An Automatic Large-scale Tool for X.509v3 Certificate Collection and Analysis

An Automatic Large-scale Tool for X.509v3 Certificate Collection and Analysis

Diana Gratiela Berbecaru, Antonio Lioy, Anuar Elio Magliari
Department of Control and Computer Engineering
Politecnico di Torino, Italy

***Abstract*—Since its invention in the 1970s, public key cryptography has undergone many changes and updates to meet security, usability, and efficiency needs in various environments and application contexts. Digital certificates, though, have started to be exploited also as attack vectors in cyberattacks. To effectively combat the threats associated with the Web Public Key Infrastructure (PKI), it is essential to employ tools or platforms capable of swiftly gathering and thoroughly analyzing large volumes of certificates, including, but not limited to, web domains across defined geographic regions, countries, or pertaining to specific application domains. We introduce DigitalCertiAnalytics, a tool for collecting and analyzing X.509v3 certificates used in the Web PKI. DigitalCertiAnalytics incorporates software tools that may carry out partial analyses pertinent to X509v3 certificates, like *Zgrab2* for certificate collection, *Zlint* for certificate format validation, and *SSLyze* for certificate chain verification. We validate the proposed tool by examining over 10 million web domains in two lists containing the most visited web domains in different geographical regions. DigitalCertiAnalytics is accurate, flexible, and fast, making it a useful tool for attack prevention, where early detection of invalid, malformed, or erroneous certificates helps to identify attacks in the early stage.**

***Index Terms*—X.509v3 certificates, certificate-related threats, public key certificate inspection**

I. Introduction

PKIs and digital certificates have been largely exploited so far to offer security services across various security applications and protocols. It has been estimated that the global market for digital certificates and PKIs reached US$6.2 Billion in 2023 and will further grow to US$22.7 Billion by 2030. Ongoing efforts are made to revise the standards [1] and requirements for certificates [2], to identify gaps between guidelines and practices, to implement tools for checking the compliance of certificates against certificate related standards and requirements, and to detect misissued certificates [4]. These efforts share a common goal: to enhance interoperability among applications that utilize digital certificates and to minimize security risks. Unfortunately, invalid public key certificates are still encountered and they can be used to carry out security attacks, such as man-in-the-middle, domain (DNS) impersonation, or (distributed) denial of service [3]. In the Web PKI, malicious users exploit malformed, misissued or even expired certificates to impersonate domains or services, or to digitally sign malicious code injected to the client applications. An old study revealed that, "*almost 88% of certificates observed across all these IPv4-wide scans were invalid*" [5]. More recent studies, such as [6], identified also which CAs have created certificates with errors or generating warnings in processing. Although the rate of certificate misissuance has declined over time, a small percentage of certificates are still non-compliant. Thus, they should be individuated as fast as possible to prevent cyberattacks. For instance, CERTainty [7] proposes a tool for analyzing TLS certificates collected through DNS resolve operations to accurately detect DNS manipulation and gather more information about the adversaries responsible for such actions.

To address the threats present in the Web PKI, it is thus crucial to develop tools or platform that enable a rapid collection and effective analysis of digital certificate sets within specific regions or nations. Such tools should be exploited both in open environments (Internet Web) as well as in closed contexts (e.g. military). Moreover, they should provide thorough reporting on potential certificate errors or discrepancies in the analyzed certificates to facilitate comparisons with other analyses of the same certificate set conducted at different time intervals.

Motivation and related works. Nowadays, many tools and platforms allows for processing and validating X509v3 certificates. Some organizations provide (typically upon registration) powerful certificate discovery platforms, such as the `Censys platform` [1] that provides detailed information about all certificates worldwide; their certificate dataset currently holds 16 billion certificates and is continuously updated. `Cloudflare Radar` is another very useful platform that allows users to visualize various information about current networking traffic and web domains, including TLS certificates [2]. However, such platforms cannot be easily customized for monitoring certificates in closed domains (such as inside an organization to protect from cyberattacks or in protected environments) or to build a keep track of a (certificate) history, which is useful in case of an internal investigation. In contract, other works, such as [8], rather than processing large quantities of certificates from global scanning, evidenced several discrepancies between the certificate format specification and their support in some widely-used web browsers.

