# Content-Agnostic Detection of Phishing Domains using Certificate Transparency and Passive DNS

Mashael AlSabah
Qatar Computing Research Institute
Qatar
msalsabah@hbku.edu.qa

Mohamed Nabeel
Qatar Computing Research Institute
Qatar
mnabeel@hbku.edu.qa

Euijin Choo
University of Alberta
Canada
euijin@ualberta.ca

Yazan Boshmaf
Qatar Computing Research Institute
Qatar
yboshmaf@hbku.edu.qa

## ABSTRACT

Existing phishing detection techniques mainly rely on blacklists or content-based analysis, which are not only evadable, but also exhibit considerable detection delays as they are reactive in nature. We observe through our deep dive analysis that artifacts of phishing are manifested in various sources of intelligence related to a domain even before its contents are online. In particular, we study various novel patterns and characteristics computed from viable sources of data including Certificate Transparency Logs, and passive DNS records. To compare benign and phishing domains, we construct thoroughly-verified *realistic* benign and phishing datasets. Our analysis shows clear differences between benign and phishing domains that can pave the way for content-agnostic approaches to predict phishing domains even before the contents of these webpages are up and running.

To demonstrate the usefulness of our analysis, we train a classifier with distinctive features, and we show that we can (1) perform content-agnostic predictions with a very low FPR of 0.3%, and high precision (98%) and recall (90%), and (2) predict phishing domains days before they are discovered by state-of-the-art content-based tools such as VirusTotal.

## CCS CONCEPTS

• **Security and privacy → Web protocol security**.

## KEYWORDS

phishing domains detection, passive DNS, certificate transparency, machine learning

**ACM Reference Format:**
Mashael AlSabah, Mohamed Nabeel, Euijin Choo, and Yazan Boshmaf. 2022. Content-Agnostic Detection of Phishing Domains using Certificate Transparency and Passive DNS. In *25th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2022), October 26–28, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3545948.3545958

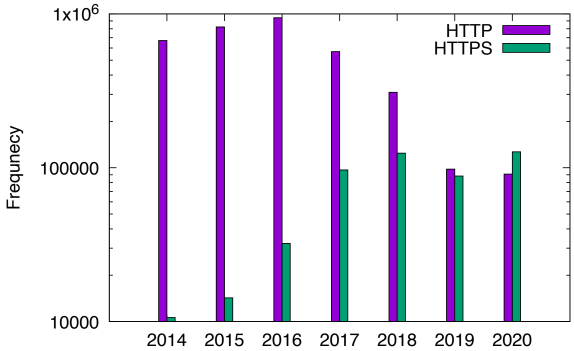


**Figure 1: Frequency count (in log scale) of HTTP (red bars) and HTTPS (green bars) phishing domains per year from 2014 to 2020**

## 1 INTRODUCTION

While web-based malware attacks have declined, phishing attacks are showing no signs of slowing down [8, 22, 65]. Previous research examined content-based analysis [62, 70, 85], and network-based or URL-based approaches [27, 28, 48, 53, 75, 81]. In practise, existing defensive techniques rely on published blacklists, reported and verified by users [16, 17], or deployed by companies or organizations, including Google Safe Browsing (GSB) [21], VirusTotal (VT) [82], and the Anti Phishing Working Group (APWG) [1]. One key requirement for such blacklists is timeliness. That is, a domain must be added to the blacklists before it has affected many victims. Currently however, a domain is added to blacklists only after it has started its campaign. Furthermore, blacklists are vulnerable to cloaking [63, 64], a technique where phishing domains evade crawlers such as VT and GSB, but are still made visible to victims. Content-based techniques can't be proactive as well and can be expensive. With the challenges facing content-based detection techniques, there is a growing need to utilize alternative defensive approaches.

*RAID 2022, October 26–28, 2022, Limassol, Cyprus*

ACM ISBN 978-1-4503-9704-9/22/10…$15.00
https://doi.org/10.1145/3545948.3545958

In this work, we observe that there are multiple sources of information before domain content is available online, and we seek to understand if these sources can indicate phishing intent. Such sources include WHOIS records, passive DNS records [1] and certificate records. While WHOIS records are increasingly hard to aggregate and query, passive DNS traces and domain certificates are readily available. Historical certificates in particular are available through Certificate Transparency (CT) logs [19]. Those are public transparent append-only servers publishing certificates almost as soon as they are issued. Phishing domains, in particular, are increasingly incentivized to be CT-compliant to (1) appear more legitimate and to (2) be reachable by victims (non CT-compliance affects reachability by browsers [14]). Oest *et al.* recently estimated that HTTPS phishing attacks were three times more[10, 52] successful than their HTTP counterparts [65].

Figure 1 compares the total number of HTTP and HTTPS reported instances in Phishtank's published list [16] of verified phishing domains between 2014 and 2020. There is a significant increase in the utilization of TLS (green bars) until recently. Generally speaking, the difference between the number of HTTP (purple bars) and HTTPS domains (green bars) has been decreasing over the years. TLS adoption will continue growing [42, 72], and with CAs providing free automated certificates [26], domain owners (phishing or benign) are increasingly getting on the TLS bandwagon.

Almost all prior research on phishing domain detection utilizes a subset of Alexa top domains to represent benign domains [54, 78]. This results in a biased sample as a vast majority of the benign domains do not appear in Alexa top lists. Further, domains appearing in Alexa are usually operational for decades, have stable hosting infrastructures and follow best security practices compared to non-Alexa benign domains. Such characteristics make it easier to distinguish Alexa top domains from phishing domains, but it is not surprising that machine learning models trained on Alexa domains do not work well with the majority of benign domains which are not present in Alexa top list as they exhibit significantly different characteristics. A key distinction in our study is that we build and analyze a realistic benign dataset that reflects the majority.

**Approach.** The event of real-time publishing of domain certificates in CT logs gives us an early peek at upcoming phishing domains. We create thoroughly-verified datasets to represent benign and phishing domains. We extract over 80 certificate-, pDNS-, and lexical-based features based on novel observations, previous literature, or available tools. This is the first work that combines all these features to perform content-agnostic phishing detection. We evaluate our features using a Random Forest classifier, and we analyse them rigorously, and show why they work.

Previous work [57] suggests that it is generally impossible to differentiate between benign sites and phishing sites based on the content of simple X.509 certificates features alone. To address this issue, we combine aggregate and historical certificate features taken from CT logs to effectively identify recurring long-term phishing domains. We also combine CT and pDNS features to effectively mark new phishing domains.

**Contributions.** Our work makes the following contributions:

- We create large thoroughly-verified datasets of phishing and benign domains, comprising thousands of domains. Unlike previous work, our benign domains are non-Alexa domains which are harder to distinguish from phishing domains than Alexa domains.
- We define and extract novel CT-, pDNS-, and lexical-based features that do not require access to page content.
- Using our features, we train a classifier to predict and identify new and long-term phishing domains without inspecting content. We evaluate our model using various attribute sets and show their importance. We show that our model is able to identify phishing domains with very high precision and recall.
- We analyse our features, and provide insights into how they separate benign from phishing.
- We carry out live experiments to show the usefulness and proactivity of our approach in practise. Indeed, we are able to identify phishing domains *days* before they are identified by the widely used GSB and VT phishing crawlers.

## 2 BACKGROUND

**Passive DNS data feed.** Passive DNS (pDNS) [83] captures traffic by cooperative deployment of sensors in various locations of the DNS hierarchy. For example, Farsight pDNS data [39] utilizes sensors deployed behind DNS resolvers and provides aggregate information about domain resolutions as well as public available zone file updates. In our research, we use this as our pDNS feed. One advantage of pDNS is that it preserves the privacy of individual Internet users as it contains only aggregated information. However, such data is not as rich in information as proxy/HTTP DNS logs, which not only contain individual DNS queries and responses, but also timing information. Despite these limitations, we are able to extract important data and characteristics as described in Section 3. Previous research utilizes this data feed to uncover the behavior of domains in the wild and also detect malicious domains [44, 45]. In this work, we use the following three record types from pDNS:

- SOA records: they contain the MNAME (the primary name server for the domain) and RNAME (email of the domain name administrator).
- A records: these records map the domain names to their IPs.
- MX (mail exchanger) and NS (name server) records: these specifiy where inbound mail for a domain should get directed and authoritative name servers respectively.

All of the above records also contain timing information (such the first and last seen timestamps for each record), and the count of DNS lookup requests.

**VirusTotal(VT) URL feed.** VT [82] is a Google-based service that provides a public querying facility to obtain intelligence on any URL by analyzing more than 70 third-party scanners and URL/domain blacklisting services, including Google Safe Browsing (GSB) [21], COMODO site inspector [5], phishtank [16] and many others. Each tool in VT categorizes a URL as clean, malicious, phishing or malware. VT provides a rate limited public API to check the status of URLs programmatically. Additionally, every hour, VT publishes a feed of URLs along with aggregated intelligence for the URLs queried by Internet users all around the world during the previous

[1] These records are primarily related to setup, configurations and testing. It should be noted that these records include publicly available zone file updates as well.

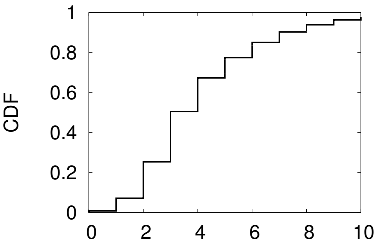


**Figure 2: Number of VT engines reporting a domain as "phishing"**

hour. Previous research [44] utilizes VT data to compile malicious ground truth for detecting or predicting malicious activities in the Internet. However, there are challenges related to the intelligence reported by VT. For example, GSB and phishtank results in VT are not always consistent with their direct results and different tools provide different labels such as phishing and malware for a given URL. A common practice is to obtain the intersection between multiple sources and also to use majority voting as the final VT label [44].

