Bachelor's Thesis Computing Science

Radboud University Nijmegen

A Study of CNAME Cloaking-based Tracking on Dutch Websites

Author:
Michiel van de Noort
s1019223

First supervisor/assessor:
Assistant Professor Gunes Acar

Second assessor:
Associate Professor Erik Poll

June 8, 2022

Abstract

In order to circumvent measures aimed to stop web tracking, tracker providers are implementing a technique named CNAME cloaking-based tracking, which violates privacy principles and comes with security risks. Most of the top visited websites world-wide have implemented these trackers, but are they also starting to catch on in the Netherlands? In this thesis we analyse the top 10.000 most popular Dutch (.nl) websites and compare our results with those obtained in the first in-depth analysis of this new tracking technique. We determine that although not yet commonplace, this tracking technique is starting to gain popularity and new tracker providers are expanding to Dutch websites.

Contents

Chapter 1

Introduction

On the modern web, there exists a constant arms race between the development of tracking techniques and anti-tracking measures. As new measures are developed to combat third-party tracking cookies[11][18], trackers are starting to implement new techniques in order to continue tracking web users. One of these new techniques gaining traction is called CNAME cloaking, a technique which allows tracking cookies to circumvent conventional anti-tracking measures by disguising their origin.

This technique is not without its risks; using CNAME cloaking-based trackers can enable Cross Site Scripting[6] or session fixation[1][5][6] attacks. These vulnerabilities negatively impact not only the users' privacy, but also their account security.

Because CNAME cloaking can impact peoples' privacy and account security, it is important to know how prevalent this new technique has become. Large-scale analyses of this tracking scheme have been performed and discovered that this scheme is gaining popularity globally[4][6]. In this thesis we focus on Dutch domains and compare our results with those gathered in the first in-depth analysis of CNAME cloaking-based tracking[4] to determine if this issue has become more prevalent and to see if new tracking agencies have expanded to the Netherlands.

In chapter 2, we present previous work and show where our studies differ and overlap. In chapter 3, we explain technical terms and knowledge necessary to understand the research subject. Then in chapter 4, we explain our research process and the motivations behind it. Following this, in chapter 5 we present and discuss our results and compare them with those obtained in previous work. In chapter 6 we discuss our results and limitations, and discuss possibilities for future work. Finally in chapter 7 we present the conclusions of our research as well as future work.

Chapter 2

Related Work

Multiple research papers have been written on the subject of CNAME cloaking-based trackers. "Characterizing CNAME Cloaking-Based Tracking on the Web"[4] written in 2020 by Dao et al. was the first in-depth analysis of this tracking technique, and is of special importance to our thesis since we compare our results with a subset of their results which contains only Dutch websites.
In their study, Dao et al. built their data set by crawling the Alexa top 300K websites using a standard OpenWPM[8] web crawler. For their blocklist, Dao et al. used sub-lists from the EasyPrivacy[7] and Adguard tracking protection[20] filter lists to construct their own CNAME cloaking-based tracker blocklist. Although we used a different list of websites, web crawler and blocklist than Dao et al., our methodology for detecting trackers is based on theirs, there is significant overlap between our blocklists and we found comparable results for Dutch domains.

In "The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion"[6] (2021) Dimova et al. perform a rigorous in-depth analysis of CNAME cloaking-based tracking. Dimova et al. used crawl data leveraged from the HTTP archive[2] in their analysis. Besides this data set, they also analysed crawl data collected using a modified Tracker Radar Collector and a Tranco list. Besides providing a large scale and longitudinal analysis of CNAME-based tracking, they also discovered and discussed multiple privacy and security issues as well as the effectiveness of recently developed countermeasures.

"Towards Understanding First-Party Cookie Tracking in the Field"[5] (2022) by Demir et al. is the most recently published research paper on this subject. In this paper Demir et al. show that CNAME cloaking-based tracking is becoming more prevalent and can be found on as many as 69% of the worlds' 15.000 most popular websites.

Demir et al. constructed their data set by crawling a Tranco list of the top 15.000 websites with OpenWPM. They also used the EasyPrivacy and AdGuard filter lists as their blocklist, but in contrast to Dao et al., at the time of this study the filter lists did contain CNAME domains belonging to the tracker providers *TraceDock* and *Plausible Analytics*.

Chapter 3

Preliminaries

In this section we discuss components of the web that are relevant to our work. We explain how web requests work, what cookies are and why they are necessary, how tracking in general works and what the differences between third party and CNAME cloaking-based tracking are.

3.1 Web requests

When a person visit a web page, their browser starts sending out requests to fetch resources needed to load the page such as images, scripts, cookies and style sheets. These resources all come from somewhere, sometimes from the same web server as the visited website is hosted on, other times from a server belonging to a third party.
Every resource falls in one of three categories, depending on where they originate from, they are either *same-origin*, *same-site* or *cross-site*. What category a resource belongs to depends on whether the resource shares its scheme (http:// or https://), host (e.g. www.website.com) and port number with the embedding website. Table I shows what categories resources belong to for the embedding site *http://www.example.com:80*.

