Finding Personally Identifiable Information Leaked via Publicly Accessible CT Logs

Frank van Mourik
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
f.g.vanmourik@student.utwente.nl

ABSTRACT

Certificate Transparency (CT) is a network security standard that includes all public-key certificates in publicly accessible logs. Major browser vendors such as Chrome require certificates to be present in CT logs before accepting them. These logs can be analysed and audited by everyone in the world, adding an extra layer of security on top of the Internet. These certificates, however, might include personally identifiable information (PII) from website creators or administrators, for instance their first and last name. Since CT logs are queryable at a large scale, possibly contain PII, and are non-optional, a privacy issue arises.

This research provides a proof-of-concept approach to find PII in these public logs by looking at the registered domain names within over one billion certificates and characteristics of these domain names in combination with commonly-used Dutch first and last names. Additionally, in this work we aim to find providers of PII in certificate's domain names, focused on the ".nl" DNS zone. Here we found several companies that potentially forward PII of their customers in CT. Finally, this research looks into the amount of PII in domain names over time in order to spot possible increasing and decreasing trends. No significant trends were observed.

Keywords

CT, Certificate Transparency, Public-key certificates, Data analysis, Privacy, Information leaks, Personally Identifiable Information, CT Logs

1. INTRODUCTION

The idea of Certificate Transparency (CT) started in 2011 after two major Certificate Authorities (CA), the authorities that hand out public-key certificates, were compromised [1] [2]. CT was started by Google to "safeguard the certificate issuance process by providing an open framework for monitoring and auditing HTTPS certificates" [2]. The general idea of CT is that all trusted certificates should be present in a publicly accessible append-only log. Using these public logs, clients and auditors can check whether a certificate is trustworthy or fraudulent. A flowchart on the working of CT can be found in figure 1.

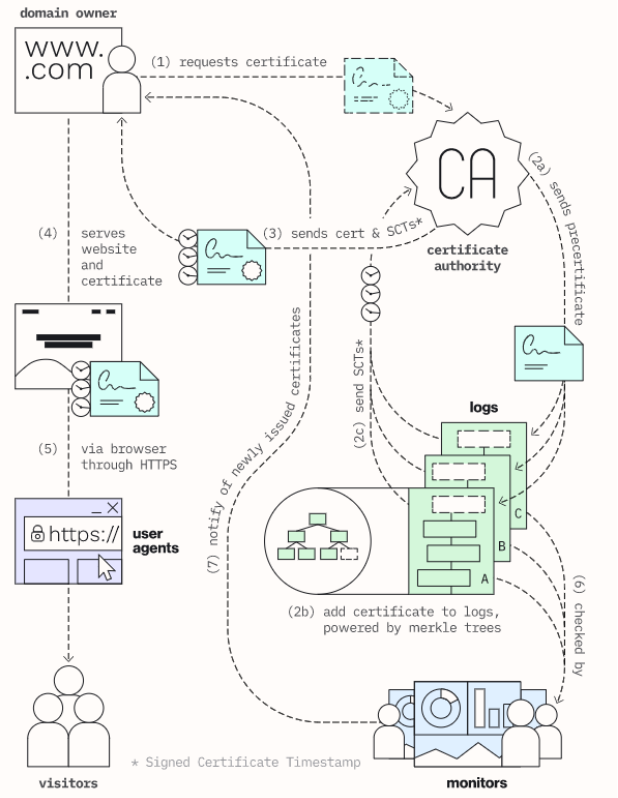

Figure 1. Flowchart on the working of Certificate Transparency. Source: https://certificate.transparency.dev/howctworks/

The public-key certificates are used to verify ownership of a public-key, which provides the end-user with a guarantee that the host of a website is the owner of the website. In order to issue a certificate, the owner of the website has to give away some private information about themselves, such as the domain name, country name, city name and possibly more. Most of this information is stored in the issued certificate.

Before CT was implemented, the public-key certificates were only stored at the host of a website. This would mean that the certificate could only be requested after knowing the domain name and initiating a connection to the web server to retrieve the certificate.. Even though

[1] https://security.googleblog.com/2011/08/update-on-attempted-man-in-middle.html
[2] https://security.googleblog.com/2011/04/improving-ssl-certificate-security.html

35^{th} *Twente Student Conference on IT* July. 2^{nd}, 2021, Enschede, The Netherlands.

the certificates were also publicly accessible, the requester had to know at least some information, namely the domain name, before accessing the certificate. One could say that the domain name functioned as a gateway to the certificate.