This work was supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

[1] https://docs.censys.com/docs/platform-certificate-dataset#/
[2] https://radar.cloudflare.com/domains/domain#tls-certificates

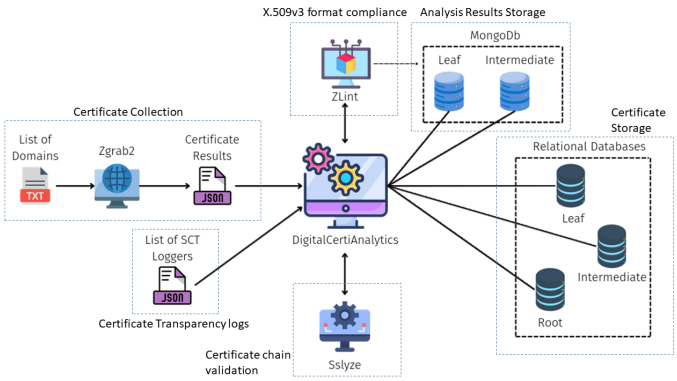

Fig. 1. Architecture of DigitalCertiAnalytics tool for large-scale collection and analysis of X.509v3 certificates.

Contribution. We propose a tool for the collection and analysis of X.509 certificates, named **DigitalCertiAnalytics**. The tool can be customized for monitoring specific domains (under loop), like web domains in specific areas or countries. For example, we have chosen (the most visited) web domains from the United States (US) and Europe. We used two lists composed of different web domains, namely DomCop Top10M [9] and another list of 10 million domains generated with BigQuery [10] for a specific geographical area under loop, i.e., Europe. We analyze the certificates collected both for format and validation by integrating several other tools like *ZLint* and *SSlyze*. The tool that can be used to perform repetitive certificate collection and analysis so as to support comparison with a previous situation. We detail the procedure and the results obtained so far, underlining the critical points.

Organization The paper is organized as follows: Sect. II details the proposed tool and its components, Sect. III shows some of the results obtained with DigitalCertiAnalytics for the web domains in the selected lists. Finally Sect. IV resumes the conclusions and indicates future work.

II. Architecture of DigitalCertiAnalytics

DigitalCertiAnalytics, illustrated in Fig. 1, performs the collection and analysis of certificates for web domains whose names are given in input lists, such as for particular regions, countries, operational domains, or application sectors. Starting from the list(s), the tool dynamically collects the certificate dataset(s), by exploiting Zgrab2 utility [3] and generates a certificate set. The certificates are analyzed for compliance with standards and requirements while also verifying their validity. Finally, the certificates (divided in leaf, Intermediate CA and Root CA certificates) are stored in designated (local) databases for archival and to facilitate subsequent comparison with other certificates sets collected at different times. The outcomes of this processing are displayed in user-friendly graphs for straightforward comprehension and to identify variations that may indicate possible risks.

[3]https://zmap.io/

The tool [4], developed in Python, integrates various software tools that have been chosen for their ability to perform partial analyses related to digital certificates, that is, Zlint [5], Zgrab2 [6], and SSLyze [7]. This architecture allows for the reuse of existing tools, making the overall process more flexible, faster and scalable. We detail further below each subpart in the DigitalCertiAnalytics architecture.

