Analyzing Certificate Transparency Logs:
Evaluation of domains and timestamps

Hannes Tuhkala
Linköping University
Linköping, Sweden

Axel Karlsson
Linköping University
Linköping, Sweden

Niklas Carlsson
Department of Computer and Information Science (IDA)
Linköping University
Linköping, Sweden

***Abstract*—Certificate transparency has been of interest to domain owners whom host websites, and for certificate authorities that issues new certificates. This paper looks into certificates that were gathered from Certificate Transparency Logs and a database of phishing domains. Analysis was then performed on these certificates by first extracting information such as domain names and the time that they are valid for. An algorithm was used on the domain name associated with the certificate, to see whether it is suspicious or not. It uses a scoring system that determines whether a certificate could be suspected of being malicious. The results show that the algorithm used is not without fault but is still usable in order to detect malicious websites, and some reoccurring keywords and patterns are found when using it.**

I. Introduction

There are billions of people using the Internet everyday. To be able to use the internet securely and trustworthy, certificates are being used. With the help of certificates, it allows browsers to detect whether a website is safe to use and increases the chances that it can be trusted. This way the user is prevented from accessing sites that may be harmful and that could possibly provide a threat to the user. These certificates are issued by Certificate Authorities (CA). Certificate Authorities are organizations that issue certificates for domains, and which browsers in most cases trust. These CA:s can be Google, Cloudflare or DigiCert, for example. The purpose of this paper is to download certificates from Certificate Transparency Logs (CTL) and extracting certain data from them to analyze. In particular, the domain name and for how long time the certificate is valid for. An algorithm is used to process the domain name to determine whether it is suspicious or not. These suspicious domains are often phishing links, and as such, attackers may now use certificates to make their websites look legitimate. Since users can be misinformed and assume that if a website has a certificate - then it is trustworthy.

II. Background

Certificate Transparency is used to provide an open auditing and monitoring system for users to be able to look up whether an TLS/SSL certificate was issued to a domain mistakenly, or worse, maliciously. CT manages this by keeping a public log of issued certificates, monitoring and auditing.

The goals of CT, according to their website [1]:

- To make it difficult for a CA to issue a TLS/SSL certificate for a domain without the certificate being visible to the owner of that domain.
- Provide an open auditing and monitoring system that lets any domain owner or CA determine whether certificates have been mistakenly or maliciously issued.
- Protect users (as much as possible) from being duped by certificates that were mistakenly or maliciously issued.

In order to fulfill these goals, CT provides an open-source framework that can monitor and audit certificates.

A. Certificate Transparency Logs

The Certificate Transparency Logs (CTL) keeps track of all issued certificates. The CTL is append-only, meaning that it is not possible to remove any issued certificates. This is because of the special structure that is used - a Merkle tree. A Merkle tree is a binary tree where every leaf node is a hash of a data block, in this case a certificate and every internal node is a hash of both its child nodes. These log servers should be operated independently, to prevent a majority of them being owned by a single entity or organization.

B. Monitoring

Certificate Transparency monitors the certificates in real time in order to detect SSL certificates that, e.g. may have been maliciously acquired. It enables the possibility to see if there have been any certificates that have been mistakenly issued by a certificate organization.

C. Auditing

Auditors verify that logs are consistent and whether a particular certificate exists in a log. When connecting to a website and you discover a certificate that does not appear in a log, then it is very suspicious and should not be approved.

III. Methodology

One of the datasets that was used in the paper was taken from the CTL *Argon 2018* [2]. This dataset will be referred to as the large dataset in the text. A python program was used to download these certificates from the CTL, called Axeman [3], with some small modifications. This dataset contains approximately 17.8% of the total amount of the logs. A total of 76,290,560 unique certificates was gathered.

Another dataset was gathered by listening in realtime using Certstream [4] during the 4th of May, 2019. A total of 1,520,618 unique certificates was downloaded. This dataset will be refered to as the small dataset.

A third dataset was downloaded from Phishtank [5] which contained 7824 confirmed, unique phishing links. This set will

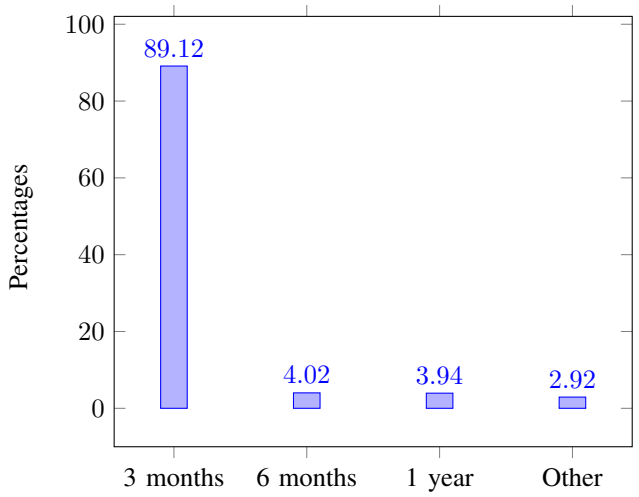

Fig. 1: A bar chart over the different timespans found in the large dataset. Shown in percentages for each of the timespans.

be refered to as the phishing dataset and will be a measure of how accurate the program mentioned below is.

