found that currently popular phishing targets often use EV certificates, and that it seems harder to copy websites using such certificates. Specifically, we found only a few instances of phishing websites where the issuer and organization of

the certificate used matched the equivalent information in the target's certificate. To assess if the differences we observed will persist in the future, we discussed how hard it would be for an attacker to obtain certificates that are more similar to their target's certificates. Unfortunately, it seems possible that at least some of the certificate features may be spoofed in the future.

Finally, we encountered instances of the particularly dangerous threat of hosting services, where user content is shown under the domain and protected by the certificate of a legitimate service. This can be abused by attackers to host their phishing websites, resulting in similar issuer and organization information as well as a similar URL on a legitimate looking top-level domain.

In future work, we plan to explore whether the observed differences between benign and phishing website certificates can be used to enhance the phishing detection capabilities of automated detection tools or users themselves. We also intend to further explore the question of how robust the subject `Organization` is against active attacks, and if subject spoofing might become more common in the future.

## Acknowledgments

This research was supported by the research training group "Human Centered Systems Security" sponsored by the state of North-Rhine Westphalia.

## References

[1] Alexa Top Sites. https://www.alexa.com/topsites. Online, accessed 26-Feb-2019.

[2] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82:69–82, 2015.

[3] Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior*, 60:185–197, 2016.

[4] CA-Browser Forum BR 1.6.3. https://cabforum.org/baseline-requirements-documents/, 2019.

[5] EV SSL Certificate Guidelines 1.6.8. https://cabforum.org/extended-validation/, 2018.

[6] Chromium Security: Marking HTTP As Non-Secure. https://www.chromium.org/Home/chromium-security/marking-http-as-non-secure. Online, accessed 27-Feb-2019.

[7] curl Website. https://curl.haxx.se/. Online, accessed 28-Feb-2019.

[8] X de Carné de Carnavalet and Mohammad Mannan. Killed by proxy: Analyzing client-end TLS interception software. In *Network and Distributed System Security Symposium*, 2016.

[9] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.

[10] Python Docs: Difflib SequenceMatcher. https://docs.python.org/3/library/difflib.html. Online, accessed 24-Feb-2019.

[11] Zheng Dong, Apu Kapadia, Jim Blythe, and L Jean Camp. Beyond the lock icon: real-time detection of phishing websites using public key certificates. In *2015 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–12. IEEE, 2015.

[12] Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala Al-Fuqaha, and Mohsen Guizani. Systematization of Knowledge (SoK): A systematic review of software-based web phishing detection. *IEEE Communications Surveys & Tutorials*, 19(4):2797–2819, 2017.

[13] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*, pages 79–90. ACM, 2006.

[14] Zakir Durumeric, Zane Ma, Drew Springall, Richard Barnes, Nick Sullivan, Elie Bursztein, Michael Bailey, J Alex Halderman, and Vern Paxson. The Security Impact of HTTPS Interception. In *Network and Distributed System Security Symposium*, 2017.

[15] Evgeniy Gabrilovich and Alex Gontmakher. The homograph attack. *Communications of the ACM*, 45(2):128, 2002.

[16] Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode*, pages 1–8. ACM, 2007.

[17] Google Transparency Report: HTTPS encryption on the web. https://transparencyreport.google.com/https/overview. Online, accessed 28-Feb-2019.

[18] Google Safe Browsing. https://safebrowsing.google.com/. Online, accessed 22-Feb-2019.

[19] Anti-Phishing Working Group. Phishing Activity Trends Report: 3rd Quarter 2018. *APWG*, 2018.

[20] Collin Jackson, Daniel R Simon, Desney S Tan, and Adam Barth. An evaluation of extended validation and picture-in-picture phishing attacks. In *International Conference on Financial Cryptography and Data Security*, pages 281–293. Springer, 2007.

[21] J Klensin. Internationalized Domain Names in Applications (IDNA): Protocol. No. RFC 5891. Technical report, 2010.

[22] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 3. ACM, 2009.

[23] Let's Encrypt. https://letsencrypt.org/. Online, accessed 26-Feb-2019.