A. Collecting digital certificates with ZGrab2. In the early development stages, for certificate collection and analysis, we employed the OpenSSL library and OpenSSL commands, such as the 's_client' command, but this approach is not efficient for collecting certificates in large datasets. ZGrab2 instead is a powerful solution, which has proven particularly useful for automating and accelerating the collection of the digital certificates on a large scale. The tool establishes TLS connections towards the hosts (in the domains under loop) and produces a JSON file containing connection details, including errors, protocol, and timestamps. ZGrab2 can also automatically parse digital certificates generating an output in JSON format, which contains the raw data of the certificate along with the information extracted from the certificate fields or extensions, such as the issuer, subject, extensions, and the complete certificate chain. Furthermore, the tool extracts the Signed Certificate Timestamps (SCTs) [11] from the certificates. The SCTs are certificate extensions which are used to check whether the certificate is listed in public logs of the Certificate Transparency (CT) system promoted by Google to detect misissued certificates [6]. We retrieved additional information from the CT system, that is the list of SCT loggers including log IDs, in order to analyze the SCT extensions contained in the collected certificates. Since the SCT extensions store information about the proof that a certificate has been registered in one or more CT logs, data about those CT logs is also registered in the database exploited by our tool. This integrated approach, which includes automatic data collection as well as thorough certificate parsing, is incredibly useful for in-depth studies of the digital certificate ecosystem.

B. Classification of certificates with a relational database. To store the collected digital certificates we used relational databases, which are logically divided into three distinct parts, namely `Leaf`, `Intermediate`, and `Root`, based on the three certificate types composing a certificate chain. This separation mirrors the typical hierarchy of a PKI and facilitates focused analysis of each certificate type. In the dedicated database for root certificates, we have saved the ones transmitted along with the chains in TLS connections. The database was created using SQLite3, a serverless solution known for its lightweight design, speed, and ease of management. This allows all data to be stored in a single, easily transferable file.

C. Integration with Zlint for certificate analysis and checking compliance with standards and regulation.

[4]https://github.com/MagliariElio/DigitalCertiAnalytics
[5]https://github.com/zmap/zlint
[6]https://github.com/zmap/zgrab2
[7]https://github.com/nabla-c0d3/sslyze

A lint (or linter) is a software tool that performs static analysis on code, identifying potential errors and suggesting improvements. For DigitalCertiAnalytics, detecting certificates thar are non-compliant with standards and recommendations — the ones defined by the CA/Browser (CAB) Forum, various RFCs, and Mozilla PKI specifications — is a crucial aspect of digital certificate analysis.

Thus, we integrated existing solutions that perform this task efficiently and that can be updated automatically as the standards or the recommendations evolve. We have exploited Zlint, part of the ZMap project, that offers several important functionalities, such as verifying compliance for the certificate (format) with standards and recommendations, checking the validity and correctness of TLS/SSL certificates, and detecting violations of policies and best practices. Additionally, Zlint checks custom guidelines defined by major operating system or browser vendors, like Apple or Mozilla. For instance, Apple defined custom guidelines for digital certificates, needed for compatibility with macOS and iOS operating systems. It performs a series of automated checks, named *linters*, that can be applied to certificates. These checks help detect various misconfigurations, insecure practices, and inconsistencies that could compromise the security of SSL/TLS-protected communications. By integrating Zlint, we obtain an analysis tool that supports multiple certificate-related standards and rules, which simplifies implementation and reduces the risks associated with manual code changes.

Initially, Zlint focused on supporting RFC 5280 [1] and the CA/B baseline requirements [2]. These documents are crucial as they define the format of certificates, including those used for web server certificates in SSL/TLS connections. In its first version, Zlint included around 220 linters that could verify 95 % of the clauses related to the CA/B baseline requirements and 90 % of those in the RFC 5280. This ensured that the certificates complied with established guidelines for security and reliability. In the current version of DigitalCertiAnalytics, we used Zlint version v3.6.4-9-g989baefd, and the total linters has expanded to 370. This growth has allowed for the inclusion of additional international standards, recommendations and custom regulations, including those from Apple, Mozilla, ETSI, and other industry bodies. Zlint categorizes the linters it employs to verify certificates into varying levels of severity, based on their degree of non-conformance with the CAB Forum documents and baseline requirements: INFO, WARNING, ERROR, FATAL. These categories enable a modular and accurate evaluation of certificate compliance with international regulations and best practices, helping to identify vulnerabilities early on.