Table I
Examples of same-origin, same-site and cross-origin for the embedding site http://www.example.com:80

URL	Category	Explanation
http://www.example.com	Same origin	Same as host site
http://www.example.com/images/image1.png	Same origin	Only path differs
https://www.example.com	Same site	Different protocol, but same domain
http://www.store.example.com	Same site	Subdomain of example.com
http://www.tracker.com	Cross origin	Different domain as host

Before your browser connects to a server containing resources, it resolves the servers' domain name to an IP address. It does this by running a DNS

query, which returns either an A record which contains the domain name's corresponding IPv4 address, an AAAA record which contains an IPv6 address, or a CNAME record which contains another domain name. In this case the new record is queried which yet again returns either an A, AAAA or CNAME record, this process can be iterative, continuing until either an A or AAAA record is returned.
When you visit a website *website.com* and this site has resources located on *x.website.com*, then you can assume that those resources are same-site. However, because a domain name can resolve to a CNAME record and another domain name, it is possible that requests sent to *x.website.com* are actually routed to *y.tracker.com*. This phenomenon is what we call *CNAME cloaking*.

3.2 Cookies

Cookies are small blocks of data that are used by web browsers to maintain state. When a web page is loaded, the web server that hosts the page sends one (or multiple) cookies to the browser to enable certain functionalities. These cookies enable everything which requires information to be stored such as: shopping baskets, remembering previously entered information, storing user information (such as pages visited and buttons pressed), or being able to create accounts and log in. Cookies used to perform these last two functionalities are known as *tracking cookies* and *authentication cookies* respectively.

Most websites embed content from other sites. This means that when you visit a website, say *example.com* which embeds content from different websites such as *a.com* and *b.com*, then all these websites can set cookies in your browser. All cookies are defined as either *first-party* or *third-party*. Cookies that are set by the host website (*example.com*) or JavaScript loaded on the website are first-party cookies and cookies that are set by web servers storing embedded content are third-party cookies.

3.3 Web tracking

Web tracking is the practice of observing and storing information about how users interact with websites by the sites themselves or third parties. Analysing user behaviour can be performed for different goals such as advertising[10], web analytics[22] or usability tests[9].

3.3.1 Third party tracking

When a user visits a website which includes resources from a third-party that employs tracking, this third party can set a cookie in the users' browser. The next time this user visits a website that also includes resources from the same third-party, the cookie is sent along in the request to the tracker. This allows the tracker to build a profile of the user, containing information such as which sites you have visited, or which advertisements you clicked. This technique is called third-party tracking and is the traditional method of tracking users on the internet. However, with major browsers phasing out third-party cookies[19], not only third-party cookies, but also third-party tracking seems to be on their way out.

3.3.2 CNAME cloaking-based tracking

With measures being implemented against third party tracking[18][19], tracking agencies are looking for solutions to keep tracking possible. One such solution is CNAME cloaking or CNAME-based tracking. With CNAME cloaking-based tracking, instead of sending HTTP cookie requests to third-parties (cross-site), the website sends these requests to a *subdomain* of itself (same-site). However, this domain name is actually an alias for another domain, possibly another website. The domain name that this alias belongs to is called the alias' *canonical name* or *CNAME*.

This 'tricks' your browser into believing that the cookie returned by the HTTP request comes from the host website, while in actuality it is sent by a different website. In other words: third-party cookies can be disguised as first-party cookies and tracking remains possible. When this 'trick' is used and the cookie received is a tracking cookie, then we speak of CNAME cloaking-based tracking.

Chapter 4

Research

4.1 The setup

4.1.1 The dataset

To analyze top Dutch websites for CNAME-cloaking based tracking, we used a Tranco List[12] generated on 20 March 2022[1]. This list aggregates the ranks from the lists provided by Alexa, Umbrella, and Majestic from 18 February 2022 to 19 March 2022, using the Chrome User Experience Report (CrUX) to collect the 100.000 most popular websites visited by users from the Netherlands.
Since we want to research Dutch websites, we filter out all URLs that do not end in .nl from this list. We then reduce this list of .nl domains to end up with our final list of the 10.000 most visited Dutch websites.

4.1.2 Crawling the data set

In order to crawl our list of websites, we need to use a web crawler. For our thesis we chose to use DuckDuckGo's Tracker Radar Collector[3], TRC is a Puppeteer[15]-based web crawler used by DuckDuckGo to generate third-party request data for their Tracker Radar[17]. We run two crawls: the first crawl is performed on 29/4/2022 and does not interact with consent management platforms. The second crawl is performed on 17/5/2022 and does interact with consent management platforms, giving each visited website consent to process personal information. The reasoning behind running two crawls is that this gives us the possibility to see if interacting with consent management platforms changes the websites' behaviour. Unless specified otherwise, the crawl data used in our research belongs to the second crawl.
The command we use to run our crawls is:
npm run crawl – -i ./crawl_lists/top10000_NL.txt -o ./data_top10000_opt_in/

[1] `https://tranco-list.eu/list/823YV/10000`

-d 'requests,cmps' –run-autoconsent
This configuration uses some options which we explain below:

- -i ./crawl_lists/top10000_NL.txt: -i specifies the input file that contains our list of urls.
- -o /data_top10000_opt_in/: -o specifies the output directory in which the crawl results are to be stored.
- -d 'request, cmps': -d specifies which data collectors are to be used, we use the parameter 'request' to collect all HTTP requests and a modified version of 'cmps'. The 'cmps' collector is used to deal with consent management platforms (cookie consent) and is used in conjunction with the *–run-autoconsent* option. We modify this collector because running it in its unedited state causes the crawler to automatically deny all cookies upon sites it visits, while we wish to accept all cookies.[2]
- –run-autoconsent: this option activates the 'cmps' collector.