After the implementation of CT, it became much easier to query a large amount of certificates since all the trusted certificates are stored in large publicly accessible logs. Querying these logs could result in receiving millions of certificates without much effort.

Even though some personal data, including the domain name, has to be given away by the issuer upon requesting a certificate, the domain name was publicly accessible anyway. Domain names are crawled and indexed by for instance search engines. By indexing the domain names, they are queryable, hence adding another path to the certificate. In this way, the domain name again functions as a gateway to the rest of the certificate, hence to more potential personal information. This indexing can however be blocked by a "robots.txt" file [3], so that the internal structure of a website is not indexed and cannot be queried using a search engine.

Since the gateway to a certificate is now no longer needed to access the certificate, the subject's sensitive information stored in the certificate is easier to request than it was before CT. The fact that possibly some personally identifiable information of the subject is stored in CT, perhaps even without them knowing, results in the subject having less control over their data than they had before. Although personally identifiable information (PII) has multiple definitions, we will define it as information that could pinpoint one specific person in a society. In this case, we focus on the combination of a first and last name.

It is possible that third parties, such as hosting providers on behalf of their customers, request the certificate for a website, which might result in the creator's PII being present in CT without them knowing.

Finally, we are interested to see if creators and administrators have become more aware of the possibility of PII in CT. We will try to aggregate the certificates over time to see whether the problem of PII in CT is becoming an increasing problem, decreasing problem or stable problem.

1.1 Problem Statement

In this research, we will provide a proof-of-concept (PoC) approach to find PII in CT logs. We will use CT logs from Google and combine this with frequently-occurring names in order to find PII in these logs. For this research we propose the following research questions:

> *RQ 1: Can we find personally identifiable information in publicly accessible CT logs?*
>
> *RQ 2: What are the characteristics of this information?*
>
> *RQ 3: What are the trends over time?*

Our main contributions are as follows. We identify almost 200k unique domain names containing first and last names, demonstrating that PII can easily be found in CT data. We characterize the structure of the domain names in question and reveal that, at times, third parties are responsible for CT inclusion. Finally, we investigate trends over time.

In the following sections, we will discuss the related work that has already been conducted in this research field. In section 3, we will describe the methodology that will be used in this research. The results are discussed in section 4. Afterwards, in sections 5, we will draw some conclusions, which we also embed in our ethical considerations in section 6. Finally, in section 7, we will describe the main discussion points and considerations of the research and suggest future work that could be conducted in this field. Additionally, we provide the codebase and data used in this research for reproducibility.

2. RELATED WORK

Since privacy is directly associated with the safety of people, it has been a relevant topic leading to much research being conducted in this topic, already since the start of the internet. In this section we will mainly focus on privacy-related research in the area of Certificate Transparency and personally identifiable information research in DNS.

Since the start of Certificate Transparency, a lot of research has been conducted into the working and effectiveness of the CT standard. Additionally, research has been conducted to find privacy issues in these CT logs. Eskandarian et al. have researched two privacy issues with CT regarding auditing and support for non-public subdomains [1]. They found a solution for both issues, where browsers can conduct CT auditing, without their vendor (for instance Google for Chrome) knowing the browsing history of a client, using a "Zero-Knowledge Proof of Exclusion". Additionally, they found a solution where a private subdomain could be included in the CT logs, without publicly revealing the inner workings of said subdomain.

Another research on privacy in CT was conducted by Kales et al. [3]. Their research related to the first issue stated in the research mentioned above: since clients directly contact CT Log APIs, the CT log owners are able to track the browsing history of clients. This is the same issue as in the previously-mentioned research, however on a different level. They provide a solution for this issue built on an approach by Lueks and Goldberg [4], using the Private Information Retrieval principle, to protect users from the CT ecosystem.

An additional research trying to find private sub-network was conducted by Roberts et al. called "When Certificate Transparency Is Too Transparent" [6]. Their conclusion is that there is indeed an information leak in CT and that that information is stored indefinitely. They provide some potential solutions, including wildcard ("*") certificates or private subdomains, which we will discuss in section 5.

Additionally, research has been conducted in the field of PII presence in Domain Name System (DNS) records, especially passive DNS data collection [7]. Even though the conclusion of this paper differs per definition of PII (this differs per legal country), it concludes that PII *is* present in DNS records, which in some countries would make passive DNS data collection a privacy violation. In this paper, PII is interpreted as an individual end-user's DNS behaviour.

