Structure and Origin of CT Based Domain Lists

Lorenz Lehle, Patrick Sattler*, Johannes Zirngibl*
*Chair of Network Architectures and Services
School of Computation, Information and Technology, Technical University of Munich, Germany
Email: lorenz.lehle@tum.de, sattler@net.in.tum.de, zirngibl@net.in.tum.de

Abstract—**The participation of Certificate Authorities in Certificate Transparency has the side effect of publicly accessible certificates from which domain lists can be derived. In this paper we analyse such domain lists and lay special focus on wildcard domains because of their ability to prevent information leakage. We find that of all domains in the domain list 18.0 % are wildcard domains, and 8.6 % are wildcard domains at the first level. Most of these immediate wildcard domains appear in a certificate with the corresponding eSLD and no other domain. Furthermore we find several patterns in the distribution of domains, such as more eSLD with an even rather than an odd number of subdomains. Additionally we find that the leftmost labels of domains correlate with used services and reveal more information about the domain holders internal structure.**

Index Terms—**domain lists, certificate transparency logs, internet scans**

1. Introduction

Domain lists are an integral part of many areas of research regarding network architectures. Researchers conduct Internet measurements to obtain information about the distribution of IP addresses, reachable hosts, and used protocols or services. Conventionally such scans require iterating through the whole IP address range, which is impossible for IPv6 due to the vast and largely unused address space. This is where domain lists prove useful. They contain domains of actually existing and used services, thereby alleviating the necessity to scan all addresses.

An alternative to these conventional domain lists are Certificate Transparency (CT) based domain lists. These are obtained by extracting the domain names noted in certificates issued by Certificate Authorities (CAs). In this paper we present an overview of Certificate Transparency and how CT based domain lists are created. Furthermore we analyse the structure and labels of domains that are contained in domain lists available to the *Chair of Network Architectures and Services*.

Firstly we present the necessary background about Certificate Transparency and introduce the terminology used for domain names. In Section 3 we present studies that examine the applicability of CT log based domain lists and security implications of CT logs. In Section 4 we analyse the structure and distribution of domains contained in a CT log based domain list. For this purpose we explore which domain labels are used in practice and how wildcards affect the obtainable information. Lastly we conclude our findings and discuss how these domain lists can be used in further research.

2. Background

This section explains the details of Certificate Transparency, outlines terminology used later in the analysis of the domains and explains the importance of wildcard domains in the context of this paper.

2.1. Certificate Transparency

Certificates are the basis of confidential and authentic communication in the modern Internet and are issued by Certificate Authorities. The obtained authenticity is based on the chain of trust which reaches from a root Certificate Authority over intermediate CAs to issued certificates. If a client trusts a root CA, it implicitly trusts all certificates issued by intermediate CA. Therefore authenticity is based solely on trust into one or only few entities without further means of verification. If adversaries gain access to key material of a CA, it is possible that illegitimate certificates are issued to entities which are not the owners of the respective domains. An instance of such a misissuance occurred in 2011 where a breach at DigiNotar resulted in fraudulently issued certificates. [1]

To combat misissuance or malpractice of CAs, Certificate Transparency was introduced by Google following the DigiNotar misissuance incident. It is defined in RFC 6962 [2] and provides a way to monitor CAs and the certificates they produce. CT logs are append-only data structures that are operated by CAs or independent organisations like Google. A participating CA appends issued certificates to one or multiple CT logs. These CT logs can then be audited by domain owners or independent monitors. This way CAs can be held accountable and misissued certificates can be detected faster [3]. Modern browsers like Google Chrome only accept certificates for web traffic if they have been recorded in at least one CT log [4]. Additionally Chrome enforces a maximum validity duration of 398 days for certificates [5]. This creates an incentive for service providers to regularly renew certificates and for CAs to append them to CT logs.

While monitoring for misissuance is the primary purpose of CT logs, these public logs can also be parsed to obtain a list of domains for which certificates have been issued. This is done by extracting the *Common Name* (CN) and *Subject Alternative Names* (SAN) from a certificate which specify the *Fully Qualified Domain Names* (FQDNs) the certificate is valid for [6].

doi: 10.2313/NET-2023-11-1_15

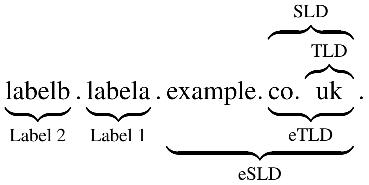

Figure 1: Structure of an FQDN with distinction between conventional TLD and SLD, and the introduced effective TLD (eTLD) and effective SLD (eSLD) with sub labels

2.2. DNS Terminology

According to RFC 8499 [7] a domain name consists of one or more labels that are separated by dots. Conventionally, the rightmost label is the Top Level Domain (TLD) and the label to its left in combination with the TLD is called the Second Level Domain (SLD) (see Figure 1).