Besides modifying how the crawler interacts with consent management platforms we use a standard implementation of TRC, this means the following:

- The crawler runs a headless Chromium browser.
- A fresh user profile is created for every crawl.
- The crawler visits only the homepage for every website.
- The crawler waits 30 seconds to allow the website to load. If the website does not (fully) load within this time frame, this visit fails.
- When a website is loaded, the crawler waits an additional 2.5 seconds to collect queries.

For each website that is successfully crawled, the crawler creates a JSON file in the output directory. These files contain all data collected by the crawler such as: the website's URL, the time at which the website was visited (in Unix time) and HTTP requests sent. At the end of our crawl 8901 websites were successfully visited and 1099 visits failed, these failed visits are further discussed in the discussion.

[2]We publicly released this modified version on Github: `https://github.com/michielvdnoort/tracker-radar-collector-consent-to-cookies`

TABLE II
An overview of the distribution of HTTP requests.

Metrics			Count	Percentage
3rd party requests			436.866	50,22%
1st party requests	domain		110.410	12,69%
	subdomain	w/o CNAME	189.313	21,76%
		w/ CNAME	133.320	15,33
Total requests			869.909	100%

4.2 Analysis of crawl data

We start by loading the crawl data into a Jupyter notebook for further processing. For each file, we extract all URLs to which HTTP requests were sent and place them in a pandas[13] dataframe. We then strip these URLs by removing http(s):// and the URLs' paths to end up with only URLs in the form: *subsite.site.tld*. We filter out all third party requests, so that only first party requests are kept. For these we check which URLs are the same as the site visited and which are subdomains of it, keeping only subdomains.

At this point, all remaining URLs are those to which same-site HTTP requests were sent. For this thesis, we decided that sites of the form: *www.example.nl* also count as subdomains of *example.nl*. An overview of the distribution of HTTP requests can be seen in table II.

Having collected a list of URLs to which same-site HTTP requests were sent, we use pydig[16] to run DNS queries on these URL names[3] and keep only those that point to a CNAME of a different site. Finally, we run the domain names in those CNAME records against the nextDNS[14] CNAME cloaking blocklist. Whenever we get a match, we flag the subdomain this CNAME belongs to.
This leaves us with only HTTP requests made to install tracking cookies via CNAME cloaking. An overview of the research process is detailed in figure 4.1.

[3]The script used to run these queries can be found in the appendix.

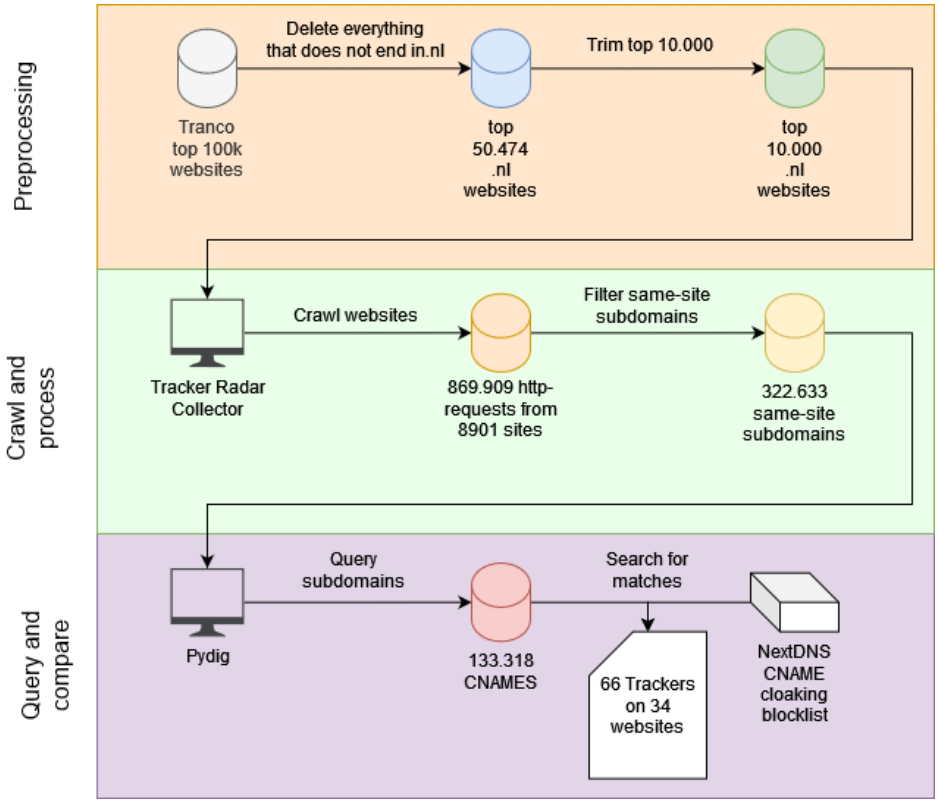

Figure 4.1: An overview of the research process

Chapter 5

Results

In this chapter we present our results. We first go over the obtained results in general in section 5.1, then in section 5.2 we compare our results with those obtained by Dao et al.[4] in 2020. In section 5.3 we present other results of note, finally in section 5.4 we present an analysis of the 10 most frequent CNAMES in our data set.

