

Reverse Engineering of Web Cookies

When is too late for your private data?

A.P. Taneva

Delft University of Technology

Reverse Engineering of Web Cookies

When is too late for your private data?

by

A.P. Taneva

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended publicly on Monday March 08, 2023 at 11:50 AM.

Student number:	4510488	
Project duration:	April 1, 2022 – March 1, 2023	
Thesis committee:	Prof. dr. ir. Georgios Smaragdakis,	TU Delft, supervisor
	Dr. S. Picek,	TU Delft
	Dr. G. Iosifidis,	TU Delft

Cover: Canadarm 2 Robotic Arm Grapples SpaceX Dragon by NASA under CC BY-NC 2.0 (Modified)

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Preface

This thesis represents the final stage of acquiring a degree of Master of Science in Computer Science at the Delft University of Technology, with a specialization in Cyber Security. The initial work on the thesis began in April 2022 and was completed in February 2023.

First and foremost, this project is only possible with the guidance, valuable advice, and motivation of prof. Georgios Smaragdakis as thesis advisor and Stjepan Picek as daily supervisor provided. My gratitude goes further to my family, that supported me through the whole process of my education. With their patience and support, I had the opportunity to complete my studies.

My passion for mathematics, logical thinking, and solving problems has been an inseparable part of my life since my early childhood. When the beautiful world of mathematics was introduced to me by my grandfather. His example and academic success were the driving force for me to start and finalise Computer Science studies at TU Delft.

I dedicate my final work and degree to my grandfather, who was selflessly supportive and never doubted my potential.

A.P. Taneva
Delft, February 2023

Abstract

Nowadays, the online industry contributes to a multi-billion dollar business, facilitates most of the population's everyday activities, and processes vast amounts of data, including personal data. Current work aims to explore the inconsistency or consistency of the content obtained by the websites to generate cookies based on various data that the user provides when visiting a web page, being by explicit consent or not. Some websites integrate with third-party companies that track users and collect their data. For this research, a custom-made Selenium-based web application (crawler) visits the Top 50 Alexa most visited websites and observes the cookies that are collected before the user's consent. After a brief data set analysis, a significant inconsistency in the cookies' count that deviates per location, device type, and operating system is detected. The results show that some websites collect private data, even though users are not informed about the collected data, and the consent for using cookies is not retrieved. These results imply that tracking persists as a serious concern. Tracking raises ethical and legal matters due to its potential adverse effect on users' data. That is why it is essential to analyse the content these trackers obtain, e.g., location, internet protocol address, and browsing history. Subsequently, to suggest possible techniques to avoid tracking.

Contents

Preface	i
Abstract	ii
1 Introduction	1
1.1 Motivation	1
1.2 Problem statement	2
1.2.1 Privacy and Privacy by Design	2
1.3 Research description	3
1.3.1 Research goal and motivation	3
1.3.2 Problem definition	3
1.3.3 Scope of the research	4
1.4 Outline thesis	4
2 Background	5
2.1 Fundamental components	5
2.1.1 OSI model and TCP/IP model	5
2.1.2 Cookies in detail	8
2.2 Related work	10
3 Methodology	14
3.1 System architecture	14
3.1.1 System overview	14
3.2 Users overview	17
3.3 Data collection overview	17
3.3.1 Data decryption	19
3.3.2 Preparation and basic analysis of the results	19
3.3.3 Visualisation of the differences	19
3.3.4 Identifying third-party cookies	20
3.4 Websites	20
3.5 User agents	23
3.6 Locations	24
3.7 Dataset and database structure	25
3.7.1 Database models and description	25
4 Results and analysis	30
4.1 General observations	30
4.1.1 Parameter: total number of cookies and variety of website categories	30
4.1.2 Parameter: number of third-party vs. first-party cookies per website category	32

4.1.3	Parameter: number of third-party vs. first-party cookies per operating system or device type	33
4.1.4	Parameter: total number of cookies per operating system and category	33
4.1.5	Parameter: expiry date	35
4.1.6	Outliers observations about first-party and third-party cookies count in location Dallas and location The Hague per website	36
4.2	Data visualization task	37
4.2.1	Data preprocessing	38
4.2.2	Data observations	38
4.2.3	Data imbalance	41
5	Mitigation	43
5.1	Solutions that require little or no action from the user	43
5.2	Solutions that require some action from the user	44
5.3	Concluding remarks	45
6	Discussion and conclusion	46
6.1	Discussion	46
6.1.1	Research sub-questions	47
6.2	Future work and limitations	48
6.3	Concluding remarks	48
	References	50
A	Source code	53
B	Figures	56
C	Database	60

List of Figures

3.1	System architecture	15
3.2	Database diagram	27
3.3	Table Cookies - design and fields	28
3.4	Table Locations - design and fields	28
3.5	Table Websites - design and fields	29
3.6	Table User Agents - design and fields	29
4.1	Distribution of the number of websites per category	31
4.2	Distribution of total number of cookies per website category for location Dallas	31
4.3	Distribution of first-party and third-party cookies count per website category (location Dallas)	32
4.4	Distribution of first-party and third-party cookies count per website category (location The Hague)	32
4.5	Distribution of total number cookies per operating system per website category (location Dallas)	33
4.6	Distribution of total number cookies per operating system per website category (location The Hague)	34
4.7	First-party and third-party cookies expiry date for different locations	35
4.8	Left: number of third-party cookies per origin and category. Right: number of first-party cookies per origin and category. Location: The Hague	38
4.9	Left: number of third-party cookies per origin and category. Right: number of first-party cookies per origin and category. Location: Dallas	39
4.10	Average duration of cookies per location Left: The Hague Right: Dallas	40
4.11	Left: number of third-party cookies per user agent and category. Right: number of first-party cookies per user agent and category. Location: The Hague	40
4.12	Left: number of third-party cookies per user agent and category. Right: number of first-party cookies per user agent and category. Location: Dallas	41
4.13	Left: K-NearestNeighbour classifier. Right: Random Forest classifier	42
B.1	Distribution of first-party and third-party cookies per operating system (Left: The Hague, Right: Dallas)	56
B.4	Left: AUC Curve K-NN classifier. Right: Precision recall curve K-NN classifier.	56
B.2	Distribution of first-party cookies expiry date per user agent	57
B.5	Left:AUC Curve Random Forest classifier. Right: Precision recall curve Random Forest classifier.	57
B.3	Distribution of third-party cookies expiry date per user agent	58
B.6	Left:AUC curve Decision Tree classifier. Right: Precision recall curve Decision Tree classifier	58
B.7	Left:AUC curve Logistic classifier. Right: Precision recall curve Logistic classifier	59

List of Tables

3.1	List of Alexa Top 50 websites, snapshot from 16th June 2022, including category and country of origin	22
3.2	List of user agents for Google Chrome 105 browser	23
3.3	List of user agents for Windows 10, desktop configuration	24
3.4	List of locations	25
4.1	Database Snapshot for Adobe.com for Dallas and The Hague	37
C.1	List of Alexa Top 50 websites, snapshot from 16th July 2022, number of cookies per website and user agent for location Dallas and The Hague.	73

1

Introduction

Nowadays, the online industry contributes to a multi-billion dollar business, facilitates most of the population's everyday activities, and processes huge amount of data, including personal data.

Data has a crucial role in our society. Many companies and processes nowadays depend entirely on online data. Some of them process and manage private data in a digitised form. If collected data is not stored, processed or treated correspondingly, it can imply a serious threat to the privacy of Internet users. It is hard to navigate the web these days without coming across a box prompting you to accept essential, if not all, cookies before proceeding. Since the creation of cookies in 1994, it has fundamentally changed how data is collected and how Internet users are recognised and distinguished online. With the help of the cookies, the Internet ecosystem provides an environment to track users and collect a vast amount of user data, such as browsing history, location, religious and political beliefs.

1.1. Motivation

The recent years online tracking mechanisms have advanced and prove to be hard to detect and control. Constantly the user actions and behaviour are being tracked and analysed, mainly with the help of web cookies. Tracking involves collecting data from the user activity across certain context and consequently processing this data in a context outside of the first one [20].

Nowadays, all online users encounter HTTP (HyperText Transfer Protocol) Cookies at a certain point when browsing websites and communicating with web servers. Cookies are represented by a text string and are saved on the client's side when a user opens a website through a browser engine. Afterward, when the user revisits the same or other websites, these cookies and data they have collected are part of the request [17].

The cookies are owned and created by the website itself, first-party, or third-party-owned and created by other websites. First-party cookies track the end-user's website activity and accelerate user experience during subsequent visits. In comparison, third-party cookies can track the end-users and are managed by external servers. Currently, the primary browser agents block partially or fully third-party

cookies [11]. In some cases, that moves the large-scale tracking onto first-party cookies, which can provide data to the tracking websites on request. Although first-party cookies are critical to providing good user experiences in multiple web services, such as social networks or commercial websites, simultaneously, they accumulate Personally Identifiable Information [11]. Cookies are agreed to be a commonly adopted technique for tracking users and their behaviour on the Internet.

Not all users are fully aware of what kind of data is transported from their IP (Internet Protocol) address to web servers every moment when they browse online [12]. Even fewer users know how this data is protected and what mechanisms websites are used to track individual visitors. Also, there needs to be more understanding of the extent to which privacy policies are applied to protect personal data. Additionally, the explicit consent requested from the user is vague and usually automatically accepted, and the users mostly ignore its content. Knowing what information websites collect through cookies is valuable, as each suggested countermeasure to avoid tracking and collecting private data will depend on security issues and mitigation techniques.

1.2. Problem statement

Currently, the entire business model of the online companies is designed towards collecting and sharing user data, sometimes referred to as the main economic asset of the company. This model triggers the progress of advanced tracking technologies. These sophisticated technologies and business strategies are hard to comprehend by a user with no technical background.

HTTP Cookies are believed to be the most popular and widely adopted technique for user tracking. Even though cookies enhance the user's experience on the websites while browsing, they may contain Personally Identifiable Information. Tracking by third-party cookies is a common practice and hard to mitigate. Countermeasures like blocking or removing are ineffective and usually only target third-party cookies [2]. The first-party cookies are ignored, as they are associated with user experience and service improvement. But it is not always the case. First-party cookies can also be used to track users and collect personal data [11].

However, while data collection and data exchange via cookies can greatly benefit online businesses, they can cause side effects, negligence, and violation of privacy for Internet users.

1.2.1. Privacy and Privacy by Design

In the present day, privacy and private data remain hot topics. And data, in particular, is a commodity treated like gold in the digital world. According to the United Nations, privacy is a human right [38] and implies freedom from (undue) surveillance [9]. With General Data Protection Regulation (GDPR) [16], the European Union (EU) seeks to harmonise the existing legal framework regarding personal data. As stated, parties responsible for storing and processing privacy-sensitive data should take specific measures to comply with (local) regulations and legislation and preserve privacy accordingly. Sometimes this can go beyond the borders of a single country or economic area and be part of multiple jurisdictions. Cookies fall under the GDPR passed in 25th of May 2018 and addressed in 2002 by ePrivacy Directive (Directive 2002/58/EC, Directive on privacy and electronic communications) of the European Union,

adapted in 2009 (2009/136/EC) [39]. The purpose of the European Union (EU) regarding GDPR is to create harmonized rules and enforce standardised measures on organisations to protect fundamental rights and freedom, particularly for personal data protection [16]. In some cases, GDPR is also applied to companies outside the European Union when they collect, process, track, and analyse sensitive data of European citizens regardless of their origin [16]. The EU acknowledges that privacy-sensitive data consists of two broad categories of data. The first category is personal data (name, postal address, social security number), and the second is sensitive data (religion, sexual orientation, biometric data). The combination of personal and sensitive data can undoubtedly identify an individual, and is often referred to as Personally Identifiable Information (PII). According to the GDPR, a data controller that processes PII should receive a the relevant consent from the user [16].

But legislation alone is not a necessary and sufficient condition to enforce the processing of privacy of sensitive data accordingly. Data privacy should be considered throughout all stages of the engineering process, design, and operation of procedures and business practices. This framework is referred to as privacy by design. Privacy by design states that a company demonstrates the ability to secure and protect the personal and confidential data of its customers, employees, and business partners.

1.3. Research description

This section will discuss the research description and problem definition regarding tracking cookies. The research questions and sub-questions will be introduced.

1.3.1. Research goal and motivation

HTTP cookies are a highly discussed feature of online pages; many policies and regulations provide a boundary that website owners must comply with. These rules deviate significantly from one economic area to another. Most users of websites need to be fully aware of all the information websites are collecting from them, sometimes even before giving consent. This research will try to reveal and address and understand the magnitude of the information has been collected from the users. This is a hot topic and has been addressed in various papers. The current work will try to provide a overview of the tracking cookies and evaluate the findings and proofs regarding the ambiguous tracking cookies. Further, this research will focus on the differences in collected data and analyse the potential consistencies or inconsistencies in this data derived from various user parameters - like location, device, operating system.

1.3.2. Problem definition

This research aims to investigate the inconsistency or consistency of the content obtained by the websites to generate cookies based on various data that the user can provide when visiting a web page, being by explicit consent or not. The main goal of current research is to analyse the content of the collected cookies and, as such, leads to the following question:

"How to reverse engineer the technique of creating cookies based on personalizing the content that is stored within them, and what model can be built that analyses the information stored in the cookies?"

Based on the main research question, the following sub-questions can be formulated:

- How information about the user profile is collected by the tracking cookies, and how does this information deviate per different user parameters (device type, location, browser, operation system)?
- How the extracted user parameters can be used to build a model and analyse the results?
- What strategies and tools can be used to mitigate the effect of the tracking cookies and protect the end-user?

This thesis project accepts an empirical research as method with accent on quantitative analysis. The first set of questions will be addressed with the help of a custom-built software system. This software system will simulate user behaviour and populate the database with collected cookie data. Next, it will run a basic analysis of this data set. Later, based on this analysis, a model will be built based on selected parameters. For this purpose, another software-based tool will be implemented to further support the results' analysis. The last set of questions will be answered with the help of the current research work, and the existing strategies and tools will be compared and evaluated.

1.3.3. Scope of the research

This research includes several components. The first part involves developing a fully functional web system that aims to collect the necessary information about the cookies in the database. The system should have a user-friendly interface to make it as easy as possible for users. The collected information is the same that the website collects and keeps in the browser when a real person opens a website in the browser. The first part is also responsible for raw data labeling, classification, and pre-processing. The next stage of the work is about data classification with the help of machine learning algorithms, including imbalanced learning that addresses specific characteristics of the collected data set. The input data set for the machine learning scheme is relatively small, and choosing a traditional machine learning classifier rather than a neural network or deep learning technique, is a better strategy. This part will compare a few machine learning classifiers, and the best-performing one will be selected. If needed, sampling of the data set will be applied too. Lastly, this research proposes and summarises different mitigation techniques for tracking online users. All parts combine different data analysis methods to answer the research questions and reach the goal of this research.

1.4. Outline thesis

The deliverable of the current research project includes six chapters. The present chapter represents an extensive version of the research proposal. The remainder of the research project is structured in the following way. Proceeding to Chapter 2, where the preliminary knowledge and the related work are discussed. This chapter builds the theoretical and research fundamentals for this thesis. Chapter 3 continues with a theoretical and practical overview of the data and introduces the system's architecture. Chapter 4 presents the results and the observations. Chapter 5 answers the last research question in a separate section. Finally, Chapter 6 completes this thesis with conclusions and a discussion of the results and future work.

2

Background

This chapter will reveal the concepts and the theoretical background related, but not exclusively, to cookies and the fundamentals of network communication. The HTTP protocols and underlying technical basis are illustrated, including associated technical details and explanations. Lastly, cookies and their function are explained, as well as what types of cookies are recognised by the theory. This section continues with background information about Cookie Syncing. This chapter is concluded by overview of the relevant academic work that studied the topic of tracking, cookies and privacy.

2.1. Fundamental components

World Wide Web (henceforth: the web), or the web for short, is one of the most popular services offered on the Internet. The significant advantage of the web is that everyone has access to information and can publish information practically for free. In the following sections, we will look at the theoretical underpinning needed to properly understand the concept of web cookies and the related technical background. First, the framework behind network communication is presented, and then a short description of the HTTP protocol follows.

2.1.1. OSI model and TCP/IP model

TCP (Transmission Control Protocol)/IP (Internet Protocol) [15] and OSI (Open System Interconnection) are the two most widely used models for network communication. There are some similarities and dissimilarities between them. One of the main differences is that OSI is a conceptual model that is practically not used for communication. In contrast, TCP/IP is used for establishing a connection and providing communication route over the network.

The OSI network model [27] is an abstract model that describes how computer networks communicate. The OSI model allows different systems to exchange messages and data seamlessly with each other. It is a standard that network equipment manufacturers use when designing hardware, operating systems, and protocols. The model is used only when data is used for transmission over a network

and not when accessing data locally on one's computer system.

It consists of seven layers, each of which is a step in the communication process. Each layer has well-defined functions - it provides an interface and services to its top layer and receives services from the layer below it. Before data is sent over the network, it passes through the individual layers; each layer adds its information to the original information. Information over the network is transmitted in package form. On reaching the receiving computer, the packets pass through the individual layers in ascending order, each layer removing the additional information added by the layer of the same name when it was sent. Thus, the data package must be unpacked after passing through all the layers to produce the original message.

TCP/IP [15] was developed by a United States Department of Defense design agency. Unlike the OSI Model, it consists of four layers, each with its own protocols, and Internet protocols. These protocols define a set of rules designed for communication over the network. The four layers are as following:

- **Network Interface Layer:** This layer acts as an interface between hosts and transmission links and is used to transmit datagrams. It also specifies what operation needs to be performed over connections such as serial and classic Ethernet to fulfill the requirements of the wireless Internet layer.
- **Internet Layer:** The purpose of this layer is to transmit an independent package across any network that travels to the destination, or maybe on another network. It includes IP, Internet Control Message Protocol (ICMP), and Address Resolution Protocol (ARP) as the standard packet format for the layer.
- **Transport Layer:** Enables the seamless end-to-end data delivery between source and destination hosts in the form of datagrams. The protocols defined by this layer are TCP and User Datagram Protocol (UDP).
- **Application Layer:** This layer allows users to access global or private Internet services. The different protocols described in this layer are virtual terminal (TELNET), e-mail (SMTP), and file transfer (FTP). Some additional protocols such as Domain Name System (DNS), HTTP, and Real-time Transport Protocol (RTP). The operation of this layer is a combination of the OSI model's application, presentation, and session layers.

TCP/IP is reliable, flexible, and tangible and offers how to send data over the network. The TCP/IP Model transport layer checks whether the data has arrived in order, whether there is an error, whether lost packages are sent or not, whether an acknowledgement is received or not, etc. In contrast, the OSI model is a conceptual framework to interpret how applications communicate over a network.

Since cookies are part of the HTTP protocol [3], the next section will pay close attention to this protocol, its characteristics and theoretical justification. The underlying concepts will be discussed to gain insight into this protocol without going into excessive detail.

An overview of HTTP and HTTPS

World Wide Web Consortium and the Internet Engineering Task Force (IETF) published a series of Request for Comment (RFC) documents, of which RFC 2616 (from June 1999) [14] is accepted as a standard and describes the initial version of HTTP/1.1. HTTP stands for HyperText Transfer Protocol.

This protocol consists of a list of rules by which computers exchange data on the Internet. It is at the top of the OSI model (at the application layer), where applications communicate with each other. Typically, HTTP uses TCP/IP protocols and uses them to transfer virtually any data, collectively called resources. HTTP is a simple text protocol that transfers data in text form. The HTTP protocol has concepts such as client (usually web browsers) and server (these are web servers). The standard HTTP port is 80, but any other free TCP port can be used. HTTP consists of a request, or a message from the client to the server, and a response, or server's response to the message from the client.

Web servers are applications that "listen" on a specific port and respond to requests received from client applications. Most often, the clients of web servers are browsers. After the client sends requests, it waits for data from the server. When the server receives a request, it processes it and returns the response to the client. All modern web servers have the ability to provide dynamically generated HyperText Markup Language (HTML) to their clients. This technology is called CGI - Common Gateway Interface. CGI is that based on the HTTP request and can be written in virtually any programming language or script.

```
1 <method> <URI> HTTP/1.1
2 <headers>
3 <empty line>
```

Listing 2.1: Syntax of HTTP 1.1 request

The HTTP Request consists of three main elements: method, URI (Unique Resource Identifier), and Header placeholder - Listing 2.1. Cookie headers consist of key/value pairs. Each request ends with an empty line. The method describes the type of request sent by the client. The most used methods are GET and POST out of 8 in total (GET, POST, HEAD, PUT, DELETE, OPTIONS, TRACE, CONNECT), whereas when using the GET method, the client requests some resources from the Web server, and with the POST method, data is transmitted to the server. Always names of the methods in HTTP requests are written in upper case. URI identifies the resource over which the query (request) method will be executed. URI is the part of the URL (Uniform Resource Locator) that follows the host (server) name. URI is followed by the HTTP protocol version that will be used for this HTTP session. The header fields of the request contain additional information pertaining to the request and that define the resource requirements expected to be returned by the server.

```
1 HTTP/1.1 <status>
2 <headers>
3 <empty line>
4 <resource>
```

Listing 2.2: Syntax of HTTP 1.1 response

One of the essential parts of the headers of every HTTP request are the UserAgents. The UserAgent is a text string that serves as an identifier of the browser, device and operating system. The information of the user agent is passed inside the HTTP header when a browser makes a request to a web server. It has a specific syntax, as shown in the syntax in Listing 2.3. Every browser, device and operating system has its own UserAgent characteristic string, including the version and additional comments.

```
1 User-Agent: <product> / <product-version> <comment>
```

Listing 2.3: Syntax of user agent

After receiving a request (valid or not) from the client, the web server returns a response. If the request is valid, the web server returns the requested resource to the client if it exists. The format of the HTTP server response can be seen at Listing 2.2. The first line of the server response contains the protocol version (HTTP/1.1), the request status code, and a phrase as a short text explanation of the code. If a problem with the requested resource exists, the response code on the status line will indicate precisely what the problem is. The headers in the response have the same format as the request, containing additional information about the resource. The first part of the response ends with an empty line, followed by the requested resource if it exists and it is available.

The current version of HTTP that most websites use, supported by most major browsers (Chrome, Opera, Firefox, Internet Explorer 11, Safari), is HTTP/2. The HTTP/2 specification was officially published as a standard in RFC 7540 (from May 2015). In February 2020, was made available RFC 8740 regarding HTTP/2 over TLS 1.3 (Transport Layer Security) protocol. Unlike previous versions, the HTTP/2 protocol is binary, and data is framed and transported differently between the client and the server. More, HTTP/2 has more efficient mechanisms for data streaming, support of request multiplexing, active server-side push notifications and prioritization, parallel loading of multiple elements, and header truncation. Nevertheless, HTTP/2 preserved all structure elements of HTTP 1.1, like methods, header fields, status codes, and URIs.

One of the most significant drawbacks of the HTTP protocol is its non-session, or stateless, nature. It means that the server is not required to track state over subsequent requests, which leads to unambiguous client identification on the web application's server side. Each request is executed independently, not knowing the previous requests. As a solution to the problem, HTTP version 1.1 introduces Keep-Alive connections, where multiple HTTP requests are executed over a single connection between the client and the server. One of the most popular ways to implement a session within HTTP is with cookies. The cookies are created within the header field of the response message with the Set-Cookie command.