For the analysis, we split a domain into a public suffix and private user-controlled part. The public suffix is in most cases equivalent to a TLD, but can also be a SLD because some registries use fixed SLDs for certain purposes. Examples for such domains are `co.uk` and `com.au`. The public suffix list [8] is a comprehensive list of these public suffixes. We call the public suffix the effective TLD and define the effective SLD, which is also called the private suffix, analogously to the SLD. Any further sub labels are the first, second, third, . . . label of a domain.

2.3. Wildcard Domains

The domain name system allows the creation of records for wildcard domains. RFC 4592 [9] states that wildcard domains are domains where the leftmost label is an asterisk (`*`). A domain matches a wildcard domain when all labels of the domain, except for the label where the the asterisk is located in the wildcard domain, are identical. This specifically means that the domain `example.com` and domains with more labels like `a.b.example.com` do not match `*.example.com`.

Wildcard domains are especially relevant in the context of CT based domain lists, because they can hide information about operated services. If individual certificates are created for all subdomains (i.e. services) of an organisation, they are logged in a CT log. This unintentionally publishes identifiers that can be used to estimate the type of operated services, assuming that sensible names are chosen for the subdomains. When creating a certificate for the wildcard domain instead, the subdomains are not published.

3. Related Work

With certificate transparency being a relatively new component of the Internet, it has just been picked up by research in recent years.

Marquardt and Schmidt [10] examined whether CT based domain lists can be a viable alternative to other common domain top lists such as the Majestic or the, now discontinued, Alexa domain lists. Their reasoning for the search of alternatives was that the acquiration of these top lists is often not clearly defined and that CT based domain lists may be a well defined alternative. They used the FQDNs obtained from logged certificates and performed active measurements to compare the created list against the conventional domain lists. They found that while there are 30 % to 50 % less responsive hosts and in general more errors in name resolution with the CT based domain list, such lists can be used as a supplement for the conventional domain lists.

Scheitle et al. [11] examined the implications on security and privacy that arise when certificates are logged in CT logs. The use of CT has the consequence that FQDNs and therefore information about the structure of services is publicly logged. They performed a static analysis of domains to find labels of commonly used services. However they did not consider the depth of the label. They have also set up a *CT honeypot* to see whether CT logs are monitored by active parties. The honeypot hosts were reachable under random domain names which were only published in CT logs. They found that the domains were queried shortly after the certificates were published in a CT log by both presumably well intentioned services like Google or DigitalOcean, but also by suspicious sources. They conclude that CT does remedy one attack vector of certificate misissuance, but argue, that it may introduce new attack vectors based on the suspicious queries.

Pletinckx et al. improved this honeypot experiment in [12]. They ran the experiment for a longer time, used a larger number of hosts, and had a control group where the hosts had self signed certificates installed that were not submitted to any CT log. They noticed that the hosts with the logged certificates received significantly more traffic than the hosts with the unrecorded certificate, especially immediately after the certificate was issued, thereby confirming the previous work of *Scheitle et al.* They performed these experiments with both IPv4 and IPv6 hosts and the IPv6 hosts with self signed certificates experienced no traffic at all. This is explained by the mostly unused IPv6 address space, which, compared to the IPv4 address space, cannot be scanned on a regular basis. Third parties are therefore bound to rely on information like domain names obtained from CT logs to facilitate scanning in the IPv6 address space.

4. Analysis

In this section we analyse domain lists that were extracted from CT logs as described in Section 2.1. The domain lists are available on a per day basis since August of 2022 and contain only unique domains. In the analysis we use domain lists that represent a full month. The month lists contain all unique domains retrieved from certificates issued on the different days of the month. Because the lists contain unique names, there is no bias through multiple certificates issued for the same domain. This analysis focuses on the distribution of registered domains across the available eSLDs and the structure of subdomains configured by the respective domain holders.

The data set used for this analysis is the domain list retrieved in March 2023. We conducted the same analysis shown below for the neighbouring months of January and February and obtained similar results. Therefore we consider one month a representative time frame.