The algorithm used in the program Phishing Catcher [6] was then used on these three sets. This program was used because it checked newly issued certificates to the CTLs, and some modifications were made to be able to run it offline. Afterwards statistics was done on the output, which includes the different criteria the algorithm used.

IV. Analysis of certificate transparency logs

The information that we focused on extracting consisted of:

- The timespan that the certificates were valid for
- The domain name

A. Timespans

The timespans indicate how long the certificates are valid for. The timespans vary in the different certificates and we will see whether there exists any certain pattern to it. A few are only a week long, but many may be up to three years. Figure 1 shows how the timespans varied in the large dataset.

The bar chart shows the percentage (y-axis) of how many percent of the timespans that were in a particular time amount (x-axis). The x-axis shows periods for three months, six months, one year and other, where other contains any other timespan. The data grouped into three months are timelapses of 90 and 91 days, for one year it is timelapses of 365 and 366 days. The vast majority of the certificates were valid for three months, and there was not a big difference between the certificates that were valid for longer than this. There were some outliers in the data that had unusual timespans, such as 17 days and 4 days, i.e. it does not match ”normal” timespans such as 1 week, 2 weeks, 1 month, 1 year, etc.

B. Phishing

To tackle the attack vector of phishing with regards to Certificate Transparency a scoring system was used based on the domain name. This is what the algorithm in Phishing catcher uses.

Criteria	Score
Suspicious TLD	20
Entropy	Shannon entropy * 50
Keyword	Depends on keyword
Levenshtein distance	70 for each similar word found
Hyphens	Total hyphens * 3
Nested subdomains	Level of depth * 3

TABLE I: Shows the different points assigned to each criteria.

- If a domain has a high entropy, that is, a very long domainname
- If it contains suspicious keywords, such as: authenticate, gmail, hotmail, paypal, etc
- If it contains many hyphens (’-’), e.g. www.authenticate-apple-com.ga
- If it contains deeply nested sub-domains, e.g. www.hotmail.com.login.authenticate.ga
- If it is a word close enough to a popular domain (Levenshtein distance), e.g. youltube, paypol, etc.
- If it uses some TLD (Top Level Domain) that might be unusual, such as .tk, .ga, etc.

By assigning different amounts of points to these criteria we can have a threshold whether a domain is suspicious or not. There were some false positives, but some can be remedied by whitelisting trusted domains, e.g. anything *.microsoft.com, *.apple.com, etc.

Table I shows how different amount of points have been assigned to each criteria.

Any domain that passes the threshold of 75 is flagged by the program as suspicious. Domains that get a score between 65 and 75 are flagged as potential to be malicious, between 80 and 90 as likely to be malicious, and finally if it gets a score over 90 then it is branded as very suspicious.

For example, the domain *www.gmail.folsom.pw* had a score of 120 and is therefore classified as very suspicious.

Criteria	Count	Percentage
Keywords	3753	98.58%
Subdomains	917	24.09%
TLD	364	9.56%
Hyphens	144	3.78%
Levenshtein	56	1.47%

TABLE II: Shows the count and percentage of the different criteria that was flagged for the suspicious domains.

1) Analysis of the small dataset: A total of 1,520,618 domains was checked against phishing catcher and the output returned 3,807 suspicious domains, or 0.25% of the total amount of domains.

Table II shows the count of each criteria, and is a measure of which criteria the suspicious domains had in common. Out of the 3807 domains, 24.1% of them had subdomains, 98.6% used hyphens, and so on.

Keyword	Count	Percentage	Category
google	672	17.65%	Website
whatsapp	671	17.63%	Website
.com.	506	13.29%	Fake TLD
amazon	343	9.01%	Website
account	280	7.36%	Action
facebook	195	5.12%	Website
confirm	158	4.15%	Action
blockchain	153	4.02%	Action
apple	150	3.94%	Website
microsoft	144	3.78%	Website

TABLE III: Most common keywords for the suspicious domains in the small dataset. The count, percentage and category of each keyword is displayed.

Table III shows the total count of the most common keywords. *Google* is used the most, with *whatsapp* and *.com.* close after.