From these related papers, we can conclude that research has been conducted in the areas of privacy issues in CT logs and PII presence in DNS records, but no research has

[3] https://developers.google.com/search/docs/advanced/robots/intro4

been conducted in the area of PII presence in CT logs. This is where our research will add.

3. METHODOLOGY

3.1 Data

3.1.1 CT logs

The CT logs are hosted by companies like Google, but also by Certificate Authorities like DigiCert. Since a CT log is hosted in a single place, it is easy to query the entire log, resulting in all certificates that are includes in the log. The most-relevant fields included in these certificates are:

- Serial number
- Issuer
- Subject
- Not before time
- Not after item
- Extensions, including a list of domains

Most of the hosted CT logs were scraped prior to this research and stored by the University of Twente. For this research, we made use of the Google Pilot CT log, which contains over 1 billion certificates stored in Apache Parquet file format.

From these certificates, we use the domain names registered by the certificate and the "not before" field, which indicates when the certificate became valid, which is roughly the same date as the certificate was requested. This latter field is used to infer time of registration in order to answer RQ3.

3.1.2 Frequently-used names

As mentioned in section 1, we defined personally identifiable information (PII) as the combination of a first and a last name. In order to find PII in the domain names of the certificates, we created a list of frequently-occurring first and last names in the Netherlands. The first names dataset consists of the top-100 used names for babies per sex per decade, from 1960 until 2010 [5]. In total, 530 unique first names were used. We combine all unique first names from all six decades. This is done, since we try to find all PII present in the domains.

Since frequently-used last names change less-frequently than frequently-used first names, the last names dataset contains the top-100 most-used last names in 2007. Some last names have multiple spellings (e.g. Bruijn / Bruyn), which results in a list of 102 unique last names.

The domain names are string-matched on the list of first names and the list of last names to include at least one of both.

3.2 Tools

Since we are dealing with vast amounts of data, we use the spark distributed computing tool in order to compute the results in reasonable time.

The computations are done by the Hadoop [4] cluster that is run on the computing cluster of the University of Twente.

[4] https://hadoop.apache.org/

3.3 Finding PII

In order to answer RQ1, we will execute queries on the certificates to find domains containing both first and last names. We chose for the combination of first and last name, since in most cases we can pinpoint an individual based on their first and last name, especially since their domain name links to their website, which could reveal more. In this way, the domain name functions as a gateway to more sensitive information about the individual. Afterwards, we will take distinct domain names in order to eliminate the chance of counting domains multiple times. At the distinct query, the first "not before" date is aggregated, since that is the point when the information was leaked.

The result will be analysed on characteristics, such as the position in the domain name where PII is present and possible providers of PII in CT logs.

3.4 Labeling

To answer RQ2, the domains will be split into labels, which means that *"john.doe.nl"* will be split into "john", "doe", and "nl". The last label is called the extension, which will be ignored in this research as we know that no PII will be present in this part. The second-from-last label is called the parent zone of the domain. Any further labels present will be called "first label", "second label" until the deepest label where we find PII After splitting the domain, we can analyse each part of the domain to see where PII occurs most often and to find possible common denominators, such as website providers that forward PII in CT. For example, we might find domains like *"johndoe.random_provider.nl"*, which means that the *random_provider* is hosting John Doe's website and registered the subdomain in CT. This might be an issue since *John Doe* might not know about their name being registered in this public log.

3.5 Inferring time of registration

To answer RQ3, we will aggregate the domains that include PII over time. The "not before" field of each certificate is used to determine when it was registered. Even though the "not before" field is not exactly the time of registration, we assume that these two moments are relatively close to each other where the difference is negligible.

The "not before" timestamps will be aggregated per month and per year in order to spot trends in the amount of PII present in the certificates. The amount of PII present will be normalized using the total number of certificates requested per month as well.

4. RESULTS

The Google Pilot CT log contains a little more than 1 billion public-key certificates. Since one certificate can contain more than one domain name, we explode [5] the certificates to create one entry per domains, resulting in a total number of almost 5 billion domains. Since certificates expire after a period of time, we distinct all domains, resulting in almost 600 million unique domains. By only taking distinct, we for instance eliminate all but one registration of *john-doe.nl*, since otherwise we would count the same domain more than one time.