5.1 General results

From our research we found a total of 66 HTTP requests to set CNAME-based tracking cookies divided over 34 websites in our data set of 8901 successfully crawled websites. These requests went out to 11 unique tracker domains belonging to eight different tracker providers. For most tracker providers we discovered one CNAME/tracker for all websites we discovered them on. For Adobe and TraceDock however, we discovered multiple tracker domains. For Adobe we found three unique tracker domains and for Trace-Dock two. The distribution of these tracker providers is shown in figure 5.1, table III shows an overview of which tracker domains we found and to which tracker provider they belong.

An overview of all websites containing CNAME-based trackers, the CNAME that installs tracking cookies, the tracker provider and number of trackers on that site can be found in the Appendix and on Github[1].

[1] `https://github.com/michielvdnoort/dutch_cname_cloaking_sites`

Table III
Legend of tracker CNAMES and corresponding tracker provider

Tracker CNAME	Tracker provider
sc.omtrdc.net	Adobe Experience Cloud
data.adobedc.net	Adobe Experience Cloud
2o7.net	Adobe Experience Cloud
actonsoftware.com	Act-On
a351fec2c318c11ea9b9b0a0ae18fb0b-1529426863.eu-central-1.elb.amazonaws.com	TraceDock
afc4d9aa2a91d11e997c60ac8a4ec150-2082092489.eu-central-1.elb.amazonaws.com	TraceDock
wt-eu02.net	WebTrekk
custom.plausible.io	Plausible Analytics
affex.org	Ingenious Technologies
hs.eloqua.com	Oracle Eloqua
at-o.net	AT Internet

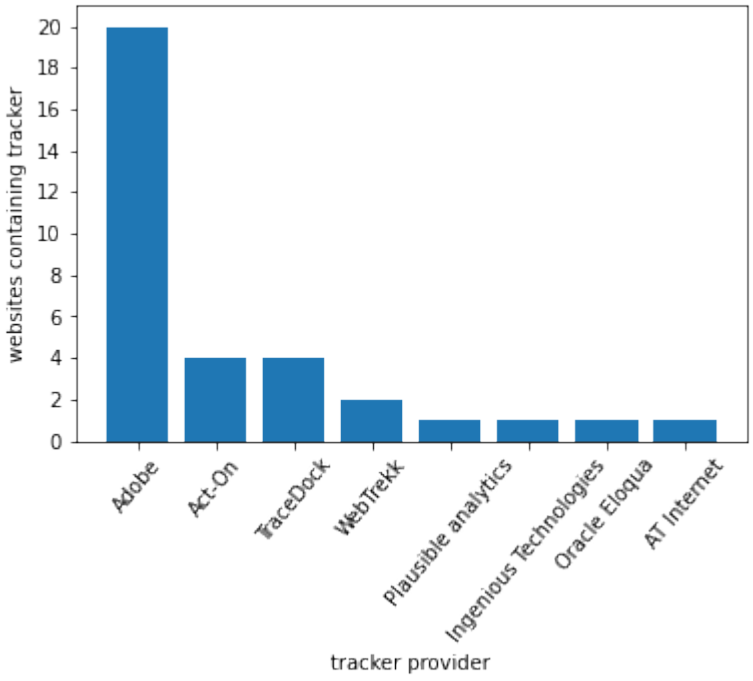

Figure 5.1: An overview of tracker providers and how many websites they were discovered on

5.2 Comparison with previous work

In January of 2020, Dao et al. performed the first in-depth analysis on CNAME cloaking-based tracking. They crawled the Alexa 300k data set and compared the CNAMEs of same-site subdomains with the EasyPrivacy[7] and Adguard tracking protection[20] blocklists. Their data set contained the top 300k websites world-wide, which included roughly 6500 Dutch websites. In chapter IV of their paper, Dao et al. went into detail about their results separated by country domain. For Dutch websites, they discovered the following:
0.3% of .nl domains contained a CNAME cloaking-based tracker, which were 19 websites in total. These 19 websites contained trackers from Adobe, Act-On, Oracle Eloqua, Webtrekk and Ingenious technologies. Table IV and figure 5.2 compare the distribution of tracker providers as discovered in 2020 with the results of this study. Since our data set of Dutch domains is larger than the one used in 2020, we also show how our results would change if we only analysed the first 6500 sites in our data set. Manual inspection showed that the only sites in the range 6501 - 8901 containing CNAME-based trackers are *maatwerkonline.nl* and *yoobi.nl*. Our results excluding these two websites are shown in the second row of table IV, as well as visualised in figure 5.2.