## 3 DATA COLLECTION AND VALIDATION

**Phishing domain dataset.** Phishtank [16] is a free community based site, where users submit and query phishing domains and URLs. We maintain a local database of Phishtank URLs that are synced with Phishtank. From this database, we constructed a list of 11K HTTPS domains randomly sampled after applying the following filtering:

- We included HTTPS domains labelled as "valid", meaning that they have been verified as phish by the community members.
- We eliminated HTTPS domains relying on web hosting (such as domains hosted by `godaddysites.com`) since querying the certificates for any of these categories does not return the intended certificates for phishing domain.

We checked each domain against VT to ascertain that it is indeed still phishing. Figure 2 shows the distribution of the number of "phishing" label positives for this data. In addition to phishtank, 50% of the domains have been red flagged by more than three engines. We maintained only the domains that have been specifically marked by VT as "phishing" (rather than "malicious" or "suspicious") by at least one VT engine. This reduced the domain list slightly to 10.8K domains. From these domains, there were roughly 7.8K domains with multiple certificates and 3K domains with only one certificate.

While the dataset contains domains that have been reported prior to November 2020, we performed our VT cross-check of their current phishing status early December 2020 to ensure that they are still considered phishing during our analysis. To further ensure accurate phishing ground truth, we actively queried these domains periodically and found that over 50% are non-existent, which is consistent with short-lived disposable phishing domains behaviour. For the phishing domains that are currently online, we took random samples and manually verified that they are indeed phishing web sites.

**Benign domain list.** Although Alexa lists have been generally used in previous work as the main source of benign domains [54, 78], we believe that relying solely on Alexa top domains can result in biased observations[2]. We craft benign non-Alexa datasets to realistically model the various types of domains that get appended daily to CT logs. Alexa domains alone cannot realistically represent mainstream benign domains.

Instead, we construct our benign domains list as follows. First, we compiled roughly 35M domains that were appended to CT logs in 5 nonconsecutive days in November 2020. Then, we filtered the domains which are never marked by VT URL feed as malicious (i.e. consistently maintained a VT score of zero) during the last year[3]. We also filtered out any domains that appeared in Phishtank in the last 10 years. We sampled down the list to 3,000 one-certificate domains, and 8,000 multiple-certificate domains to match the number of domains in our phishing dataset.

Likewise to the phishing dataset, verification against VT of malicious domain status for the benign dataset was carried out early December 2020.

**Alexa domains list.** For completeness and comparison purposes, we also compiled a list of top 10K Alexa domains by randomly picking domains from top 20K Alexa domains.

For each domain in our lists, we query COMODO's `crt.sh` server [13] to get all its certificates using the `pycrtsh` Python module. We use the certificate serial numbers (for each domain) so as not to double count certificates due to the existence of pre-certificates and leaf certificates. We successfully downloaded most of the certificates in our domain lists[4].

## 4 RETROSPECTIVE ANALYSIS

The goal of our analysis is to understand and compare the characteristics exhibited by the different datasets in order to derive effective features with a clear understanding of why they work in identifying phishing domains.

### 4.1 CT logs temporal characteristics

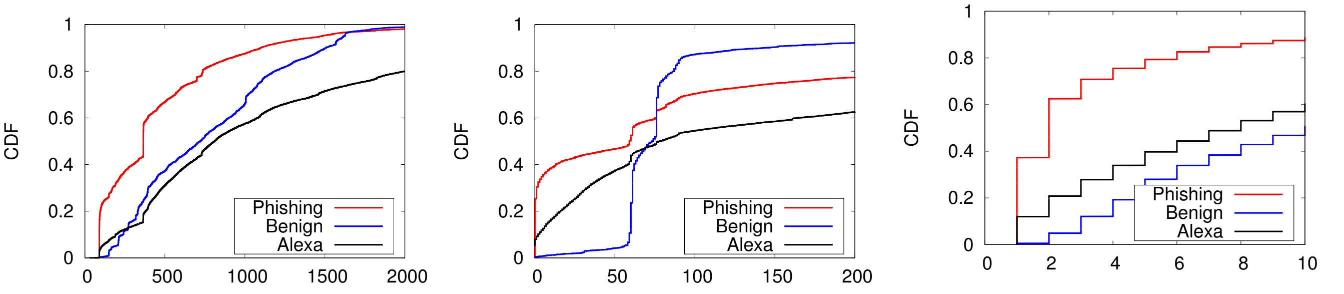
CT logs are a rich repository containing historical records of certificates for each domain. We first examine if phishing domains exhibit distinguishing temporal characteristics. In particular, we compare the distributions of the CT lifetime between the benign and the phishing datasets. We also compare the number of historical obtained certificates and frequency of issued certificates.

**Lifetime.** We extract the *CT lifetime* of each domain. The lifetime is the difference between the expiration date of the last certificate and the issuance date of the first certificate. We show the lifetime for domains with multiple certificates. For single-certificate domains, we show the certificate validity period distributions later in the section (in Figure 6).

[2] Because various Alexa domains are well-provisioned e-commerce, social networks or localised versions of top corporations, data can be severely biased.

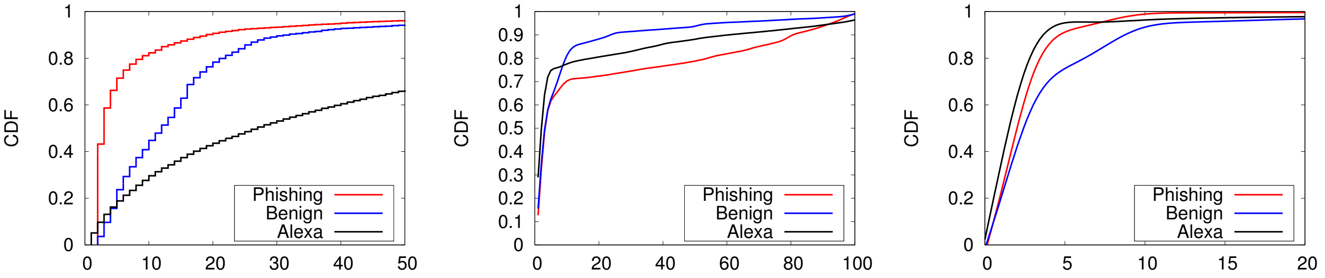
[3] We collect VT hourly URL feed that include all the URLs queried by the community in each hour of the day since Nov 2017.

[4] Only 688 phishing domains were not successfully downloaded due to server or network errors.



**(a) CT log domain lifetimes (days)** **(b) Certificates inter-arrival times (days)** **(c) Number of unique (month,year) of certificates issuance**

**Figure 3: CT logs temporal characteristics**



**(a) Total number of certificates** **(b) SAN list size** **(c) SAN matches**

**Figure 4: Issuer and SAN-based characteristics**

Figure 3a shows the distribution of the CT lifetime (in days) of each domain as it appears in CT logs. As expected, Alexa domains clearly have larger lifetimes than our benign and phishing domains. Furthermore, benign and phishing domains clearly exhibit different distributions. Our benign domains have a lifetime of 720 days, whereas phishing domains have a lifetime of 365 days at the median, though 20% of phishing domains have lifetime ranges from three to several years. This certificate lifetime is possibly different from the active phishing campaign lifetime, as it has been shown previously that the active lifetime of the majority of phishing domains does not exceed a few days [56, 80]. However, contrary to popular belief, some phishing domains have a longer time span over multiple years, as we also observe in our dataset. For example, Le Page *et al.* [49] shows that there are many URL-shortened phishing URLs that are active for more than one year. Indeed, Oest *et al.* point out that while some campaigns may be short lived, organized criminals carry out successive deployment of persistent and sophisticated attacks to improve their profits and longevity [65]. Our observations are consistent with these findings.

Our close examination showed that some of the longer lived phishing domains are actually "parked" or "revived" after their prior use. For example, `bestbuy-us.com` is a domain from our phishing dataset that was initially parked and then turned phishing after a while. Also, the domain `paypal-secure-limited.com` is an example domain from our phishing dataset that was initially registered ten years ago, blacklisted shortly afterwards, and then re-registered in April 2020, a few months after the expiration.

Of the 20% long-lived phishing domains, we observed that 13% were parked one or more times during their lifetime and about 30% of the phishing domains exhibit the "revived" behavior. To verify this further using `WHOIS`, we extracted a subset of phishing domains that made their first CT appearance in 2020, and tracked the dates of when they were first registered. We found that 50% of the domains have been registered between 1 and 300 days in advance. These observations suggest that parking and revived behavior of domains are a good discriminator of benign and phishing.

**Inter-arrival times.** While manually examining issuance dates of domains, we were surprised to observe that it was common for domains to obtain certificates with very close issuance dates. To quantify this behaviour, we compute the *inter-arrival times* between certificates for multi-certificate domains. This is the time gap between the issue dates of every two consecutive certificates of a domain. We expect benign domains to acquire new certificates near the expiry date of their existing certificates so as not to risk using expired certificates. Phishing domains inter-arrival times are expected to depend on their longevity goals.

Figure 3b depicts the distribution of the mean inter-arrival times between certificates of each domain. Roughly 30% of phishing domains had close to zero days inter-arrival times, and only 0.3% of benign domains had certificates with such short inter-arrival times. We were surprised to see that certificates were being pushed frequently, and almost on a daily basis for some cases, even though their earlier certificates are still valid. For example, the phishing domain `fakeid.top` obtained 225 Let's Encrypt certificates between

2019-03-02 and 2020-09-24. Many of these certificates were obtained on consecutive days. Alexa domains showed larger inter-arrival times compared to phishing domains.