Since we will filter the domains on Dutch first and last names, we will only use the domains that have the ".nl" extension in order the number of false positives. This results

[5] https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.functions.explode.html

Property	Number
Total certificates in Google Pilot	1,077,200,000
Total domains in certificates	4,951,000,000
Total unique domains	578,000,000
Total unique domains in ".nl" space	6,400,000
Total unique domains in ".nl" space with first and last name	187,196

Table 1. The dataset in numbers

Location in domain	Number of domains with PII
Parent zone	174,329
First label	5,173
Second label	541
Third label	300
Fourth label	12
Fifth label	2

Table 2. The number of domains where PII is present split per label

in 6.4 million domains. After filtering these on first and last names, roughly 190.000 domains remain: the number of domains with potential PII present. See table 1 for an overview of the exact numbers.

After splitting the domains in parent zone and labels, we see that the vast majority of PII is present in the parent zone of the domain (174,329). The first to fifth label have resp. 5,173, 541, 300, 12, and 2 domains with PII present, see table 2. Deeper labels did not include PII. Note that these numbers do not add up to 187,196 since in some domains the first and last names are present in separate labels. Those cases are not included in the table.

Since we try to find patterns in the PII-containing domains, we analyse the most-frequent occurring parent zone. The top-5 can be found in table 3. Here we can see that the parent zone *"vpweb.nl"* has significantly more subdomains registered than the others. The second-most registered parent zone, *"cas.ms"*, is the Cloud Access Security from Microsoft [6], which means that Microsoft too forwards PII in CT. This has not been analysed further in this research, partially due the fact that "Cas" is also in the list and hence only a last name is included in these domains.

In order to observe trends in PII presence in domains over time, we aggregated the domains' "not-before" date per month per year. When filtering the domains to only include distinct domains, see section 3.5 , we took the minimum value of the "not before" date, meaning that if a domain has two certificates in the CT log for two different periods, i.e. 01-01-2018 until 31-12-2018 and 01-01-2019 until 31-12-2019, the first date is taken, in this example 01-01-2018. We explicitly made this decision since, as mentioned earlier, the first occurrence of the domain in CT is the moment that the PII of a person is leaked. The result of this aggregation can be found in figure 2.

[6] https://docs.microsoft.com/en-us/cloud-app-security/what-is-cloud-app-security

Parent zone	# Registered subdomains
vpweb.nl	636
cas.ms	401
amsterdam.nl	224
fleurglansbeek.nl	218
vriesencooutdoorliving.nl	195

Table 3. Top-5 most occurring zones with PII

We see that the absolute number of certificates with PII presence has increased over time. When we take a look at the total number of certificates issued over time, we see the same increase. This can be observed in figure 3. After normalizing these two figures, we end up with the percentage of certificates with PII present. This can be observed in figure 4.

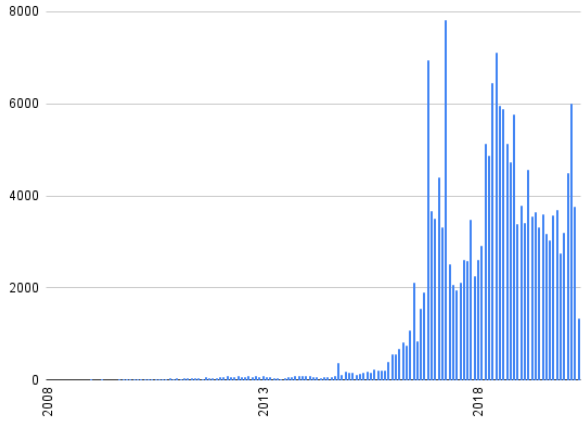

Figure 2. Number of certificates with PII in Google Pilot

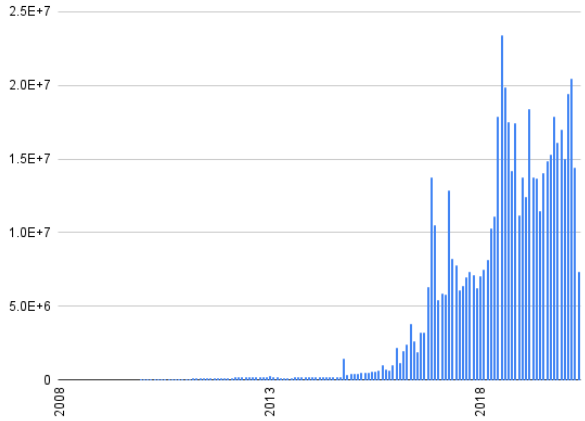

Figure 3. Number of certificates in Google Pilot

5. CONCLUSION

To answer RQ1 whether we can find PII in publicly accessible CT logs, we queried and analysed the Google Pilot CT log. Our results show that 190.000 unique domains include personally identifiable information. This demonstrated that we indeed can find personally identifiable information in CT logs.