A few years after HTTP was created, HTTPS was introduced as an extension of the HTTP protocol. It stands for HyperText Transfer Protocol Secure. HTTPS implements the packaging of transmitted data in cryptographic protocols such as SSL (Secure Sockets Layer) or TLS. HTTPS is not a separate protocol but an HTTP protocol. HTTPS allows encryption of HTTP protocol over an encrypted SSL/TLS connection. Here, including the URL of the requested page, request parameters, headers, and cookies (which usually contain Personally identifiable information). The security of HTTPS relies on TLS, which encrypts the data flow between client and server using long-term public and secret keys to exchange a short-term session key. HTTPS creates a secure channel over an insecure network and protects information exchange from interception and eavesdropping (or man-in-the-middle attacks). Certificate Authorities (CA) and public key certificates verify the certificate and its owner, its contents. Further, CAs generate, sign, and administer the validity of the certificates. HTTPS uses X.509 certificates to authenticate the server. HTTPS connections usually use TCP port 443, and their URLs start with https://. Currently, HTTPS is supported by all major web browsers.

2.1.2. Cookies in detail

HTTP Cookies [5, 25], or just cookies, are small text files generated by websites while browsing. They are saved locally on the user's hard drive (local machine) for a certain amount of time, together with

other browser information, in a dedicated folder structure [26]. This folder structure differs per browser, device, and operating system. The purpose of the cookies is to store data about individuals that will improve their browsing experience and further used by advertisers to track users' activity online and offer different content based on that. Subsequently, this data can be used by other websites, shared between parties, and can contain data that can identify an individual.

The amount of time each cookie is kept locally is determined by the expiry date of each cookie by the time of its creation. The expiry period is set by the websites and cannot be changed. Cookies are created when users browse online. Few types of cookies can be recognised, but the most popular distinction is between first-party and third-party cookies. We will focus on these categories further in this subsection.

Depending on various parameters, like purpose, duration, and provenance, the type of cookies can be organised into different groups [24]. Concerning the duration, there are recognised session cookies and persistent cookies. Session cookies remain alive as long as the browser window stays open. After closing the browser, session cookies are permanently deleted. Persistent cookies remain on the user's device until they reach their expiry date or even longer. The expiry date is set during their creation. They are kept from being deleted once the session has closed. They remain locally until deleted by the user or by the browser itself.

From origin point of view, we distinguish two types of cookies: first-party and third-party cookies. First-party cookies are created by the website user visits, whereas third-party cookies are set by third-party websites and used to track users across the web.

First-Party Cookies are stored directly on users' local machines by the website they visit. The website collects analytical data and valuable information from users' behaviour to improve user experience. When a user visits a website, this website will send a request to the user's computer that creates a unique cookie in the specific domain. If necessary cookies are not created or deleted later, websites will not automatically log in and remember the user's preferences from past sessions.

Third-party cookies are created by domains other than the one users visit straightforward. Third-party cookies, commonly used for tracking, usually persist on the user's local machine after closing the browser window.

According to their purpose, the following types of cookies are being recognised, as introduced in [24]:

- Strictly necessary cookies - These are compulsory because they enable you to browse the website and use basic functions - such as the shopping cart. These cookies are required for the basic functionality of the website. Usually, these cookies are first-party session cookies and are obligatory; their purpose should be explained to the user.
- Preferences cookies, also known as functionality cookies - allow the website to remember your choices (e.g., username, language, or region) and thus provide more personalised features
- Statistics cookies, also known as performance cookies - cookies for collecting analytical data. These are statistics on how a user interacts with a particular Google service. These cookies collect information about users' experiences during browsing the website in an anonymous and

aggregated form. They are used to analyse and improve the website's performance.

- Marketing cookies - usually third-party persistent cookies that adapt the online content (mainly advertisements and offers) relevant to the user's online activity. The advertising cookies allow some websites to display relevant advertisements on certain third-party websites based on the data they receive from the third-party cookies.

A special category of cookies not present in the above categorisation are the evercookies [23]. Evercookies, also known as supercookies or zombie cookies, are a special category of cookies [1]. They actively reproduce previously deleted cookie data by using ambiguous techniques to retrieve user data from the browser's storage, different from the data stored in the cookies. Evercookies identify the deleted cookies with the help of a JavaScript API (Application Programming Interface). This makes users' intentional attempts to delete the data (stored in the cookies) and start with a fresh profile ineffective [1]. In fact, in California, USA, the use of evercookies is considered an illegal invasion of privacy.

Cookie syncing

Cookie syncing [12] is a process that surrounds the Same-Origin Policy and allows trackers to exchange user data despite the restrictions of the browsers [1]. SameOrigin policy prohibits third-parties to read the cookies set by first-party websites or other third-party websites [4]. The method of cookie synchronising overcomes the impediment when two third-party tracker websites want to share the information stored in each other's cookies. Not only is it hard to detect, but cookie syncing triggers severe concerns about data leakage and users' privacy. The reported tracking mechanisms appear to be advanced and successful mitigating techniques ineffective and insufficient while limiting legitimate functionality [1, 4].

2.2. Related work

The following section focuses on the recent literature that examines the academic work relevant to the selected problem regarding tracking cookies. The goal is to build theoretical and historical foundations of cookies and related online tracking and to reveal the possible issues and concerns associated with this topic. Online companies are believed to have adopted many strategies to track users and their behavior. Amongst the reasons for monitoring these users, is to gain economic interest and an advantage against their business opponents. This does not represent an extensive list, but it includes overview of the essential papers in the field.

HTTP cookies were created back in 1994 by Lou Montulli from Netscape Communications as a technique that preserves the state of the websites between two consecutive visits of a user. Since the purpose and mechanism behind the cookies have evolved, their function has become essential, but their effect on users' privacy still needs to be investigated. Various academic papers indicate of lack of transparency and violation of the privacy norms when using cookies [28].

In this section, the related work will be discussed in the field of tracking cookies, and related security and privacy issues will be identified that are relevant to build the basis of this thesis.

With respect to online tracking, many papers address privacy and related concerns arising from digital

advertising and web tracking in particular in attempt to measure, comprehend and circumvent such tracking [1, 13, 41, 43, 29, 5, 30]. While other authors devote their work to analyse tracking in the current setting of legal measures and regulations in a specific geographical region [9, 10, 21, 19, 22, 40].

Even though the focus is on the desktop configurations, Lerner et al. [28] of reveal how since 1996 the web tracking and third-party cookies have evolved and "increased in prevalence and complexity". The authors suggest further that it is essential to understand tracking fundamentally, in order to set ground for stable and effective policy dialogue.

Specific characteristics of HTTP cookies have been studied in various papers amongst academic community. One of the fundamental studies regarding web tracking, is Englehardt and Narayanan [12]. Cookies are proved to be one of the mechanisms for collecting user data. That involves tracking users' data and actions across the web and raises specific ethical concerns, including privacy violations. In their study, Englehardt and Narayanan [12] handle the most comprehensive survey regarding web tracking up to that moment (January 2016). They focus the research on one million popular websites according to Alexa's ratings. Their work is believed to validate the first large-scale evidence of web tracking. Authors [12] use a web crawler (bot) to automate user actions, scrape website information, and conduct analysis and search collected from all vectors. They developed a tool called OpenWPM, later made open-source and extensively used by other researchers, developers, and regulators in human dreads of research projects for data collection, privacy measurements, privacy-preserving implementations, and solutions. In the report, the authors show that OpenWPM can resolve significant problems in the current state of research in the topics like web privacy and security measurements.

A similar approach was used previously by the authors Mikians et al. [37]: by simulating the user's behavior, followed by data collection of the observations, and eventually evaluation and analysis of the findings. The developed tool is designed to complete the first two tasks - behavior simulation and data collection and facilitate the researchers in executing the research. One of their findings is that trackers are primarily found on news sites. And as one of their main concerns, they point out the web tracking and subsequent effect on user privacy. Further, the authors address that most browsers' privacy features are not as effective as they should be. The authors indicate that cookie syncing, performed by most third-party websites, raises serious privacy concerns, as syncing is practiced behind the scenes and beyond the knowledge of the average user.

In "The cookie recipe: Untangling the use of cookies in the wild" [17], authors execute a research regarding the parsing and unpacking content of HTTP cookies with high accuracy. As a result authors Gonzalez et al. [17] succeed to parse 86% of the content with accuracy of 92%. Their goal is to understand better what information cookies contain. Knowing this, it will help to address the gap between the users' desire of protection of their privacy on one hand, and advertisers and business that aim to offer better services and user experience on the other.

It is reported in various papers that the development of anti-tracking tools is growing while more aggressive measures to track users are being developed as well. Examples of such techniques appear to be the ever cookies and cookie syncing. The available tools to block cookies, such as browser plugins or network proxies, are reported to be not so effective as expected, as they prevent between 6% and

21% the execution of the website's functional scripts and they fail to block between 37% and 78% of the tracking scripts. Other sources reveal that due to ad-blockers used by the users, the estimated revenue loss of companies is nearly 21 billion US dollars and that resulted in different websites demanding users to deactivate the ad-blocker in order to access their content [17].

Gonzalez et al. [17] address that in order to reach the balance between the user and the business and to develop more effective tools, a comprehensive understanding of cookies content is needed. In the study 5.6 billion HTTP connections are analysed. It analyses several thousand real user connections. Authors indicate that all concerns, regarding the privacy of the user has been addressed by data anonymisation and complying with European regulations.

Gonzalez et al. [17] reveal further how cookies are used through statistical analysis. Their findings reveal that 60% of the content includes complex information and that the name-value pair often contains multiple values that change in time. Authors conclude that existing tools that aim at protection of user privacy require improvement and addressing the complexity of cookies content.

Similar problems identify the authors of "Identifying sensitive urls at web-scale" [29]. Even though the legal framework of European Union regarding GDPR gives definition regarding the protection of sensitive data in categories of data like: ethnic origin, sex orientation, political, religious or philosophical beliefs, including personal and biometric data that that undoubtedly to identify natural person, it is not clear how content can be identified as sensitive.

The measures set by the GRPD and its derivative laws in the member states are rather reactive measures - the measures come into force after an incident occurs. In the event that measures must be aimed at protecting sensitive personal information, they must be proactive and targeted at the tools that expose the information coming from users on the internet.

This is probably why the authors Matic et al. [29] use a classification project to train a classifier that recognizes sensitive web addresses. The accuracy they achieved in classifying web addresses is over 88%, and that in identifying sensitive categories even reaches 90%. Subsequently, this classifier performs a search across 1 billion web addresses and domains. One of their findings is that the classification of web addresses as containing sensitive information or not based on category or content is not accurate. Consequently, blocking certain addresses from NSA (National Security Agency) servers is not effective.

Another seminal and groundbreaking work was published in June 2022. The authors Gotze et al. [18] examined the extent to which cookies are a privacy threatening factor when visiting certain websites. In their work, they pay attention to the cookies that are created when a user visits government websites. Cookies on such sites are a sign of extra care as expectations and trust in such sites are particularly high. The availability of electronic citizen services is attractive and convenient for a large number of users. Some of the services are even offered solely online and should then be used with little or no risk to users. The starting point for collecting information in the form of cookies from users has been shown to be prior to the existence of consent and receipt of informed consent from users. The authors show the results of the cookies and relevant details around their lifetime, who created them and from which websites. The target group is impressive and represented by over 5,500 websites and containing over 118 thousand web urls. They use Pythia - framework that is used to visit predefined web addresses,

as well as using Chrome protocols to access the collected cookies.

To summarize - in their paper, the authors [18] present the results of a large-scale study on the extent of the number of cookies on government websites and other public websites that offer administrative and information services to the public. Their results show that tracking users using cookies is a serious problem. In this work, they provide evidence of quite worrying problems that would arise from the presence of such a number of trackers, the presence of which has been found in over 90% of websites. These trackers aim to collect not only anonymous but also personal information from users. This is why serious attention must be paid to the protection and prevention of these practices in connection with the implementation of the established legal frameworks. In this case, we are also talking about practices whose existence is contrary to the established legal framework.

3

Methodology

Data is an essential commodity for the current work. The current section describes all requirements regarding the data structure and the data used in the data collection methodology. Data in this project provides an environment for analysis and a basis for answering the research questions. The system uses an algorithm to gather all the required data and to compose a dataset. However, data about websites, locations, and user agents must be present in the database before collecting the main dataset. The dataset will later be used as a source for the model needed for the machine learning algorithm. Therefore, the algorithm must provide large volumes of consistent raw data, as incorrect results in the analysis part must be avoided.

3.1. System architecture

This section describes the system's architecture and how different elements are combined. The main goal of the current work is to create a design as simple as possible and to simplify the process and data collection as optimally as possible, consequently decreasing the handwork as much as possible.

3.1.1. System overview

As it can be seen from Figure 3.1. System architecture, the system consists of three main blocks that form the architecture: data collection system (web application), data storage component (database), and data analysing module (software application)¹. Together with the supplementary blocks, they form complete framework for collecting cookie data. Detailed description of each main block, along with functional characteristic of the features, is provided further in the chapter. The arrows represent the flow of information between the blocks.

Database block

One of the main blocks is the database. The main function of this block is to store and retrieve data

¹All source code is provided in TU Delft online repository <https://gitlab.tudelft.nl/gsmaragdakis/aleksandra-taneva-msc-thesis>

requested by the web application. The database is organised into tables that include Locations, Websites, UserAgents, Trackers, and Cookies. The names of the tables describe the content. A detailed overview of the database, an overview of the data sets, and a database scheme can be found in Chapter 3.7. Dataset and database structure. Data manipulation operations (like edit, delete, add) in the database can be preformed manually by the system administrator directly in the database, or via the web application interface by the users.

Microsoft SQL (Structured Query Language) Server [36] is used as a relational database management system (RDBMS). Developed by Microsoft, and it is one of the three market-leading database solutions. It has become one of the best alternatives for building a database system because of the high speed and flexibility with which it operates. Microsoft SQL Server is a client-server platform. The only way to communicate with the database is to send a request, which will contain commands for the database. The protocol used to communicate between the application and the database is called TDS (Tabular Data Stream) and is described in the technical documentation on the MSDN (Microsoft Developer Network) portal [33]. The main language used to query the database is T-SQL (Transact-SQL), which is a variant of SQL. It consist of instructions (keywords) that the server interprets, executes the requested operation and returns result to the client application.

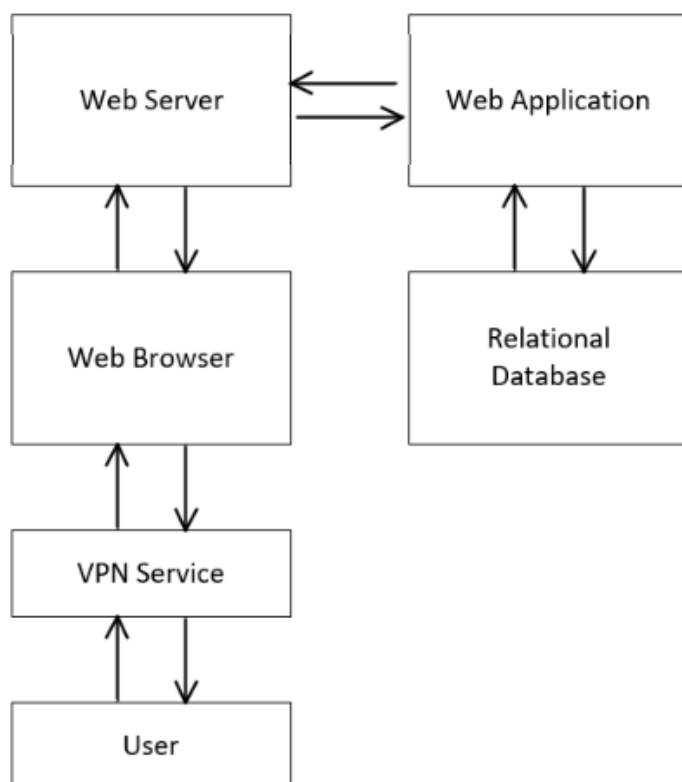


Figure 3.1: System architecture

Web application block

The web application is developed using C# and ASP.NET Core [32]. Microsoft .NET Framework [31] is a platform for developing, deploying, and running web services, desktop applications and websites. It provides a highly productive, standardised, multilingual environment to integrate existing applications with next-generation applications and web services and to solve the challenges of deploying and running web applications. The .NET Framework is part of the .NET platform that combines libraries, tools and programming languages for building and running applications. Applications created with .NET can be written in many different programming languages. It consists of two components: the common language runtime (CLR) and the .NET Framework Class Library. CLR represents the execution engine, responsible for running the applications. The Class Library supports a library (a collection of APIs) of common functionality that developers can use straight away.

This block contains the fundamental functionality of the web application. It is accessed through a user interface that is simplified and straightforward to use. The system allows user to collect cookie data based on different parameters - per location, operational system, device and website. All these parameters are editable in the current system by the users. The collection of data is executed after a user requests it. This block is the connecting module between the user and the database. This stage allows simple analysis of the collected data.

Virtual Private Network (VPN) block

A Virtual Private Network (VPN) provider is used. It hides the current IP (Internet Protocol) address by replacing it with another IP address from the list of the available VPN servers. Even though the primary purpose of VPN services is protection and additional encryption of sensitive user data, here it is more simply used to resemble the actual location of different personas and to provide more reliable results. VPN service runs in the background, and it does not disturb data collection and analysis. It was made sure that in the settings of the VPN application, the option to block ads, third-party cookies, etc., is turned off. As additional benefit to virtually changing the location, the VPN service provides protection and Internet Service Providers (ISP) are not able to track many characteristics of the users.

Web server block

IIS (Internet Information Services) [35] is a web server developed by Microsoft for its operating systems. The product is fully proprietary and comes bundled with Windows Operating System. It works the following way: a user's request goes to the server and is sent to IIS for processing. The main module of the web server is the WWW service. The service processes user requests via HTTP/HTTPS protocols. One web server works in parallel with several sites and when an application crashes, the whole server does not crash. For example, one server with one IP address handles requests on one TCP port from multiple sites. Within IIS, DNS records are created to identify each site. Inside the web server, home directories are created for each site with a differentiation of access rights for directories, or Active Directory service.

Browser block

The user interacts with the system via a Web browser. Through the browser, the user sends a request

to the server to access a specific page. This is done by specifying the communication protocol, the server name, the path to the desired page, and the page's name in the following form protocol://host-name/path/site-name. The server, in turn, returns the selected page to the user, and the page is rendered through the browser. If there is C#.NET code on the page, it is executed on the server, and the final result is returned to the client. If necessary, when the C#.NET code is executed, it is possible to make several requests to the database to retrieve the data, store or update certain information. Accessing the database is achieved through the MSSQL database communication interface built into C#.NET by specifying a server name, database name, username, and password.

3.2. Users overview

The web application has two types of users. They are defined in advance and, in a significant sense, determined by the specifics of the technologies used. One of the roles is the administrator with the corresponding right and access scope. The administrator has the right to add, delete, approve and edit all fields in the database. He has full administration rights to the system, such as possibility to modify user content, add new pages, etc.. He also has access as a user for testing purposes. The system administrator has full access and can resolve any problems if necessary. It is essential to note that the term system administrator is related to any user that has access to the database and the system's source code; the administrator can deploy the system on the server as well. With this, special privileges and credentials are needed. The system administrator is responsible for the backup of the database.

The second type is the ordinary user. The user can access the web application and start the process of data collection, also he can use the analysis part.

3.3. Data collection overview

A web application developed for the purposes of this project is responsible for the generating of the dataset. This web application runs locally and uses a local database that is populated by the web application when the browser engine visits one by one all individual data nodes. This task is executed by the Selenium tool [42] integrated in the application. Selenium tool helps to simulate user behavior and automates the data collection process. This tool imitates automatic opening a browser window and visiting a certain web address. Private browsing mode is not permitted. Before browsing a particular website, all possible old cookies are deleted to guarantee the correctness and freshness of the results. The platform Selenium uses for testing purposes is a WebDriver and in particular a Chrome Driver [8]. Each website is visited, and the corresponding cookies are scanned, decrypted, and sent to the database. The browser engine does not interact with the website in any form - no actions like clicking on cookie consent or anywhere else is executed. It only visits once the home page of each website stays there for a couple of seconds and then closes the browser window. With every start of the automatic data collection, the data related to cookies is saved in a dedicated table in the database. Every database record about a cookie contains a unique data node, that is a combination of website, location, operating system and device type.

For ensuring accurate input and conservation of limited resources of the local machine, the amount of time each website is visited is 10 seconds. This is the optimal amount of time-based on a few experi-

ments with various periods (1 second, 5 seconds, 15 seconds, and 20 seconds). The time length of ten seconds ensures that all available cookies are gathered and no excessive time is spent on each page.

Each time a data node (a unique website from a unique user agent at a unique location) is visited, the browser usually collects first-party and third-party cookies. When a single cookie is discovered; certain additional information is being collected per cookie like cookie name, the cookie value, date and time of the visit, domain where the cookie originated, the expiration date of the cookie, the creation date of the cookies, value if the cookies is HTTP only, a value indicating how the web browser handles third-party cookies, value showing if the cookie is secure or not.

The real life data related to cookies (first-party or third-party) is collected by crawling the Alexa Top 50 websites. Each instance (window) of the Chrome browser collects a separate cluster of cookie data. For the purposes of the analysis, a fresh instance of Chrome browser is launched for every user agent, location, and website. Each browser instance is run only once for every combination of vector data (user agent, location, website). To avoid some websites issuing a ban and blocking the IP address, browser windows are not run in parallel.

As an advantage of the application, it can be pointed out that the extraction of the results is into .csv files to make future analysis more compatible with data engineers. The application has a web interface, and all functionality available to the user is accessible via the browser. The results are analysed by external module that is custom developed to serve the purposes of the current project. From collecting the tracking data and editing the parameters to analysing the data and exporting the results: none of these components require any manual operation from the user. An important requirement of the system is a stable internet connection. All data processing elements (download of raw data and data visualization) of the system are separated in compliance with the design principle of single responsibility (every module has a single purpose). Data processing and data visualization are discussed in detail in Chapter 3.7 Dataset and database structure.

After the user has initiated a request to collect data, the system automatically starts downloading the browser data. This process begins with opening a new browser window and simulating the following user's actions: visiting a website, ignoring (not taking any action) the cookies consent message, and then closing the browser window. This repeats for every website from the list. A combination of different locations and different user agents visits every website. The overall number of visits is a permutation of locations and user agents. Before starting the system, a location must be selected from the VPN service. That way, the reliability of the results is guaranteed. The Google driver can mock the browser's location based on longitude, latitude, and accuracy, but this is unreliable. Some websites can use the IP address as a location source rather than the web browser settings. The IP address gives approximately the location and other data. At the same time, the device information (User Agent parameter) is read from the browser and can be mocked successfully. All the tables from the database are represented by models that directly map all columns to properties at the backend side of the system. With the help of database context, we can directly access and manipulate the database without additional layers.

Cookie data is used in Chapter 4 Results and analysis to discuss and analyse the results is collected over two periods. The first data collection was completed on the 16th of June 2022, whereas the second was between the 19th and 21st of September 2022.

3.3.1. Data decryption

After the (automated) browser collects the relevant cookies data, it gets saved in SQLite file placed in a dedicated folder on the local machine. This is a standard functionality of the browser engine and occurs every time a user opens a browser and visits a website. The browser cookie data is placed in the file named "Cookie". This file contains SQLite3 database records. A deviation from this standard functionality is that the folder's default location is changed for the convenience and purposes of the current project.