A related metric is shown in Figure 3c, which compares the distributions of the number of certificates that were issued in unique (month, year) pairs. Approximately, 60% of phishing domain certificates were issued in two or one distinct (month, year) pairs. In comparison, benign domains showed more than ten unique pairs at the median. This shows that benign certificates issuance is more spaced out than phishing domains. Alexa domains appear to have more distinct months than phishing but less than benign domains.

One explanation for the high frequency of certificates is that domain owners may have a misconfigured Certbot client [6], a tool allowing clients to automate the process of renewing Let's Encrypt certificates. The interarrival times between benign certificates were generally larger than phishing certificates with more than 90% exceeding 60 days (as shown in Figure 3b).

Another reason for the high frequency of certificates for Alexa, benign and phishing domains is the use of Content Delivery Networks (CDNs) such as Cloudflare, which use one certificate to support multiple domains by using the Subject Alternative Names (SAN) certificate extension. Those are also known as *cruise certificates.* The frequency of certificate updates increases with such use of multi-domain (SAN) certificates. While the impact of misconfigured certbots and cruise certificates is present in all datasets, we believe the phishing datasets is the mostly affected.

**Number of certificates.** While phishing domains had a shorter certificate inter-arrival times, benign and Alexa domains have a significantly larger number of certificates. Figure 4a shows the distributions of the total number of certificates obtained by multi-certificate domains in our datasets. As can be seen, 80% of phishing domains obtain between two and ten certificates. The same percentage of benign domains obtain more certificates that range between three and 25. The distributions also exhibit a long tail where a very small fraction of domains own significantly more certificates.

We inspected domains more closely to find out if some domains have "uncertified" periods during their lifetime. In other words, for domains that have more than one certificate $C_1, C_2, .., C_n$, we checked if the issue date of $C_i$ is greater than the expiry date of $C_{i-1}$. We found that 12% of benign domains showed uncertified periods with a mean gap of 133 days. In comparison, 21% of the total phishing dataset had such uncertified periods with a mean gap duration of 181 days. Some of these phishing domains appear to be squatting domains which contained substrings like "paypal", "netflix", etc. This could be an indication of phishing domain reuse or revival, possibly after a takedown operation, either by the same operators or not. This observation is consistent with previous studies [58] of domain drop-catching suggesting that malicious domains are more likely to be caught after they are dropped.

## 4.2 Certificate-based characteristics

**SAN list.** From the certificate extensions, we extracted the Subject Alternative Names (SAN) field when available. This field lists all domains that are authenticated using the same certificate. Some organizations with multiple domain names, aliases, and subdomains, also use this field for wildcard certificates. For each domain, we compute the *SAN list size*, which is the number of domains in its SAN field for one-certificate domains, or the average number of domains in its SAN field from all certificates for multi-certificate domains.

Figure 4b compares the distributions of the mean SAN list size for Alexa, phishing and benign domains. For 50% of benign and phishing domains, there are at most three domains in their SAN lists. At the 75th percentile, benign domains have eight domains in the SAN list, whereas phishing domains have 31. In general, it appears that phishing domains have larger SAN lists than benign or Alexa domains. We believe that the reason for the large SAN list size for some phishing and benign domains is the reliance on CDNs which generate the multi-domain certificates.

In some cases, when a CDN issues a multi-domain certificate, the SAN field usually contains arbitrarily unrelated domains. For example, one of the certificates for the domain `mainlinehometeam.com` has a large SAN list which includes unrelated domains like `alabamacoastliving.org` and `mariottrealestate.com`. Domains like `cnn.com` have more related domains `money.cnn.com`, and `moneystream.cnn.com`, which are controlled by the same apex domain `cnn.com`.

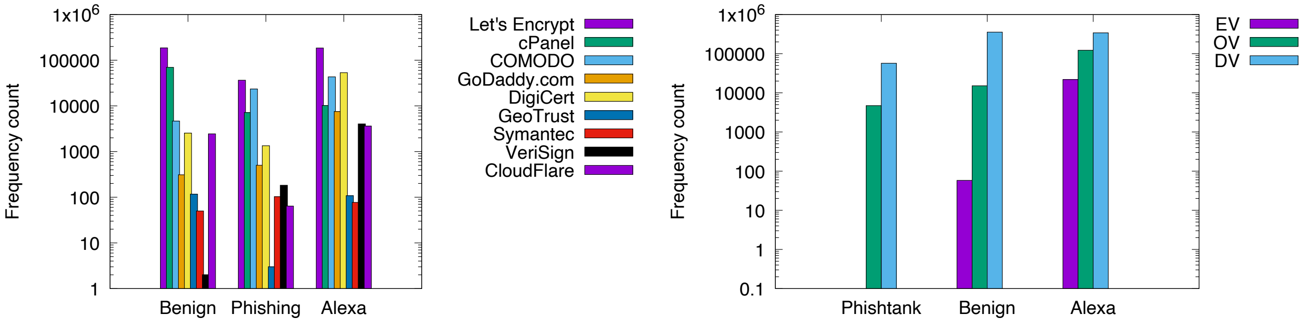
Based on the above observation, we expect the domains in the second SAN list to have a higher similarity compared to the first SAN list. To quantify similarity here, we compute the *SAN matches*, which is the number of second-level domain matches between a domain and the domains in its certificate's SAN list. The rationale is that a higher number of matches indicates a related group of domains.

Figure 4c compares the number of SAN matches for benign and phishing (one- and multi-certificate) domains. As expected, 50% of domains from both datasets had at most two matches as those are the domains that have a SAN list of size one or two (Figure 4b). At the 80th percentile, benign domains had eight matches whereas phishing domains had only two despite having the larger SAN sets.

**Issuers.** Intuitively, one would expect that phishing domains would rely more on free certificates issued from Let's Encrypt or COMODO to minimize the cost of their operations. Phishing catcher [15] is a CT- and heuristics-based phishing domain detection tool that listens to incoming certs and assigns a higher phishing score based on several metrics, one of which is if the issuer is specifically Let's Encrypt. To assess the usefulness of this assumption, we extract all certificates issued to each domain, and from each certificate, we extract the *issuer* field. The issuer field consists of several subfields, including the Organization (O), which indicates the name of the issuing organization (e.g. COMODO CA Limited), and the Common Name (CN), which indicates the server name within the organization (e.g. COMODO RSA Domain Validation Secure Server CA2 or COMODO ECC Extended Validation Secure Server CA). Since organizations can have a large number of common names, we restrict our analysis to organizations.

Figure 5a shows the frequency count of the most common CAs (in log scale) in our datasets. The top four Organizations in all datasets belong to COMODO, Let's Encrypt, Digicert and cPanel[5]. While there are slight differences in the frequencies of some CAs

[5] Note that cPanel is a web-hosting control panel provided by many hosting providers to website owners to facilitate managing their pages. By default, cPanel uses COMODO certificates, but the issuer organization field will still show as cPanel Inc.



**(a) Frequency count of CA organizations**

**(b) Frequency count of certificate validation options**

**Figure 5: Issuers and certificate validation**

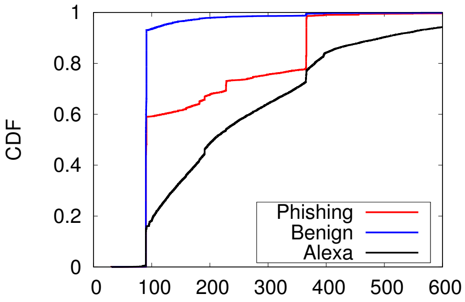


**Figure 6: Mean certificate duration (days)**

(e.g. Geotrust and Verisign), the dominance of the top issuers makes such differences insignificant. One takeaway is that the issuer alone is insufficient to infer if the domain is suspicious. Contrary to popular belief [74], the domain association with Let's Encrypt doesn't necessarily mean malicious use.

We also checked if certain pairs of organizations occur more frequently together for the phishing dataset. In general, we observed that utilizing CloudFlare and Let's Encrypt together by one domain is more frequent for phishing domains with a count of 977 compared to 147 in benign domain. Note that both CloudFlare and Let's Encrypt provide free certificates, which makes them more attractive for phishing operators trying to minimize the cost of their operations.

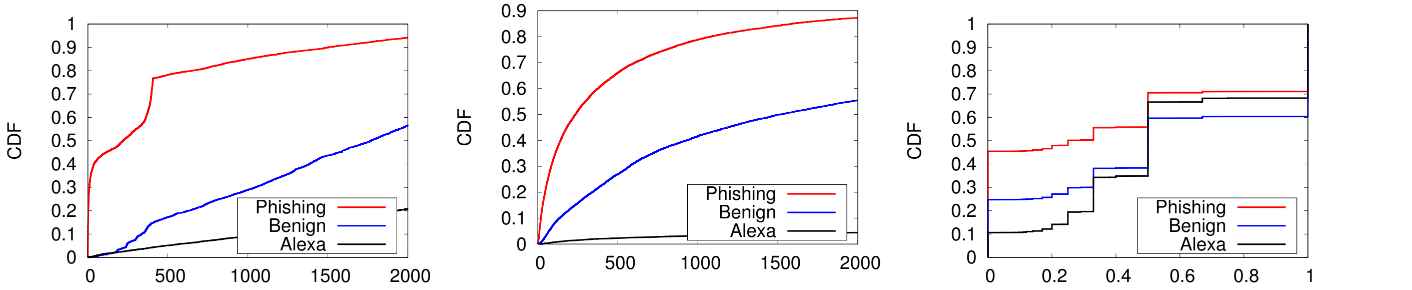
**Validation.** Certificate *validation* has been used in previous work as a feature to identify phishing certificates [80]. The rationale is that benign domains opt for higher validation, while phishing domains tend to reduce their costs by using lower validation options. Each certificate has a validation option that is performed by the CA. The most common validation options are Domain Validation (DV), Organization Validation (OV), and Extended Validation (EV).