[24] Pernelle Mensah, Gregory Blanc, Kazuya Okada, Daisuke Miyamoto, and Youki Kadobayashi. AJNA: Anti-phishing JS-based Visual Analysis, to Mitigate Users' Excessive Trust in SSL/TLS. In *2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, pages 74–84. IEEE, 2015.

[25] Mozilla CA Certificate Storage. https://www.mozilla.org/en-US/about/governance/policies/security-group/certs/. Online, accessed 24-Feb-2019.

[26] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–12. IEEE, 2018.

[27] OpenSSL s_client manual page. https://www.openssl.org/docs/man1.1.1/man1/openssl-s_client.html. Online, accessed 24-Feb-2019.

[28] OpenSSL Official website. https://www.openssl.org/. Online, accessed 24-Feb-2019.

[29] Phishtank: Phishing Database. https://www.phishtank.com/. Online, accessed 26-Feb-2019.

[30] Swapan Purkait. Phishing counter measures and their effectiveness–literature review. *Information Management & Computer Security*, 20(5):382–420, 2012.

[31] Eric Rescorla. Http over tls, RFC 2818. Technical report, 2000.

[32] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373–382. ACM, 2010.

[33] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 88–99. ACM, 2007.

[34] Jennifer Sobey, Robert Biddle, Paul C Van Oorschot, and Andrew S Patrick. Exploring user reactions to new browser cues for extended validation certificates. In *European Symposium on Research in Computer Security*, pages 411–427. Springer, 2008.

[35] Ivan Torroledo, Luis David Camacho, and Alejandro Correa Bahnsen. Hunting Malicious TLS Certificates with Deep Neural Networks. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 64–73. ACM, 2018.

[36] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. Phinding phish: Evaluating anti-phishing tools. *Network and Distributed System Security Symposium*, 2007.

## A Additional Results of Certificate Collection and Analysis

In the following we include several tables that contain additional details on our certificate collection results and its subsequent analysis. Tables 6 and 7 show the exact number of certificates issued by the most popular issuers for benign and phishing certificates.

| Issuer CN | Count |
|---|---|
| COMODO ECC Domain Validation Secure Server CA 2 | 7189 |
| Let's Encrypt Authority X3 | 6854 |
| COMODO RSA Domain Validation Secure Server CA | 4027 |
| CloudFlare Inc ECC CA-2 | 2564 |
| Amazon | 1908 |
| DigiCert SHA2 Secure Server CA | 1744 |
| Go Daddy Secure Certificate Authority - G2 | 1722 |
| GeoTrust RSA CA 2018 | 1426 |
| RapidSSL RSA CA 2018 | 1015 |
| DigiCert SHA2 Extended Validation Server CA | 1001 |
| GlobalSign Organization Validation CA - SHA256 - G2 | 825 |
| GlobalSign CloudSSL CA - SHA256 - G3 | 624 |
| cPanel, Inc. Certification Authority | 612 |
| DigiCert SHA2 High Assurance Server CA | 571 |
| COMODO RSA Organization Validation Secure Server CA | 523 |

Table 6: Number of benign certificates for the 15 most popular issuers.

| Issuer CN | Count |
|---|---|
| Let's Encrypt Authority X3 | 3259 |
| cPanel, Inc. Certification Authority | 2103 |
| RapidSSL TLS RSA CA G1 | 862 |
| COMODO RSA Domain Validation Secure Server CA | 502 |
| COMODO ECC Domain Validation Secure Server CA 2 | 489 |
| CloudFlare Inc ECC CA-2 | 474 |
| DigiCert SHA2 Secure Server CA | 321 |
| Go Daddy Secure Certificate Authority - G2 | 272 |
| Google Internet Authority G3 | 188 |
| RapidSSL RSA CA 2018 | 128 |
| Microsoft IT TLS CA 1 | 88 |
| GlobalSign CloudSSL CA - SHA256 - G3 | 74 |
| Actalis Domain Validation Server CA G1 | 70 |
| Amazon | 63 |
| DigiCert SHA2 High Assurance Server CA | 58 |

Table 7: Number of phishing certificates for the 15 most popular issuers.