Table IV
Breakdown of tracker providers found in 2020 by Dao et al. versus this study

Year	Adobe	Act-On	Webtrekk	Oracle	Ingenious	TraceDock	Plausible Analytics	AT Internet
2020 (Dao et al.)	13	2	2	1	1	0	0	0
2022 (same-size data set as Dao et al.)	20	3	2	1	1	3	1	1
2022 total results	20	4	2	1	1	4	1	1

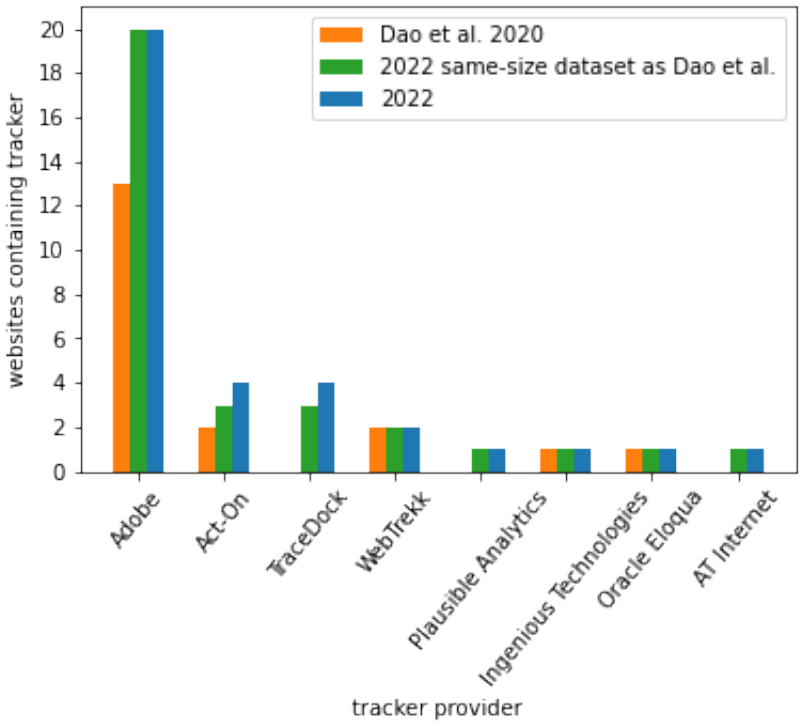

Figure 5.2: A comparison of tracker providers found on websites

As the table and figure show, there is a clear increase in Dutch websites containing CNAME-based trackers. Where 0.3% of the top 6500 Dutch websites contained CNAME-based trackers in 2020, this number has now grown to 0.49%.
We also discovered tracker providers that previously were not detected on Dutch domains, namely: *TraceDock*, *Plausible Analytics* and *AT Internet*. It should be noted that TraceDock and Plausible Analytics were not present in the final blocklist[2] used by Dao et al. and it is therefore entirely possible that these trackers were already active, but undiscovered back then. AT Internet however, is almost surely a new addition to the CNAME-based tracker providers active in the Netherlands, seeing as we found it on *npo.nl* which is in the top 30.000 most popular websites worldwide and was therefore almost certainly one of the websites analysed in 2020.

5.3 Other results of note

As mentioned in chapter 4.1.2 we crawled our list of websites two times: first without interacting with consent management platforms and then a second

[2] `https://github.com/fukuda-lab/cname_cloaking/blob/master/updated/tracking_provider.txt`

time in which we accepted all cookies. The first crawl was performed on 29/4/2022 and the second on 17/5/2022. In this subsection, we compare the results of these crawls to determine which websites in our data set respect cookie consent and show two websites that stopped using CNAME-based trackers during this time period, as well two websites that started using these trackers.

5.3.1 Websites that stopped using CNAME-based trackers

While the initial intent behind performing two crawls was purely to study how websites handle cookie consent, we noticed during the inspection of our second crawl's results that some URLs were missing and others had appeared. During the first crawl the websites *mazda.nl* and *eurosport.nl* were discoverd to have sent out HTTP requests to domains associated with CNAME-based tracker providers, these providers were Oracle Eloqua and Adobe respectively. However, during the second crawl we found no such requests on either site, suggesting they possibly have stopped using CNAME-based tracking. We manually inspected these websites and verified that no more requests went out to domains related to CNAME-based trackers.

5.3.2 Websites that started using CNAME-based trackers

During the second crawl, we found two websites with trackers that were not detected in the first crawl: these websites were *blgwonen.nl* and *cz.nl*. The tracker provider for blgwonen.nl is Adobe, which is the most prevalent CNAME-based tracker. The tracker provider for cz.nl is Oracle eloqua, which we detected on no other sites in our data set.

5.3.3 CNAME cloaking-based tracking and cookie consent

After comparing the results of our two crawls we can conclude that nearly all websites in our data set that utilise CNAME cloaking-based trackers ignore whether the crawler consents to the processing of personal information or not and sends requests to CNAME-based tracker domains regardless of interaction with the consent management platform. The only website for which this is not the case is cz.nl, which is a healthcare insurer. After additional manual testing we can confirm that this website only attempts to load a CNAME-based tracker when consent is given. According to cz's cookie policy[3], this is a deliberate choice, health-related information is not collected and personal data is collected with the intent to create personalised emails for the user.