Note that certificates don't explicitly encode their validation type. To identify the validation type of each certificate, we rely on (1) the Object Identifier (OID) field in the certificate if it exists (not all certificates include it), and (2) other assumptions and heuristics since there is no deterministic way to identify the certificate validation type. First, if the certificate is issued by Let's Encrypt, we assume it is a DV certificate, since Let's Encrypt only provides this category of certificates [18]. Next, we extract the OID field from each certificate, and check if it is EV by checking if it belongs to a list of known EV OIDs predefined in browsers [33, 36]. If not, we check the policy against aggregated OIDs for DV and OV certificates which we compiled from tools such as Censys [35] or manually from the CA websites [61, 76]. If the OID does not belong to any of our predefined lists, we follow (1) heuristical-based approaches that have been proposed by experts [43], and (2) the documentation of baseline requirements set by the CA/Browser forum [32]. For example, if a certificate has organizationName, localityName, stateOrProvinceName, and countryName that are set, we assume it is OV. In most cases, the organizationName field of the subject field contains the strings "Domain Control Validated" or "Domain Validated", which indicates a DV certificate.

Figure 5b compares the frequency count (in log scale) of the validation code of all certificates belonging to the domains in our datasets. DV is clearly the most common form of validation, as it comes by default with basic and even free certificates. However, OV appears to be common in both benign and phishing and more common in Alexa domains. Note that only 4,725 phishing certificates have OV validation compared to 15,135 benign certificates[6]. Note that our benign datasets didn't include many EV certificates as we deliberately selected non-Alexa benign domains with no history of maliciousness/phishing. Note that EV certificates may be irrelevant soon [4]. As expected, Alexa domains obtained significantly more EV and OV (21,910 and 122,631, respectively) domains than benign or phishing domains, but overall, DV validation is still mainstream in all datasets, including Alexa domains which had 341,330 DV certificates.

**Validity period.** Each certificate has a different *validity period* that can range from a few months to a few years. Long-term paid certificates are recommended for e-commerce due to their longer validity, enhanced validation options, and customer support. One would expect that benign domains would use more premium certificates compared to phishing domains. Note that some CAs, such as Let's Encrypt [24] and COMODO provide free certificate options

[6] We confirmed the OV certificate validation using COMODO's SSL analyzer service [12].

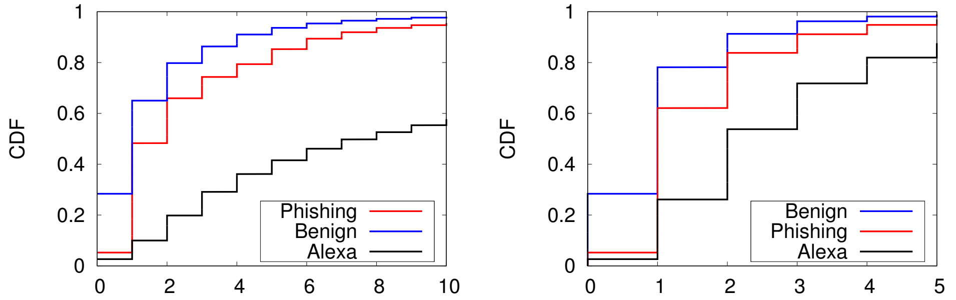


**(a) pDNS SOA durations (days)**

**(b) Number of DNS queries from pDNS SOA records**

**(c) MNAME fields Jaccard distances**

**Figure 7: Distributions of pDNS characteristics**



**(a) Number of IPs that hosted a domain**

**(b) Number of ASNs that hosted a domain**

**Figure 8: Distributions of IPs and ASNs obtained from pDNS A records**

which have a duration of 3 months [5, 26], which can be attractive for both benign and phishing domains. More than 90% of our benign domains have a validity of 90 days, which is the default validity period for various DV certificates. While 60% of phishing domains have a similar validity period, about 30% acquired certificates with up to one year of validity. For example, `lloydsbank.com.login-review-976.info` is a phishing domains that has a certificate with one year validity from 02-10-2020 to 02-10-2021, whereas `paypals-securitys.com` is a phishing domain that has a validity of one month from 07-09-2019 to 07-10-2019.

## 4.3 pDNS characteristics

Phishing domain revival can manifest itself in pDNS traces. It has been previously observed that revived domains often exhibit bursty DNS lookup behaviours where they show a sudden increase followed by a sudden decrease in the number of requests [30]. These patterns of increases and decreases correlate with revival attempts.

For example, the domain *paypal-verify-secure.com* is a phishing domain that made its first appearance in May, 2017 with a lookup volume of 329. Two years later (a gap of exactly 729 days), it made its second appearance in May, 2019, with a small lookup volume of 8. The latest and final appearance occurred after a gap of 358 days in May, 2020, with a lookup volume of 31. We believe that profiling such bursty request behavior could help discriminate revived or long lived phishing domains from benign ones.

**Traffic.** Using the Farsight API, we downloaded the Start of Authority (SOA), A, and MX records for each domain. From each of these records, we extract the *count*, which is the number of DNS queries or resolutions recorded for a domain. Each record also maintains timestamps of when a domain was first and last seen. We use these fields to compute the *pDNS lifetime*, which is the difference between the last time and the first time a domain was seen in the pDNS traces. In general, we use the timestamps of all records to calculate their respective durations and request counts.

Figures 7a and 7b show the distributions of pDNS lifetime (in days) and DNS lookup counts taken from SOA records for benign and phishing domains. Benign domains have significantly longer lifetimes and DNS lookup requests compared to phishing domains. Approximately, 30% of phishing domains have duration periods that are close to zero and the remaining 70% range between a few days to several years.

**Domain ownership changes.** Inspired by previous research that aims to detect domain ownership changes [50], we compute three components that indicate ownership changes based on: *SOA differences*, *infrastructure changes*, and *lookup volume*. To compute potential SOA differences, we utilize the MNAME (the primary name server for the domain) and RNAME (specifies the email of the domain name administrator). This is achieved by dividing each field (RNAME and MNAME) into two halves of a temporal window, and computing the Jaccard similarity between them. A similar process is done to detect infrastructure changes by also computing host similarity hosts (from A records) between two halves of a window. The idea is that substantial changes between the first and second halves of a temporal window can indicate ownership changes. Finally, we compute a statistical t-test between the lookup distribution of the first and second halves of a temporal window as well.

Figure 7c plots the distribution of the Jaccard distances computed based on SOA fields[7]. For benign domains, 25% showed no similarity between their SOA halves, 40% showed perfect similarity, and the rest are in between with roughly 20% showing 50% similarity. In comparison, 45% of phishing domains have no similarity and 29% have perfect similarity. Clearly, benign domains have more similarities which indicates less ownership changes. Distances based on RNAME showed similar distributions. For Alexa domains, 10% showed no similarity, and more than 30% had perfect similarity.

**Infrastructure changes.** Rapid changes in the hosting infrastructure can be a sign of fast fluxing or maliciousness. To capture domains that are either likely to be re-registered or hosted on multiple hosting providers over their lifespan, we extract the *number of name servers*, and the *number of administrative servers* related to the domain. We also extract the total *number of IPs* that have hosted a domain during its lifetime. Since CDN hosting might cause a domain to be related to multiple IPs, we also extracted the *number of unique Autonomous System Numbers (ASNs)* that hosted a domain.

We have not observed differences in the distributions of the numbers of pDNS nameservers or the numbers of administrative servers between benign and phishing domains. However, we observed different distributions of the number of IPs and ASNs used. This observation is consistent with the jumping hosting provider behavior of many malicious domains [60]. Figures 8a, and 8b compare the number of IPs and ASNs (owning the IPs) that host each domain in our dataset. At the 80th percentile, benign domains are hosted on two IPs, whereas phishing domains are hosted on up to four IPs. Phishing domains are hosted on more IPs possibly due to CDN use. In terms of the number of hosting ASNs, the difference between the two distributions shrinks with one ASN at the 75th percentile for benign and two ASNs for phishing. Alexa domains use more IPs and ASNs than both benign and phishing domains as they are more long term and utilize multiple CDN providers.

## 4.4 Lexical characteristics

We examine common lexical-based characteristics which utilize domain name strings, some of which have been used previously as features that can identify phishing domains [3, 15]. Such features include domain name entropy, domain length, number of special characters or digits, and squatting-based features.

**Entropy.** Randomized domain names generated by algorithms (e.g. DGAs) can possibly be indicative of maliciousness. Relative entropy has been used as a measure of randomness in domain names in previous work [15, 29, 86]. To compute the relative entropy, we compute the character entropy based on Alexa top 10K domains. The idea is that characters in domain names should not be equally probable, but should follow Alexa top domains in terms of character probabilities as a baseline. We also computed a *dictionary-based entropy*, where in each domain name, base words are matched against English dictionary words, or patterns. The entropy is then calculated based on non overlapping matches[8].