Additionally, we analysed several characteristics of the certificates with PII present. We observed that the PII is most-often present in the root label, but to a lesser extent in the first label too.

Furthermore, we saw that there are some root labels that registered multiple subdomains. The most-registered root label is "vpweb.nl" with more than 600 subdomains. When we take a look at what vpweb is, we find out it is part of Vistaprint where customers can easily create and manage a website. Even though the customers of Vistaprint determine the domain name of their website themselves, they may not be aware that Vistaprint requests certificates on

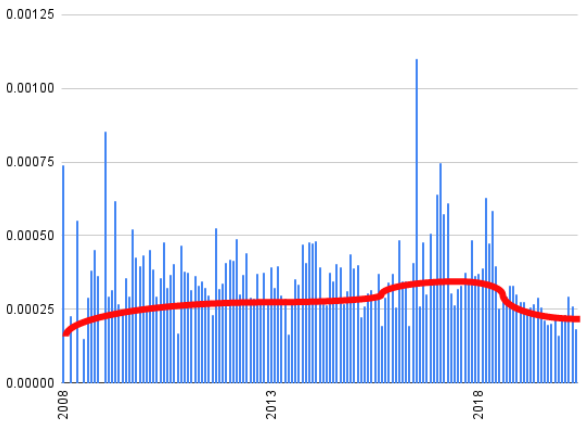

Figure 4. Percentage of certificates with PII in Google Pilot, including a trend line

their behalf, let alone that the inclusion of these certificates in CT makes it much easier for their first and last name to be found.

We also looked into the "robots.txt" file [7] of multiple "vpweb" subdomains. In these files, information is stored that tells web crawlers which areas of the website should or should not be scanned. This also tells search engines like Google whether or not to index the website. The robots.txt file of most "vpweb" websites did *not* prohibit web crawlers from scanning the website, hence the websites are indexed by Google. This would mean that the domain names are also included into the search algorithms of Google and are not private subdomains.

Another interesting fact about the certificates, is that we observed approximately 1 million certificates which a "not before" date that dated from before CT started. Most of these certificates where issued in 2010 or 2009, meaning that they had not yet expired before CT started and thus were included in CT. However, there are around 300 certificates that have a "not before" that of 1999 or 1998. After looking into these certificates, we found out that these are the root certificates of the Certificate Authorities [8], which of course also need to be trusted, hence need to be present in CT. Some of these certificates expire in 2036, meaning that they are probably still used as root certificate.

Finally, in order to answer RQ3, we plotted the number of PII-containing domains with their "not before" date. We can clearly see an increase over time in the total number of domains with PII presence. However, since the total number of issued certificates also increased over that time, we have plotted a normalized time-series which shows the percentage of domains containing PII. In figure 4, we see a somewhat constant percentage of around 0.025% of the certificates containing personally identifiable information. At the end of the plot, we see a slight decrease, but not enough to draw a solid conclusion.

To conclude, with this work we revealed a potential privacy-related issue in the Certificate Transparency infrastructure by conducting an at-scale analysis of all certificates in the Google Pilot CT log. Additionally, we laid down a proof-of-concept for future research to expand upon.

[7] https://developers.google.com/search/docs/advanced/robots/intro

[8] https://docs.microsoft.com/en-us/previous-versions/windows/it-pro/windows-server-2003/cc778623(v=ws.10)?redirectedfrom=MSDN

6. ETHICAL CONSIDERATIONS

Since this paper creates a proof of concept for future work, we aim to publish as much as possible of our work. This includes this paper, the code base and additional data that we gathered for this research. We chose *not* to publish the raw results of these results, since those include personally identifiable information.

7. DISCUSSION

Even though we could clearly answer the set of research questions, there are still several parts of this research that need to be discussed and taken into account.