After the "Cookie" file is populated with data collected by the browser, the web application immediately reads the data. Then, the data is decrypted with the help of the user's decryption key. This key is stored locally in the standard browser folder structure, and every user has access to it. The system reads this key, decrypts the data, and sends the decrypted data to the local MSSQL database. Different libraries of the programming language support all steps of this process. A complete list of the used libraries is to be found in Appendix A, where reference to the source code is also included. In the next paragraph follow details regarding the technical details of the encryption of this data.

The records in the SQLite3 database are encrypted and need to be decrypted before being transferred to the relevant table in the database. Since version v80.0 of Chrome browser, Google Chrome uses an AES-256-GCM algorithm to encrypt cookies data and the current user's password on the local machine as a seed. It uses the AES-256-GCM algorithm with the Master key and 12-byte random IV data. Before the cookies are stored in the file, a "v10" signature is inserted. Google Chrome uses Master key-based encryption to store passwords. Subsequently, it generates 32-byte random data; then, it is encrypted using Windows Data Protection API ("CryptProtectData") function. Windows Data Protection API (DPAPI) must be used to decrypt the stored information from the Chrome browser. DPAPI allows developers to encrypt keys using a symmetric key derived from the user's logon secrets or using the system's domain authentication secrets in the case of system encryption. This makes it very easy for developers to save encrypted data on the computer without worrying about how to protect the encryption key. In the next step, Google Chrome inserts the signature "DPAPI" at the beginning of the key for identification. Eventually, this key is encoded using Base64 and stored in the "Local State" file in the corresponding folder structure with browser files.

3.3.2. Preparation and basic analysis of the results

The web application has functionality that provides possibility to execute basic analysis of the results and make initial visualisation of detected differences. Later, these differences could be used in the extended analysis by the machine learning tool. The results can be compared per location, per browser and per user agent individually. Whereas, the delta results show the difference per two of the vectors. The results as quantitative and qualitative measurement are discussed in the following Chapter 4 Results and analysis.

3.3.3. Visualisation of the differences

Two different visualisation approaches are chosen to support the basic analysis of the results. The

results appear in a list or in a table. In a list are presented the whole dataset that is filtered between two cookie identifiers - as combination of location, user agent or browser. The initial analysis is implemented with the help of colour scheme. Colours represent the proportion of the number of the cookies that deviate compared to the minimum count of the cookies for the selected location, browser, operating system.

The following colours are chosen to distinguish differences in the results:

- Green - by green colour of the number of the cookies, little or no difference is detected. The difference of the number of the cookies is less than 10%.
- Yellow - represents that this specific number of cookies is more than 10% and not more than 25% than the minimum number for the group
- Orange - represents that this specific number of cookies is more than 25% and not more than 50% than the minimum number for the group
- Red - represents that this specific number of cookies is more than 50% than the minimum count for the group

When the results are displayed in a table form, only the websites that has different distribution among cookies per data vector, are displayed. As basis is always taken the results with minimum count of the cookies, then the percentage difference is calculated as the difference between two values divided by the average of the two values, measured in percentage. The following Formula 3.1 calculates the percentage difference within collected cookies per user agent per website:

$$PercentageDifference = \left| \frac{MinValue - CurrentValue}{(MinValue + CurrentValue)/2} \right| \cdot 100\% \quad (3.1)$$

Where the MinValue is the minimum value of the number of the cookies collected for given website. The CurrentValue is the value for which the difference will be calculated. If the minimum value is zero, then as minimum value is taken the next non-zero one.

3.3.4. Identifying third-party cookies

Once the data is collected in a file on the local machine, the information is transferred from there to the database by the web application automatically and without any additional action from the user. During the cookie transfer, a categorisation of every cookie record is also performed - whether they are third-party or first-party cookies. The domain address of the creator of the cookie determines the difference between the two types. Each cookie contains information by which domain it was created, among other details. Suppose the cookie was created by a website with a domain similar to the current one (for example: ".youtube.com" or "www.youtube.com"), then the web application recognises that cookie as a first-party cookie. Otherwise - the cookie's domain (for example: "doubleclick.net") and that of the visited website do not match - then the cookie is identified as a third-party cookie. Cookies created by another website can usually be assumed to be tracking cookies. More about the types of cookies and theoretical details can be found in Chapter 2 Background.

3.4. Websites

The data collection of website cookies data is essential component for the analysis of data of the cur-

rent work. The web application visits one by one pre-selected list of websites. The selection is based on Alexa's Top 50 most popular websites (<https://www.alex.com>) according to the number of visits. Every website record in the database has a category and country of origin. All components and characteristics of the website are editable and can be adapted according to the requirements and purposes of the analysis.

Unfortunately, on May 1st, 2022, Alexa.com announced that they would retire the global web ranking service for all subscribers after more than 25 years. Fortunately, Technical University of Munchen distributes a list of Alexa's top 1 million websites (source: <https://toplists.net.in.tum.de/archive/alex.com/>) on a daily basis. As stated in their README.txt, the data they provided can be used only for scientific purposes: "The original copyright takes prevalence, but we operate this mirror under CC BY-NC 4.0 [1] for non-commercial purposes only". Therefore, for the purposes of the current work falls under this purpose and explicit approval is not needed. The precise extract that populates the website's table is from 16th June 2022.

Table 3.1 List of Alexa Top 50 websites, snapshot from 16th June 2022, including category and country of origin gives overview of the selection of the websites. As it can be seen from Table 3.1, website data include information about the website's web address (or Uniform Resource Locator (URL) address), category and country of origin, and domain. The categorisation of the websites is borrowed from <https://www.similarweb.com/top-websites/> (last accessed: 2 Oct 2022). Any website that did not appear in the freely distributed list was categorized based on content or domain similarity.

Except for URL address, Category and Country, the Websites table has another column - Domain Name. The column Domain Name is used in the analysis phase. It determines whether the cookies originate from a website with a domain name different from the current website. Not always does the created cookies' field domain match the current website domain name. For example: from the collected cookies, the domain field can be ".youtube.com" or "www.youtube.com" and not all of them match the website web address: "www.youtube.com". The dissimilarity in this field makes it hard to use techniques like regular expressions. If the URL address and the domain do not match, then the collected cookies record is considered a third-party cookie, otherwise is marked as first-party cookie.

Website	Category	Country	Domain Name
https://www.youtube.com	Arts and Entertainment	USA	youtube
https://www.google.com	Search Engine	USA	google
https://www.amazon.com	Marketplace	USA	amazon
https://www.facebook.com	Social Network	USA	facebook
https://www.twitter.com	Social Network	USA	twitter
https://www.wikipedia.org	Encyclopedia	USA	wikipedia
https://www.instagram.com	Social Network	USA	instagram
https://www.baidu.com	Search Engine	China	baidu
https://www.yahoo.com	News and Media	USA	yahoo
https://www.whatsapp.com	Social Network	USA	whatsapp
https://www.reddit.com	Social Network	USA	reddit
https://www.naver.com	News and Media	SouthKorea	naver

https://www.bing.com	Search Engine	USA	bing
https://www.qq.com	News and Media	China	qq
https://www.msn.com	News and Media	USA	msn
https://www.aliexpress.com	Marketplace	China	aliexpress
https://www.taobao.com	Marketplace	China	taobao
https://www.zoom.us	Social Network	USA	zoom
https://www.office.com/	Software	USA	office
https://www.yandex.ru	Search engine	Russia	yandex
https://www.mail.ru	Mail	Russia	mail
https://www.linkedin.com	Social Network	USA	linkedin
https://www.ebay.com	Marketplace	USA	ebay
https://www.bilibili.com	Arts and Entertainment	China	bilibili
https://www.live.com	Mail	USA	live
https://www.zhihu.com	Social Network	China	zhihu
https://www.vk.com	Social Network	Russia	vk
https://www.xvideos.com	Adult	USA	xvideos
https://www.pornhub.com	Adult	USA	pornhub
https://www.github.com	Cloud Service	USA	github
https://www.csdn.net	Social Network	China	csdn
https://www.tiktok.com	Social Network	China	tiktok
https://www.fandom.com	Arts and Entertainment	USA	fandom
https://www.yahoo.co.jp	News and Media	Japan	yahoo
https://www.canva.com	Arts and Entertainment	Australia	canva
https://www.netflix.com	Arts and Entertainment	USA	netflix
https://www.163.com	Mail	China	163
https://login.microsoftonline.com	Software	USA	microsoftonline
https://www.paypal.com	Payment Platform	USA	paypal
https://sina.com.cn	News and Media	China	sina
https://www.amazon.in	Marketplace	India	amazon
https://www.weibo.com	Social Network	China	weibo
https://www.jd.com	Marketplace	China	jd
https://www.t.co	Social Network	USA	twitter
https://www.stackoverflow.com	Social Network	USA	stackoverflow
https://www.xhamster.com	Adult	Cyprus	xhamster
https://www.apple.com	Marketplace	USA	apple
https://www.myshopify.com	Marketplace	USA	myshopify
https://www.google.com.hk	Search Engine	Hong Kong	google
https://www.adobe.com	Software	USA	adobe

Table 3.1: List of Alexa Top 50 websites, snapshot from 16th June 2022, including category and country of origin

3.5. User agents

The web application uses a simulation of different personas by visiting every website. Each persona has a different combination of device, operating system and browser. User agents help to simulate particular characteristic of different users. Using different combinations of platforms (desktop or mobile), web browsers (Google Chrome, Mozilla Firefox, Safari, Edge, Opera) or operating systems (Windows, Apple Operating System, Android, Linux) on each individual website visit, the system can help detect any tendency in the count of collected cookies based on these attributes.

Every user agent visits every website for every location, and then all the corresponding browser cookies data is collected and consequently stored in the database. Hence, this setup allows user sessions to be distinguished and each individual website visit to be examined later. The user agent string (User Agent Value) is assigned to each browser in the corresponding HTTPS header. The string values for the user agents are retrieved from the following website:

<https://www.whatismybrowser.com/guides/the-latest-user-agent/chrome>

During the different data collection series, the system recognises two configurations. This allows possibility for comprehensive data analysis and evaluation.

The first setup of settings adopts six different user agents, whereas the browser is unique parameter in the configuration. Table 3.2 shows all combinations of user agents for the browser Google Chrome 105 [7]. In such way, the system can detect any possible inconsistencies in collected number of cookies with regards to used device and respectively operating system.

User Agent Value	Platform	Operating System	Browser
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/105.0.0.0 Safari/537.36	Desktop	Windows 10	Chrome 105
Mozilla/5.0 (Macintosh; Intel Mac OS X 12_5_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/105.0.0.0 Safari/537.36	Desktop	macOS (Monterey)	Chrome 105
Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/105.0.0.0 Safari/537.36	Desktop	Linux	Chrome 105
Mozilla/5.0 (iPhone; CPU iPhone OS 15_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) CriOS/105.0.5195.100 Mobile/15E148 Safari/604.1	Apple iPhone	iOS 15.6	Chrome 105
Mozilla/5.0 (iPad; CPU OS 15_6 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) CriOS/105.0.5195.100 Mobile/15E148 Safari/604.1	Apple iPad	iOS 15.6	Chrome 105
Mozilla/5.0 (Linux; Android 10; SM-A205U) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/105.0.5195.79 Mobile Safari/537.36	Samsung Mobile	Android 10	Chrome 105

Table 3.2: List of user agents for Google Chrome 105 browser

The second settings setup combines user agents with constant parameter operating system (Windows 10) and device type (desktop), as shown in Table 3.3. The experiments with these user agents show any potential deviation regarding number of collected cookies for the different browsers on the same hardware specification. Web application provides interface to change the current list with user agents.

User Agent Value	Platform	Operating System	Browser
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/105.0.0.0 Safari/537.36	Desktop	Windows 10	Chrome 105
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/106.0.0.0 Safari/537.36 Edg/106.0.1370.34	Desktop	Windows 10	Edge 106
Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:105.0) Gecko/20100101 Firefox/105.0	Desktop	Windows 10	Firefox 105
Mozilla/5.0 (Macintosh; Intel Mac OS X 12_6) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/16.0 Safari/605.1.15	Desktop	Windows 10	Safari 16.0
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/106.0.0.0 Safari/537.36 OPR/91.0.4516.20	Desktop	Windows 10	Opera 91

Table 3.3: List of user agents for Windows 10, desktop configuration

3.6. Locations

Locations in the web application refer to the location of the user that uses the application. Every time a user initiates data collection through the application, the location of the user is recorded as identifier in the relevant field of the database and helps distinguish the different runs of the system. Since the application is only installed on the local machine, it was only possible to run it for one location. To overcome this flaw and gather more inquiry data for more objective results, a VPN service is used to simulate users from different locations. VPN virtually changes the location of the user from the list, specified by the VPN provider. As additional benefit to virtually changing the location, the VPN service provides protection. The benefits and qualities of couple VPN providers were compared and Surfshark was selected based on price and functionality that provides.

The locations were selected randomly, but equally distributed across the globe and under various economic and legal regions. The data for these six locations include city name and Global Positioning System (GPS) coordinates of the corresponding city. With the help of the user interface of the system, users can edit this list (Table 3.4 List of locations). Within the user interface of the web application is possible to change the current list with locations.

Location	Latitude	Longitude
The Hague	52,057499	4,49306
Delft	52,006672	4,35556
Dallas	32,783058	-96,806671
Hong Kong	22285521	114157692
Sydney	-33867851	151207321
Johannesburg	-26202271	28043631

Table 3.4: List of locations

3.7. Dataset and database structure

This section describes the components of the database, how different objects are connected, and outlines the tables and their characteristics. The main goal is to show the logical connections and how the data is prepared to suit this project's purposes. For the current research project that relies entirely on the information and its correct visualization, it is essential to have a structured database.

The current work generates the data set exclusively by the web application. With every start of the automatic data collection, the relevant data is saved in a dedicated table in the database. The data are collected in several tables, with a table corresponding to each data category. The information is collected per website, location, and user agent. Each table column describes a specific variable, and each row represents a single record of the data set. The tables of data set list values for each information vector, such as location, website, and user agent. A custom database was designed for the current work, and the tables are populated using the web application, described earlier in detail in the current Chapter 3. Methodology.

Data collection in the application relies entirely on the web application and its processes. The web application is the primary source of data for all the tables. Collection of the tracking data is the initial step of the process, and the user initiates it. The amount of information is paramount to the success of chosen machine learning algorithm. The combination of the complexity of the problem and the desired algorithm can determine the volume of data needed in a certain way. The right amount is hard to estimate, but it is always advisable to have as many relevant examples as possible for machine learning to produce accurate results as output. The collected data will provide the basis for analysis and support addressing the research questions.

As data collection covers all activities around data like: gathering data, pre-processing, assessing and analysing, it is spread in the next sections. Additional details about the used mechanisms and techniques regarding data gathering can be found in Chapter 3. Methodology.

3.7.1. Database models and description

Database models represent an abstract method to illustrate the organisation of the data in the database. The database represents a collection of data allocated in N-dimensional distinct tuples. These tuples are represented as table rows, and each row is unique. Every row has a unique identifier (or primary

key) to grant the definition of relationships.

The database is called CookieApp and contains the following types of records: Cookies, Websites, UserAgents, and Locations. The diagram in Figure 3.2.Database diagram provides a graphic representation of the relationships between the tables and the system's overall architecture in terms of the basic functionality of the database. The structure of the database is normalised. Normalisation ensures that redundant or repetitive data is avoided, larger tables are divided into smaller ones, and information is organised logically.

The connection between the tables follows the "primary key-foreign key" relationships concept. This concept is characterised by a one-to-many relationship between two tables in a relational database. A foreign key is a column (field) or a set of columns (fields) in one table that references the primary key columns (fields) in another table. The primary key is a column (or set of columns) where each value is unique and identifies a single table row. This column becomes a foreign key in the second table. The primary keys enforce the entity integrity of the tables and guarantee the uniqueness of the records (table rows).

In current database three primary key-foreign key relationships exist as following:

- The table Cookies and table UserAgents are connected as follows: the column UserAgentID from table Cookies is a foreign key for table Cookies, pointing at the same field at table UserAgents, where UserAgentID is a primary key.
- The table Cookies and table Websites are connected as follows: the column WebsiteID from table Cookies is a foreign key for table Cookies, pointing at the same field at table Websites, where WebsiteID is a primary key.
- The table Cookies and table Locations are connected as follows: the column LocationID from table Cookies is a foreign key for table Cookies, pointing at the same field at table Locations, where LocationID is a primary key.

Cookies table

The most important table is Cookies. It includes all data collected from the system. Figure 3.3 reveals the fields and the values the table cells can hold. The table is filled during the data collection phase of the system, and the regular user cannot edit the data. Otherwise, the results will be compromised. Each row stores information about one cookie record.

Most columns represent the same data (cookies attributes) that browser stores locally in the folder. The IsHttpOnly flag (optional) is part of the HTTP response header. It prevents the cookie from being accessed through the client-side script. The SameSite flag prevents the browser from sending cookies with cross-site requests. The Secure flag prevents the cookie from being accessed by unauthorized parties. Path and Domain attributes of the cookies play a significant role when an n HTTP request is made - the values of these fields must be compared and matched. Additional technical and theoretical details about cookies can be found in Chapter 2. Background

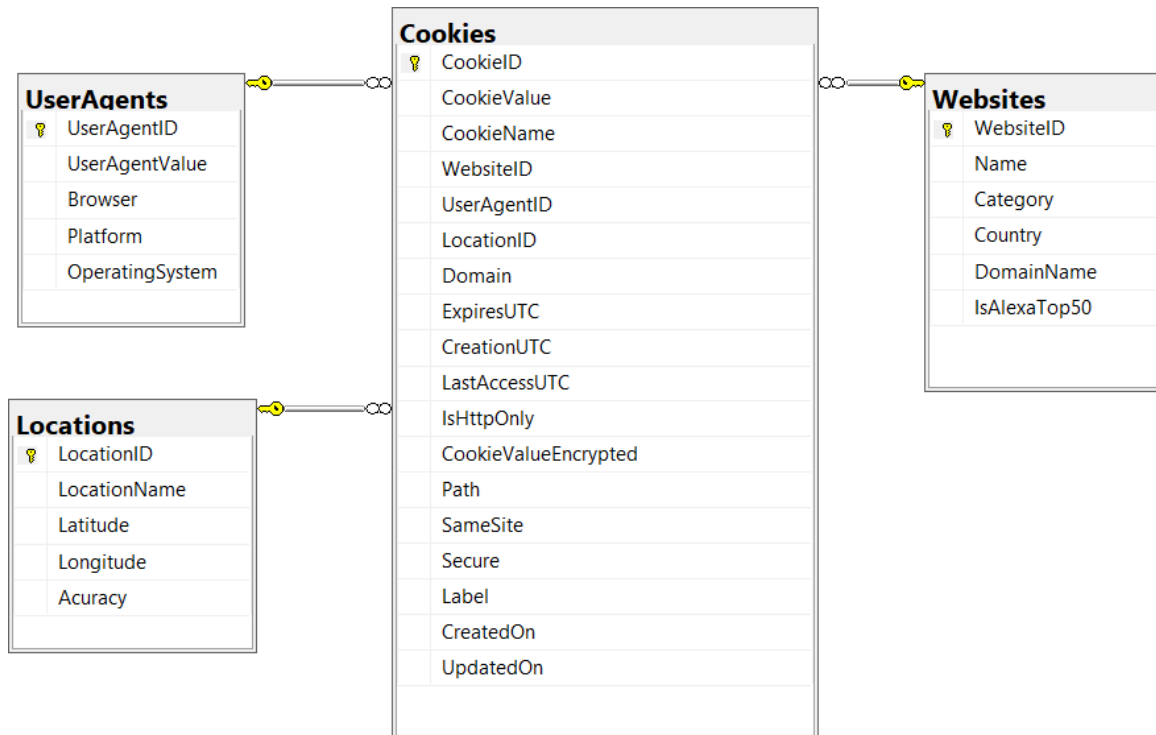


Figure 3.2: Database diagram

Two additional columns are added for administrative purposes: "CreatedOn" and "UpdatedOn". These fields are standard for certain design patterns and hold the date and time when the record is created. Another additional column - 'Label', is also custom added and does not get collected by the browser. This is a text field that is used as an alias to help distinct consequent data requests if needed. The value for this column is not obligatory.

Locations table

Table Locations is used to store data about the location of the user. The web application records the location of the user that uses the system. The web application requires at least one record to exist in the database. This initial record can be edited or deleted by the administrator in the database or the user through the web application. An additional number of records can be added and further deleted and edited. Every record (location) has a name (column LocationName) and coordinates in decimal degrees format (column Latitude and Longitude). All the data in table Locations is static - the data is not populated during the cookies collection process.

Websites table

Table Websites stores data about the websites that the web application will visit. The web application requires at least one record to exist in the database. This initial record can be edited or deleted by the administrator in the database or the user through the web application. An additional number of records can be added and further deleted and edited. Every record (website) has a name (column

	Column Name	Data Type	Allow Nulls
🔑	CookieID	int	<input type="checkbox"/>
	CookieValue	nvarchar(2048)	<input type="checkbox"/>
	CookieName	nvarchar(512)	<input type="checkbox"/>
	WebsiteID	int	<input type="checkbox"/>
	UserAgentID	int	<input type="checkbox"/>
	LocationID	int	<input type="checkbox"/>
	Domain	nvarchar(1048)	<input type="checkbox"/>
	ExpiresUTC	datetime	<input checked="" type="checkbox"/>
	CreationUTC	datetime	<input checked="" type="checkbox"/>
	LastAccessUTC	datetime	<input checked="" type="checkbox"/>
	IsHttpOnly	bit	<input type="checkbox"/>
	CookieValueEncrypted	nvarchar(512)	<input type="checkbox"/>
	Path	nvarchar(256)	<input type="checkbox"/>
	SameSite	bit	<input type="checkbox"/>
	Secure	bit	<input type="checkbox"/>
	Label	nvarchar(128)	<input checked="" type="checkbox"/>
	CreatedOn	datetime	<input type="checkbox"/>
	UpdatedOn	datetime	<input type="checkbox"/>

Figure 3.3: Table Cookies - design and fields

	Column Name	Data Type	Allow Nulls
🔑	LocationID	int	<input type="checkbox"/>
	LocationName	nvarchar(50)	<input type="checkbox"/>
	Latitude	float	<input checked="" type="checkbox"/>
	Longitude	float	<input checked="" type="checkbox"/>
	Acuracy	int	<input type="checkbox"/>

Figure 3.4: Table Locations - design and fields

Name), category of the purpose of the website, country (of origin) where the website is registered, and the domain name of the website. the domain name is the string that identifies a website. The boolean value 'IsAlexaTop50' shows if the website appears in the Alexa Top50 list.

User Agents table

Table User Agents is used to store data about the websites that the web application will visit. The web application requires at least one record to exist in the database. This initial record can be edited or deleted by the administrator in the database or the user through the web application. An additional number of records can be added and further deleted and edited. The User Agent string that appears to be part of the HTTP request consist of elements that discover the used browser, platform and operating system. Therefore, the corresponding fields are presented as string values in the columns of this table.


	Column Name	Data Type	Allow Nulls
	WebsiteID	int	<input type="checkbox"/>
	Name	nvarchar(256)	<input type="checkbox"/>
	Category	nvarchar(50)	<input type="checkbox"/>
	Country	nvarchar(50)	<input type="checkbox"/>
	DomainName	nvarchar(50)	<input type="checkbox"/>
	IsAlexaTop50	bit	<input type="checkbox"/>

Figure 3.5: Table Websites - design and fields


	Column Name	Data Type	Allow Nulls
	UserAgentID	int	<input type="checkbox"/>
	UserAgentValue	nvarchar(MAX)	<input type="checkbox"/>
	Browser	nvarchar(32)	<input type="checkbox"/>
	Platform	nvarchar(16)	<input type="checkbox"/>
	OperatingSystem	nvarchar(32)	<input type="checkbox"/>

Figure 3.6: Table User Agents - design and fields

4

Results and analysis

This chapter describes the initial analysis of the results and how qualitative and quantitative differences are evaluated. The main goal of the current work is to analyse the current results, simplify the data classification process, and optimise the future analysis of the collected data by machine learning as much as possible. The system consists of three main blocks that form the architecture: data collection system (software web application), data storage component (database), and data analysing module (software application). A detailed description of each module is presented in Chapter 3. Methodology.