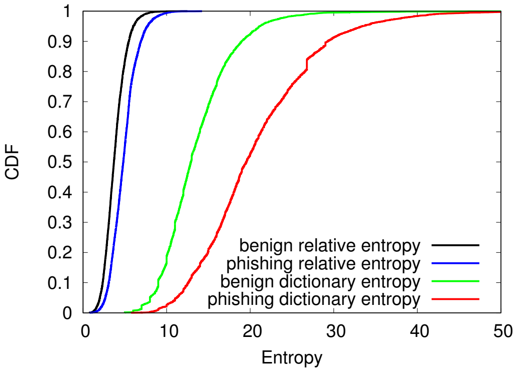


**Figure 9: Comparison between relative and dictionary entropy distributions for the benign and the phishing datasets**

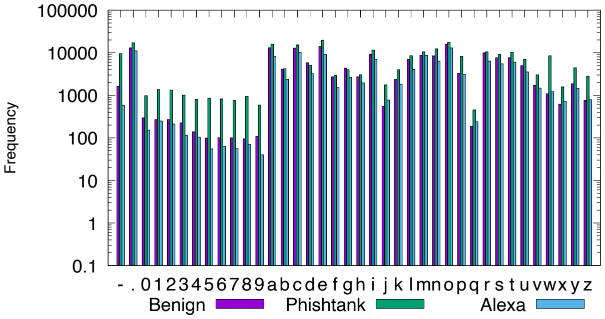


**Figure 10: Occurrences of characters**

Figure 9 compares the relative and dictionary entropy distributions for benign and phishing datasets. In both approaches, phishing domains result in more entropy. However, dictionary-based entropy produce a significantly more distinguishable distribution between benign and phishing domains compared to the relative entropy approach. For example, at the median, benign domains have a dictionary-entropy of 12.9, whereas the dictionary phishing entropy is 19.5 bits.

**Characters.** Figure 10 compares the occurrences of characters used in the domains in our datasets. In general, benign and Alexa domains show similar frequencies compared to phishing domains, which specifically use more digits and dashes. Alphabet usage is similar for benign and phishing domains except for letters j, p, q, x, z, and y, which appear to be used more frequently by phishing domains. Most of these letters are also known to be the least commonly used in English writing.

**Squatting.** Some phishing domains are also known to use squatting techniques [45, 69] to trick more victims by mimicking legitimate domains by embedding known popular “brand” names such as `paypal` or `apple` in the domain name. To understand the relevance

[7] Other components based on the RNAME field and host similarities showed similar differences between phishing and benign domains.

[8] We utilized *zxcvbn* [84], a state-of-the-art password strength estimator tool to compute this entropy.

of squatting techniques in our datasets, we use *squatphish* [78] to detect squatting domains in our datasets. For each given domain name, squatphish scans hundreds of popular brands for different squatting types including typo-, combo-, homograph-, bits-squatting, or wrong-TLD-squatting.

One challenge is that many benign domains exhibit the same behaviour. For instance, Squatphish red flagged 17.7% of our phishing domains, categorizing 97% of which as combosquatting aiming to look like apple.com or paypal.com (e.g. apple.com.icloudto.cn). The remaining 3% (of the 18%) were marked as the other squatting types (homographs, typo squatting, etc)[9]. Squatphish also marked 8% of our benign domains as squatting (mostly combosquatting). For example, `cpcontacts.premiersurgoogle.ca` appeared to resemble `google.com` is a benign SEO website.

## 5 WHY CT LOGS AND PDNS TO DETECT PHISHING?

The advantage of using CT logs and pDNS records as a source to detect phishing domains early is twofold. First, both are accessible sources that anyone can tap into and get an early peek of upcoming phishing domains (new, or old and revived), which can result in *improved detection latency.* In some cases (as we'll show in Section 6.2), pDNS and CT footprints of phishing domains becomes available before their content is which can improve detection latency compared to content-based approaches and tools. Other sources to identify phishing include domain registration data. While domain registration can also provide an earlier peak, in practise, it is increasingly hard to acquire a complete set of registration data [50] that can allow early detection.

Previous work [57] finds that it is generally impossible to differentiate between benign sites and phishing sites based on the content of simple X.509 certificates features alone. However, such work looked at individual certificates of phishtank domains without using aggregate or temporal features extracted from CT logs that can indicate longer lived campaigns. In this work, we combine aggregate and historical certificate data taken from CT to derive our observations and insights.

Second, CT logs is a free and accessible data source that anyone can tap into and get an early peek of upcoming phishing domains (new, or old and revived). Other sources to identify phishing include WHOIS and DNS records. While WHOIS records can provide an earlier peak, in practise, it is increasingly hard to acquire a complete set of registration data that can allow for early detection. On the other hand, pDNS data is widely available with some services providing free access to their APIs [11].

## 6 CAN WE PREDICT PHISHING DOMAINS?

As we have seen in our analysis, CT, certificate and pDNS traces can provide distinguishing characteristics that can help in content agnostic prediction of phishing domains. We use these characteristics and insights to derive machine learning features to evaluate the usefulness of our observations. In this section, we present our evaluation of the features, and in Section 6.2, we perform live experiments on newly added domains to CT, and show that we are able to predict phishing domains before VT.

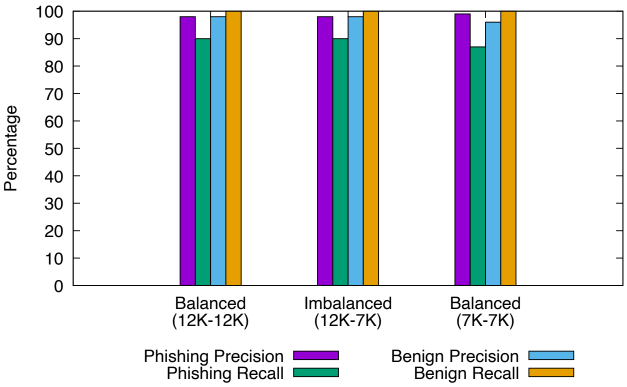


**Figure 11: Summary of results**

| Feature category | FPR |
|---|---|
| Lexical | 24% |
| pDNS | 5.9% |
| CT | 9.8% |
| All | 0.3% |

**Table 1: Performance of different feature categories**

### 6.1 Experiments

**Datasets.** We compiled fresh benign, and phishing datasets in April 2021 in a similar data collection and cleaning processes presented in Section 3. We used them to create benign and phishing datasets, both comprising of 12K domains. It should be noted that both datasets are disjoint from the datasets we presented in our analysis indicating the generalizability of features across different datasets. **Features and classifiers** Table 2 summarizes the features we used based on the insights derived from our analysis. As for classifiers, we use Random Forests (RF) [31] for classifying benign and phishing domains. Random forests are an ensemble learning algorithm that is an improvement over decision trees. Instead of performing splits on one feature as in decision trees, it uses a subset of features on each split. We also experimented with several classification algorithms including decision trees [68], SVMs [77], and regression models, and found that RF outperform other algorithms.

We used the `scikit-learn` module (`python 3.9.0`) to carry out the classification experiments. For training and testing, we use 10-fold cross validation, where all datasets are divided into 10 subsets, 9 of which are used for training and 1 is kept for testing. The process is iterated 10 times and the average is taken so that all subsets are tested. As evaluation metrics, we use the *precision*, and *recall.* Precision measures the proportion of the identified positives that is actually correct ($TP/(TP + FP)$). Recall measures what proportion of actual positives was identified correctly ($TP/(TP+FN)$). For completeness, we also present the False Positive Rate ($FPR = FP/(FP + TN)$).

[9] While the small percentage of squatting discovered may seem surprising, we note that previous work [79] observe that less than 1% of phishtank domains utilize squatting.

**Results.** Figure 11 summarizes our results for 3 experiments where we vary the number of benign and phishing domains. We get identical results when we use 12K phishing and 12K benign, or 12K benign and reduce phishing to 7K. For both cases, we get phishing precision and recall of 98%, and 90%, respectively. We also achieve higher benign precision and recall at 98% and 100% respectively. While this indicates that a small percentage of phishing domains are undetected, the high precision of phishing indicates that what is predicated as phishing is very likely phishing. This increases the confidence of the prediction. A smaller dataset of 7K benign and 7K phishing slightly reduces the phishing recall to 87% and the benign precision to 96% , but phishing precision is still very high at 99%.

Table 1 demonstrates the importance of each feature category on the classification FPR. Using our lexical features alone results in 24% FPR. Using a pure pDNS or CT classifiers results in 5.9% and 9.8%, respectively. The lowest FPR is achieved when all features are used together resulting in 0.3% FPR.

## 6.2 Live experiment

To explore the feasibility of content agnostic prediction, we implemented a proof of concept prototype system to classify domains in CT log stream from CertStream [34] near real-time and also cross validate a subset of detected predicted phishing domains periodically against VT/GSB. On average, our system process 4.4 million domains from CT per day. We process the latest set of documents from CT logs roughly every 10 minutes. Each batch contains on average 32K domains. We follow the following process to classify them.