[3]https://www.cz.nl/cookies/overzicht

5.4 Most frequent CNAMES

Of the 133.318 same-site HTTP requests that turned out to be CNAMEs, only 66 matched CNAME-based tracker domains in our blocklist. Since our blocklist might not be complete, we ran the 10 most occurring CNAMEs in our data set against the EasyPrivacy, Adguard Tracking Protection and Tracker Radar filter lists. We also manually searched the internet to match these CNAMEs with the company they belong to.
Although we did not discover new CNAME-based trackers, we did find one domain -privacy.mgmt.com- which corresponds to a consent management platform and was to the best of our knowledge not yet categorised online. An overview of these CNAMEs can be found in table VI in the appendix, or on Github[4].

[4] https://github.com/michielvdnoort/dutch_cname_cloaking_sites/blob/main/possible_trackers_autoconsent.csv

Chapter 6

Discussion

In this chapter we discuss our results and their implications, the limitations of our study and possible future work.

6.1 Discussion of results

Some of our results were more surprising/interesting than others. In this section, we present and discuss some of the more interesting websites on which we discovered CNAME-based tracking.

6.1.1 cz.nl

As mentioned in chapter 5.3.3, cz.nl is the website of a Dutch healthcare insurer with the same name. According to their privacy statement, they only utilise CNAME-based tracking to personalise e-mails and do not track you on pages containing information regarding the user's health.

6.1.2 consumentenbond.nl

The consumentenbond is a non-profit organisation which protects the interests of consumers. According to their cookie policy[1] they use CNAME-based as well as regular trackers to advertise their own services to potential customers on other websites. While they do share this data with Google, Microsoft, Facebook and Adobe; they claim that they have given none of these organisations permission to use the data for other ends.

6.1.3 npo.nl and jeugdbibliotheek.nl

The npo stands for 'Nederlandse Publieke Omroep' or 'Dutch Foundation for Public Broadcasting' and is a governmental organisation. The jeugdbib-

[1] `https://www.consumentenbond.nl/over-ons/voorwaarden-en-privacy/cookies?icmp=footer_tekstlink_cookiebeleidonderaan`

liotheek or 'Youth Library' is a website belonging to the 'Royal Libary of the Netherlands'. Both of these organisations belong to the Dutch ministry of Education, Culture and Science.
What else is interesting about these websites is that in their cookie policies[2][3] both websites state that the cookies set via CNAME cloaked domains are not used for advertising, but analytics.

6.1.4 asnbank.nl, blgwonen.nl, regiobank.nl and snsbank.nl

All these websites belong to banks that are part of 'de Volksbank' or 'the People's Bank' and therefore have the same privacy and cookie policy[4]. They state that they use analytic cookies, but no cookies for advertising purposes and that data collected on users is only shared between these four banks, the Dutch Central Bank, the Financial Market Authority and the European Central Bank (ECB).

The websites above show that cookies set via CNAME-cloaking are used for more than just advertising. Some of these websites state they use analytic cookies because they are required by law to make reports to the government, and they need analytic cookies to do so. It is possible that with measures being implemented against third party tracking, these websites started using CNAME cloaking to keep their analytic cookies functional.

6.2 Limitations of our study

As chapter 4.2 shows, CNAME-based tracking is more prevalent on Dutch websites at the time of this study than it was in 2020. Where Dao et al. discovered CNAME-based trackers on 19 out of the 6500 Dutch websites they analysed, we discovered these on 32 out of 6500 websites. Although this is a clear increase, we discovered less trackers than anticipated. It is possible that we could have found more trackers if we configured our crawler differently. The crawler used by Demir et al.[5] used multiple extra measures to find as many cookies as possible, these measures included:

- Using a longer wait time for every visit.
- Faking user interactions (i.e. random scrolling and mouse clicks).
- Loading a base profile on each page visit (because websites tend to use more cookies if the cookie jar and local storage of a browser instance is populated)[21].

[2] `https://www.jeugdbibliotheek.nl/cookies.html`

[3] `https://cookies.npo.nl/sites/NPO/npo.nl/settings.html?version=v3.1.14-ebllf&referrer=https`

[4] `https://www.devolksbank.nl/over-ons/privacyreglement`

Modifying our crawler or using OpenWPM with a similar configuration could allow us to find more trackers in the data set.

As mentioned in chapter 4.1.2, 1099 out of 10.000 or 10.99% of page visits failed. Most likely, these failed visits happened due to time-outs, but because no destination path for log files was specified in the configuration of our crawler we do not posses the log files required to verify this. It is therefore unknown why these page visits failed and whether (some of) these fails could have been prevented.

6.3 Future work

For future work, we can try multiple website lists or crawlers/crawler configurations to get a more complete data set. In this study we used only websites with a Dutch top-level domain (.nl), but not all Dutch websites use .nl. Extending our list with Dutch websites using other domains and crawling it with Duckduckgo's Tracker Radar Collector as well as OpenWPM could result in a more representative data set.