D. Integration with MongoDB for results management. Once the certificates have been collected and saved in the corresponding relational databases, DigitalCertiAnalytics performs the following steps in order: 1) Parses the relational databases containing the collected certificates, retrieving each leaf X.509v3 certificate for analysis; 2) Invokes Zlint to perform checks of the X.509v3 certificate format; 3) Receives the results from Zlint in JSON format; 4) Stores the Zlint results in the MongoDB database(s).

The decision to use a non-relational database to manage the results generated by Zlint was based on the format of the data produced. Zlint generates results in JSON format, which contains numerous fields, each representing the outcome of various checks performed. This complex and dynamic structure poses challenges for management in a traditional relational database, which relies on a rigid schema and is less adaptable to changes or the addition of new fields.

For these reasons, we selected a non-relational document-based database, namely MongoDB, as storage solution. MongoDB provides greater flexibility due to its lack of a fixed schema and the use of collections and documents, which aligns naturally with the JSON format of Zlint's results. This flexibility enabled us to store the Zlint outputs in DigitalCertiAnalytics tool with no need to change the database(s) in response to updates or changes in the Zlint data format. Additionally, MongoDB delivers high performance in terms of both write and read speeds, making it especially beneficial for quick data queries and prompt responses during subsequent analyses or verification. From a practical perspective, we used pymongo version 4.10.1 to render efficient the communication between the local database and the MongoDB database.

E. Certificate Chain Validation. A digital certificate is considered valid if the validation of the certification chain (composed of the leaf certificate, one or more intermediate CA certificates and a root CA certificate) is successful. Additionally, it is essential to verify whether the root CA certificate is present in the local certificate trust store, so as to guarantee trust in the entire chain. The root CA certificates are typically pre-installed in the applications or operating systems according to strict policy rules, e.g., Mozilla Root Store Policy Microsoft Trust Root Program or Apple Trust Root Store program. To perform certificate (chain) validation, the applications may implement a custom module or library following the algorithm specified in RFC 5280 [1] and CAB BR [2], as well as specific options that might apply depending on the platform. Other applications may rely on open-source crypto libraries, like OpenSSL, which contains dedicated functions for certificate management and chain verification. Alternatively, the certificate validation could be delegated to an external service or process.

We have selected SSLyze, an open source tool in Python, which allows to verify the validity of the certificate chains, ensuring compliance with the RFC 5280 certificate standard even though other alternative certificate validation services or libraries could be used for this purpose. SSLyze can automatically check the validity of certificate chains and periodically updates its integrated trusted root CA certificate store[8]. More precisely, SSLyze verifies the certification chains against different certificate trust stores, including Android CA Store (15.0.0_r5), Apple CA Store (iOS 17, iPadOS 17, macOS 14, tvOS 17, and watchOS 10), Java CA Store (jdk-13.0.2),

[8]https://nabla-c0d3.github.io/

Mozilla CA Store (2024-09-01), and Windows CA Store (2023-12-11). The tool uses the hosts whose certificates are stored in the relational database(s) to query SSLyze and obtain the results regarding the chain validation. The results, provided in JSON format, are stored in the database, keeping only the data strictly necessary for further analysis and reporting. In addition, SSLyze provides results related to the control of additional features, such as support for OCSP Must-Staple and checking whether and how many SCTs are present in the certificate.

III. Results

To validate the tool we used two customized lists: DomCop Top10M list [9] and a list containing the most visited European domains. The DomCop Top10M list was obtained from the DomCom platform, specialized in searching for and purchasing expired web domains. DomCop features a list of 10 million domains obtained from the Open PageRank initiative, which was created by DomCop itself. This initiative aims to provide metrics similar to Google's PageRank using open-source data from Common Crawl [9] and Common Search. The goal is to enable cross-domain comparisons through free authority metrics that are accessible to everyone.