4.1. General observations

In this section, general observations of the results are organised in subsections, where a different parameter is discussed in each section. This initial measurement and data analysis provides a basis for more detailed analysis and visualisation. The websites are divided according to their country of origin and category. Additional details on how data is categorised and organised can be found in Chapter 3. Methodology.

From initial analysis, it can be observed that total number of cookies for location Dallas is 6190, and for location The Hague is 3791. This suggests a difference of nearly 63% more cookies collected in Dallas than in The Hague. The first-party cookies have a greater number for location Dallas than for location The Hague, respectively 3397 and 2593, or nearly 31% more first-party cookies collected at location Dallas than at location The Hague. The third-party cookie results are more severe and show that Dallas collected 130% more than The Hague, with a count of 2793 and 1198, respectively.

4.1.1. Parameter: total number of cookies and variety of website categories

The initial observation clearly shows considerable dominance of specific website categories like "Social Network", "Marketplace", and "News and Media" in Alexa's Top 50, Table 3.1. They form approximately 66% of all websites, as seen in Figure 4.1. The dominance of the three categories will affect the number of collected cookies. The categories with only one website are categories "Payment Platform", "Cloud

Service”, and ”Encyclopedia”.

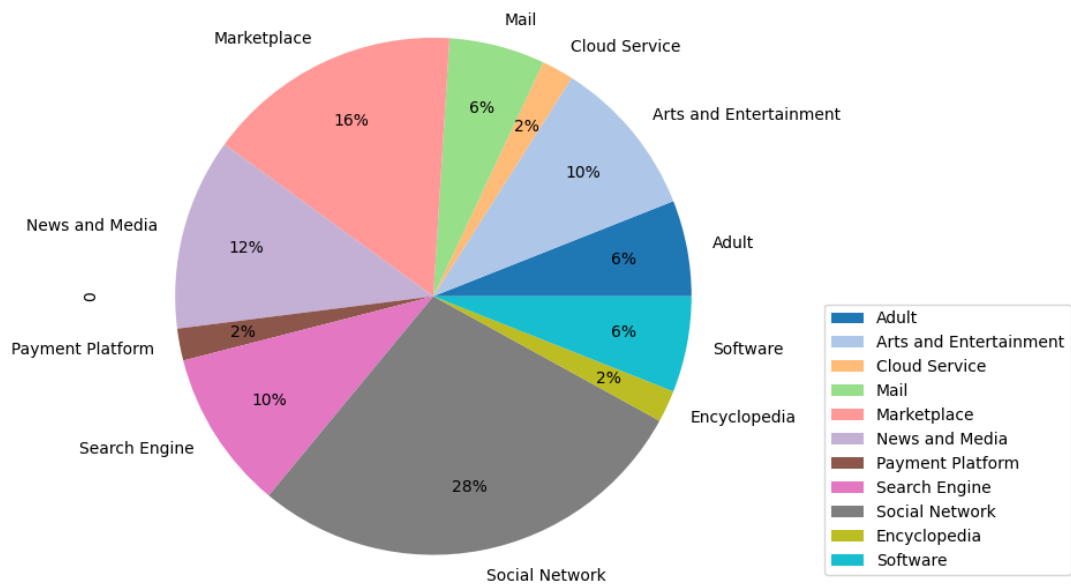


Figure 4.1: Distribution of the number of websites per category

Further, as it can be noticed from Figure 4.2, the websites in the categories ”Marketplace”, ”Social Network” and ”News and Media” form not more than 57% of the total number of cookies (first-party and third-party) for this location as expected. The distribution of the percentages for location The Hague is similar.

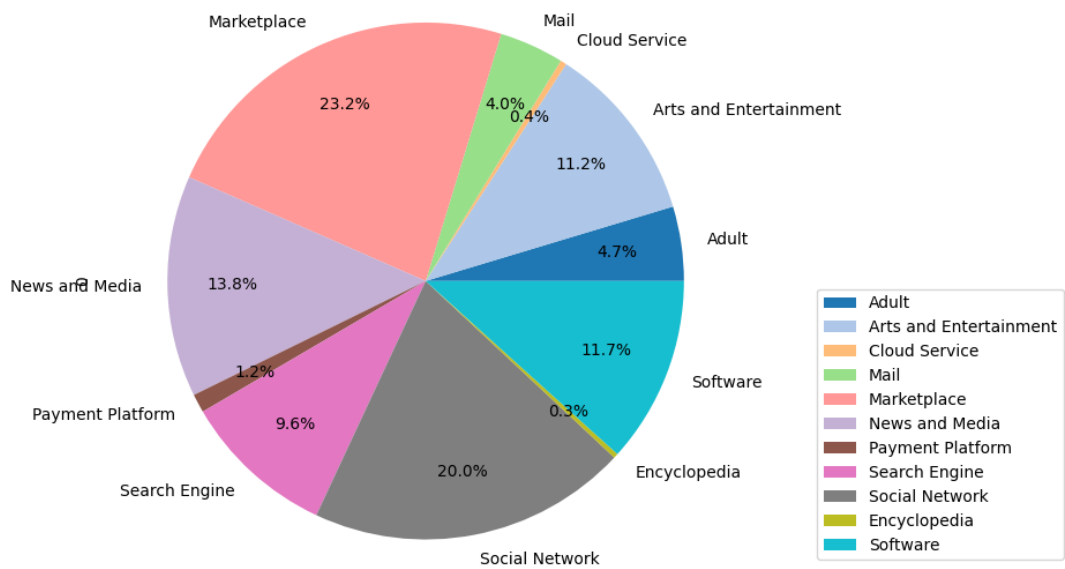


Figure 4.2: Distribution of total number of cookies per website category for location Dallas

4.1.2. Parameter: number of third-party vs. first-party cookies per website category

Figure 4.3 and Figure 4.4 show the distribution of the total number of cookies (first-party and third-party) in each category for both locations. From the results on Figure 4.3 and Figure 4.4, it is obvious that for location Dallas there are significant difference in the collected first-party and third-party cookies for categories: "Arts and Entertainment", "News and Media", "Social Network" and "Software" compared to location The Hague.

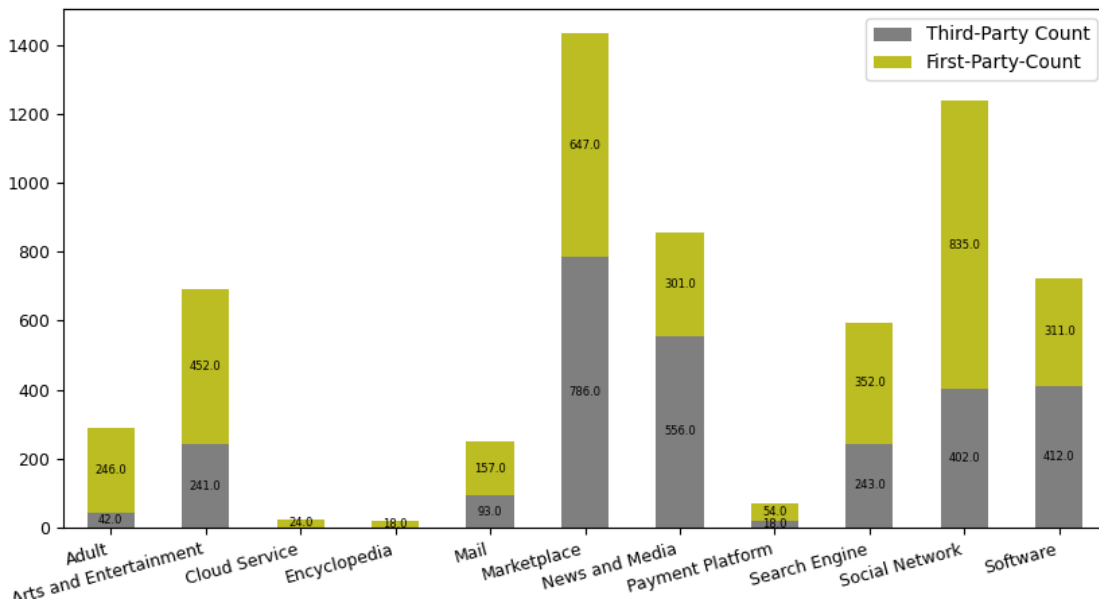


Figure 4.3: Distribution of first-party and third-party cookies count per website category (location Dallas)

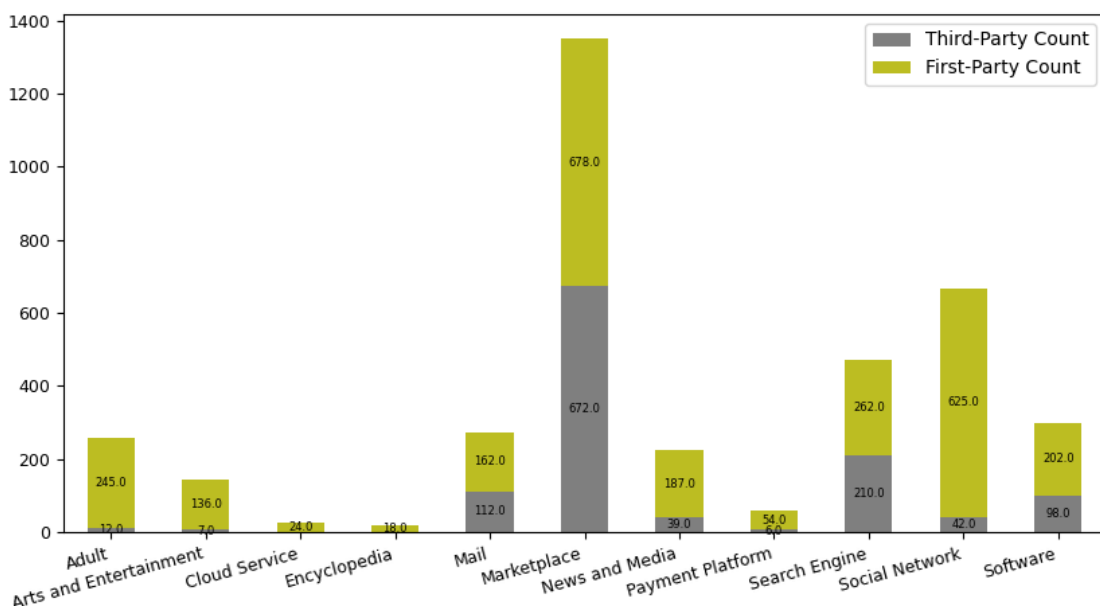


Figure 4.4: Distribution of first-party and third-party cookies count per website category (location The Hague)

For the category "Arts and Entertainment" the collected third-party cookies are 34 times more if the user lives in Dallas, and the first-party one is only three times more than the location The Hague. For the category "News and Media" the collected third-party cookies are 14 times more if the user lives in Dallas, and the first-party one is only 1,5 times more than the location The Hague. For the category "Social Network" the collected third-party cookies are nine times more if the user lives in Dallas, and the first-party one is only 30% more than the location The Hague. For the category "Software" the collected third-party cookies are four times more if the user lives in Dallas, and the first-party one is only 50% more compared to the location The Hague.

Observation (1): The results per website category show that the number of collected first-party and third-party cookies depends straightforwardly on the location and deviate significantly. It concerns the following categories: "Arts and Entertainment", "News and Media", "Social Network" and "Software".

4.1.3. Parameter: number of third-party vs. first-party cookies per operating system or device type

The difference between the number of collected first-party cookies for both locations does not deviate from the general observation – an average difference of 30% per operating system. The same is valid for third-party cookies - the average difference remains around 130% per operating system. The charts are to be found in the Appendix B - Figure B.1. Similar results are shown concerning the device type.

4.1.4. Parameter: total number of cookies per operating system and category

Here, we compare only the four categories that have already demonstrated deviation regarding location (Figure 4.3 and Figure 4.4).

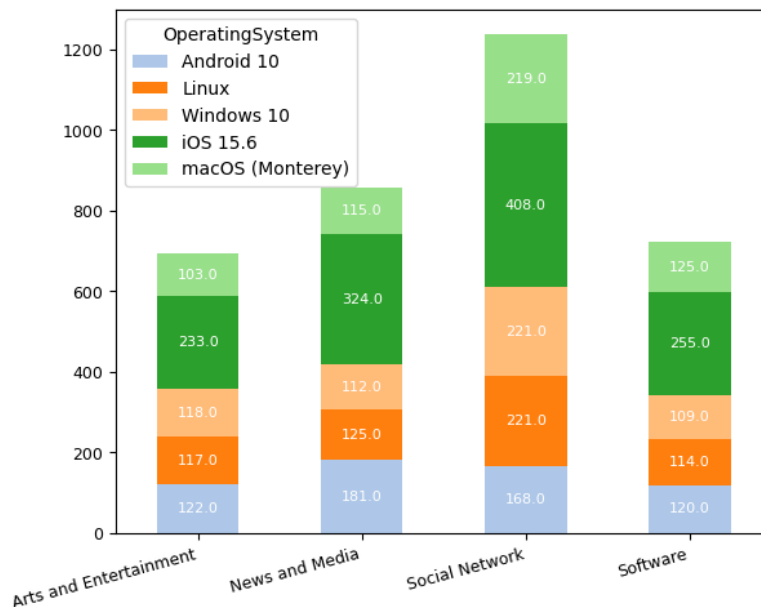


Figure 4.5: Distribution of total number cookies per operating system per website category (location Dallas)

The other categories were also analysed, but they did not show any other discrepancy than the general one. Therefore, they stayed outside the scope of the current analysis. Figure 4.5 and Figure 4.6 show that the ratio between the different operating system in both locations stays relatively the same.

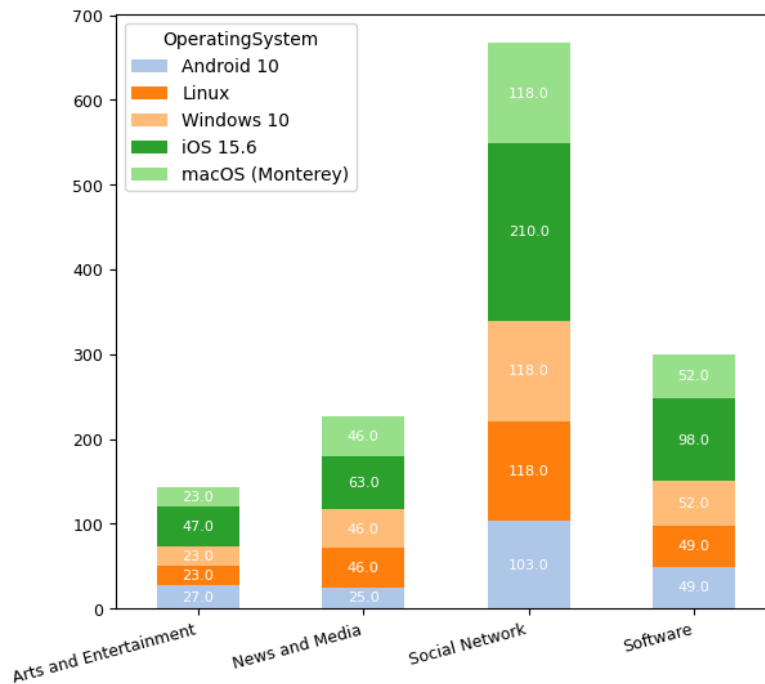


Figure 4.6: Distribution of total number cookies per operating system per website category (location The Hague)

The results show that the total number of cookies in category "Arts and Entertainment" and "News and Media" deviates significantly from the general ratio. The general proportion of the total number of cookies has shown that the total count for location Dallas is nearly 63% more than that for location The Hague. And this ratio is expected to be seen, but this is not the case for the four categories - "Arts and Entertainment", "News and Media", "Social Network" and "Software". In their case, the ratio exceeds 63%. For the category "News and Media" the collected total number of cookies for Dallas is more than seven times for "Android 10" operating system and five times more for "iOS 15.6", compared to location The Hague, and that is 2-3 times more than the ratio for the other operating systems in the same location. For the category "Arts and Entertainment" the collected total number of cookies for Dallas is approximately five times more than for location The Hague. Still, this ratio is only 3.8 for the "Android 10" operating system.

Observation (2): The operating system per website category result shows that the total number of collected cookies depends, in some cases, directly on the operating system type. It concerns the following categories: "Arts and Entertainment", "News and Media", and mainly mobile operating systems. In "Arts and Entertainment" at location Dallas, the "Android 10" operating system is detected, and therefore fewer cookies are collected than expected and, respectively, more for the other systems. In "News and Media", the proportion of collected cookies for Dallas is more significant for mobile devices (operation system "iOS 15.6" and "Android 10" for mobile devices), respectively less for desktop devices ("Windows 10" and "macOS Monterey").

Observation (3): Looking at iOS devices, a certain inconsistency can be seen from the results - they seem to collect fewer third-party cookies than the standard desktop machines and Android devices and relatively the same amount of first-party cookies per location.

4.1.5. Parameter: expiry date

As can be seen from the results presented in Figure 4.7, between 17-23% of the first-party cookies have a NULL expiry date, whereas third-party cookies show between 8-17%. The trend one can see is that there is a significant difference in the expiry date of the cookies for different locations. That means that the cookies expire only after the user closes the browser. Approximately 37-48% of all cookies have an expiration date of more than a year, and 41% of all cookies with an expiry date of more than a year are categorised as third-party cookies. Since Google introduced a security feature after Chrome 104 (release date August 2022), the maximum life of the cookies was limited to 400 days.

A significant difference cannot be seen if one looks at the distribution of the expiry date per user agent per device and operation system, both for first-party cookies and third-party cookies (Figure B.2 and Figure B.3). As it concerns the distribution per location (Figure 4.7), 17-23% of the first-party cookies have an expiry date set to NULL, while third-party only between 8-17%. 37-43% of the first-party cookies have an expiry date of more than a year, and third-party up to 40-48%.

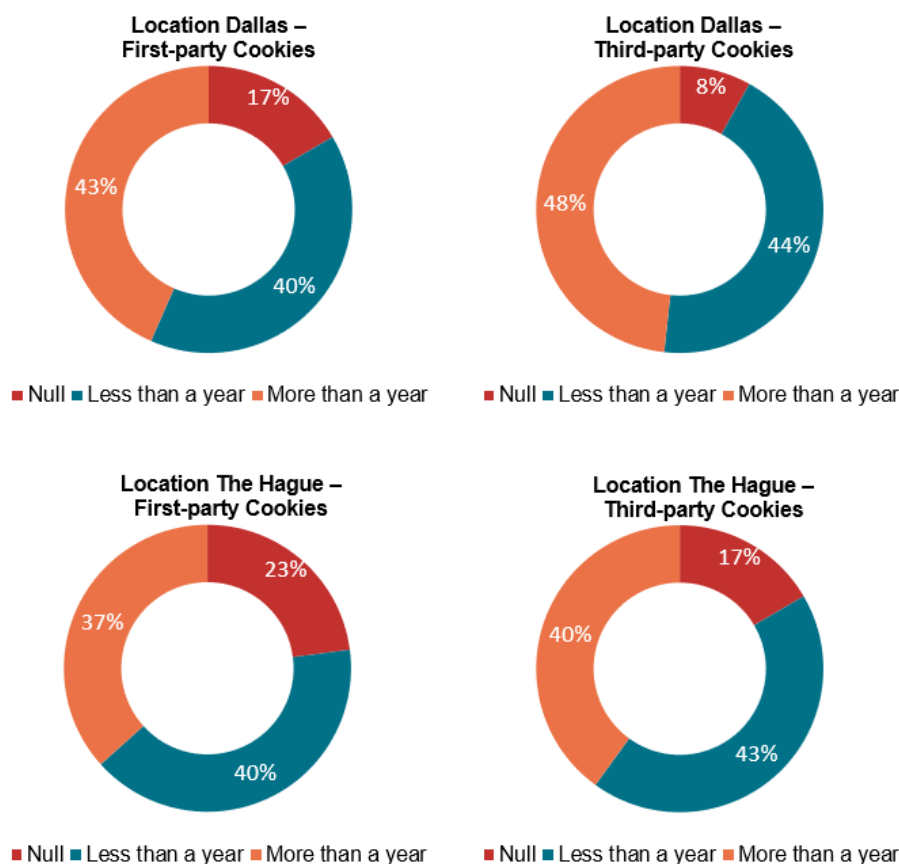


Figure 4.7: First-party and third-party cookies expiry date for different locations

4.1.6. Outliers observations about first-party and third-party cookies count in location Dallas and location The Hague per website

In this section, the results of the analysis performed by the web application are presented. Partial database snapshots are shown in Table 4.1. The complete data set used for the analysis in the current section can be found in the Appendix C in Table C.1. The analysis below contains only the websites where inconsistencies have been found; the websites with an equal number of cookies per device are not selected. The websites where the difference is one or two cookies are also not selected. The following basic observations about the distribution of the cookies across different user agents can be noted:

The following websites are observed to show significant differences in collected third-party and first-party cookies per device type (operating system) of the user:

- Amazon.com collects third-party cookies only for mobile devices in Dallas; the first-party cookies are ten times more for mobile than desktop.
- Baidu collects approximately two times more first-party cookies for mobile phone devices for both locations.
- Yahoo collects 3-10 times more third-party cookies for desktop devices and iPad than for mobile phone devices for Location Dallas.
- Bing does not collect any third-party cookies for mobile phone devices for both locations.
- QQ collects fewer first-party cookies for mobile phone devices for both locations.
- MSN collects 3-5 times more third-party cookies for mobile devices than for desktop for location Dallas; at the same time, the collected third-party cookies are 80-100 times more for mobile devices and 11-17 times more for desktop than for location The Hague
- Aliexpress collects 50% more first-party cookies and 200-800% more third-party cookies for desktops than for mobile devices for location Dallas. Observations for location The Hague are similar: 50% more first-party cookies for desktops and 50-70% more third-party cookies for desktops and Samsung devices.
- Yandex collects four times more third-party cookies for desktops than for mobile devices for both locations.
- Xhamster collects 300-600% more third-party cookies for mobile phone devices than for the rest for location Dallas; same is valid for location The Hague - more third-party cookies for mobile phone devices
- Adobe collects 25% less third-party cookies for Windows 10 and Linux than for mobile devices and macOS for location Dallas.

The following websites are observed to show a significant deviation in collected number of third-party and first-party cookies per location of the user:

- Zoom.us does not collect third-party cookies for The Hague and collects twice more first-party cookies for Dallas
- Office.com collects 100% more third-party cookies for location Dallas than for location The Hague
- LinkedIn collects eight times more third-party cookies in Dallas than in The Hague

- Fandom collects 20-40 times more cookies for location Dallas than for location The Hague
- Canva collects 18 third-party cookies per device for Dallas, whereas for The Hague, there are no third-party cookies collected
- Adobe collects 30-47 times more third-party cookies for location Dallas than for location The Hague.