For the domains that used subdomains, 344 (37.5%) used four subdomains, 289 (31.52%) used three subdomains and 131 (14.29%) used five subdomains. 153 (16.68%) domains used six or more subdomains.

Criteria	Count	Percentage
Keywords	124,233	97.0%
TLD	31,530	24.62%
Subdomains	13,204	10.31%
Hyphens	7,946	6.20%
Levenshtein	6270	4.90%

TABLE IV: Shows the count of the different criteria that was flagged for the suspicious domains.

2) Analysis of the large dataset: Out of the 76,590,260 certificates, 128,080 (0.17%) were found as suspicious, as in having a score over 75.

The data from Table IV shows the count of each criteria, and is a measure of which criteria the suspicious domains had in common. Out of the 128,080 suspicious domains, 97.0% of them used keywords, 24.6% used a suspicious TLD, 10.3% used subdomains, 6.2% used hyphens, and 4.9% had words similar to popular words.

Keyword	Count	Percentage	Category
account	18,588	14.51%	Action
apple	16,752	13.08%	Website
amazon	15,891	12.41%	Website
appleid	9,755	7.62%	Service
.com.	8,530	6.66%	Fake TLD
service	8,136	6.35%	Action
.com-	8,032	6.27%	Fake TLD
google	7,867	6.14%	Website
support	7,457	5.82%	Action
paypal	7,383	5.76%	Website

TABLE V: Most common keywords for the suspicious domains in the large dataset. The count, percentage and category of each keyword is displayed.

Table V shows the total count of the most common keywords in this dataset. *Account, apple,* and *amazon* are used the most, they account for 30% of the most common keywords.

Out of these 128,080 suspicious domains, 13,204 used subdomains. From these 13,204 domains, 8,268 (62.62%) of them used three subdomains, 2,770 (20.98%) used four, 1,138 (8.62%) used five, 411 (3.12% used six and 617 (4.67%) used seven or more. This can be seen in Figure 2.

For the domains that flagged the hyphen criteria, 4,940 (62.12%) domains used four, 1,869 (23.52%) used five, 814 (10.24%) used six, 323 (4.06%) used seven or more.

3) Analysis of the phishing dataset: Out of the 7,819 bad domains from the phishing dataset, only 346 (4.4%) were marked as suspicious by the algorithm.

Criteria	Count	Percentage
Keywords	341	98.55%
Subdomains	99	28.61%
TLD	32	9.24%
Hyphens	18	5.2%
Levenshtein	9	2.6%

TABLE VI: Shows the count and percentage of the different criteria that was flagged for the suspicious domains.

The data from Table VI shows the counts of each criteria, and is a measure of which criteria the suspicious domains had in common. Out of the 346 suspicious domains, 98.55% of them used keywords, 9.2% used a suspicious TLD, 28.61% used subdomains, 5.2% used hyphens and 2.6% had words similar to popular words.

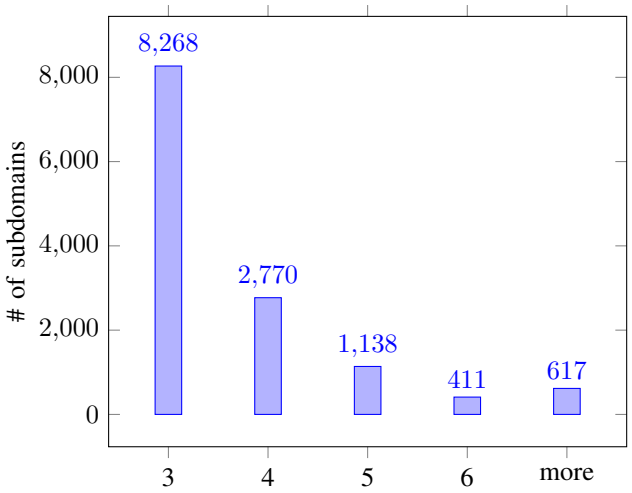

Fig. 2: Shows the number of subdomains that was used for the suspicious domains.

Keyword	Count	Percentage	Category
paypal	93	27.27%	Website
.com.	78	22.87%	Fake TLD
account	44	12.90%	Action
apple	43	12.61%	Website
login	37	10.85%	Action
appleid	34	9.97%	Service
.com-	28	8.21%	Fake TLD
support	25	7.33%	Action
tumblr	25	7.33%	Website
microsoft	23	6.74%	Website

TABLE VII: Most common keywords for the suspicious domains in the phishing dataset. Also shown is which category the keyword fits into and the percentage it had on the total keywords found.