The second list was generated by using BigQuery [10], a fully managed platform for analyzing large amounts of data available on Google Cloud. BigQuery allows users to perform SQL queries on extensive datasets in a scalable and efficient manner. In our case, we used the Chrome User Experience Report (CrUX) dataset, which provides real-world measurements of user experience on websites through performance metrics such as Largest Contentful Paint, First Input Delay, and Cumulative Layout Shift. This data, collected from Chrome browser users, offers valuable insights into website performance from the end-user perspective. The integration of BigQuery and the CrUX dataset facilitates advanced queries on real-world website performance data, enabling detailed analyses at scale. Users can segment data based on variables such as device type, geolocation, and operating system. We executed a query to extract a list of domains located in Europe, based on the column *experimental.popularity.rank*, which indicates the popularity ranking of websites. The goal was to identify the most visited European domains to analyze their corresponding certificates.

Certificate Collection in DomCop dataset. By utilizing the dataset provided by DomCop, starting on 10 October 2024, we collected the X.509v3 digital certificates for 10 million web domains. The data collection operation, running on an Intel Core i7-8550U platform equipped with an CPU quad-core, 1.80 GHz - 2.00 GHz, 8 GB of DDR4 RAM at 2400 MHz and a 510 GB SSD, lasted 7 hours and 33 minutes and produced as result a 82 GB JSON file. In this phase, we encountered three error types: a) Connection timeout, likely due to latency issues or unreachable domains; b) I/O timeout (14.61%): timeout during input/output operations, indicating potential problems with data transfer; c) Unknown error (2.04%): errors of indeterminate nature, which may arise from anomalies in the protocol or unforeseen circumstances.

The certificate collection yielded 7,045,024 valid leaf certificates (70.45%), while for the rest of 29.55% we encountered connection errors or other anomalies. Initially, only leaf certificates were scrutinized in detail, examining all aspects and fields to comprehend the configurations employed in the final segment of the trust chain. Subsequently, intermediate and root certificates—distinct from the trust roots pre-installed in the testing system—were obtained during the TLS/SSL negotiation. All collected certificates were imported and analyzed by DigitalCertiAnalytics in 19 hours and 46 minutes. The primary limitation encountered pertains to the required memory space for the substantial amounts of data generated. Furthermore, a part of the total leaf certificates, amounting to 7,045,024, contained also duplicate certificates. These duplicates may arise from the inclusion of multiple Subject Alternative Names (SANs) within a single certificate. Consequently, the unique leaf certificates amounted to 5,000,264. This number is however notably greater than the 442,331 distinct domains that have been successfully queried. We detail further below some of the results pertaining the leaf certificates.

The analysis of Leaf certificates in DomCop dataset showed that 99% of the certificates were in X.509v3 format, whereas a mere 1% were still utilizing outdated formats (versions 1 and 2), which are no longer recommended. The geographical analysis of the leaf certificates revealed that 89% of the valid certificates were issued by CAs in the United States, based on the 'C=' value in the Distinguished Name. The distribution of the remaining certificates was as follows: 4.7% were issued by CAs in the United Kingdom, 2.9% by CAs in Belgium, 0.6% by CAs in Japan, 0.5% by CAs in the Netherlands, and the remaining 2.3% were issued by other CAs. We observe that the issuer landscape is significantly dominated by a few major players. In particular, Let's Encrypt has been confirmed as the leading CA, responsible for the issuance of 46.33% of the certificates in the analyzed dataset. Google Trust Services accounts for 22.06% of the total certificates issued, reflecting the increasing trust web operators and applications place in issuers who are part of established ecosystems like Google's. Additionally, DigiCert Inc emerges as a significant CA, having issued 8.09% of the certificates. Finally, Sectigo Limited and Amazon, with respective shares of 4.49% and 3.60%, complete the landscape of major certificate issuers. Meanwhile, the involvement of Amazon underscores the trend of cloud operators venturing into certificate issuance, thereby integrating their service offerings to ensure secure connections. Other issuers (CAs) encountered were: GoDaddy.com, GlobalSign, Cloudflare, cPanel, Starfield Technologies, Internet2, ZeroSSL, GEANT Vereniging, DigiCert, Entrust, Actalis S.p.A, testexample (!), Japan Registry Services Co., Ltd, Unizeto Technologies S.A., others (!) with few issued certificates.