- We maintain a list of Alexa top one million domains consistently appeared in the last year and filter these domains from the above mentioned batch. The rationale is that these domains are highly likely to be benign. On average, this filtering reduces the target domain set size by 2% from the original set.
- We maintain a list of public apex domains compiled from publicly available lists such as browser public suffix list [10] [7], CDN lists [2, 9], popular webhosting domains or proxy services. We filter out the domains whose apex domain is in the public apex list. The rationale for doing this step is that these domains are not under the control of the public apex domain and the certificate features of the public apex domains are not specifically related to these domains. This filtering reduces the dataset by 8% on average from the original set. This results in around 25K domains per batch to classify.
- We then collect certificates for the target domains and extract features for them. It takes roughly 25 seconds on average to process 25K domains. Extracting their features is done in parallel. We measured the time to extract our features for 25K domains. It takes 0.050 seconds at the median, and 0.0564 seconds at the $75^{th}$ percentile to extract features using our python implementation on a commodity virtual machine environment.
- Finally, using the extracted features, we utilized our trained classifier to predict the label of these domains.[11]

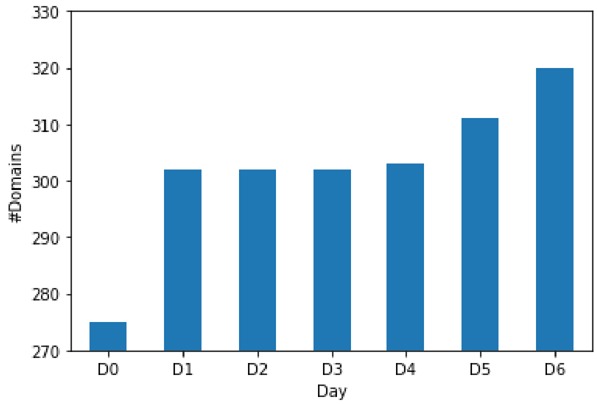


**Figure 12: Cross checking results against VT/GSB.**

To measure the effectiveness of the real-time classifier, we selected all 492 phishing domains predicted by our classifier for one batch of domains on Jan. $28^{th}$ 2021. The reason we restrict our cross checking to one batch is that we first manually verify predicted phishing domains are in fact phishing, which is a time consuming task. Our manual inspection identified 21 false positives. We excluded them from the validation task. In order to validate the proactiveness of our classification, we scanned and compared remaining predicted phishing domains (471) against VT and GSB daily. Figure 12 shows the number of phishing domains detected daily by VT and GSB from Jan. $28^{th}$ to Feb. $4^{th}$ 2021. Each bar depicts the number of phishing domains marked by VT/GSB from our predicted list. The spike in the second scan is likely due to the fact that VT scanners update their engines based on the domains seen on the previous day. We observe a steady increase of 45 new domains being detected over the seven day period. Table 3 shows a sample of phishing domains predicted by our classifier that are failed to be detected by VT/GSB during the seven day study period.

We observe several interesting results from this experiment. After seven days of cross checking, as shown in Figure 12, VT/GSB was able to detect only 320 domains from our predicted set, indicating the concerning fact that VT/GSB are still lagging in detecting phishing domains as they rely on content to make the decision. Our manual inspection of predicted phishing domains that are not detected by VT/GSB (151 domains) show that most of these domains either utilize evasive techniques or impersonate popular brands such as Paypal, Amazon, and Apple but they are still at their early stage of web hosting. For example, as shown in Table 3 some websites are either under construction or parked. These inspections indicate that they are highly likely to be used for phishing attacks in the future. We also observe that a number of predicted phishing domains are detected by VT/GSB only after days, demonstrating the proactiveness of our approach and the improved detection latency we achieve compared to VT/GSB. A key reason for this is that VT/GSB rely on content and/or user activity in order to detect phishing domains whereas our approach can predict using content-agnostic features which are available even before the content is published or widely accessed.

[10] We select only the effective second level domains.

[11] For the live experiment, we trained the classifier with the balanced datasets.

| Feature Name | Description | Type |
|---|---|---|
| **CT Logs Temporal Features** | | |
| Lifetime | The difference between the expiration date of the last certificate and the issuance date of the first certificate | Numerical |
| Inter-arrival time | The time gap between the issue dates of every two consecutive certificates of a domain | Numerical |
| Number of certificates | The number of certificates obtained by a domain during its lifetime | Numerical |
| **CT Logs Certificate-Based Features** | | |
| SAN list size | The number of domains in the SAN field of a certificate for one-certificate domains, or the average number of domains in its SAN field from all certificates for multi-certificate domains | Numerical |
| Issuers | The distinct set of issuers associated with the certificates related to a domain | Categorical |
| Validation type | The most common validation type of the certificates associated with the domain including EV, OV and DV | Categorical |
| Validity period | The average validity period of the certificates associated with the domain | Numerical |
| **pDNS Features** | | |
| Duration | The time gap between the first seen and last seen records for a domain in the pDNS repository | Numerical |
| Number of IPs | The number of IPs on which the domain is hosted during the study period | Numerical |
| Number of queries | The number of times the domain is recorded in the passive pDNS repository, which is proportional to the popularity of the domain | Numerical |
| Number of name servers | The number of authoritative name servers associated with the domain | Numerical |
| Name server match | Does at least one name server domain matches with the domain name? | Boolean |
| Number of SOAs | Number of SOA domains associated with the domain | Numerical |
| SOA match | Does at least one SOA domain matches with the domain name? | Boolean |
| Number of domains | The average number of domains hosted on all hosting IPs of a domain | Numerical |
| **Lexical Features** | | |
| Relative entropy | The entropy of the characters in the domain name based on the Alexa top 10K domains | Numerical |
| Dictionary entropy | The dictionary-based entropy of the domain name | Numerical |
| Domain length | Length of the domain name | Numerical |
| Number of subdomains | Number of subdomains in the domain name | Numerical |
| Number of dashes | Number of dashes in the domain name | Numerical |

**Table 2: Summary of features used**

| Domain | Owner | Issuer | Hosting Provider | Brand | Status |
|---|---|---|---|---|---|
| vyoutube.com | Privacy Protected | Let's Encrypt | Confluence Networks | Youtube | Parked |
| instagramfor-fb.ml | Not Available | Let's Encrypt | OC1-Mochahost | Facebook | Under construction |
| l0gcahsbc.com | Privacy Protected | cPanel | VenomDC | HSBC | cgi-bin exposed |
| cuenta-netfiix.ru | Privacy Protected | Let's Encrypt | Ihor Hosting | Netflix | NX domain |
| supportt-paypal.com | Privacy Protected | Let's Encrypt | Amazon | PayPal | Up for sale |

**Table 3: A sample of predicted phishing domains not detected by VT/GSB as of Feb. $4^{th}$ 2021**

## 7 LIMITATIONS

One key challenge is acquiring clean groundtruth datasets. For the benign and phishing datasets, we cross checked our collected domains with external services such as VT, which uses a large number of blacklisting and scanning services (including GSB, and COMODO site inspector, etc). However, consistent with previous findings [67], we noticed that sometimes those sources maintain stale information, as sometimes benign domains are misreported. We address this issue by analyzing historical VT data in addition to the current online VT results. We also performed a significant load of manual checking when in doubt regarding some domains.

Another limitation is detecting phishing URLs. When a domain is hosted on a popular hosting service, or a popular public domain such as github.io, where anyone may create their own subdomains, querying certificates will return those of the popular public domain. Another related issue is URL shortening, which is increasingly popular for phishing campaigns. This opens the door for future work direction on how to identify phishing for such cases.

As with other machine learning approaches, periodically retraining detection classification models is mandatory to maintain accuracy and performance due to concept drifts. Since our feature set is small, and is not computationally expensive, retraining can be performed periodically utilizing an active learning approach. For example, as we confirm more phishing domains using our approach and content-based analysis, we can use the newly discovered domains as well as those available from external sources for retraining, and thus improve the classifier performance over time.

# 8 RELATED WORK

## 8.1 Content-based phishing detection

Approaches based on content analysis are vast. Whittaker et al. [85] developed a classifier that analyzes millions of pages a day, examining the URL and the contents of a page to determine whether or not a page is phishing. These methods [78, 85, 87] utilize features from the web page content itself to train a machine learning model to detect phishing URLs. For example, phishpedia [51] identifies logos on webpage screenshots and matches logo variants with the corresponding brand to identify phishing pages. While they are quite accurate, it is quite time consuming and resource intensive to train classifiers based on the content of web pages. Also, cloaked phishing domains can escape content-based analysis [25]. Therefore, a content agnostic approach based on other available data sources can aid previous work. Tian et al [78] proposed an approach to first detect squatting domains such as combosquatting, bitsquatting, TLD squatting and homophgraphs from passive DNS data and then train a content based classifier to identify phishing websites from the detected squatting domains. We have used this work to measure the percentage of squatting used in our datasets. While useful and can red flag squatting domains, a higher percentage of phishing domains are not detected (as they don't use squatting), but can be otherwise caught using pDNS- and CT- based features.

## 8.2 Non-content based phishing detection

These methods utilize features other than content based features such as URL/domain lexical features, WHOIS information, DNS information and hosting information [27, 28, 40, 47, 48, 53, 75, 81]. Most of these prior approaches try to classify URLs, whereas we focus only on domain names themselves which makes the problem harder as we have limited information in domains compared to URLs. Further, a key concern on prior work is that the type of datasets used are inadvertently biased with respect to the features based on the website URLs. We now describe some of the recent work in this area to make the related work complete. Bahnsen et al. [27] proposes machine learning models to predict phishing sites given URLs. They utilize two million known phishing and benign URLs to train a random forest classifier and a LSTM network to predict phishing domains. Shirazi et al. [75] uses features mostly related to domain names themselves, but they utilize a feature that checks if the domain name matches the web page title. Kumar et al. [47] proposes a context-free grammar based algorithm that models inconsistencies in domain names of banking websites and use it to generate potentially impersonating domains, which are later classified into defensive, malicious, suspicious and unrelated. Transparent phish [46] proposes a machine learning classifier that utilizes network-level features to detect MITM phishing web pages.

## 8.3 CT-based malicious domain detection

With the browser endorsement towards HTTPS on all sites, even phishing sites are forced to obtain certificates to present pages unsuspecting to general users. Hence, CT logs should be useful to detect phishing domains. There are several proprietary solutions such as Facebook [20] and CertSpotter [23], open source solutions such as phishing catcher [15] and solutions from academia such as CT-Honeypot [73] that provides notification services to users to get information about primarily combosquatting domains (e.g. paypal-mysite.com), Sakurai et al.'s approach [71] that clusters certificates with similar CNs to identify phishing domains and Drichel et al.'s approach [38] trains a random forest classifier utilizing direct certificate features and lexical features. However, it is not clear what the detection classification performance for the clustering based approach and the TPR of Drichel et al.'s approach is quite low. While CT-Honeypot relies on content based classification, and the certificate clustering and Drichel et al.'s approaches rely on the certificate features to train classifiers. The remaining approaches mentioned above utilize a rule based approach to filter squatting domains. While they are useful to spot likely phishing domains early, they are also inundated with many false positive notifications that make security administrators life harder. Further, rule based approaches are difficult to maintain as one needs to manually change as Internet miscreants change their phishing tactics.