4.1. Number of Domains

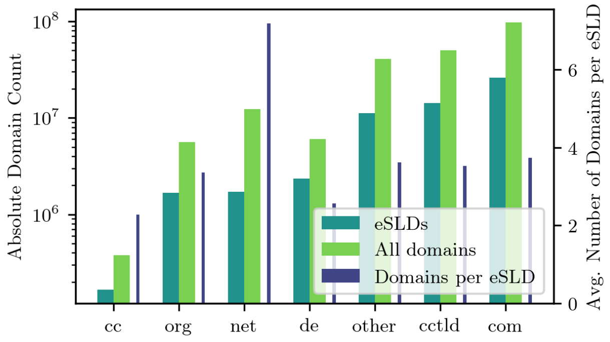

Figure 2: Absolute distribution of both unique eSLDs and all domains across various eTLDs.

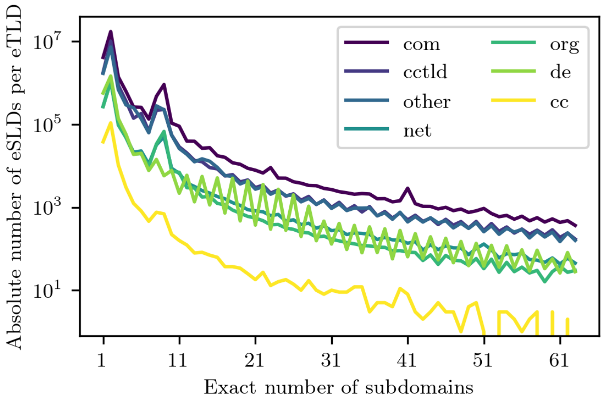

Figure 3: Number of exactly n subdomains per eSLD and eTLD

We begin the analysis by inspecting how many domains are in the data set and how they are distributed across eTLDs and eSLDs. The public suffix list [8] used here contains roughly 6800 unique public suffixes. Because of this large number of suffixes we limit our analysis to the prevalent generic Top Level Domains (gTLDs) com, net, and org and selected Country Code Top Level Domains (ccTLDs) which are de as a regular country code domain and cc. We chose the latter one because it is a popular open ccTLD that can be registered by anyone. These openTLDs are often used as replacement for the more conventional TLDs like com [13]. All other ccTLDs are grouped and represented with ccTLD; all remaining public suffixes are represented with other.

In Figure 2 we see that most of the domains read from the CT log are under the com eTLD. The number of eSLDs under org and net is by a factor of 10 to 15 less than of the com eTLD and in the same magnitude as the country code domain de. We observe that the single com eTLD has more eSLDs and domains than all country code TLDs or all gTLDs combined. In addition to the absolute number of domains, the graph shows the average number of domains under a single eSLD per eTLD. Here we see that the net eTLD has the most logged domains per eSLD with an average of 7.2, while the eSLDs in other eTLDs have an average of 2 to 3 domains.

Figure 3 shows the number of eSLDs in an eTLD that have exactly a specific number of domains. This includes the eSLD itself, conventional subdomains, and wildcard subdomains. Here we see an exponential drop, where most eSLDs do not have more than 10 domains. There are three noticeable deviations from the curve:

- In all cases there are slightly more eSLDs with exactly two domains than with exactly one domain.
- There is a burst of domains at 8 and 9 domains. An inspection of the eTLDs and eSLDs where 8 and 9 domains were present did not yield any pattern that would explain this phenomenon.
- Additionally there is a repeating zig-zag pattern that indicates that there are more eSLDs that have an even number of domains rather than an odd number. This pattern begins to emerge at a number of about 10 domains per eSLD and is especially pronounced in the de eTLD.

4.2. Wildcard Certificates

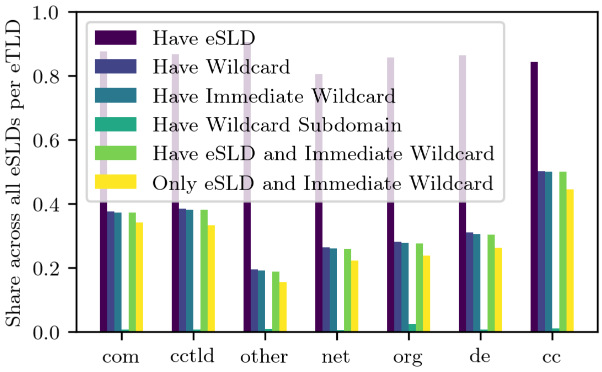

Figure 4: Relative amount of eSLDs per eTLD that fulfil the corresponding criterion

We pay special attention to wildcard domains, because of their ability to hide information as explained in Section 2.3. In this section we analyse how wildcard domains are distributed and used.