From all Alexa Top 50 websites, 18 (36%) collect different cookies per device. That includes 33% of the websites in the category Adult, 20% of Entertainment, 66% Mail, News and Media and Software, 38% Marketplace, 60% of Search Engine, and only 14% of Social Network websites categories. As it concerns the origin of the website: 50% of all China websites, 100% of the websites from Cyprus and India, 66% from Russia, and 28% of the websites from the USA (Reference: Appendix C in Table C.1).

Name	OperatingSystem	Platform	TP	FP	Location
https://www.adobe.com	Android 10	Samsung Mobile	45	35	Dallas
https://www.adobe.com	Android 10	Samsung Mobile	1	21	The Hague
https://www.adobe.com	iOS 15.6	Apple iPad	47	39	Dallas
https://www.adobe.com	iOS 15.6	Apple iPad	1	21	The Hague
https://www.adobe.com	iOS 15.6	Apple iPhone	47	39	Dallas
https://www.adobe.com	iOS 15.6	Apple iPhone	1	21	The Hague
https://www.adobe.com	Linux	Desktop	35	39	Dallas
https://www.adobe.com	Linux	Desktop	1	21	The Hague
https://www.adobe.com	macOS (Monterey)	Desktop	48	37	Dallas
https://www.adobe.com	macOS (Monterey)	Desktop	1	21	The Hague
https://www.adobe.com	Windows 10	Desktop	30	35	Dallas
https://www.adobe.com	Windows 10	Desktop	1	22	The Hague

Table 4.1: Database Snapshot for Adobe.com for Dallas and The Hague

Observation (4): As seen in the last results, some of the websites distinguish the type of device or the operating system and demonstrate preference in collected cookies.

Observation (5): Some websites do not collect cookies for location The Hague (Facebook, Whatsapp, Instagram from Meta Inc., as well as Yahoo.jp, and Shopify) or collect significant more third-party cookies for location Dallas.

Observation (6): Three of the website categories, "News and Media", "Marketplace" and "Encyclopedia", collect cookies containing "Geo" in the cookie name. And the value contains the city and/or the user's country.

4.2. Data visualization task

For the visualization of the data, Python in combination with the scikit-learn library has been used as data analysis tool and to generate plots as heatmaps. Heatmaps serve as graphical representation of data and each individual values that are contained in a matrix is represented by a different color. As

a result of comparing the data, a few notable relationships between different variables are worth to be mentioned and used in the classifier.

4.2.1. Data preprocessing

Initially, the cookie data set was prepared for processing. The labels 'Domain' and 'UserAgent' were encoded and labeled with value between 0 and $n_classes-1$. The rest of the data is already in a form that allows further visualization and classification. A new column 'Duration' (value in hours) is added that is used to indicate the life span of each cookie. Value 0 represents a cookie that gets deleted when browser session gets closed.

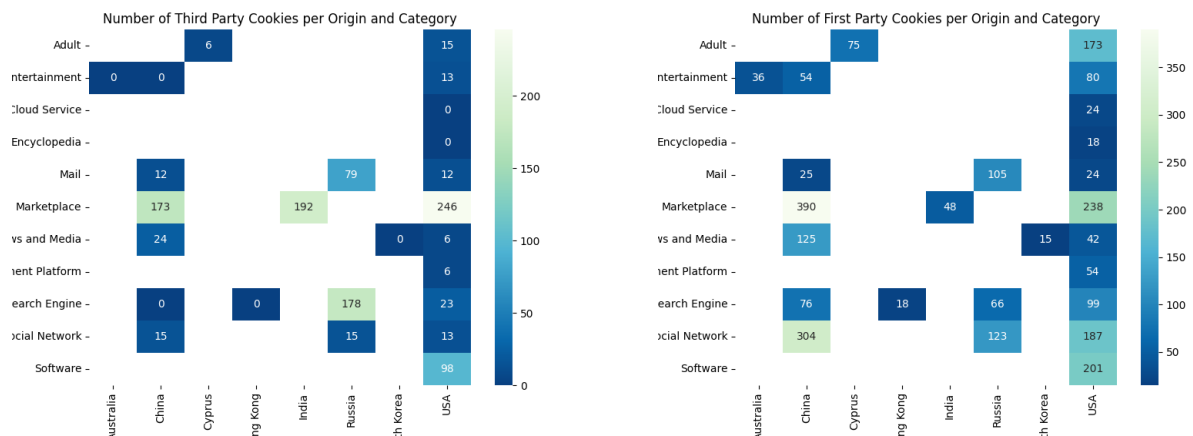


Figure 4.8: Left: number of third-party cookies per origin and category. Right: number of first-party cookies per origin and category. Location: The Hague

4.2.2. Data observations

This section presents a comparison of several variables and their interactions to demonstrate observed differences in the collected cookie data.

Accumulative count of first-party and third-party cookies per location (Category-Origin).

An interesting observation from Figure 4.8 and Figure 4.9 is that both third- and first-party cookies count are significant for websites from category 'Marketplace' and 'Social Network'. If the count of the first-party is compared to the count of the third-party, both for locations The Hague and Dallas, a significant difference can be seen. For example, the third-party count for the category 'Search Engine' is more than three times more for location The Hague than the first-party cookies for websites with origin 'Russia'; for location Dallas, the difference is only 60%. A similarity can be observed for the category 'Marketplace' and origin 'India' - the count of third-party cookies compared to first-party ones is five times more for location The Hague and almost six times for location Dallas. For location Dallas and websites with origin 'USA', categories 'News and Media' have 4,5 times more third-party cookies and for category 'Software' - 30% more third-party cookies are detected than in location The Hague.

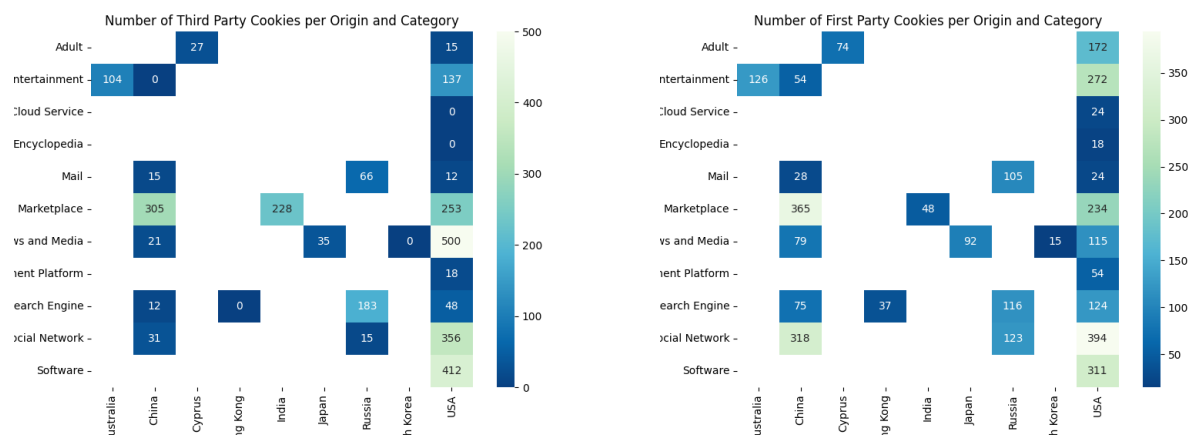


Figure 4.9: Left: number of third-party cookies per origin and category. Right: number of first-party cookies per origin and category. Location: Dallas

If the first-party cookies are compared in both locations: The Hague and Dallas, a few remarks are interesting. Category 'Entertainment' and origin 'USA' have 3,5 times more cookies collected in Dallas, compared to The Hague, same can be noted for category 'News and Media'. In categories 'Social Network' and 'Software' is the difference respectfully 100% and 50%, and category 'Search Engine' only 25% more first-party cookies for location Dallas than for location The Hague.

The outliers seem more outrageous when looking at heatmaps of the third-party cookies for both locations. For instance, websites from the category 'Entertainment' and origin 'Australia' collect in location Dallas 104 third-party cookies, whereas no cookies are collected for location The Hague. Same category, but with the origin 'USA'; they collect 137 cookies in Dallas but only 13 in The Hague. This is more than ten times difference. Another notable example is the websites with the origin 'USA' in the category 'News and Media' - the difference between the two locations is in favor of location Dallas 166 times more cookies are collected there (500) and only 6 for The Hague. Similar observations are in the category 'Social Network' with origin 'USA' - 15 cookies if the websites are visited from The Hague and 356 if they visited from Dallas; for the category 'Software' - 98 cookies in The Hague and more than four times more in Dallas (412).

Average duration of first-party and third-party cookies per location

As it can be seen from Figure 4.10, some interesting observations can be derived. The results on average duration of the cookies per category and website origin are similar. But two outliers are standing out. Category 'Search Engine' from origin 'Hong Kong' that has significantly longer duration at location The Hague than location Dallas - 7865 hours and 2485 hours respectively. Less difference is detected from websites from category 'Software' and origin 'USA' - the duration for location Dallas is approximately 30% more than for location The Hague. The rest of the websites appear to store the cookies relatively same amount of hours.

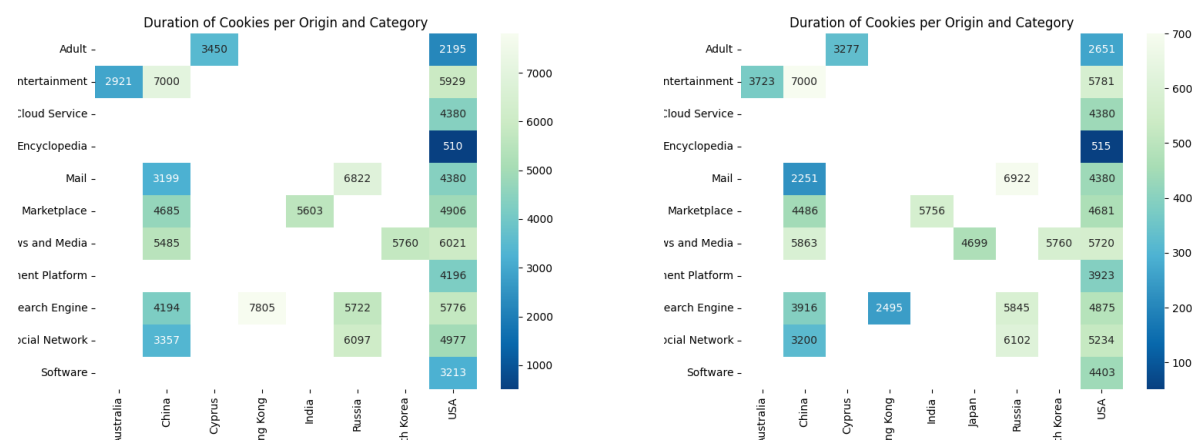


Figure 4.10: Average duration of cookies per location Left: The Hague Right: Dallas

Accumulative count of first-party and third-party cookies per location (category - user agent).

As can be seen from Figure 4.11 and Figure 4.12, specific categories tend to collect more first-party and third-party cookies for both locations. For location The Hague, the categories are 'Marketplace' and 'Search Engine' for third-party cookies and 'Marketplace' and 'Social Network' for first-party cookies. For location Dallas, the heatmap looks slightly different. Here, more categories collect a significant number of first-party and third-party cookies. For the third-party, the categories are: 'Marketplace', 'News and Media', 'Social Network', and 'Software', whereas for the first-party cookies: 'Entertainment', 'Marketplace', and 'Social Network'.

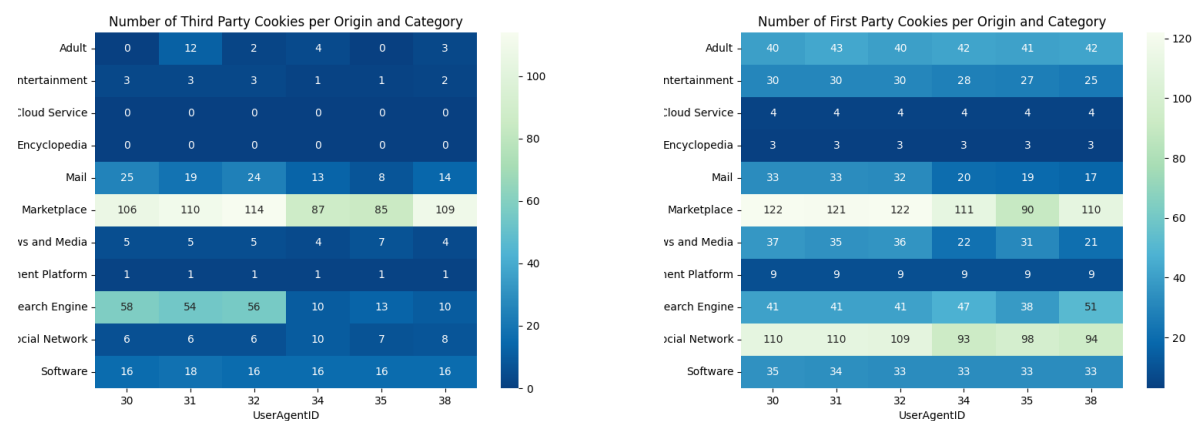


Figure 4.11: Left: number of third-party cookies per user agent and category. Right: number of first-party cookies per user agent and category. Location: The Hague

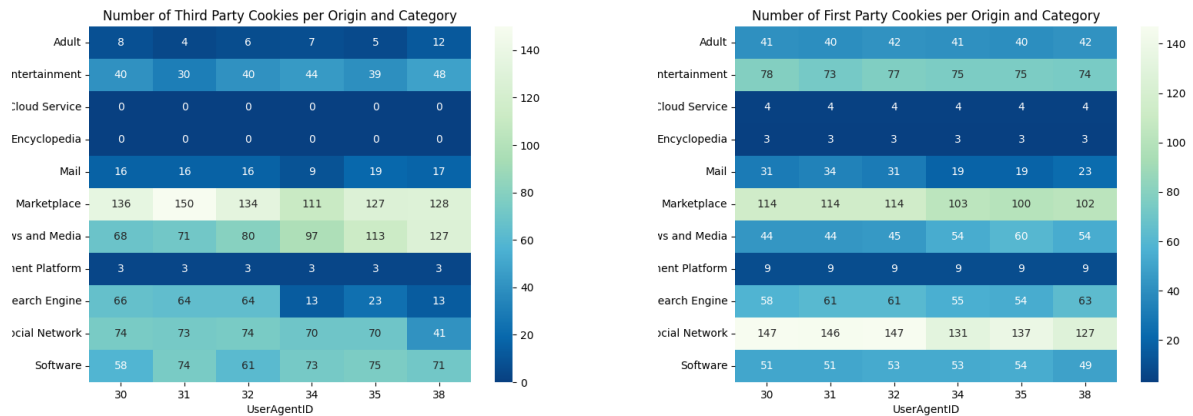


Figure 4.12: Left: number of third-party cookies per user agent and category. Right: number of first-party cookies per user agent and category. Location: Dallas

4.2.3. Data imbalance

This section uses a technique called SMOTE (Synthetic Minority Over-sampling Technique) to enhance the performance of three chosen classifiers. This algorithm is used to address the problem of the imbalanced data set. The goal is to identify the websites from a specific category and origin that is likely to collect more third-party than first-party cookies. All data categories that do not provide meaningful information will be removed.

First of all, the data has to be made ready for processing. The data set should only contain numerical values. To get more algorithms' features, categorical data columns must be transformed into numerical ones. At first, every category and origin country is mapped to a number. After this, a different approach was considered because the chosen method would give the idea of a linear relationship between all the categorical values. So instead of giving every category its value in one feature, now every class is a separate feature containing a boolean value, as zero if that row does not have that category, one if it does.

After getting all the required numerical features, the categorical features can be filtered out (except for simple journals since this is needed for the algorithm). Also, the booking date and creation date are removed. Booking date and creation date are not an allowed feature, so it is not helpful to keep them.

For this task, four classifiers were selected to be trained with the cookie data set: Random Forest, K-NearestNeighbours (K-NN), Logistic, and Decision Tree Classifier. The four algorithms were compared to see how each performs and which performs best. To visualise the results as a plot, an additional Python library was used *scikit-learn* in combination with *imblearn* to include SMOTE in the analysis. All the source code is to be found in the script *data-ML.py* in the Appendix A. The other ROC (receiver operating characteristic) curves and the corresponding Precision-Recall curves can be found in Appendix B.

Results

To address the issue with the imbalanced data set, SMOTE is used. Based on the results for the current data set, a conclusion that SMOTE increases the accuracy of the results versus the results from the original data can not be made. As seen from the results from Figure 4.13, SMOTE does not significantly increase the algorithm's performance compared to just using the original data. The following classifiers have close results and have shown nearly the same performance:

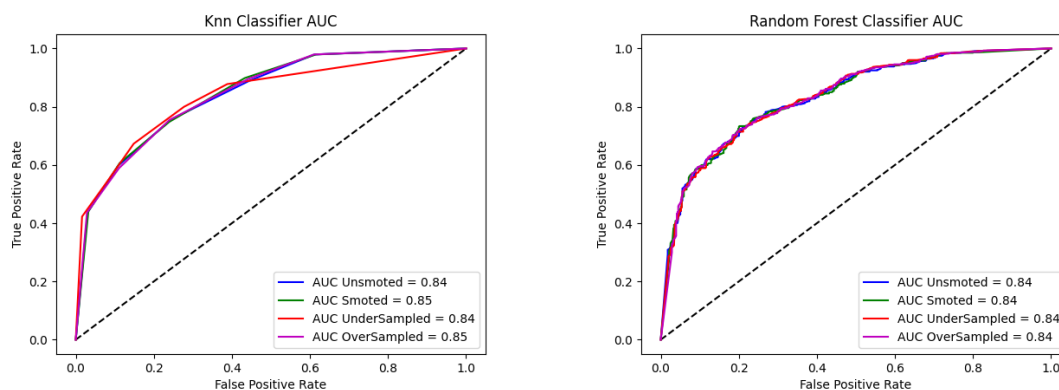


Figure 4.13: Left: K-NearestNeighbour classifier. Right: Random Forest classifier

According to the results, K-NearestNeighbour Classifier and Random Forest classifier perform the best among the four selected and show a similar curve. The Decision Tree classifier performed slightly similarly - with a score of 83. In contrast, the Logistic classifier demonstrated a score of 64, with or without SMOTE, as can be seen in Figure B.7 Appendix B.

Concluding, SMOTE is a good idea because it can increase the number of minority samples for some data sets. Still, SMOTE only works in some cases, and others perform poorly, like in this case, Decision Trees. Therefore, SMOTE should be used when the context of the data is clear, and the objective is known.

5

Mitigation

This chapter discusses possible ways to mitigate the effect of tracking cookies and aims to answer the last research question of the current work. It addresses the following sub-question: *"What strategies and tools can be used to reduce the impact of the tracking cookies and protect the end-user?"*. In the process of answering this question, various methods and techniques for mitigation will be addressed. Apart from the popular methods like deleting cookies (and cache) from the browser, less-popular techniques among users will be discussed. The suggested solutions are not rated by importance, effectiveness, or any other attribute, but they are grouped according to the actions required from the users.

5.1. Solutions that require little or no action from the user

This section suggests solutions to mitigate the effect of first-party and third-party cookies and can be grouped as browser-related measures. These solutions require no or minimum action from the user. Tracking cookies can be disabled within the standard functionality of every main browser. Google Chrome, Mozilla Firefox, Safari, and Microsoft Edge have similar steps to block third-party cookies. Even as an initial step, the user is advised to delete all previously collected cookies and empty the browser's cache - this must be performed at least once a month anyway. The first approach prevents cookies from tracking users and possibly exposing private data and is integrated into the browser. In addition to this first method of limiting third-party and tracking cookies, it can be suggested using the browser's private mode. But this has a downside - the websites will not "remember" the user, and every time settings must be chosen again, or login credentials typed, among many others.

One of the features that Google Chrome released with their browser version 104 is the limit of the expiry date of the cookies to 400 days (source link: [Feature: Cookie Expires/Max-Age attribute upper limit](#)). That prevents future tracking cookies from being stored for years on the file system unless deleted. Further, Google announced in 2019 a project called "Privacy Sandbox" (source link: [Privacy Sandbox](#)). As described, this project aims to enhance and protect users' privacy online and reduce cross-site and cross-app tracking without decreasing the functionality of the online content. The project owners promise that it will deliver new web standards and are currently in the testing stage of the process.

The second group of solutions within this category that prevent users' private data from possibly leaking, is relying on the legal framework of the region of the user. Every country should take measures to protect the privacy of its citizens. The right to privacy is recognised as one of the primary human rights, according to the "Universal Declaration of Human Rights" [38]. This document states that each citizen has the right to protection by the law in case of interference or attack against privacy, among other rights. The two active documents related to the protection of privacy in the digital environment are the General Data Protection Regulation (GDPR) [16] and ePrivacyDirective [39]. The GDPR states the obligatory measure that each state member of the European Union should force and regulations overrule national law. In contrast, the ePrivacyDirective represents a set of goals each member should achieve, and each member is free to decide how to apply those requirements. In some cases, GDPR is also used by companies outside the European Union when they collect, process, track, and analyse sensitive data of European citizens regardless of their origin. An equivalent to GDRP in the United States is California Consumer Privacy Act (CCPA), but this document only applies to the state of California and differs from the GDRP. The GDPR affects prior concerns of the citizens regarding collecting their private data, whereas the CCPA reflects on the withdrawal and transparency of already collected data.

Unfortunately, legislation alone is not a necessary and sufficient condition to enforce the processing of privacy of sensitive data accordingly. "Privacy by Design" is a term retrieved from the GDPR. Privacy by Design is developed into a framework that guarantees that privacy will be considered throughout all stages of engineering and design process. It was initially suggested in 2009 by Ann Cavoukian [6]. Privacy by Design certificate can be obtained by organisations and by that they prove they follow the seven foundational principles and that a company demonstrates the ability to secure and protect the personal and confidential data of its customers, employees, and business partners.

5.2. Solutions that require some action from the user

As we have seen earlier in Chapter 4. Results and analysis, browsers collect first-party and third-party cookies even before the user's consent. These cookies can contain private data. The file with collected cookies is saved on the user's local folder structure. Some cookies disappear by closing the browser window, but others persist locally until the expiry date. A possible workaround to prevent cookie data from remaining locally for a longer period is to delete the browser folder containing cookie data by hand. That should happen at least once a month or, even better - once a week. But users tend to forget to do it regularly.

An alternative to the deletion of cookies manually by the user is the following small script instructions. This solution prevents users from deleting the wrong folder or forgetting to do it regularly. It requires one-time actions from the user to set up the standard functionality in Windows. The following script (Listing 5.1) is written in PowerShell [34] and is used to delete all the files from a directory. A common command adapted for the current use case is deleting collected cookies from a browser. The browser, in this case, is Google Chrome.

```
1 $FileName = "%LocalAppData%\Google\Chrome\User Data\Default"
2 if (Test-Path $FileName) {
3     Remove-Item $FileName
```

```
4 }
```

Listing 5.1: Powershell script that deletes local files

Another option for deleting files, is to use Batch (Listing 5.2). It can be executed within the standard Windows command line functionality:

```
1 if exist "%LocalAppData%\Google\Chrome\User Data\Default" rmdir "%LocalAppData%\Google\Chrome
   \User Data\Default" /q /s
```

Listing 5.2: Batch script that deletes local files

Both variants of script commands execute the same logic steps - first check if the folder exists and then delete the folder. This deletes all temporary files, stored by the Chrome browser, including all cookies. This script file needs to be stored on user local file storage.

Next, this piece of code needs to be scheduled for execution at least once a week. It depends on the frequency of browsing of each user, but the recommended is once a week. For this purpose, a standard functionality of Windows can be used. The application is called "Task Scheduler". Setting up the scheduler is not complicated task and a user can find good instructions of how to schedule execution of the PowerShell script online.