Besides improving upon our methods, we can also further analyse our results. One such analysis would be to categorise the websites on which we discovered CNAME-based tracking. Categorising these websites would enable us to determine if certain website categories are more likely to use CNAME-based tracking or whether websites in the same category tend to use the same tracker providers. Another analysis would be to categorise to which goals (advertising, analytics, usability test) CNAME-based trackers are used by the websites in our results.

Chapter 7

Conclusions

In our thesis we collected a list of the top 10.000 websites with a Dutch domain (.nl), then crawled this list using a version of Duckduckgo's Tracker Radar Collector that we modified for this study. Subsequently we analysed the crawl data gathered by TRC to detect CNAME cloaking-based trackers.

We discovered 66 HTTP requests to CNAME-based trackers on 34 websites in total.

We compared our results (the top 6500) with previous work from Dao et al. and discovered five trackers from three tracker providers that were not present in 2020, these trackers providers are TraceDock, Plausible Analytics and AT Internet. Furthermore, we confirmed Adobe is still the largest CNAME cloaking-based tracking provider for Dutch websites, but has lost some ground. Where Adobe trackers were found on $68,4\%$ of websites containing CNAME-based trackers in 2020, they are now found on $62,5\%$ of those websites. We also discovered that the amount of CNAME-based trackers found has increased from 19 to 32 websites and that therefore the percentage of websites in the top 6.500 Dutch domains on which CNAME-based trackers have been found has grown from $0,3\%$ to $0,49\%$.

Additionally, after crawling our data set twice, where in one crawl we consented to all cookies and in the other opted out of all cookies. We analysed the data from both crawls and concluded that only cz.nl does not install a CNAME-based tracker unless given explicit consent.

Finally, we discovered that a number of websites in our data set use CNAME-based tracking not for advertisements, but analytic goals.

Bibliography

[1] Assel Aliyeva and Manuel Egele. *Oversharing Is Not Caring: How CNAME Cloaking Can Expose Your Session Cookies*, page 123–134. Association for Computing Machinery, New York, NY, USA, 2021.

[2] HTTP Archive. State of the web report. `https://httparchive.org/`, 2020.

[3] DuckDuckGo's Tracker Radar Collector. Modular, multithreaded, puppeteer-based crawler used to generate third party request data for the duckduckgo tracker radar. `https://github.com/duckduckgo/tracker-radar-collector`.

[4] Ha Dao, Johan Mazel, and Kensuke Fukuda. Characterizing cname cloaking-based tracking on the web, 2020. Research Paper, `https://tma.ifip.org/2020/wp-content/uploads/sites/9/2020/06/tma2020-camera-paper66.pdf`.

[5] Nurullah Demir, Daniel Theis, Tobias Urban, and Norbert Pohlmann. Towards Understanding First-Party Cookie Tracking in the Field. *arXiv e-prints*, page arXiv:2202.01498, February 2022.

[6] Yana Dimova, Gunes Acar, Lukasz Olejnik, Wouter Joosen, and Tom Van Goethem. The cname of the game: Large-scale analysis of dns-based tracking evasion. *Proceedings on Privacy Enhancing Technologies*, 2021(3):394–412, 2021.

[7] EasyPrivacy. Available:. `https://easylist.to/easylist/easyprivacy.txt`.

[8] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of ACM CCS 2016*, 2016.

[9] Interaction Design Foundation. What is usability testing? `https://www.interaction-design.org/literature/topics/usability-testing`.

[10] GFC global. Internet safety: understanding browser tracking. https://edu.gcfglobal.org/en/internetsafety/understanding-browser-tracking/1/.

[11] Georgios Kontaxis and Monica Chew. Tracking protection in firefox for privacy and performance, 2015.

[12] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019.

[13] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[14] nextDNS cnamecloaking blocklist. A list of domains used by tracking companies as cname destination when disguising third-party trackers as first-party trackers. https://github.com/nextdns/cname-cloaking-blocklist, 2019.

[15] Puppeteer. Puppeteer is a node library which provides a high-level api to control chrome or chromium over the devtools protocol. https://github.com/puppeteer/puppeteer.

[16] pydig0.4.0. Python wrapper library for the 'dig' command line tool. https://pypi.org/project/pydig/, 2021.

[17] DuckDuckGo's Tracker Radar. A data set of the most common third party domains on the web with information about their behavior, classification and ownership. https://github.com/duckduckgo/tracker-radar.

[18] Privacy Sandbox. https://www.privacysandbox.com/intl/en_us/.

[19] the Keyword. Building a more private web. https://www.blog.google/products/chrome/building-a-more-private-web/, 2019.

[20] Adguard tracking protection filter. Available:. https://raw.githubusercontent.com/AdguardTeam/FiltersRegistry/master/filters/filter_3_Spyware/filter.txt.

[21] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. *Beyond the Front Page:Measuring Third Party Dynamics in the Field*, page 1275–1286. Association for Computing Machinery, New York, NY, USA, 2020.

[22] Wikipedia. Wikipedia: Web analytics. https://en.wikipedia.org/wiki/Web_analytics.