## 8.4 X.509 certificate based phishing detection

There are a number of past research that utilizes X.509 certificate datasets to detect phishing URLs/domains [37, 59]. While these approaches do utilize X.506 certificate features, our work significantly differs from these approaches due to several reasons: (1) Most of previous work rely on Alexa top list to create the benign data set, which can be biased as we have seen in our analysis, (2) In addition to X.509 certificate features, we utilize temporal aggregate CT logs as well as pDNS and lexical features in order to classify domains with high accuracy.

## 8.5 Malicious domain detection

There have been many research efforts in the past to detect malicious domains using DNS data [44], HTTP logs [30, 54], and enterprise logs [66]. They broadly fall under two categories: classification based [30] and inference based [44, 54]. In a classification based approach, features related to domains are extracted from the above mentioned sources and a binary classifier is trained to mark each domain as malicious or benign. Inference based approaches, on the other hand, build a graph of domains under consideration and infer the maliciousness of domains based on known malicious and benign domains. While each approach has pros and cons, their goal is to find any type of malicious domains, whereas we focus on phishing domains in this study. Further, these schemes are enriched using auxiliary information such as Whois information [41], IP ASN information [55], IP geolocation information, and hosting information. In our study we mainly focus CT log based features and our goal is to detect phishing domains as early in their domain life cycle as possible.

## 9 CONCLUSION

In this paper, we create, and thoroughly verify phishing and realistic benign domain datasets. We provide an extensive analysis and comparison using their pDNS, CT traces and lexical characteristics. Our analysis provides various insights and sheds light on how phishing domains can be distinguished from benign domains, even when they are not Alexa top domains. Our analysis paves the way for content agnostic prediction approaches. To evaluate the usefulness of our observations, we extract features that we use to train a random forest classifier and we indeed show high precision and recall for both benign and phishing domains. We also demonstrate the possibility of building proactive detection solutions based on CT logs which can detect phishing domains even before mainstream tools like VT.

## REFERENCES

[1] 2019. Anti-Phishing Working Group. https://apwg.org.

[2] 2019. CDN Planet CDN List. https://www.cdnplanet.com/cdns/. [Online; accessed 24-05-2021].

[3] 2019. Certificate Transparency. https://developers.facebook.com/docs/certificate-transparency/. Accessed April 2022.

[4] 2019. Chrome and Firefox Changes Spark the End of EV Certificates. https://www.bleepingcomputer.com/news/software/chrome-and-firefox-changes-spark-the-end-of-ev-certificates/.

[5] 2019. Comodo Free SSL Certificate. https://www.comodo.com/e-commerce/ssl-certificates/free-ssl-certificate.php.

[6] 2019. Getting Started. https://letsencrypt.org/getting-started/.

[7] 2019. Public Suffix List. https://publicsuffix.org/. [Online; accessed 24-05-2021].

[8] 2019. Wombat Security The State of the Phish Report 2019. https://www.wombatsecurity.com/state-of-the-phish/. Accessed April 2022.

[9] 2019. WPO Foundation CDN List. https://github.com/WPO-Foundation/webpagetest/blob/master/agent/wpthook/cdn.h. [Online; accessed 24-05-2021].

[10] 2021. https://gdpr.eu.

[11] 2021. CIRCL Passive DNS. https://www.circl.lu.

[12] 2021. COMODO SSL Analyzer. https://sslanalyzer.comodoca.com. Accessed April 2022.

[13] 2021. crt.sh Certificate Search. https://crt.sh. Accessed April 2022.

[14] 2021. CT Enforcement in Google Chrome. https://tinyurl.com/y2nyyjtm. Accessed February 2021.

[15] 2021. Phishing catcher. https://github.com/x0rz/phishing_catcher.

[16] 2021. Phishtank. Out of the Net, into the Tank. https://www.phishtank.com. Accessed April 2022.

[17] 2021. The Domain Block List (DBL). https://www.spamhaus.org/dbl/. Accessed April 2022.

[18] 2021. What Services Does Let's Encrypt Offer? https://letsencrypt.org/docs/faq/. Accessed May 2021.

[19] 2022. Certificate Transparency. https://www.certificate-transparency.org/. Accessed April 2022.

[20] 2022. Facebook Certificate Transparency Tool. https://developers.facebook.com/docs/certificate-transparency/. Accessed April 2022.

[21] 2022. Google Safe Browsing: Making the world's information safely accessible. https://safebrowsing.google.com. Accessed April 2022.

[22] 2022. Mcafee Labs Threat Report December 2018. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-dec-2018.pdf. Accessed April 2022.

[23] 2022. SSL Mate Certspotter. https://sslmate.com/certspotter/. Accessed April 2022.

[24] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J. Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, Seth Schoen, and Brad Warren. 2019. Let's Encrypt: An Automated Certificate Authority to Encrypt the Entire Web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. ACM, New York, NY, USA, 2473–2487. https://doi.org/10.1145/3319535.3363192

[25] Bhupendra Acharya and Phani Vadrevu. 2021. PhishPrint: Evading Phishing Detection Crawlers by Prior Profiling. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 3775–3792. https://www.usenix.org/conference/usenixsecurity21/presentation/acharya

[26] Maarten Aertsen, Maciej Korczyński, Giovane C. M. Moura, Samaneh Tajalizadehkhoob, and Jan van den Berg. 2017. No Domain Left Behind: Is Let's Encrypt Democratizing Encryption?. In *Proceedings of the Applied Networking Research Workshop* (Prague, Czech Republic) *(ANRW '17)*. ACM, New York, NY, USA, 48–54. https://doi.org/10.1145/3106328.3106338

[27] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González. 2017. Classifying phishing URLs using recurrent neural networks. In *2017 APWG Symposium on Electronic Crime Research (eCrime)*. 1–8.

[28] A. C. Bahnsen, U. Torroledo, D. Camacho, and S. Villegas. 2018. DeepPhish: Simulating Malicious AI. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*. 1–8.

[29] BEN DOWNING. 2021. Using Entropy in Threat Hunting: a Mathematical Search for the Unknown. https://redcanary.com/blog/threat-hunting-entropy/. Accessed February 2021.

[30] Leyla Bilge, Sevil Sen, Davide Balzarotti, Engin Kirda, and Christopher Kruegel. 2014. Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains. *ACM Transactions on Information and System Security* 16, 4 (apr 2014), 14:1–14:28.

[31] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32.

[32] CA Browser Forum. 2021. Baseline Requirements. https://cabforum.org/wp-content/uploads/CA-Browser-Forum-BR-1.6.6.pdf. Accessed Jan 2021.

[33] CA / Browser Forum. 2022. Object Registry of the CA / Browser Forum. https://cabforum.org/object-registry/. Accessed April 2022.

[34] CaliDog. 2022. CertStream Python. https://github.com/CaliDog/certstream-python. Accessed April 2022.

[35] Censys. 2022. See Your Entire Attack Surface in Real Time. https://censys.io. Accessed April 2022.

[36] Chromium. 2021. EV OID list. https://chromium.googlesource.com/chromium/src/net/+/master/cert/ev_root_ca_metadata.cc/. Accessed February 2021.

[37] Zheng Dong, Apu Kapadia, Jim Blythe, and L Camp. 2015. Beyond the lock icon: Real-time detection of phishing websites using public key certificates. *eCrime Researchers Summit, eCrime* 2015 (06 2015). https://doi.org/10.1109/ECRIME.2015.7120795

[38] Arthur Drichel, Vincent Drury, Justus von Brandt, and Ulrike Meyer. 2021. Finding Phish in a Haystack: A Pipeline for Phishing Classification on Certificate Transparency Logs. In *The 16th International Conference on Availability, Reliability and Security* (Vienna, Austria) *(ARES 2021)*. Article 59, 12 pages.

[39] Farsight Security, Inc. 2022. DNS Database. https://www.dnsdb.info/. Accessed April 2022.

[40] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. 2007. A Framework for Detection and Measurement of Phishing Attacks. In *Proceedings of the 2007 ACM Workshop on Recurring Malcode*. ACM, New York, NY, USA, 1–8.

[41] J. Gargano and K. Weiss. 1995. *Whois and Network Information Lookup Service, Whois++*. RFC 1834. RFC Editor. http://www.rfc-editor.org/rfc/rfc1834.txt.

[42] Josef Gustafsson, Gustaf Overier, Martin F. Arlitt, and Niklas Carlsson. 2017. A First Look at the CT Landscape: Certificate Transparency Logs in Practice. In *PAM*.

[43] Ryan Hurst. 2012. How to Tell DV and OV Certificates Apart. http://unmitigatedrisk.com/?p=203.

[44] Issa M. Khalil, Bei Guan, Mohamed Nabeel, and Ting Yu. 2018. A Domain is Only As Good As Its Buddies: Detecting Stealthy Malicious Domains via Graph Inference. In *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy* (Tempe, AZ, USA) *(CODASPY '18)*. ACM, New York, NY, USA, 330–341. https://doi.org/10.1145/3176258.3176329

[45] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Rosa Romero-Gómez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. 2017. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 569–586.

[46] Brian Kondracki, Babak Amin Azad, Oleksii Starov, and Nick Nikiforakis. 2021. Catching Transparent Phish: Analyzing and Detecting MITM Phishing Toolkits. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (Virtual Event, Republic of Korea) *(CCS '21)*. Association for Computing Machinery, New York, NY, USA, 36–50.