5.3. Concluding remarks

To conclude, the answer to this research question does not have one universal solution. The presented solutions represent an extended but not complete list. Some of the suggested techniques alone have proven insufficient compared to the sophisticated algorithms that some websites may use. But combining a few of the measures mentioned above should decrease the effect of tracking significantly and increase personal data protection.

6

Discussion and conclusion

This chapter completes the current work. Moreover, it reflects on the related research work and summarises the answers of the research questions. The limitations of the current work are discussed, as well as suggestions and ideas for future development and expanding the functionality and integrity of the web application and the solution as such.

6.1. Discussion

This thesis started with defining the project's goals and introducing the research questions in Chapter 1 Introduction. Chapter 2 Background continued with setting the theoretical background of the topic and a review of the related research work regarding HTTP Cookies was executed in Chapter 2.2 Related work. The next Chapter 3.7 Dataset and database structure made an overview of the database's structure and the data set's main characteristics. Next, in Chapter 3 Methodology the conceptual model of the system was introduced, together with a description of the data collection procedure for this research. The results of the current work are presented in Chapter 4 Results and analysis database's structure and the data set's main characteristics. Chapter 5 Mitigation focused on the mitigation of the threat of tracking cookies.

Here in this concluding chapter, an overview of this research will be given and how it relates to the problem of tracking cookies. Moreover, the results and analysis of the current research will be associated with answers to the questions defined in the Introduction.

Over the last few years, web tracking has matured and evolved into a billion-dollar industry. Even though numerous solutions to avoid tracking and protect users' privacy exist, it is still little known how this tracking affects the users and what the actual impact on the user is. This impact seems hard to measure and the negative effect on users' confidence is hard to prove. As seen in the current research and from the state-of-the-art, online tracking appears to be hard to define and hard to circumvent. This practice has raised privacy concerns and resulted in users adopting mitigating techniques. In this work, the effect of online tracking and its impact on users' private data is evaluated by the following research question:

"How to reverse engineer the technique of creating cookies based on personalizing the content that is stored within them, and what model can be built that analyses the information stored in the cookies?"

6.1.1. Research sub-questions

To address the first sub-question, the current work focused on the features and user characteristics that websites collect as first-party and third-party cookies. A literature survey was conducted to compare previous research in the area and to recognise the main parameters. The discussed study showed examples of custom-made applications that collect data from websites. Browser cookies, being first-party or third-party, are the fundamental technique for tracking users. Cookies are used as a place to store user settings and personal identification. A small amount of data inside cookies contains identifiers that can be recognised as personal data. When a website requests a cookie, the previously saved data is returned. After this initial investigation, proof of concept was built as a basis. Next, a web application was developed to serve the purpose of HTTP cookie collection. Usually, the small amount of data in cookies contains identifiers that can be recognised as personal data. Practically, how websites comply with legal requirements concerning storing and processing users' private information needs to be clarified.

The second sub-question identifies the user's characteristics and parameters that can be used to build a machine-learning model that contextualises websites that are likely to collect tracking cookies. These tracking cookies could contain private data that can identify the user. Analysis of the data exploration results demonstrates that a model can be trained to recognise the features that users create with their online activity and characteristics that might lead to leaking this content to trackers. To prepare a model, it is essential to identify the biggest possibility of a website to collect tracking cookies based on its characteristics or other features.

Based on comparisons made between the current information system and other similar ones, the advantages of the current one stand out. Unlike other systems, there is improved functionality in terms of cookie collection, initial information analysis, simple use, and the possibility to adapt the system to the user's needs. It is possible to keep available information as up-to-date as possible without an administrator's need for continuous intervention. This makes the applications preferred and valuable for users. Moreover, the user vectors that were investigated included simulation of several devices, several operating systems, several browser applications and two different geographical locations were compared. The results collected from this multi-vector environment were analysed and presented in Chapter 4 Results and analysis.

The answer of the third research sub-question will not be discussed here, as it has been answered and discussed in a separate chapter - Chapter 5 Mitigation.

As it can be seen from the results and observations in Chapter 4, it can be concluded that there exists a pattern in collecting cookies by some websites. This pattern, in some cases, is based on the user's location, while others on the device type or the operating system. A possible explanation for this anomaly could be the different legal frameworks of the country of origin of the websites or location of the users at the moment of the browsing. The reason for more collected first-party or third-party cookies for some mobile devices can be the responsive design of the websites or different structure

of the operating system. It can be explained as more cookies are needed to render the websites for mobile browsers. On the other hand, fewer cookies for mobile devices can derive from the different folder structure compared to desktop machines. Online community and legal parties should stimulate systems and frameworks that circumvent web tracking and reduce privacy risk.

6.2. Future work and limitations

Even though current research showed sufficient results to build an answer to the research question, it would be practical to extend the functionality of the web application and develop a bigger system. After the completion of current work, the following improvement points are recognised:

- To investigate if there is and to what extent exchange of information between specialised trackers and other mainstream advertisers and marketers.
- To research how GDPR relates to tracking of sensitive content - religion, race, political beliefs and sexual orientation.
- How cross-border tracking of sensitive content can be detected?
- Web application should to be tested with more websites and more locations on a machine with better hardware configuration.
- Further development could include simulating user navigation in a website - with clicking on cookies consent or refusal and detect the difference in the collected cookies
- Develop an API or framework that can be used by other developers or researchers

This is not extensive list but these features can add additional functionality and value to the system and make it more attractive for future re-use in other research work.

As a limitation of this work can be recognised the limited capacity of the hardware used. More extensive test cases can be executed on a machine with better memory size and productivity performance.

6.3. Concluding remarks

This paper analysed the scale of possible first-party and third-party tracking among Alexa's Top 50 websites. It focused on tracking cookies and websites that obtain (private) data extracted from users' online activity. The current work demonstrated the scope of the presence of collected cookies in the selection of the most visited websites in the world. The collection of the cookies by the majority of the websites appeared to be happening before receiving the user's compliance and consent. Hence, the data collected in the cookies was not strictly limited to third-party cookies. And since, in some cases, it was shown that some of the cookies contain private data and personal identifiers, while others have unclear content, it should be alarming evidence for the online community and legal authorities to take relevant measures. After a brief analysis of the data set, a significant inconsistency in the cookies' count that deviates per location, device type, and operating system is detected. The results show that some websites collect private data, and users are not informed about the collected data, even though the consent for using cookies is not retrieved. These results imply that tracking persists as a serious concern, regardless of legal and technical countermeasures. Despite the fast-developing tracking technologies and the constant interplay between the users and the business, the question of

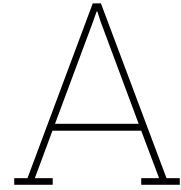
what data is collected from the users and how this data is further processed remains. The intention of the users to be protected from unintentionally sharing private data has been primarily ignored by the digital organisations. The transparency and reduced unknown risk for the user while browsing online will result in trust and confidence in a particular website. The combination of legal, economic, and ethical concerns results in an adverse user reaction. That is why businesses and online communities should focus on reducing privacy risks. Legal measures should strategically support the circumventing technical efforts. It seems like a never-ending race between the tracking websites on one side and the users on the other.

References

- [1] Gunes Acar et al. “The web never forgets: Persistent tracking mechanisms in the wild”. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 2014, pp. 674–689.
- [2] Jason Bau et al. “A promising direction for web tracking countermeasures”. In: *Proceedings of W2SP* (2013).
- [3] Tim Berners-Lee, Roy Fielding, and Henrik Frystyk. *Hypertext transfer protocol–HTTP/1.0*. Tech. rep. 1996.
- [4] Reuben Binns et al. “Tracking on the Web, Mobile and the Internet of Things”. In: *Foundations and Trends® in Web Science* 8.1–2 (2022), pp. 1–113.
- [5] Aaron Cahn et al. “An empirical study of web cookies”. In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 891–901.
- [6] Ann Cavoukian et al. “Privacy by design: The 7 foundational principles”. In: *Information and privacy commissioner of Ontario, Canada* 5 (2009), p. 12.
- [7] Google Chrome. *Chrome Browser 105*. <https://developer.chrome.com/blog/new-in-chrome-105/>.
- [8] Open Source project of Chromium and WebDriver teams. *Chrome Driver*. <https://chromedriver.chromium.org/>.
- [9] Adrian Dabrowski et al. “Measuring cookies and web privacy in a post-gdpr world”. In: *International Conference on Passive and Active Network Measurement*. Springer. 2019, pp. 258–270.
- [10] Martin Degeling et al. “We value your privacy... now take some cookies: Measuring the GDPR’s impact on web privacy”. In: *arXiv preprint arXiv:1808.05096* (2018).
- [11] Nurullah Demir et al. “Towards Understanding First-Party Cookie Tracking in the Field”. In: *arXiv preprint arXiv:2202.01498* (2022).
- [12] Steven Englehardt and Arvind Narayanan. “Online tracking: A 1-million-site measurement and analysis”. In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 1388–1401.
- [13] Steven Englehardt et al. “Cookies that give you away: The surveillance implications of web tracking”. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 289–299.
- [14] Roy Fielding et al. *RFC2616: Hypertext Transfer Protocol–HTTP/1.1*. 1999.
- [15] Behrouz A Forouzan. *TCP/IP protocol suite*. McGraw-Hill Higher Education, 2002.
- [16] *General Data Protection Regulation (GDPR)*. European Commission. May 25, 2018.
- [17] Roberto Gonzalez et al. “The cookie recipe: Untangling the use of cookies in the wild”. In: *2017 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE. 2017, pp. 1–9.

- [18] Matthias Gotze et al. "Measuring Web Cookies in Governmental Websites". In: *14th ACM Web Science Conference 2022*. 2022, pp. 44–54.
- [19] Samuel Greengard. "Weighing the impact of GDPR". In: *Communications of the ACM* 61.11 (2018), pp. 16–18.
- [20] W3C Working Group. *Tracking Compliance and Scope*. Accessed: 16-02-2022.
- [21] Xuehui Hu and Nishanth Sastry. "Characterising third party cookie usage in the EU after GDPR". In: *Proceedings of the 10th ACM Conference on Web Science*. 2019, pp. 137–141.
- [22] Costas Iordanou et al. "Tracing cross border web tracking". In: *Proceedings of the internet measurement conference 2018*. 2018, pp. 329–342.
- [23] Samy Kamkar. "Evercookie". In: *URI: http://samy.pl/evercookie* (2010).
- [24] Richie Koch. *Cookies, the GDPR, and the ePrivacy Directive*. Accessed: 27-12-2022.
- [25] D Kristol and L Montulli. *RFC 2109-HTTP State Management Mechanism*, 1997.
- [26] David M Kristol. "HTTP Cookies: Standards, privacy, and politics". In: *ACM Transactions on Internet Technology (TOIT)* 1.2 (2001), pp. 151–198.
- [27] Sumit Kumar, Sumit Dalal, and Vivek Dixit. "The OSI model: Overview on the seven layers of computer networks". In: *International Journal of Computer Science and Information Technology Research* 2.3 (2014), pp. 461–466.
- [28] Adam Lerner et al. "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016." In: *USENIX Security Symposium*. Vol. 16. 2016.
- [29] Srdjan Matic et al. "Identifying sensitive urls at web-scale". In: *Proceedings of the ACM Internet Measurement Conference*. 2020, pp. 619–633.
- [30] Georg Merzdovnik et al. "Block me if you can: A large-scale study of tracker-blocking tools". In: *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2017, pp. 319–333.
- [31] Microsoft. *.NET Framework*. <https://dotnet.microsoft.com/en-us/download/dotnet-framework>.
- [32] Microsoft. *ASP.NET Core*. <https://learn.microsoft.com/en-us/aspnet/core/introduction-to-aspnet-core?view=aspnetcore-7.0>.
- [33] Microsoft. *MSDN Platforms*. <https://visualstudio.microsoft.com/de/msdn-platforms/>.
- [34] Microsoft. *Powershell*. <https://learn.microsoft.com/en-us/powershell/>.
- [35] Microsoft. *Selenium*. <https://www.iis.net/>.
- [36] Microsoft. *SQL Server*. <https://www.microsoft.com/en-us/sql-server/sql-server-2019>.
- [37] Jakub Mikians et al. "Detecting price and search discrimination on the internet". In: *Proceedings of the 11th ACM workshop on hot topics in networks*. 2012, pp. 79–84.
- [38] United Nations. *Universal Declaration of Human Rights*. Accessed: 11-12-2022.
- [39] European Parliament. *DIRECTIVE 2002/58/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2002*. Accessed: 11-12-2022.
- [40] Abbas Razaghpanah et al. "Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem". In: *The 25th Annual Network and Distributed System Security Symposium (NDSS 2018)*. 2018.

- [41] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. “Detecting and Defending Against {Third-Party} Tracking on the Web”. In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. 2012, pp. 155–168.
- [42] ThoughtWorks. *Selenium*. <https://www.selenium.dev/>.
- [43] Thomas Vissers et al. “Crying wolf? on the price discrimination of online airline tickets”. In: *7th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2014)*. 2014.



Source code

Adding source code to your report/thesis is supported with the package listings. An example can be found below. Files can be added using `\lstinputlisting[language=<language>]{<filename>}`.

```
1 import numpy as np
2 import seaborn as sns
3 import matplotlib.pyplot as plt
4 import pandas as pd
5
6 from imblearn.over_sampling import SMOTE, RandomOverSampler
7 from imblearn.combine import SMOTETomek
8 from imblearn.under_sampling import RandomUnderSampler
9
10 from sklearn.model_selection import train_test_split
11 from sklearn.linear_model import LogisticRegression
12 from sklearn.metrics import confusion_matrix, precision_recall_curve, auc, roc_auc_score,
    roc_curve, recall_score, classification_report
13 from sklearn.neighbors import KNeighborsClassifier
14 from sklearn.ensemble import RandomForestClassifier
15 from sklearn.model_selection import KFold
16 from sklearn.metrics import precision_recall_curve
17 from sklearn import decomposition, tree
18 from sklearn.preprocessing import LabelEncoder
19
20 #Load dataset
21 data_set = pd.read_csv("datacsv3-d.csv", sep=';')
22
23 data_set['label'] = np.where(data_set['FirstParty']==1, 1, 0)
24
25 category = LabelEncoder()
26 category.fit(data_set['Category'])
27 data_set['Category'] = category.transform(data_set.Category)
28
29 country = LabelEncoder()
30 country.fit(data_set['Country'])
31 data_set['Country'] = country.transform(data_set.Country)
32
```

```

33 df = pd.DataFrame(data_set, columns=['WebsiteID', 'Category', 'Country', 'UserAgentID', '
    FirstParty', 'ThirdParty', 'Duration'])
34
35 #Select specific features from the entire dataset
36 #selected_features = ['WebsiteID', 'Category', 'Country', 'UserAgentID', 'FirstParty', '
    ThirdParty', 'Duration', 'label']
37 selected_features = ['Category', 'Country', 'UserAgentID', 'WebsiteID', 'Duration']
38 new_data=data_set[selected_features]
39
40 print(new_data.head(5))
41
42 #Create dummies dataset
43 new_data=pd.get_dummies(new_data)
44 #Split dataset into train and test
45 X_train, X_test, y_train, y_test = train_test_split(new_data, data_set['label'],test_size
    =0.2,random_state=42, stratify=data_set['label'])
46
47 # Replace all nan values with 0
48 X_train=X_train.fillna(0)
49 X_test=X_test.fillna(0)
50
51 #List of all classifiers
52 classifiers = {}
53 classifiers['Logistic Classifier']=LogisticRegression()
54 classifiers['Decision Tree Classifier'] = tree.DecisionTreeClassifier()
55 classifiers['Random Forest Classifier'] = RandomForestClassifier(n_estimators=50, criterion='
    gini')
56 classifiers[ 'Knn Classifier'] = KNeighborsClassifier(n_neighbors=5)
57
58 for classifier in classifiers:
59
60     #Fit Unsmoted data to classifier
61     classifiers[classifier].fit(X_train, y_train)
62     unsmoted_probs = classifiers[classifier].predict_proba(X_test)[: , 1]
63     unsmoted_predicts= classifiers[classifier].predict(X_test)
64     False_Positive_Rate_unsmoted, True_Positive_Rate_unsmoted, threshold_unsmoted =
        roc_curve(y_test, unsmoted_probs)
65     auc_unsmoted = roc_auc_score(y_test, unsmoted_probs)
66     recall_unsmoted_score=recall_score(y_test, classifiers[classifier].predict(X_test))
67
68     # Fit Smoted Data to classifier
69     smt=SMOTE(random_state=42, sampling_strategy=0.85)
70     new_X_train, new_y_train=smt.fit_resample(X_train,y_train)
71     classifiers[classifier].fit(new_X_train, new_y_train)
72     smoted_probs = classifiers[classifier].predict_proba(X_test)[: , 1]
73     smoted_predicts=classifiers[classifier].predict(X_test)
74     False_Positive_Rate_Smoted, True_Positive_Rate_Smoted, threshold_Smoted = roc_curve(
        y_test, smoted_probs)
75     auc_Smoted = roc_auc_score(y_test, smoted_probs)
76     recall_smoted_score=recall_score(y_test, classifiers[classifier].predict(X_test))
77
78     #Fit undersampling data to classifier
79     und=RandomUnderSampler(random_state=42, sampling_strategy=0.85)
80     undersampled_X_train ,undersampled_y_train=und.fit_resample(X_train,y_train)
81     classifiers[classifier].fit(undersampled_X_train, undersampled_y_train)
82     undersampled_probs = classifiers[classifier].predict_proba(X_test)[: , 1]
83     undersampled_predicts=classifiers[classifier].predict(X_test)

```

```

84     False_Positive_Rate_undersampled, True_Positives_Rate_undersampled,
      thresholds_undersampled = roc_curve(y_test, undersampled_probs)
85     auc_undersampled = roc_auc_score(y_test, undersampled_probs)
86     recall_undersampled_score=recall_score(y_test, classifiers[classifier].predict(X_test
      ))
87
88     #Fit oversampling data to classifier
89     over=RandomOverSampler(random_state=42, sampling_strategy=0.85)
90     oversampled_X_train ,oversampled_y_train=over.fit_resample(X_train,y_train)
91     classifiers[classifier].fit(oversampled_X_train, oversampled_y_train)
92     oversampled_probs = classifiers[classifier].predict_proba(X_test)[:, 1]
93     oversampled_predicts=classifiers[classifier].predict(X_test)
94     False_Positive_Rate_oversampled, True_Positives_Rate_oversampled,
      thresholds_oversampled = roc_curve(y_test, oversampled_probs)
95     auc_oversampled = roc_auc_score(y_test, oversampled_probs)
96     recall_oversampled_score=recall_score(y_test, classifiers[classifier].predict(X_test
      ))
97
98     #ROC CURVE
99     plt.title('%s AUC' % classifier)
100    plt.plot([0, 1], [0, 1], 'k--')
101    plt.plot(False_Positive_Rate_unsmoted, True_Positive_Rate_unsmoted, color='blue',
      label='AUC Unsmoted = %0.2f' % auc_unsmoted)
102    plt.plot(False_Positive_Rate_Smoted, True_Positive_Rate_Smoted, color='green', label=
      'AUC Smoted = %0.2f' % auc_Smoted)
103    plt.plot(False_Positive_Rate_undersampled, True_Positives_Rate_undersampled, color='
      red',label='AUC UnderSampled = %0.2f' % auc_undersampled)
104    plt.plot(False_Positive_Rate_oversampled, True_Positives_Rate_oversampled, color='m',
      label='AUC OverSampled = %0.2f' % auc_oversampled)
105    plt.xlabel('False Positive Rate')
106    plt.ylabel('True Positive Rate')
107    plt.legend(loc="lower right")
108    plt.savefig('%s ROC' % classifier)
109    plt.show()
110
111    #PRECISION-RECALL Curve
112    precision_unsmoted, recall_unsmoted ,_ = precision_recall_curve(y_test, unsmoted_probs
      )
113    precision_smoted, recall_smoted,_ = precision_recall_curve(y_test, smoted_probs)
114    plt.plot(recall_unsmoted, precision_unsmoted, color="blue",label='Precision-Recall
      Unsmoted')
115    plt.plot(recall_smoted, precision_smoted, color="green",label='Precision-Recall
      Smoted ')
116    plt.xlabel('Recall')
117    plt.ylabel('Precision')
118    plt.title('Precision Recall Curve - %s'%classifier)
119    plt.legend(loc="upper right")
120    plt.savefig('Precision Recall Curve - %s'%classifier)
121    plt.show()
122
123    print("FINISH")

```


B

Figures

Here, all additional figures can be found.

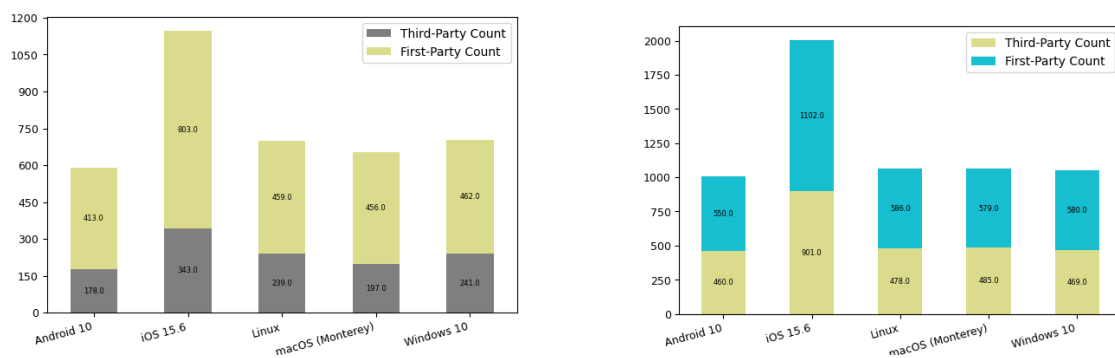


Figure B.1: Distribution of first-party and third-party cookies per operating system (Left: The Hague, Right: Dallas)

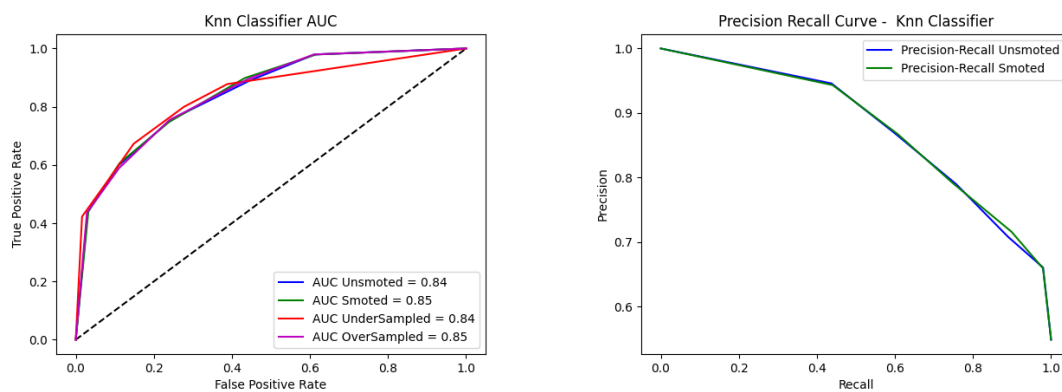


Figure B.4: Left: AUC Curve K-NN classifier. Right: Precision recall curve K-NN classifier.