Appendix A

Appendix

TABLE V
An overview of websites on which CNAME-based trackers were discovered

website	CNAME with subdomain	tracker provider
asadventure.nl	asadventure.nl.ssl.sc.omtrdc.net.	Adobe
asnbank.nl	asnbank.nl.ssl.sc.omtrdc.net.	Adobe
beterhoren.nl	beterhoren.nl.data.adobedc.net.	Adobe
betterplaces.nl	custom.plausible.io.	Plausible Analytics
bever.nl	bever.nl.ssl.sc.omtrdc.net.	Adobe
blgwonen.nl	blgwonen.nl.ssl.sc.omtrdc.net.	Adobe
consumentenbond.nl	consumentenbond.nl.102.112.2o7.net.	Adobe
cz.nl	p06g.hs.eloqua.com.	Oracle Eloqua
disney.nl	disney.nl.ssl.sc.omtrdc.net.	Adobe
douglas.nl	douglas.nl.data.adobedc.net.	Adobe
dyson.nl	dyson.nl.ssl.sc.omtrdc.net.	Adobe
esprit.nl	esprit.nl.ssl.sc.omtrdc.net.	Adobe
esri.nl	esri.nl.ssl.sc.omtrdc.net.	Adobe
essent.nl	essent.nl.ssl.d2.sc.omtrdc.net.	Adobe
home24.nl	lb1.affex.org.	Ingenious technologies
hornbach.nl	hornbach-02.wt-eu02.net.	WebTrekk
jeugdbibliotheek.nl	jeugdbibliotheek.nl.ssl.sc.omtrdc.net.	Adobe
kijk.nl	kijk.nl.ssl.sc.omtrdc.net.	Adobe
large.nl	emp01.wt-eu02.net.	WebTrekk
liander.nl	liander.nl.102.122.2o7.net.	Adobe
maatwerkonline.nl	afc4d9aa2a91d11e997c60ac8a4ec150-2082092489.eu-central-1.elb.amazonaws.com.	TraceDock
milieudefensie.nl	adepci3.actonsoftware.com.	Act-On
npo.nl	atconnect-npo-nl-cddc.at-o.net.	AT Internet
nti.nl	aiepci6.actonsoftware.com.	Act-On
overstappen.nl	a351fec2c318c11ea9b9b0a0ae18fb0b-1529426863.eu-central-1.elb.amazonaws.com.	TraceDock
qwic.nl	afc4d9aa2a91d11e997c60ac8a4ec150-2082092489.eu-central-1.elb.amazonaws.com.	TraceDock
regiobank.nl	regiobank.nl.ssl.sc.omtrdc.net.	Adobe
robeco.nl	robeco.nl.ssl.sc.omtrdc.net.	Adobe
sdim.nl	afc4d9aa2a91d11e997c60ac8a4ec150-2082092489.eu-central-1.elb.amazonaws.com.	TraceDock
snsbank.nl	snsbank.nl.102.122.2o7.net.	Adobe
stepstone.nl	stepstone.nl.data.adobedc.net.	Adobe
unive.nl	unive.nl.data.adobedc.net.	Adobe
webinar.nl	aiepci2.actonsoftware.com.	Act-On
yoobi.nl	adepci4.actonsoftware.com.	Act-On

Code excerpt:
The python script with which we queried same-site subdomains

```
import pydig

with open('subdomains.txt', 'r') as filehandle:
    data = filehandle.read()

data_intermediate = data.split("\n")
for i in range(len(data_intermediate)):
    data_intermediate[i] = data_intermediate[i].split("\t")

for i in range(len(data_intermediate)):
    for j in range(len(data_intermediate[i])):
        if not (data_intermediate[i][j] == ""):
            CNAME = pydig.query(data_intermediate[i][j], 'CNAME'
                )
            if (CNAME == []):
                data_intermediate[i][j] = "no_CNAME"
            else:
                loop = True
                while loop:
                    data_intermediate[i][j] = CNAME[0]
                    CNAME = pydig.query(CNAME[0], 'CNAME')
                    if (CNAME == []):
                        loop = False

with open('cnames.txt', 'w') as filehandle:
    for rows in data_intermediate:
        for value in rows:
            filehandle.write('%s\t' % value)
        filehandle.write('\n')
```

Table VI
An overview of the 10 most frequent CNAMEs in our data set

CNAME	Count	Company	Function	In blocklist?
cloudfront.net.	3178	Amazon Cloudfront	CDN	yes
b-cdn.net.	532	bunny CDN	CDN	yes
kxcdn.com.	491	proInity GmbH	Swiss data processing and hosting company	yes
privacy-mgmt.com.	291	unknown	Consent Management Platform	no
akamai.net.	288	akamai	CDN	yes
nodes.hypernode.io.	278	hypernode	Hosting platform	no
dscb.akamaiedge.net.	198	akamai	CDN	yes
scaliacdn.nl.	180	scalia	CDN and hosting	no
t-0009.t-msedge.net.	154	Microsoft Azure	CDN	no
autointeractiveb.netdna-cdn.com.	135	stackpath	Cloud computing and service provider	yes