We label eSLDs depending on whether the eSLD itself, any wildcard domain, the immediate wildcard domain (in the form *.eSLD), or a wildcard subdomain are contained in the domain list. Figure 4 shows that, with a share of 80.5 % to 90.5 %, nearly all eSLDs have a certificate for the eSLD itself and about 19 % to 50 % have one for the wildcard domain, which is in most cases the immediate wildcard. There are very little eSLDs that have a certificate for a wildcard subdomain in the form *.label.eSLD. The most significant set of domains are these where there is a certificate for both the eSLD and the immediate wildcard: The number of eSLDs where this is the case is only marginally higher than the number of domains where *only* these domains had a certificate. In total 10.5 % of domains were immediate wildcard domains.

This phenomenon generalises across all eTLDs and has an apparent reason. It is a widely adopted use case to obtain a certificate for both the eSLD and the immediate wildcard domain and no other domains. This approach is suitable if there is no need for subdomains with more than one label and makes the deployment of new services under additional first level subdomains very simple because no new certificates have to be generated and the wildcard certificate is sufficient. However, this does hide the internal structure and information about existing hosts reachable under that domain. While this is a privacy improvement, it hinders the usage of CT based domain lists for host reconnaissance.

4.3. Leftmost Label

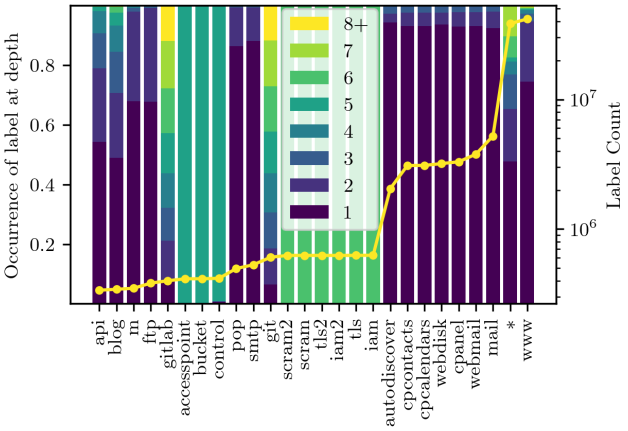

Figure 5: Distribution of most prevalent leftmost labels across different depths in all read domains including the absolute occurrence count of the respective label

The leftmost label of a domain allows for an estimation of the service available on the target host. Figure 5 shows the leftmost labels that were most prevalent across all evaluated domains along with their depth distribution. Labels that were found at a depth of 8 or more are grouped into a single section, however there are almost no popular labels at a depth higher than 7.

We distinguish three different types of distributions that emerge in the most popular labels:

- Distribution that decreases exponentially with increasing depth as seen with, among others, the labels `*` (wildcard), `www`, and `blog` and also less pronounced with e.g. `mail` and `cpanel`.
- Occurrence almost exclusively at a certain depth as seen with, among others, the `iam`, `tls`, and `bucket` labels.
- Mostly equal distribution across all depths as seen with the `git` and `gitlab` labels.

4.3.1. Exponential Decrease. The exponentially decreasing distribution is what one would normally expect for user facing domains. Domains with more labels are not that common for this use case and instead mostly used in the context of deployments or deep internal hierarchies of large organisations. Conventionally `www` is used for web services and makes up for 19.7 % of all read domains alone. Compared to this we see that the wildcard label `*` occurs equally as often as the `www` label with 18.0 %. Combined with the results from Section 4.2, this means that a large number of wildcard labels at deeper levels are concentrated towards a low number of eSLDs.

The labels `mail` or `webmail` hint at the use for a web interface to manage an email inbox. Other labels like `cpanel`, `webdisk`, `cpanelcalendars`, and `cpanelcontacts` also allow for a specific service estimation. These labels are commonly used by instances of the web hosting management software *cpanel* [14]. We see that at least 95 % of user targeting domains such as `www`, `mail` or `cpanel` are only one or two labels long. Of all read domains 30.2 % we such user facing domains.

4.3.2. Equal Distribution. The equal distribution across the depths 1 through 7 is only present with `git` and `gitlab`. There is a pattern in the data set, where the labels `git` and `gitlab` are permuted for up to 6 labels like `(gitlab|git).(gitlab|git)....` There is no apparent correlation to any eTLD or eSLD in the domains with that behaviour. In fact the distribution across the public suffixes was comparable to the distribution of all domains.