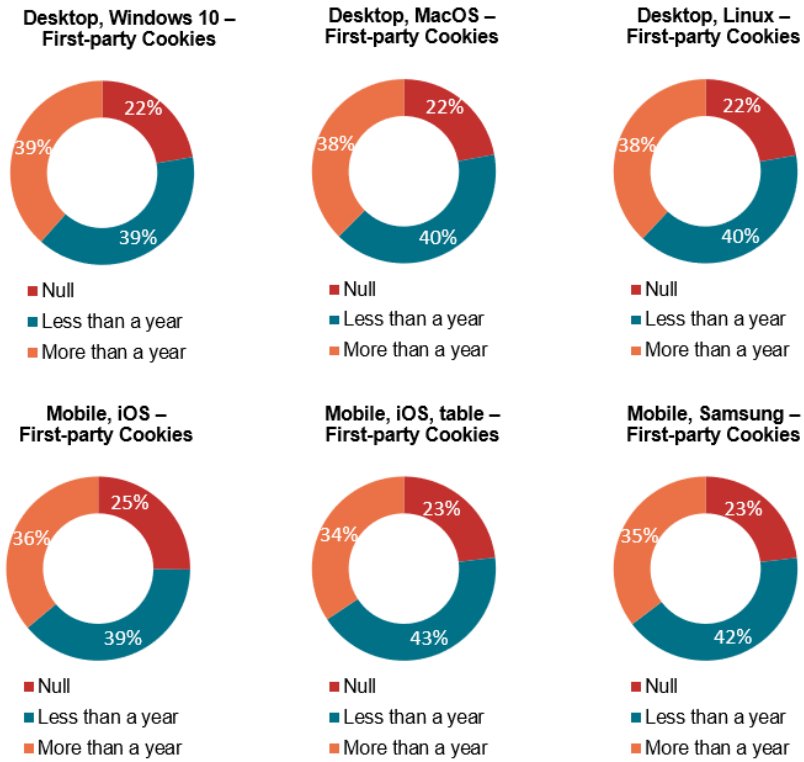


Figure B.2: Distribution of first-party cookies expiry date per user agent

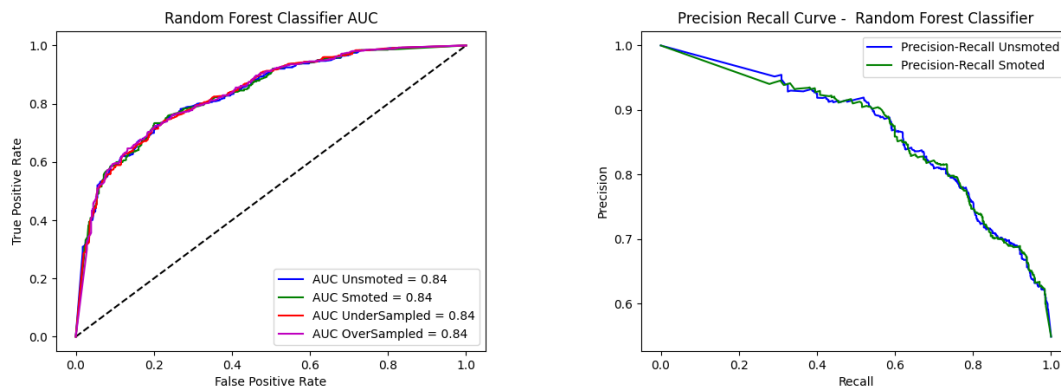


Figure B.5: Left: AUC Curve Random Forest classifier. Right: Precision recall curve Random Forest classifier.

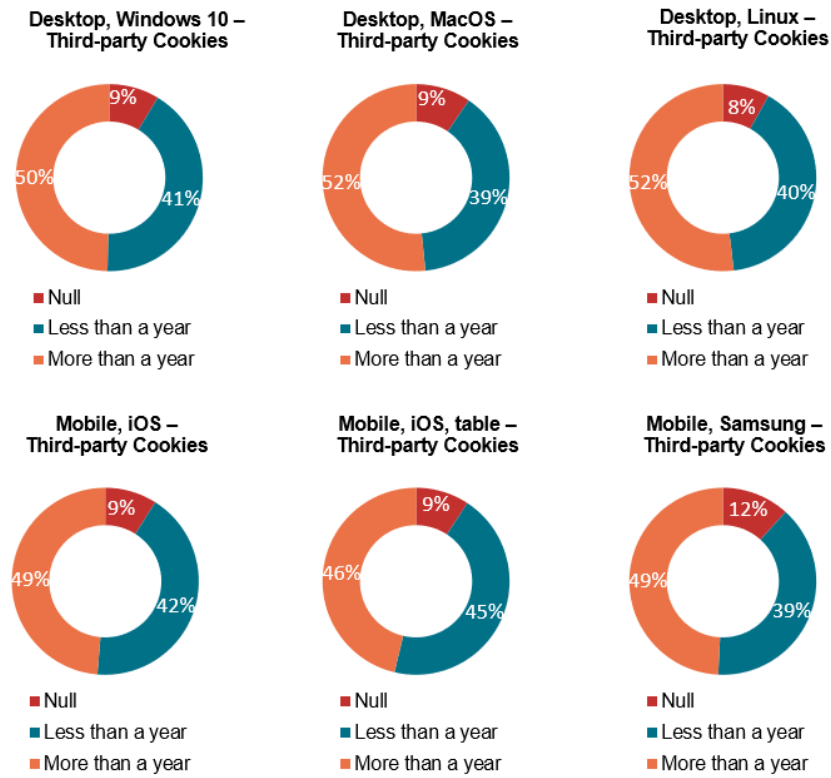


Figure B.3: Distribution of third-party cookies expiry date per user agent

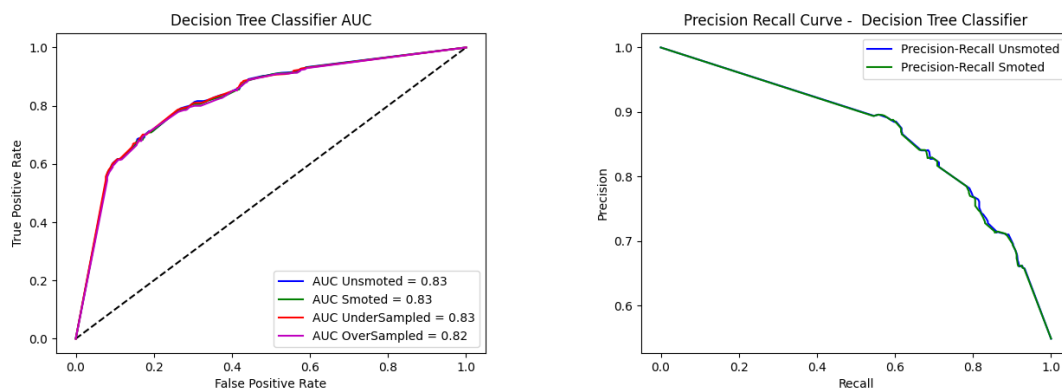


Figure B.6: Left:AUC curve Decision Tree classifier. Right: Precision recall curve Decision Tree classifier

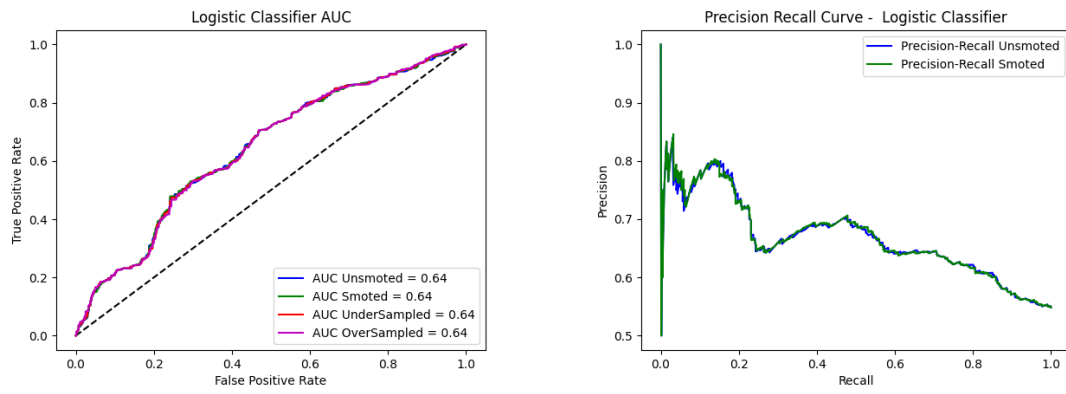
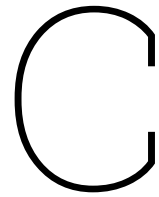


Figure B.7: Left:AUC curve Logistic classifier. Right: Precision recall curve Logistic classifier



Database

Name	OperatingSystem	Platform	TP	FP	Location
https://login.microsoftonline.com	Android 10	Samsung Mobile	9	7	Dallas
https://login.microsoftonline.com	Android 10	Samsung Mobile	7	7	The Hague
https://login.microsoftonline.com	iOS 15.6	Apple iPad	11	8	Dallas
https://login.microsoftonline.com	iOS 15.6	Apple iPad	7	7	The Hague
https://login.microsoftonline.com	iOS 15.6	Apple iPhone	9	7	Dallas
https://login.microsoftonline.com	iOS 15.6	Apple iPhone	7	7	The Hague
https://login.microsoftonline.com	Linux	Desktop	9	7	Dallas
https://login.microsoftonline.com	Linux	Desktop	7	7	The Hague
https://login.microsoftonline.com	macOS (Monterey)	Desktop	9	7	Dallas
https://login.microsoftonline.com	macOS (Monterey)	Desktop	9	8	The Hague
https://login.microsoftonline.com	Windows 10	Desktop	9	9	Dallas
https://login.microsoftonline.com	Windows 10	Desktop	7	9	The Hague
https://sina.com.cn	Android 10	Samsung Mobile	6	9	Dallas
https://sina.com.cn	Android 10	Samsung Mobile	4	10	The Hague
https://sina.com.cn	iOS 15.6	Apple iPad	9	13	Dallas
https://sina.com.cn	iOS 15.6	Apple iPad	7	14	The Hague
https://sina.com.cn	iOS 15.6	Apple iPhone	6	9	Dallas
https://sina.com.cn	iOS 15.6	Apple iPhone	4	10	The Hague
https://sina.com.cn	Linux	Desktop	6	14	The Hague
https://sina.com.cn	macOS (Monterey)	Desktop	6	14	The Hague
https://sina.com.cn	Windows 10	Desktop	6	14	The Hague
https://www.163.com	Android 10	Samsung Mobile	8	7	Dallas
https://www.163.com	Android 10	Samsung Mobile	0	2	The Hague
https://www.163.com	iOS 15.6	Apple iPad	2	3	Dallas
https://www.163.com	iOS 15.6	Apple iPad	0	3	The Hague
https://www.163.com	iOS 15.6	Apple iPhone	2	3	Dallas

https://www.163.com	iOS 15.6	Apple iPhone	6	7	The Hague
https://www.163.com	Linux	Desktop	1	5	Dallas
https://www.163.com	Linux	Desktop	2	7	The Hague
https://www.163.com	macOS (Monterey)	Desktop	1	5	Dallas
https://www.163.com	macOS (Monterey)	Desktop	2	7	The Hague
https://www.163.com	Windows 10	Desktop	1	5	Dallas
https://www.163.com	Windows 10	Desktop	2	7	The Hague
https://www.adobe.com	Android 10	Samsung Mobile	45	35	Dallas
https://www.adobe.com	Android 10	Samsung Mobile	1	21	The Hague
https://www.adobe.com	iOS 15.6	Apple iPad	47	39	Dallas
https://www.adobe.com	iOS 15.6	Apple iPad	1	21	The Hague
https://www.adobe.com	iOS 15.6	Apple iPhone	47	39	Dallas
https://www.adobe.com	iOS 15.6	Apple iPhone	1	21	The Hague
https://www.adobe.com	Linux	Desktop	35	39	Dallas
https://www.adobe.com	Linux	Desktop	1	21	The Hague
https://www.adobe.com	macOS (Monterey)	Desktop	48	37	Dallas
https://www.adobe.com	macOS (Monterey)	Desktop	1	21	The Hague
https://www.adobe.com	Windows 10	Desktop	30	35	Dallas
https://www.adobe.com	Windows 10	Desktop	1	22	The Hague
https://www.aliexpress.com	Android 10	Samsung Mobile	22	28	Dallas
https://www.aliexpress.com	Android 10	Samsung Mobile	24	28	The Hague
https://www.aliexpress.com	iOS 15.6	Apple iPad	12	27	Dallas
https://www.aliexpress.com	iOS 15.6	Apple iPad	17	27	The Hague
https://www.aliexpress.com	iOS 15.6	Apple iPhone	10	26	Dallas
https://www.aliexpress.com	iOS 15.6	Apple iPhone	17	28	The Hague
https://www.aliexpress.com	Linux	Desktop	71	42	Dallas
https://www.aliexpress.com	Linux	Desktop	30	42	The Hague
https://www.aliexpress.com	macOS (Monterey)	Desktop	81	42	Dallas
https://www.aliexpress.com	macOS (Monterey)	Desktop	30	42	The Hague
https://www.aliexpress.com	Windows 10	Desktop	68	42	Dallas
https://www.aliexpress.com	Windows 10	Desktop	27	41	The Hague
https://www.amazon.com	Android 10	Samsung Mobile	43	9	Dallas
https://www.amazon.com	Android 10	Samsung Mobile	41	11	The Hague
https://www.amazon.com	iOS 15.6	Apple iPad	56	10	Dallas
https://www.amazon.com	iOS 15.6	Apple iPad	41	11	The Hague
https://www.amazon.com	iOS 15.6	Apple iPhone	49	10	Dallas
https://www.amazon.com	iOS 15.6	Apple iPhone	41	11	The Hague
https://www.amazon.com	Linux	Desktop	0	1	Dallas
https://www.amazon.com	Linux	Desktop	39	11	The Hague
https://www.amazon.com	macOS (Monterey)	Desktop	0	1	Dallas
https://www.amazon.com	macOS (Monterey)	Desktop	0	6	The Hague
https://www.amazon.com	Windows 10	Desktop	0	1	Dallas
https://www.amazon.com	Windows 10	Desktop	41	11	The Hague

https://www.amazon.in	Android 10	Samsung Mobile	39	8	Dallas
https://www.amazon.in	Android 10	Samsung Mobile	37	8	The Hague
https://www.amazon.in	iOS 15.6	Apple iPad	38	8	Dallas
https://www.amazon.in	iOS 15.6	Apple iPad	39	8	The Hague
https://www.amazon.in	iOS 15.6	Apple iPhone	36	8	Dallas
https://www.amazon.in	iOS 15.6	Apple iPhone	41	8	The Hague
https://www.amazon.in	Linux	Desktop	37	8	Dallas
https://www.amazon.in	Linux	Desktop	35	8	The Hague
https://www.amazon.in	macOS (Monterey)	Desktop	39	8	Dallas
https://www.amazon.in	macOS (Monterey)	Desktop	37	8	The Hague
https://www.amazon.in	Windows 10	Desktop	39	8	Dallas
https://www.amazon.in	Windows 10	Desktop	37	8	The Hague
https://www.apple.com	Android 10	Samsung Mobile	2	12	Dallas
https://www.apple.com	Android 10	Samsung Mobile	0	12	The Hague
https://www.apple.com	iOS 15.6	Apple iPad	2	12	Dallas
https://www.apple.com	iOS 15.6	Apple iPad	0	12	The Hague
https://www.apple.com	iOS 15.6	Apple iPhone	2	12	Dallas
https://www.apple.com	iOS 15.6	Apple iPhone	0	12	The Hague
https://www.apple.com	Linux	Desktop	2	12	Dallas
https://www.apple.com	Linux	Desktop	0	12	The Hague
https://www.apple.com	macOS (Monterey)	Desktop	2	12	Dallas
https://www.apple.com	macOS (Monterey)	Desktop	0	12	The Hague
https://www.apple.com	Windows 10	Desktop	2	12	Dallas
https://www.apple.com	Windows 10	Desktop	0	12	The Hague
https://www.baidu.com	Android 10	Samsung Mobile	2	22	Dallas
https://www.baidu.com	Android 10	Samsung Mobile	0	22	The Hague
https://www.baidu.com	iOS 15.6	Apple iPad	2	9	Dallas
https://www.baidu.com	iOS 15.6	Apple iPad	0	9	The Hague
https://www.baidu.com	iOS 15.6	Apple iPhone	2	18	Dallas
https://www.baidu.com	iOS 15.6	Apple iPhone	0	18	The Hague
https://www.baidu.com	Linux	Desktop	2	9	Dallas
https://www.baidu.com	Linux	Desktop	0	9	The Hague
https://www.baidu.com	macOS (Monterey)	Desktop	2	9	Dallas
https://www.baidu.com	macOS (Monterey)	Desktop	0	9	The Hague
https://www.baidu.com	Windows 10	Desktop	2	8	Dallas
https://www.baidu.com	Windows 10	Desktop	0	9	The Hague
https://www.bilibili.com	Android 10	Samsung Mobile	0	9	Dallas
https://www.bilibili.com	Android 10	Samsung Mobile	0	9	The Hague
https://www.bilibili.com	iOS 15.6	Apple iPad	0	9	Dallas
https://www.bilibili.com	iOS 15.6	Apple iPad	0	9	The Hague
https://www.bilibili.com	iOS 15.6	Apple iPhone	0	9	Dallas
https://www.bilibili.com	iOS 15.6	Apple iPhone	0	9	The Hague
https://www.bilibili.com	Linux	Desktop	0	9	Dallas

https://www.bilibili.com	Linux	Desktop	0	9	The Hague
https://www.bilibili.com	macOS (Monterey)	Desktop	0	9	Dallas
https://www.bilibili.com	macOS (Monterey)	Desktop	0	9	The Hague
https://www.bilibili.com	Windows 10	Desktop	0	9	Dallas
https://www.bilibili.com	Windows 10	Desktop	0	9	The Hague
https://www.bing.com	Android 10	Samsung Mobile	0	13	Dallas
https://www.bing.com	Android 10	Samsung Mobile	0	12	The Hague
https://www.bing.com	iOS 15.6	Apple iPad	7	15	Dallas
https://www.bing.com	iOS 15.6	Apple iPad	7	15	The Hague
https://www.bing.com	iOS 15.6	Apple iPhone	0	13	Dallas
https://www.bing.com	iOS 15.6	Apple iPhone	0	12	The Hague
https://www.bing.com	Linux	Desktop	13	20	Dallas
https://www.bing.com	Linux	Desktop	7	15	The Hague
https://www.bing.com	macOS (Monterey)	Desktop	13	20	Dallas
https://www.bing.com	macOS (Monterey)	Desktop	7	15	The Hague
https://www.bing.com	Windows 10	Desktop	15	20	Dallas
https://www.bing.com	Windows 10	Desktop	9	15	The Hague
https://www.canva.com	Android 10	Samsung Mobile	18	21	Dallas
https://www.canva.com	Android 10	Samsung Mobile	0	6	The Hague
https://www.canva.com	iOS 15.6	Apple iPad	16	21	Dallas
https://www.canva.com	iOS 15.6	Apple iPad	0	6	The Hague
https://www.canva.com	iOS 15.6	Apple iPhone	16	21	Dallas
https://www.canva.com	iOS 15.6	Apple iPhone	0	6	The Hague
https://www.canva.com	Linux	Desktop	18	21	Dallas
https://www.canva.com	Linux	Desktop	0	6	The Hague
https://www.canva.com	macOS (Monterey)	Desktop	18	21	Dallas
https://www.canva.com	macOS (Monterey)	Desktop	0	6	The Hague
https://www.canva.com	Windows 10	Desktop	18	21	Dallas
https://www.canva.com	Windows 10	Desktop	0	6	The Hague
https://www.csdn.net	Android 10	Samsung Mobile	2	20	Dallas
https://www.csdn.net	Android 10	Samsung Mobile	2	20	The Hague
https://www.csdn.net	iOS 15.6	Apple iPad	2	20	Dallas
https://www.csdn.net	iOS 15.6	Apple iPad	2	20	The Hague
https://www.csdn.net	iOS 15.6	Apple iPhone	2	20	Dallas
https://www.csdn.net	iOS 15.6	Apple iPhone	2	20	The Hague
https://www.csdn.net	Linux	Desktop	1	21	Dallas
https://www.csdn.net	Linux	Desktop	1	21	The Hague
https://www.csdn.net	macOS (Monterey)	Desktop	1	21	Dallas
https://www.csdn.net	macOS (Monterey)	Desktop	1	21	The Hague
https://www.csdn.net	Windows 10	Desktop	1	21	Dallas
https://www.csdn.net	Windows 10	Desktop	1	21	The Hague
https://www.ebay.com	Android 10	Samsung Mobile	12	22	Dallas
https://www.ebay.com	Android 10	Samsung Mobile	11	21	The Hague

https://www.ebay.com	iOS 15.6	Apple iPad	13	20	Dallas
https://www.ebay.com	iOS 15.6	Apple iPad	11	21	The Hague
https://www.ebay.com	iOS 15.6	Apple iPhone	10	22	Dallas
https://www.ebay.com	iOS 15.6	Apple iPhone	8	8	The Hague
https://www.ebay.com	Linux	Desktop	15	22	Dallas
https://www.ebay.com	Linux	Desktop	12	20	The Hague
https://www.ebay.com	macOS (Monterey)	Desktop	16	22	Dallas
https://www.ebay.com	macOS (Monterey)	Desktop	12	21	The Hague
https://www.ebay.com	Windows 10	Desktop	15	22	Dallas
https://www.ebay.com	Windows 10	Desktop	12	21	The Hague
https://www.facebook.com	Android 10	Samsung Mobile	0	5	Dallas
https://www.facebook.com	iOS 15.6	Apple iPad	0	5	Dallas
https://www.facebook.com	iOS 15.6	Apple iPhone	0	5	Dallas
https://www.facebook.com	Linux	Desktop	0	5	Dallas
https://www.facebook.com	macOS (Monterey)	Desktop	0	5	Dallas
https://www.facebook.com	Windows 10	Desktop	0	5	Dallas
https://www.fandom.com	Android 10	Samsung Mobile	29	35	Dallas
https://www.fandom.com	Android 10	Samsung Mobile	1	1	The Hague
https://www.fandom.com	iOS 15.6	Apple iPad	23	36	Dallas
https://www.fandom.com	iOS 15.6	Apple iPad	1	1	The Hague
https://www.fandom.com	iOS 15.6	Apple iPhone	28	36	Dallas
https://www.fandom.com	iOS 15.6	Apple iPhone	1	1	The Hague
https://www.fandom.com	Linux	Desktop	21	36	Dallas
https://www.fandom.com	Linux	Desktop	1	1	The Hague
https://www.fandom.com	macOS (Monterey)	Desktop	11	32	Dallas
https://www.fandom.com	macOS (Monterey)	Desktop	1	1	The Hague
https://www.fandom.com	Windows 10	Desktop	21	37	Dallas
https://www.fandom.com	Windows 10	Desktop	1	1	The Hague
https://www.github.com	Android 10	Samsung Mobile	0	4	Dallas
https://www.github.com	Android 10	Samsung Mobile	0	4	The Hague
https://www.github.com	iOS 15.6	Apple iPad	0	4	Dallas
https://www.github.com	iOS 15.6	Apple iPad	0	4	The Hague
https://www.github.com	iOS 15.6	Apple iPhone	0	4	Dallas
https://www.github.com	iOS 15.6	Apple iPhone	0	4	The Hague
https://www.github.com	Linux	Desktop	0	4	Dallas
https://www.github.com	Linux	Desktop	0	4	The Hague
https://www.github.com	macOS (Monterey)	Desktop	0	4	Dallas
https://www.github.com	macOS (Monterey)	Desktop	0	4	The Hague
https://www.github.com	Windows 10	Desktop	0	4	Dallas
https://www.github.com	Windows 10	Desktop	0	4	The Hague
https://www.google.com	Android 10	Samsung Mobile	0	3	Dallas
https://www.google.com	Android 10	Samsung Mobile	0	3	The Hague
https://www.google.com	iOS 15.6	Apple iPad	0	4	Dallas