Certificate types. The distribution of validation levels indicates that 95.35% of leaf certificates were Domain Validation

[9]https://commoncrawl.org/

(DV), 3.62% were Organisation Validation (OV), and a mere 0.28% were Extended Validation (EV) type. These statistics align with expectations, as DV certificates, which necessitate solely the verification of domain control, are generally easier to obtain, thereby accounting for the majority. Conversely, OV and EV certificates, which entail additional scrutiny of the organisation's identity, are considerably less prevalent, with EV certificates, subject to particularly rigorous verification processes, appearing at a marginal percentage. We note that 0.74% of certificates lacked a recognized validation indication. This anomaly may be attributed to errors in the issuance processes or to configurations that do not adhere to the directives set forth by the CAB forum, which delineate the validation criteria for publicly trusted certificates in its baseline requirements.

Time Validity. The analysis of the certificate validity periods within the dataset showed significant diversity, which provides valuable insights regarding adherence to existing best practices and potential irregularities in the certificate issuance and management. We examined the validity of certificates by scrutinizing the *Not Before* and *Not After* fields, which denote the certificate's starting validity and expiration dates, respectively, and complemented this examination with an assessment of the certificate validation. The analysis shows that the majority of certificates, 81.76%, are valid for less than one year. Of these, 98% are DV certificates and 1.04% OV certificates, while the remainder are EV certificates. Less common, but still encountered, we have found 0.13% of certificates with a validity period of up to 10 years. Of these, 9.63% were DV certificates, 3.92% OV certificates and 86.43% did not fall into any recognised validation category. We remind that certificates with long lifetimes are generally discouraged, because too long validity periods can increase the security risk in the event of compromise and reduce the frequency with which the identity of the subject is verified. We have found however some anomalies: about 639 certificates reported a time validity between 10 and 273 years (!), and about 11 certificates showed negative validity values, that is, the date in the Not After field in the certificate precedes the one in the Not Before field. Furthermore, 4,971,980 certificates (99.43% or the total amount analyzed) issued after 1 September 2020, had an overall validity of 398 days, in line with the CAB requirements in 2024, while 3,444 certificates (0.07%) violated these limits, indicating that they exceeded the permitted validity. Similarly, for certificates issued between April 2015 and 1 September 2020, 19,903 certificates (86.60%) complied with the 39-month limit set by the previous CAB requirements, while 2,309 certificates (10.40%) exceeded this threshold.

We observe that a significant number of certificates expire between October and February (2025). This uniformity could result from renewal policies set up to facilitate automation and centralised management of certificates. However, this concentration of expiry dates may result in significant simultaneous demand for renewal and revocation operations, which requires a proactive and robust management system to prevent any delays or inefficiencies in update procedures, as well as prompt revocation checkings (which typically do not occur in many applications). So, that time period is potentially more vulnerable, as attackers could try to exploit (old) revoked certificates that have been replaced by new ones for valid domains.

Serial Number. Identifying certificates is relevant in cyber-attacks investigation. According to [1], the serial number must be a unique positive integer for each certificate (issued by one CA). The standard does not impose a fixed minimum length, making even a single byte value valid, but the CA/Browser forum recommends a maximum length of 20 bytes (160 bits). In our analysis, we observed that each certificate has a serial number, but only 99.98% (4,999,241 certificates) adhered to the aforementioned specifications. We have also detected certificates that do not conform to standards: 34 of the analyzed certificates had a negative serial number, while 989 certificates exhibited duplicate serial numbers, with 23.66% of these bearing a value of 0. None of the examined certificates had a serial number exceeding 20 bytes in length.

Chain length. We measured that 69.67% of leaf certificates (3,483,611) were accompanied were part of a chain composed a leaf, an intermediate, and a root certificates. However, approximately 26.96% of the leaf certificates (1,348,129) were part of a chian with two intermediate certificates, while 2.03% (101,260 certificates) has a chain with three intermediate certificates. A small percentage however, 1.34% of the certificates (67,264) were issued directly by the root CAs.