4.3.3. Single Depth Labels. The labels `iam`, `tls`, and `scram` and the duplicates with suffix 2 appear only at depth 6 and the labels `bucket` and `accesspoint` similarly only appear at depth 5. This differs from labels such as `cpanel` which, even though they occur at multiple depths, predominantly occur at depth 1. An inspection of the domains with these labels showed that the corresponding eSLD is `amazonaws.com` in 95 % of all cases for `bucket` and `accesspoint` and in 92 % of all cases for `iam`, `tls` and `scram`. The correlation with amazon is backed up by the fact that amazon uses the term *bucket* for its cloud object storage S3 web service [15] and *iam* for *Identity and Access Management* [16]. In these cases, the intermediate labels of the domain encode location information. These domains related to Amazon cover a total of 2.44 % of all domains in the used domain list.

5. Conclusion and Future Work

From our analysis we can see that the `com` eTLD is still the most prevalent one, despite the rising number of other gTLDs. We also find patterns like an accumulation at 8 and 9 subdomains or the preference for an even number of domains in the distribution of subdomains. Further research might clarify the underlying cause of these patterns. The evaluation of wildcard domains yields that certificates are widely used in the common configuration where the eSLD and the immediate wildcard domain are covered by the certificate. Inspecting the leftmost labels of the domains also makes it possible to identify the associated services as we have seen with the *Amazon* and *cpanel* services.

The combination of these findings might make it possible to work around the information hiding ability of wildcard domains. Future work may experiment with substituting the wildcard label with other common and concrete leftmost labels from non-wildcard domains. This could yield further host names that resolve and complete a CT based domain list to make it more comparable to other domain lists.

References

[1] H. Hoogstraaten, “Black Tulip Report of the Investigation into the DigiNotar Certificate Authority Breach,” August 2012.

[2] B. Laurie, A. Langley, and E. Kasper, “Certificate Transparency,” RFC 6962, Jun. 2013. [Online]. Available: https://www.rfc-editor.org/info/rfc6962

[3] O. Gasser, B. Hof, M. Helm, M. Korczynski, R. Holz, and G. Carle, “In Log We Trust: Revealing Poor Security Practices with Certificate Transparency Logs and Internet Measurements,” in *Passive and Active Measurement*, R. Beverly, G. Smaragdakis, and A. Feldmann, Eds. Cham: Springer International Publishing, 2018, pp. 173–185.

[4] Chromium, “Google Chrome Certificate Transparency Policy.” [Online]. Available: https://github.com/GoogleChrome/CertificateTransparency/blob/master/ct_policy.md

[5] ——, “Google Chrome Certificate Lifetimes.” [Online]. Available: https://github.com/chromium/chromium/blob/main/net/docs/certificate_lifetimes.md

[6] S. Boeyen, S. Santesson, T. Polk, R. Housley, S. Farrell, and D. Cooper, “Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile,” RFC 5280, May 2008. [Online]. Available: https://www.rfc-editor.org/info/rfc5280

[7] P. E. Hoffman, A. Sullivan, and K. Fujiwara, “DNS Terminology,” RFC 8499, Jan. 2019. [Online]. Available: https://www.rfc-editor.org/info/rfc8499

[8] Mozilla Foundation, “Public Suffix List,” 2022. [Online]. Available: https://publicsuffix.org/

[9] E. P. Lewis, “The Role of Wildcards in the Domain Name System,” RFC 4592, Jul. 2006. [Online]. Available: https://www.rfc-editor.org/info/rfc4592

[10] F. Marquardt and C. Schmidt, “Don’t Stop at the Top: Using Certificate Transparency Logs to Extend Domain Lists for Web Security Studies,” in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, 2020, pp. 409–412.

[11] Q. Scheitle, O. Gasser, T. Nolte, J. Amann, L. Brent, G. Carle, R. Holz, T. C. Schmidt, and M. Wählisch, “The Rise of Certificate Transparency and Its Implications on the Internet Ecosystem,” ser. IMC ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 343–349. [Online]. Available: https://doi.org/10.1145/3278532.3278562

[12] S. Pletinckx, T.-D. Nguyen, T. Fiebig, C. Kruegel, and G. Vigna, “Certifiably Vulnerable: Using Certificate Transparency Logs for Target Reconnaissance,” 2005. [Online]. Available: https://hdl.handle.net/21.11116/0000-000C-F940-3

[13] “Country Code Top-Level Domain.” [Online]. Available: https://icannwiki.org/Country_code_top-level_domain

[14] cPanel, L.L.C., “cPanel Products.” [Online]. Available: https://www.cpanel.net/products/

[15] Amazon Web Services, Inc., “Amazon S3.” [Online]. Available: https://aws.amazon.com/s3/

[16] ——, “Amazon IAM - AWS Identity and Access Management.” [Online]. Available: https://aws.amazon.com/de/iam/