The total count for the most common keywords in this dataset is shown in Table VII. *paypal* and *.com* are used the most, with several more with a quite high count for this small dataset.

The domains that were flagged by the subdomains criteria had 40 (40.40%) domains using three subdomains, 15 (15.15%) domains using four subdomains, 14 (14.14%) using five subdomains, 30 (30.30%) using six or more subdomains.

C. Evaluation and Comparison

The time that a certificate is valid for seems for the most part hold a minimum length of three months. This appears to be the standard for most certificates with a few exceptions where certificates where valid for a longer time. The results do not show exactly why this specific timespan is common, but it could be because these certificates were issued by Let's Encrypt [7]. However, a correlation between keeping the hashing key valid and the timeframe for the certificate to be valid may exist.

The scoring system that was used is not flawless. It finds some false positives and misses some domains. The only way to be completely accurate is still by manually analyzing each domain. Although a tool that may be used for this task would be of significant help. The actual points assigned to each criteria does not have any ground to it except for personal experience. A study looking at actual confirmed phishing links and finding common denominators, similar to what has been done in this paper, may provide better results.

The criteria that all of the datasets have in common the most, is keywords. It got the highest count in each of the sets. *apple, account* and *.com.* appears in all. Some that appeared in two of the sets are *support, paypal, appleid, .com-* and *amazon*. However, these keywords that were found are only for domains that scored high enough to be flagged as suspicious. As such, there may be phising domains that slipped by, which can be seen when running the algorithm with the phishing dataset as input. The Levenshtein criteria should give mostly correct results since it looks for words similar to paypal, apple, etc. which legit domains would not use.

Since the data from the phishing dataset are confirmed phishing links, the algorithm could be improved by looking at the domains that does not get flagged.

V. Related Work

This section provides papers that are related to the project and also contains similar topics to the work that has been done in this paper.

Gustafson et al. [8] aims to give an insight and a characterization of CT-logs. The paper goes through and describes the overview of CT-logs. It also describes how they in the paper used CT-monitors to collect and characterize CT-logs that have their similarities compared to each other.

Dowling et al. [9] looks into the security of certificate transparency and also the security of the CT-logs. It analyzes the cryptographic mechanisms of the CT-logs and shows how these mechanisms work to prevent undetected misbehaviour of a log server and also how it it prevents honest loggers to be falsely accused.

VI. Conclusions

The valid timespan for a certificate appears for the most part be three months, otherwise six months respectively one year long where the most common lengths of a certificates validity. The certificates that were valid for three months might have been issued by Let's Encrypt, which is the default time set.

By analyzing the suspicious domains found by Phishing catcher, some keywords have been found such as *apple, .com., account*. Many of the keywords appeared in more than one of the datasets. It also showed that keywords, subdomains and unusual TLDs was widely used in the suspicious domains

found. Most of the suspicious domains found used many subdomains instead of many hyphens.

The algorithm that was used is not completely accurate. It finds false positives and misses actual phishing links. The only way to be accurate is still by manually analyzing domains which in turn could be very time consuming. However, it could still be of use, for example by lowering the treshold and/or improving the algorithm and then manually analyze the output.

REFERENCES

[1] Certificate-transparency.org. (n.d.). Certificate Transparency. [online] Available at: https://www.certificate-transparency.org/ [Accessed 4 May 2019].

[2] Google, Argon 2018 CTL. [online] Available at: ”ct.googleapis.com/logs/argon2018/” [Accessed 6 May 2019]

[3] Calidog, Axeman. [online] Available at: https://github.com/CaliDog/Axeman [Accessed 20 April 2019]

[4] Certstream, Cali Dog Security. [online] Available at: https://certstream.calidog.io/ [Accessed 20 April]

[5] Phishtank.com. (n.d.). PhishTank — Join the fight against phishing. [online] Available at: https://www.phishtank.com/ [Accessed 6 May 2019].

[6] Xorz, Phishing catcher. [online] Available at: https://github.com/x0rz/phishing_catcher [Accessed 20 April 2019]

[7] 1. Why ninety-day lifetimes for certificates? - Let’s Encrypt - Free SSL/TLS Certificates [Internet]. Letsencrypt.org. 2019 [cited 20 May 2019]. Available from: https://letsencrypt.org/2015/11/09/why-90-days.html

[8] Gustafsson, Josef and Overier, Gustaf and Arlitt, Martin and Carlsson, Niklas, “A first look at the CT landscape: Certificate transparency logs in practice” International Conference on Passive and Active Network Measurement, Springer, 2017.

[9] Dowling, Benjamin and Günther, Felix and Herath, Udyani and Stebila, Douglas, “Secure logging schemes and certificate transparency”, European Symposium on Research in Computer Security, 2016