Collection of certificates in Google CrUX Dataset. We collected a separate set of certificates for web domains obtained by using a query on Google's big data system [10], more precisely by querying the Chrome User Experience Report (CrUX) [12] available through Google Cloud. Specifically, an SQL query was run by selecting the top 10 million most popular European domains, sorted by the experimental.popularity.rank field. The query focused primarily on the major European countries and we obtained a dataset that showed a good representation of major domains of the continent. This separate query could further focus on (just) one country, such as Italy, and explore regional dynamics which can vary considerably. The result of the query was downloaded from Google Cloud in a .txt file containing all selected domains. Subsequently, at the end of October 2024, that list was used as input for ZGrab2, the tool used to download the associated digital certificates, generating a 132.1 GB JSON file. A timeout of 30 seconds was set for the ZGrab2 collection operation, allowing to complete the downloading of the corresponding certificates in 4 hours and 56 minutes.

The process returned 9,440,561 leaf certificates, or 94.41% of the total, while 559,440 certificates (5.59%) produced connection errors or other anomalies. In this context, the number of unique leaf certificates, deduced from the presence of multiple SANs per certificate, turns out to be 3,164,063, showing a reduction of 36.72% compared to the total number of leaf certificates in the DomCop dataset. These results could point to substantial differences in the certificate issuance land-

scape between the two datasets and suggest possible variations due to the different geographic composition of the analyzed domains.

Analysis of certificates in Google CrUX Dataset. All the certificates analyzed were X.509 version 3, so we have not found certificates of old versions. The distribution of validation level reflects what was expected seeing a large majority (94.11%) of DV certificates followed by 0.61% of OV certificates, and almost negligible percentage of EV certificates. Upon analyzing European domains, the certificates reported the same Issuer CA entities, albeit with slightly varying percentages compared to the DomCop dataset. Once more, however, Let's Encrypt dominates the field, accounting for 45.02% of the certificates in the dataset, followed by Google Trust Services at 22.81% and DigiCert Inc. at 7.16%. Thus, there is not a significant disparity when compared to the DomCop dataset.

Time Validity. We measured that 99.96% of certificates were timely valid. However, an analysis of the timeframe from April 2015 to September 1, 2020, highlights that only 82.36% of certificates adhered to the 39-month limit, whereas the remaining 17.64% surpassed this threshold. The scenario changes for certificates that were issued on or after September 1, 2020. Almost all of them, like 99.97%, were valid for 398 days or less, which was totally in line with the CAB recommendation (BR 7.1.2.7.6) in October 2024. Just a very small percentage, 0.03%, were not respecting the rules. A comparison with the data provided by the DomCop dataset reveals some similarities and a couple of significant differences. For certificates issued after September 1, 2020, the DomCop dataset indicates that 99.43% comply with the 398-day limit, while 0.07% of certificates had a longer validity period. Similarly, for certificates issued between April 2015 and September 1, 2020, DomCop reports that 86.60% of the certificates conform to the 39-month limit, while 10.40% of the certificates exceed the permitted time validity. The validation dates showed that the certificates tend to expire between October 2024 and November 2025, with especially pronounced peaks in December 2024 (39.21%) and January 2025 (35.47%). This temporal aggregation implies a near cascaded scheduling of certificate issuance, likely associated with the expiration of numerous certificates that were issued in prior periods.

Serial number. We note that while both datasets largely conform to the established recommendations, certain anomalies persist. Notably, in the Google CrUX dataset, nearly 100% of the certificates possess a valid serial number, with no instances of length exceeding 20 bytes. However, there were 10 certificates identified with negative serial numbers and 133 duplicate instances, of which 64 certificates display the value of 0. In contrast, an analysis of the DomCop dataset indicates a marginally lower level of compliance, with 99.98% of the certificates fulfilling the requirements. Notably, there were 34 certificates with negative serial numbers and 989 certificates had duplicate serial numbers, of which 23.66% had a (non-conformant) value set to 0.