7.1 Wildcard registration

PII provider in CT, such as Vistaprint, could chose to use wildcard ("*") domain name registration, which registers all subdomains one label deeper. For instance, "*.example.com" registers "foo.example.com" and "bar.example.com" but not "foo.bar. example.com". Since Vistaprint by-default registers "builder.person.vpweb.nl", "preview.person.vpweb.nl" and more subdomains, this wildcard will not catch all subdomains. A possible solution for this would be nested wildcards, like "*.*.vpweb.nl", which would cover the example cases above.

A possibility why wildcard registration is not used, is the fact that Vistaprint might not want to share their private key if subdomains are hosted on virtual private servers, for instance. By registering all subdomains on separate certificates, Vistaprint does not have to share their private key. No further investigation into this has been done in this research.

7.2 Domains without PII in results

Not all domains that are included in our list of domains that include PII actually include PII. These domains, such as "tarievenlijst.postnl.nl", are included in the results, because it in fact includes a first and last name that is in our frequently-used names, in the example "Arie" (first name) and "Ven" (last name), but the domain semantically does not have anything to do with persons. There is no straightforward way to exclude these domains.

7.3 String matching

Within the frequently-occurring first and last names, there are some examples where a first name includes a last name or vice versa. Examples of this are (first & last name resp.):

- Bo & Bos
- Jan & Jansen / Janssen
- Noor & Noord
- Peter & Peters
- Adam & Dam

Even though in these cases a domain might only include a first or last name, but is registered as including both, it is still not a domain that includes PII. A possible solution for this is to only include domains where the first and last name do not have any overlap. Due to time constraints, this was not implemented for this research.

7.4 Reproducibility

To facilitate reproducibility and for other to build on this work, we published the codebase and extra datasets used in this research. This can be found at https://github.com/FrankvanMourik/finding_pii_in_CT_logs

7.5 Future work

Since this research provides a proof of concept for future work to extend upon, we provide future researchers with possible topics to look into.

7.5.1 Better name matching

First of all, the name matching used in this research is far from perfect. Even though it does include all domains with possible PII, it also includes a large amount of *false positives*: domains without PII that do end up in the results. Better name matching can effectively reduce false positives, for instance by excluding matches that are non-names.

7.5.2 Use PII in domains as a gateway to more PII the certificates

As mentioned in section 3.1.1, only a small part of the certificates was used in this research. Future work could look into finding personal information in other fields, such as the organization, country/state/location or common name of the certificate's subject,while using the PII in domains as gateway .

7.5.3 Splitting names per decade

Future researchers could look into splitting the names dataset per decade and then try to find differences per decade. Possible privacy-awareness could be tested using this approach.

7.5.4 Compare cultures

This research aimed at Dutch domains and Dutch first and last names. Further research could be conducted into extending this principle to other cultures throughout the world. Possible differences between cultures could be spotted.

7.5.5 Microsoft CAS

As mentioned earlier in the paper, Microsoft CAS might also be a provider of PII in CT. This would mean that a world-leading IT company leaks private information in public logs. This could be an interesting topic to dive into.

References

[1] S. Eskandarian, E. Messeri, J. Bonneau, and D. Boneh. Certificate transparency with privacy. 2017. URL https://arxiv.org/pdf/1703.02209.pdf.

[2] Google. Https encrpytion on the web: Certificate transparency. 2011. URL https://transparencyreport.google.com/https/certificates?hl=en.

[3] D. Kales, O. Omolola, and S. Ramacher. Revisiting user privacy for certificate transparency. 2019. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8806754&tag=1.

[4] W. Lueks and I. Goldberg. Sublinear scaling for multi-client private information retrieval. In *Financial Cryptography and Data Security - 19th International Conference, FC 2015, San Juan, Puerto Rico, January 26-30, 2015, Revised Selected Papers*, pages 168–186, 2015. doi: 10.1007/978-3-662-47854-7_10. URL http://dx.doi.org/10.1007/978-3-662-47854-7_10.

[5] t. Meertens Instituut. De nederlandse voornamenbank. 2021. URL https://www.meertens.knaw.nl/nvb/.

[6] R. Roberts and D. Levin. When certificate transparency is too transparent: Analyzing information leakage in https domain names. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*, WPES'19, page 87–92, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368308. doi: 10.1145/3338498.3358655. URL https://doi.org/10.1145/3338498.3358655.

[7] J. M. Spring and C. L. Huth. The impact of passive dns collection on end-user privacy. *Securing and Trusting Internet Names*, 2012.