[47] Neeraj Kumar, Sukhada Ghewari, Harshal Tupsamudre, Manish Shukla, and Sachin Lodha. 2021. When Diversity Meets Hostility: A Study of Domain Squatting Abuse in Online Banking. In *2021 APWG Symposium on Electronic Crime Research (eCrime)*. 1–15. https://doi.org/10.1109/eCrime54498.2021.9738769

[48] Anh Le, Athina Markopoulou, and Michalis Faloutsos. 2011. PhishDef: URL names say it all. *2011 Proceedings IEEE INFOCOM* (2011), 191–195.

[49] S. Le Page, G. Jourdan, G. V. Bochmann, J. Flood, and I. Onut. 2018. Using URL shorteners to compare phishing and malware attacks. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*. 1–13. https://doi.org/10.1109/ECRIME.2018.8376215

[50] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. 2016. Domain-Z: 28 Registrations Later Measuring the Exploitation of Residual Trust

in Domains. In *2016 IEEE Symposium on Security and Privacy (SP)*. 691–706. https://doi.org/10.1109/SP.2016.47

[51] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. 2021. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 3793–3810. https://www.usenix.org/conference/usenixsecurity21/presentation/lin

[52] Chaoyi Lu, Baojun Liu, Yiming Zhang, Zhou Li, Fenglu Zhang, Haixin Duan, Y. Liu, J. Chen, Jinjin Liang, Z. Zhang, S. Hao, and Min Yang. 2021. From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR. In *NDSS*.

[53] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2009. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. In *Proceedingsof theSIGKDD Conference. Paris,France.*

[54] Pratyusa K. Manadhata, Sandeep Yadav, Prasad Rao, and William Horne. 2014. Detecting Malicious Domains via Graph Inference. In *Proceedings of the 19th European Symposium on Research in Computer Security,*, Mirosław Kutyłowski and Jaideep Vaidya (Eds.). Springer International Publishing, Cham, 1–18.

[55] MaxMind. 2022. GeoLite2 Databases. http://www.maxmind.com. Accessed April 2022.

[56] D. Kevin McGrath and Minaxi Gupta. 2008. Behind Phishing: An Examination of Phisher Modi Operandi. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats* (San Francisco, California) *(LEET'08)*. USENIX Association, Berkeley, CA, USA, Article 4, 8 pages. http://dl.acm.org/citation.cfm?id=1387709.1387713

[57] Ulrike Meyer and Vincent Drury. 2019. Certified Phishing: Taking a Look at Public Key Certificates of Phishing Websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/soups2019/presentation/drury

[58] Najmeh Miramirkhani, Timothy Barron, Michael Ferdman, and Nick Nikiforakis. 2018. Panning for gold.com: Understanding the Dynamics of Domain Dropcatching. 257–266.

[59] Mishari Al Mishari, Emiliano De Cristofaro, Karim M. El Defrawy, and Gene Tsudik. 2012. Harvesting SSL Certificate Data to Identify Web-Fraud. *I. J. Network Security* 14, 6 (2012), 324–338.

[60] Mohamed Nabeel, Issa M. Khalil, Bei Guan, and Ting Yu. 2020. Following Passive DNS Traces to Detect Stealthy Malicious Domains Via Graph Inference. *ACM Trans. Priv. Secur.* 23, 4, Article 17 (July 2020), 36 pages. https://doi.org/10.1145/3401897

[61] Network Solutions, LLC. 2022. Network Solutions Certification Practice Statement. https://assets.web.com/legal/English/CertificationPracticeStatement.pdf. Accessed April 2022.

[62] Amirreza Niakanlahiji, Bei-Tseng Chu, and Ehab Al-Shaer. 2018. PhishMon: A Machine Learning Framework for Detecting Phishing Webpages. 220–225. https://doi.org/10.1109/ISI.2018.8587410

[63] A. Oest, Y. Safaei, A. Doupé, G. Ahn, B. Wardman, and K. Tyers. 2019. PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques against Browser Phishing Blacklists. In *2019 IEEE Symposium on Security and Privacy (SP)*. 1344–1361. https://doi.org/10.1109/SP.2019.00049

[64] Adam Oest, Yeganeh Safaei, Penghui Zhang, Brad Wardman, Kevin Tyers, Yan Shoshitaishvili, and Adam Doupé. 2020. PhishTime: Continuous Longitudinal Measurement of the Effectiveness of Anti-phishing Blacklists. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 379–396. https://www.usenix.org/conference/usenixsecurity20/presentation/oest-phishtime

[65] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. 2020. Sunrise to Sunset: Analyzing the End-to-end Life Cycle and Effectiveness of Phishing Attacks at Scale. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 361–377. https://www.usenix.org/conference/usenixsecurity20/presentation/oest-sunrise

[66] A. Oprea, Z. Li, T. F. Yen, S. H. Chin, and S. Alrwais. 2015. Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data. In *Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*. 45–56.

[67] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. 2019. Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines. In *Proceedings of the Internet Measurement Conference* (Amsterdam, Netherlands) *(IMC '19)*. Association for Computing Machinery, New York, NY, USA, 478–485. https://doi.org/10.1145/3355369.3355585

[68] J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (01 Mar 1986), 81–106. https://doi.org/10.1007/BF00116251

[69] 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)*. ACM, New York, NY, USA, 2489–2504. https://doi.org/10.1145/3319535.3363188

[70] A. P. E. Rosiello, E. Kirda, 2. Kruegel, and F. Ferrandi. 2007. A Layout-Similarity-Based Approach for Detecting Phishing Pages. In *SecureComm*. 454–463.

[71] Yuji Sakurai, Takuya Watanabe, Tetsuya Okuda, Mitsuaki Akiyama, and Tatsuya Mori. 2020. Discovering HTTPSified Phishing Websites Using the TLS Certificates Footprints. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*. 522–531. https://doi.org/10.1109/EuroSPW51379.2020.00077

[72] Quirin Scheitle, Oliver Gasser, Theodor Nolte, Johanna Amann, Lexi Brent, Georg Carle, Ralph Holz, Thomas C. Schmidt, and Matthias Wählisch. 2018. The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. ACM, New York, NY, USA, 343–349. https://doi.org/10.1145/3278532.3278562

[73] Quirin Scheitle, Oliver Gasser, Theodor Nolte, Johanna Amann, Lexi Brent, Georg Carle, Ralph Holz, Thomas C. Schmidt, and Matthias Wählisch. 2018. The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem. *CoRR* abs/1809.08325 (2018). arXiv:1809.08325 http://arxiv.org/abs/1809.08325

[74] Scott Helme. march 06, 2017. Let's Encrypt are enabling the bad guys, and why they should. https://scotthelme.co.uk/lets-encrypt-are-enabling-the-bad-guys-and-why-they-should/. Accessed April 2022.

[75] Hossein Shirazi, Bruhadeshwar Bezawada, and Indrakshi Ray. 2018. "Kn0W Thy Doma1N Name": Unbiased Phishing Detection Using Domain Name Based Features. In *Proceedings of the 23Nd ACM on Symposium on Access Control Models and Technologies* (Indianapolis, Indiana, USA) *(SACMAT '18)*. ACM, New York, NY, USA, 69–75. https://doi.org/10.1145/3205977.3205992

[76] Statcounter GlobalStats. 2022. Browser Market Share Worldwide. https://certs.securetrust.com/CA/twcps2_9.pdf. Accessed April 2022.

[77] J.A.K. Suykens and J. Vandewalle. 1999. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9, 3 (01 Jun 1999), 293–300. https://doi.org/10.1023/A:1018628609742

[78] Ke Tian, Steve T. K. Jan, Hang Hu, Danfeng Yao, and Gang Wang. 2018. Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. In *Proceedings of the Internet Measurement Conference 2018, IMC 2018, Boston, MA, USA, October 31 - November 02, 2018*. 429–442.

[79] Ke Tian, Steve T. K. Jan, Hang Hu, Danfeng Yao, and Gang Wang. 2018. Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. ACM, New York, NY, USA, 429–442. https://doi.org/10.1145/3278532.3278569

[80] Ivan Torroledo, Luis David Camacho, and Alejandro Correa Bahnsen. 2018. Hunting Malicious TLS Certificates with Deep Neural Networks. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (Toronto, Canada) *(AISec '18)*. ACM, New York, NY, USA, 64–73. https://doi.org/10.1145/3270101.3270105

[81] Rakesh Verma and Keith Dyer. 2015. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy* (San Antonio, Texas, USA) *(CODASPY '15)*. ACM, New York, NY, USA, 111–122. https://doi.org/10.1145/2699026.2699115

[82] VirusTotal, Subsidiary of Google. 2022. VirusTotal – Free Online Virus, Malware and URL Scanner. https://www.virustotal.com/. Accessed April 2022.

[83] Florian Weimer. 2005. Passive DNS Replication. In *FIRST Conference on Computer Security Incident*. 98.

[84] Daniel Lowe Wheeler. 2016. zxcvbn: Low-Budget Password Strength Estimation. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, 157–173. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/wheeler

[85] Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-Scale Automatic Classification of Phishing Pages. In *NDSS '10*. http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf

[86] Sandeep Yadav, Ashwath Kumar Krishna Reddy, A.L. Narasimha Reddy, and Supranamaya Ranjan. 2010. Detecting Algorithmically Generated Malicious Domain Names. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (Melbourne, Australia) *(IMC '10)*. ACM, New York, NY, USA, 48–61. https://doi.org/10.1145/1879141.1879148

[87] Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. 2007. Cantina: A Content-based Approach to Detecting Phishing Web Sites. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, New York, NY, USA, 639–648.