IV. Conclusions and Future Work

This paper introduced DigitalCertiAnalytics, a tool for collecting, analyzing, and storage of X.509v3 certificates for a subset of web domains of interest. Such certificates must be detected as early as possible because they could be exploited in security attacks. The proposed tool is flexible as it integrated several pre-existing software tools exploited for certificates collection, analysis and reporting of deviations. It is also accurate because it uses consolidated products, like Zlint, for checking the certificate format and SSLyze for validating the certificate chains. We tested DigitalCertiAnalytics by using two lists created with DomCop and BigQuery, that have been customized for web domains in different geographical regions. Even though the majority of the analyzed certificates adhered to the prescribed certificate format and requirements, we have encountered a small percentage of non-conformant certificates, such as those issued by unknown or test CAs, as well as certificates with invalid or duplicate serial numbers, or expired. Future efforts will focus on expanding the tool functionalities, such as the support for hybrid or post-quantum certificates.

References

[1] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk, "Internet x.509 public key infrastructure certificate and certificate revocation list (crl) profile," RFC-5280, May 2008, doi: 10.17487/RFC5280.

[2] CA/Browser Forum: Baseline Requirements for the Issuance and Management of Publicly-Trusted Certificates. Version 1.8.6, 14 Dec. 2022, https://cabforum.org/wp-content/uploads/CA-Browser-Forum-BR-1.8.6.pdf

[3] B. Shi, W. Li, Y. Wang, X. Bai, and L. Xing, "X.509DoS: exploiting and detecting denial-of-service vulnerabilities in cryptographic libraries using Crafted X.509 certificates," In Proc. of the 34th USENIX Conference on Security Symposium (SEC '25). USENIX Association, USA, Article 27, 509–528.

[4] 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," Internet Measurement Conference 2018 (IMC '18). ACM, New York, NY, USA, 343–349, doi: 10.1145/3278532.3278562.

[5] T. Chung, Y. Liu, D. Choffnes, D. Levin, B. MacDowell Maggs, A. Mislove, and C. Wilson, "Measuring and Applying Invalid SSL Certificates: The Silent Majority," In Proc. of the 2016 Internet Measurement Conference (IMC '16), 2016, ACM, New York, NY, USA, pp. 527–541, doi: 10.1145/2987443.2987454.

[6] D. Kumar et al., "Tracking Certificate Misissuance in the Wild," 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2018, pp. 785-798, doi: 10.1109/SP.2018.00015.

[7] E. Tsai, D. Kumar, R. S. Raman, G. Li, Y. Eiger, R. Ensafi, "CERTainty: Detecting DNS Manipulation at Scale using TLS Certificates," In Proc. of Privacy Enhancing Technologies 2023 (3), Lausanne, Switzerland, July 10–15, 2023, pp. 122-137, doi: 10.56553/popets-2023-0073.

[8] D.G. Berbecaru and A. Lioy, "An Evaluation of X.509 Certificate Revocation and Related Privacy Issues in the Web PKI Ecosystem," IEEE *Access*, vol. 11, pp. 79156-79175, 2023, doi: 10.1109/ACCESS.2023.3299357.

[9] Top 10 million Websites. https://www.domcop.com/top-10-million-websites, Accessed: May 06, 2025.

[10] From data warehouse to a unified, AI-ready data platform. https://cloud.google.com/bigquery/?hl=en#from-data-warehouse-to-a-unified-ai-ready-data-platform, Accessed: 2025/05/06

[11] J. Amann, O. Gasser, Q. Scheitle, L. Brent, G. Carle, and R. Holz, "Mission accomplished? HTTPS security after diginotar," In Proc. of the 2017 Internet Measurement Conference (IMC '17), ACM, New York, NY, USA, 325–340, doi: 10.1145/3131365.3131401.

[12] Google, "Chrome user experience report (crux)." https://developer.chrome.com/docs/crux, Accessed: May 9, 2025