https://www.google.com	iOS 15.6	Apple iPad	0	3	The Hague
https://www.google.com	iOS 15.6	Apple iPhone	0	3	Dallas
https://www.google.com	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.google.com	Linux	Desktop	0	5	Dallas
https://www.google.com	Linux	Desktop	0	3	The Hague
https://www.google.com	macOS (Monterey)	Desktop	0	5	Dallas
https://www.google.com	macOS (Monterey)	Desktop	0	3	The Hague
https://www.google.com	Windows 10	Desktop	0	3	Dallas
https://www.google.com	Windows 10	Desktop	0	3	The Hague
https://www.google.com.hk	Android 10	Samsung Mobile	0	5	Dallas
https://www.google.com.hk	Android 10	Samsung Mobile	0	3	The Hague
https://www.google.com.hk	iOS 15.6	Apple iPad	0	6	Dallas
https://www.google.com.hk	iOS 15.6	Apple iPad	0	3	The Hague
https://www.google.com.hk	iOS 15.6	Apple iPhone	0	5	Dallas
https://www.google.com.hk	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.google.com.hk	Linux	Desktop	0	7	Dallas
https://www.google.com.hk	Linux	Desktop	0	3	The Hague
https://www.google.com.hk	macOS (Monterey)	Desktop	0	7	Dallas
https://www.google.com.hk	macOS (Monterey)	Desktop	0	3	The Hague
https://www.google.com.hk	Windows 10	Desktop	0	7	Dallas
https://www.google.com.hk	Windows 10	Desktop	0	3	The Hague
https://www.instagram.com	Android 10	Samsung Mobile	0	4	Dallas
https://www.instagram.com	iOS 15.6	Apple iPad	0	4	Dallas
https://www.instagram.com	iOS 15.6	Apple iPhone	0	4	Dallas
https://www.instagram.com	Linux	Desktop	0	4	Dallas
https://www.instagram.com	macOS (Monterey)	Desktop	0	4	Dallas
https://www.instagram.com	Windows 10	Desktop	0	4	Dallas
https://www.jd.com	Android 10	Samsung Mobile	2	10	Dallas
https://www.jd.com	Android 10	Samsung Mobile	0	17	The Hague
https://www.jd.com	iOS 15.6	Apple iPad	2	10	Dallas
https://www.jd.com	iOS 15.6	Apple iPad	1	13	The Hague
https://www.jd.com	iOS 15.6	Apple iPhone	2	10	Dallas
https://www.jd.com	iOS 15.6	Apple iPhone	0	17	The Hague
https://www.jd.com	Linux	Desktop	2	12	Dallas
https://www.jd.com	Linux	Desktop	1	12	The Hague
https://www.jd.com	macOS (Monterey)	Desktop	2	12	Dallas
https://www.jd.com	macOS (Monterey)	Desktop	1	12	The Hague
https://www.jd.com	Windows 10	Desktop	2	12	Dallas
https://www.jd.com	Windows 10	Desktop	1	12	The Hague
https://www.linkedin.com	Android 10	Samsung Mobile	8	9	Dallas
https://www.linkedin.com	Android 10	Samsung Mobile	1	8	The Hague
https://www.linkedin.com	iOS 15.6	Apple iPad	7	9	Dallas
https://www.linkedin.com	iOS 15.6	Apple iPad	1	7	The Hague

https://www.linkedin.com	iOS 15.6	Apple iPhone	7	9	Dallas
https://www.linkedin.com	iOS 15.6	Apple iPhone	1	7	The Hague
https://www.linkedin.com	Linux	Desktop	8	9	Dallas
https://www.linkedin.com	Linux	Desktop	1	8	The Hague
https://www.linkedin.com	macOS (Monterey)	Desktop	8	9	Dallas
https://www.linkedin.com	macOS (Monterey)	Desktop	1	8	The Hague
https://www.linkedin.com	Windows 10	Desktop	8	9	Dallas
https://www.linkedin.com	Windows 10	Desktop	1	8	The Hague
https://www.live.com	Android 10	Samsung Mobile	2	4	Dallas
https://www.live.com	Android 10	Samsung Mobile	2	4	The Hague
https://www.live.com	iOS 15.6	Apple iPad	2	4	Dallas
https://www.live.com	iOS 15.6	Apple iPad	2	4	The Hague
https://www.live.com	iOS 15.6	Apple iPhone	2	4	Dallas
https://www.live.com	iOS 15.6	Apple iPhone	2	4	The Hague
https://www.live.com	Linux	Desktop	2	4	Dallas
https://www.live.com	Linux	Desktop	4	4	The Hague
https://www.live.com	macOS (Monterey)	Desktop	2	4	Dallas
https://www.live.com	macOS (Monterey)	Desktop	2	4	The Hague
https://www.live.com	Windows 10	Desktop	2	4	Dallas
https://www.live.com	Windows 10	Desktop	2	4	The Hague
https://www.mail.ru	Android 10	Samsung Mobile	7	12	Dallas
https://www.mail.ru	Android 10	Samsung Mobile	11	11	The Hague
https://www.mail.ru	iOS 15.6	Apple iPad	15	12	Dallas
https://www.mail.ru	iOS 15.6	Apple iPad	8	12	The Hague
https://www.mail.ru	iOS 15.6	Apple iPhone	5	12	Dallas
https://www.mail.ru	iOS 15.6	Apple iPhone	9	10	The Hague
https://www.mail.ru	Linux	Desktop	13	22	Dallas
https://www.mail.ru	Linux	Desktop	22	24	The Hague
https://www.mail.ru	macOS (Monterey)	Desktop	13	25	Dallas
https://www.mail.ru	macOS (Monterey)	Desktop	15	24	The Hague
https://www.mail.ru	Windows 10	Desktop	13	22	Dallas
https://www.mail.ru	Windows 10	Desktop	21	24	The Hague
https://www.msn.com	Android 10	Samsung Mobile	110	14	Dallas
https://www.msn.com	Android 10	Samsung Mobile	0	3	The Hague
https://www.msn.com	iOS 15.6	Apple iPad	84	14	Dallas
https://www.msn.com	iOS 15.6	Apple iPad	0	3	The Hague
https://www.msn.com	iOS 15.6	Apple iPhone	81	14	Dallas
https://www.msn.com	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.msn.com	Linux	Desktop	22	12	Dallas
https://www.msn.com	Linux	Desktop	2	11	The Hague
https://www.msn.com	macOS (Monterey)	Desktop	35	12	Dallas
https://www.msn.com	macOS (Monterey)	Desktop	2	11	The Hague
https://www.msn.com	Windows 10	Desktop	32	12	Dallas

https://www.msn.com	Windows 10	Desktop	2	11	The Hague
https://www.myshopify.com	Android 10	Samsung Mobile	2	0	Dallas
https://www.myshopify.com	iOS 15.6	Apple iPad	2	0	Dallas
https://www.myshopify.com	iOS 15.6	Apple iPhone	2	0	Dallas
https://www.myshopify.com	Linux	Desktop	2	0	Dallas
https://www.myshopify.com	macOS (Monterey)	Desktop	2	0	Dallas
https://www.myshopify.com	Windows 10	Desktop	2	0	Dallas
https://www.naver.com	Android 10	Samsung Mobile	0	3	Dallas
https://www.naver.com	Android 10	Samsung Mobile	0	3	The Hague
https://www.naver.com	iOS 15.6	Apple iPad	0	3	Dallas
https://www.naver.com	iOS 15.6	Apple iPad	0	3	The Hague
https://www.naver.com	iOS 15.6	Apple iPhone	0	3	Dallas
https://www.naver.com	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.naver.com	Linux	Desktop	0	2	Dallas
https://www.naver.com	Linux	Desktop	0	2	The Hague
https://www.naver.com	macOS (Monterey)	Desktop	0	2	Dallas
https://www.naver.com	macOS (Monterey)	Desktop	0	2	The Hague
https://www.naver.com	Windows 10	Desktop	0	2	Dallas
https://www.naver.com	Windows 10	Desktop	0	2	The Hague
https://www.netflix.com	Android 10	Samsung Mobile	0	6	Dallas
https://www.netflix.com	Android 10	Samsung Mobile	0	6	The Hague
https://www.netflix.com	iOS 15.6	Apple iPad	0	6	Dallas
https://www.netflix.com	iOS 15.6	Apple iPad	0	6	The Hague
https://www.netflix.com	iOS 15.6	Apple iPhone	0	6	Dallas
https://www.netflix.com	iOS 15.6	Apple iPhone	0	6	The Hague
https://www.netflix.com	Linux	Desktop	0	6	Dallas
https://www.netflix.com	Linux	Desktop	0	6	The Hague
https://www.netflix.com	macOS (Monterey)	Desktop	0	6	Dallas
https://www.netflix.com	macOS (Monterey)	Desktop	0	6	The Hague
https://www.netflix.com	Windows 10	Desktop	0	6	Dallas
https://www.netflix.com	Windows 10	Desktop	0	6	The Hague
https://www.office.com/	Android 10	Samsung Mobile	17	7	Dallas
https://www.office.com/	Android 10	Samsung Mobile	8	5	The Hague
https://www.office.com/	iOS 15.6	Apple iPad	17	7	Dallas
https://www.office.com/	iOS 15.6	Apple iPad	8	5	The Hague
https://www.office.com/	iOS 15.6	Apple iPhone	17	7	Dallas
https://www.office.com/	iOS 15.6	Apple iPhone	8	5	The Hague
https://www.office.com/	Linux	Desktop	17	7	Dallas
https://www.office.com/	Linux	Desktop	8	5	The Hague
https://www.office.com/	macOS (Monterey)	Desktop	17	7	Dallas
https://www.office.com/	macOS (Monterey)	Desktop	8	5	The Hague
https://www.office.com/	Windows 10	Desktop	19	7	Dallas
https://www.office.com/	Windows 10	Desktop	8	5	The Hague

https://www.paypal.com	Android 10	Samsung Mobile	3	9	Dallas
https://www.paypal.com	Android 10	Samsung Mobile	1	9	The Hague
https://www.paypal.com	iOS 15.6	Apple iPad	3	9	Dallas
https://www.paypal.com	iOS 15.6	Apple iPad	1	9	The Hague
https://www.paypal.com	iOS 15.6	Apple iPhone	3	9	Dallas
https://www.paypal.com	iOS 15.6	Apple iPhone	1	9	The Hague
https://www.paypal.com	Linux	Desktop	3	9	Dallas
https://www.paypal.com	Linux	Desktop	1	9	The Hague
https://www.paypal.com	macOS (Monterey)	Desktop	3	9	Dallas
https://www.paypal.com	macOS (Monterey)	Desktop	1	9	The Hague
https://www.paypal.com	Windows 10	Desktop	3	9	Dallas
https://www.paypal.com	Windows 10	Desktop	1	9	The Hague
https://www.pornhub.com	Android 10	Samsung Mobile	0	24	Dallas
https://www.pornhub.com	Android 10	Samsung Mobile	1	25	The Hague
https://www.pornhub.com	iOS 15.6	Apple iPad	3	24	Dallas
https://www.pornhub.com	iOS 15.6	Apple iPad	0	24	The Hague
https://www.pornhub.com	iOS 15.6	Apple iPhone	0	23	Dallas
https://www.pornhub.com	iOS 15.6	Apple iPhone	0	26	The Hague
https://www.pornhub.com	Linux	Desktop	2	24	Dallas
https://www.pornhub.com	Linux	Desktop	0	25	The Hague
https://www.pornhub.com	macOS (Monterey)	Desktop	2	24	Dallas
https://www.pornhub.com	macOS (Monterey)	Desktop	2	25	The Hague
https://www.pornhub.com	Windows 10	Desktop	6	24	Dallas
https://www.pornhub.com	Windows 10	Desktop	2	25	The Hague
https://www.qq.com	Android 10	Samsung Mobile	0	4	Dallas
https://www.qq.com	Android 10	Samsung Mobile	0	4	The Hague
https://www.qq.com	iOS 15.6	Apple iPad	0	10	Dallas
https://www.qq.com	iOS 15.6	Apple iPad	0	10	The Hague
https://www.qq.com	iOS 15.6	Apple iPhone	0	4	Dallas
https://www.qq.com	iOS 15.6	Apple iPhone	0	4	The Hague
https://www.qq.com	Linux	Desktop	0	10	Dallas
https://www.qq.com	Linux	Desktop	0	10	The Hague
https://www.qq.com	macOS (Monterey)	Desktop	0	10	Dallas
https://www.qq.com	macOS (Monterey)	Desktop	0	10	The Hague
https://www.qq.com	Windows 10	Desktop	0	10	Dallas
https://www.qq.com	Windows 10	Desktop	0	10	The Hague
https://www.reddit.com	Android 10	Samsung Mobile	3	6	Dallas
https://www.reddit.com	Android 10	Samsung Mobile	1	5	The Hague
https://www.reddit.com	iOS 15.6	Apple iPad	2	5	Dallas
https://www.reddit.com	iOS 15.6	Apple iPad	0	5	The Hague
https://www.reddit.com	iOS 15.6	Apple iPhone	2	5	Dallas
https://www.reddit.com	iOS 15.6	Apple iPhone	1	6	The Hague
https://www.reddit.com	Linux	Desktop	3	8	Dallas

https://www.reddit.com	Linux	Desktop	1	7	The Hague
https://www.reddit.com	macOS (Monterey)	Desktop	3	7	Dallas
https://www.reddit.com	macOS (Monterey)	Desktop	1	7	The Hague
https://www.reddit.com	Windows 10	Desktop	3	8	Dallas
https://www.reddit.com	Windows 10	Desktop	1	7	The Hague
https://www.stackoverflow.com	Android 10	Samsung Mobile	2	1	Dallas
https://www.stackoverflow.com	Android 10	Samsung Mobile	0	1	The Hague
https://www.stackoverflow.com	iOS 15.6	Apple iPad	2	1	Dallas
https://www.stackoverflow.com	iOS 15.6	Apple iPad	0	1	The Hague
https://www.stackoverflow.com	iOS 15.6	Apple iPhone	2	1	Dallas
https://www.stackoverflow.com	iOS 15.6	Apple iPhone	0	1	The Hague
https://www.stackoverflow.com	Linux	Desktop	2	1	Dallas
https://www.stackoverflow.com	Linux	Desktop	0	1	The Hague
https://www.stackoverflow.com	macOS (Monterey)	Desktop	2	1	Dallas
https://www.stackoverflow.com	macOS (Monterey)	Desktop	0	1	The Hague
https://www.stackoverflow.com	Windows 10	Desktop	2	1	Dallas
https://www.stackoverflow.com	Windows 10	Desktop	0	1	The Hague
https://www.t.co	Android 10	Samsung Mobile	2	0	Dallas
https://www.t.co	iOS 15.6	Apple iPad	2	0	Dallas
https://www.t.co	iOS 15.6	Apple iPhone	2	0	Dallas
https://www.t.co	iOS 15.6	Apple iPhone	1	0	The Hague
https://www.t.co	Linux	Desktop	2	0	Dallas
https://www.t.co	macOS (Monterey)	Desktop	2	0	Dallas
https://www.t.co	Windows 10	Desktop	2	0	Dallas
https://www.taobao.com	Android 10	Samsung Mobile	6	13	Dallas
https://www.taobao.com	Android 10	Samsung Mobile	6	15	The Hague
https://www.taobao.com	iOS 15.6	Apple iPad	2	13	Dallas
https://www.taobao.com	iOS 15.6	Apple iPad	1	13	The Hague
https://www.taobao.com	iOS 15.6	Apple iPhone	0	15	Dallas
https://www.taobao.com	iOS 15.6	Apple iPhone	0	15	The Hague
https://www.taobao.com	Linux	Desktop	5	17	Dallas
https://www.taobao.com	Linux	Desktop	7	17	The Hague
https://www.taobao.com	macOS (Monterey)	Desktop	8	17	Dallas
https://www.taobao.com	macOS (Monterey)	Desktop	7	17	The Hague
https://www.taobao.com	Windows 10	Desktop	8	17	Dallas
https://www.taobao.com	Windows 10	Desktop	7	17	The Hague
https://www.tiktok.com	Android 10	Samsung Mobile	2	11	Dallas
https://www.tiktok.com	Android 10	Samsung Mobile	0	9	The Hague
https://www.tiktok.com	iOS 15.6	Apple iPad	1	11	Dallas
https://www.tiktok.com	iOS 15.6	Apple iPad	0	9	The Hague
https://www.tiktok.com	iOS 15.6	Apple iPhone	1	11	Dallas
https://www.tiktok.com	iOS 15.6	Apple iPhone	0	9	The Hague
https://www.tiktok.com	Linux	Desktop	0	8	Dallas

https://www.tiktok.com	Linux	Desktop	0	9	The Hague
https://www.tiktok.com	macOS (Monterey)	Desktop	0	8	Dallas
https://www.tiktok.com	macOS (Monterey)	Desktop	0	9	The Hague
https://www.tiktok.com	Windows 10	Desktop	0	8	Dallas
https://www.tiktok.com	Windows 10	Desktop	0	9	The Hague
https://www.twitter.com	Android 10	Samsung Mobile	1	8	Dallas
https://www.twitter.com	Android 10	Samsung Mobile	0	3	The Hague
https://www.twitter.com	iOS 15.6	Apple iPad	0	8	Dallas
https://www.twitter.com	iOS 15.6	Apple iPad	0	3	The Hague
https://www.twitter.com	iOS 15.6	Apple iPhone	0	8	Dallas
https://www.twitter.com	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.twitter.com	Linux	Desktop	1	8	Dallas
https://www.twitter.com	Linux	Desktop	0	3	The Hague
https://www.twitter.com	macOS (Monterey)	Desktop	1	8	Dallas
https://www.twitter.com	macOS (Monterey)	Desktop	0	3	The Hague
https://www.twitter.com	Windows 10	Desktop	1	8	Dallas
https://www.twitter.com	Windows 10	Desktop	0	3	The Hague
https://www.vk.com	Android 10	Samsung Mobile	3	19	Dallas
https://www.vk.com	Android 10	Samsung Mobile	3	19	The Hague
https://www.vk.com	iOS 15.6	Apple iPad	3	19	Dallas
https://www.vk.com	iOS 15.6	Apple iPad	3	19	The Hague
https://www.vk.com	iOS 15.6	Apple iPhone	3	19	Dallas
https://www.vk.com	iOS 15.6	Apple iPhone	3	19	The Hague
https://www.vk.com	Linux	Desktop	2	22	Dallas
https://www.vk.com	Linux	Desktop	2	22	The Hague
https://www.vk.com	macOS (Monterey)	Desktop	2	22	Dallas
https://www.vk.com	macOS (Monterey)	Desktop	2	22	The Hague
https://www.vk.com	Windows 10	Desktop	2	22	Dallas
https://www.vk.com	Windows 10	Desktop	2	22	The Hague
https://www.weibo.com	Android 10	Samsung Mobile	2	6	Dallas
https://www.weibo.com	Android 10	Samsung Mobile	0	6	The Hague
https://www.weibo.com	iOS 15.6	Apple iPad	2	12	Dallas
https://www.weibo.com	iOS 15.6	Apple iPad	0	11	The Hague
https://www.weibo.com	iOS 15.6	Apple iPhone	2	6	Dallas
https://www.weibo.com	iOS 15.6	Apple iPhone	0	6	The Hague
https://www.weibo.com	Linux	Desktop	2	12	Dallas
https://www.weibo.com	Linux	Desktop	0	11	The Hague
https://www.weibo.com	macOS (Monterey)	Desktop	2	12	Dallas
https://www.weibo.com	macOS (Monterey)	Desktop	0	11	The Hague
https://www.weibo.com	Windows 10	Desktop	2	12	Dallas
https://www.weibo.com	Windows 10	Desktop	0	11	The Hague
https://www.whatsapp.com	Android 10	Samsung Mobile	0	3	Dallas
https://www.whatsapp.com	iOS 15.6	Apple iPad	0	3	Dallas

https://www.whatsapp.com	iOS 15.6	Apple iPhone	1	3	Dallas
https://www.whatsapp.com	Linux	Desktop	0	3	Dallas
https://www.whatsapp.com	macOS (Monterey)	Desktop	0	3	Dallas
https://www.whatsapp.com	Windows 10	Desktop	0	3	Dallas
https://www.wikipedia.org	Android 10	Samsung Mobile	0	3	Dallas
https://www.wikipedia.org	Android 10	Samsung Mobile	0	3	The Hague
https://www.wikipedia.org	iOS 15.6	Apple iPad	0	3	Dallas
https://www.wikipedia.org	iOS 15.6	Apple iPad	0	3	The Hague
https://www.wikipedia.org	iOS 15.6	Apple iPhone	0	3	Dallas
https://www.wikipedia.org	iOS 15.6	Apple iPhone	0	3	The Hague
https://www.wikipedia.org	Linux	Desktop	0	3	Dallas
https://www.wikipedia.org	Linux	Desktop	0	3	The Hague
https://www.wikipedia.org	macOS (Monterey)	Desktop	0	3	Dallas
https://www.wikipedia.org	macOS (Monterey)	Desktop	0	3	The Hague
https://www.wikipedia.org	Windows 10	Desktop	0	3	Dallas
https://www.wikipedia.org	Windows 10	Desktop	0	3	The Hague
https://www.xhamster.com	Android 10	Samsung Mobile	12	13	Dallas
https://www.xhamster.com	Android 10	Samsung Mobile	3	14	The Hague
https://www.xhamster.com	iOS 15.6	Apple iPad	2	11	Dallas
https://www.xhamster.com	iOS 15.6	Apple iPad	0	11	The Hague
https://www.xhamster.com	iOS 15.6	Apple iPhone	7	13	Dallas
https://www.xhamster.com	iOS 15.6	Apple iPhone	4	13	The Hague
https://www.xhamster.com	Linux	Desktop	2	13	Dallas
https://www.xhamster.com	Linux	Desktop	0	11	The Hague
https://www.xhamster.com	macOS (Monterey)	Desktop	2	11	Dallas
https://www.xhamster.com	macOS (Monterey)	Desktop	0	11	The Hague
https://www.xhamster.com	Windows 10	Desktop	2	13	Dallas
https://www.xhamster.com	Windows 10	Desktop	0	12	The Hague
https://www.xvideos.com	Android 10	Samsung Mobile	0	5	Dallas
https://www.xvideos.com	Android 10	Samsung Mobile	0	4	The Hague
https://www.xvideos.com	iOS 15.6	Apple iPad	0	5	Dallas
https://www.xvideos.com	iOS 15.6	Apple iPad	0	4	The Hague
https://www.xvideos.com	iOS 15.6	Apple iPhone	0	5	Dallas
https://www.xvideos.com	iOS 15.6	Apple iPhone	0	4	The Hague
https://www.xvideos.com	Linux	Desktop	2	5	Dallas
https://www.xvideos.com	Linux	Desktop	0	4	The Hague
https://www.xvideos.com	macOS (Monterey)	Desktop	0	5	Dallas
https://www.xvideos.com	macOS (Monterey)	Desktop	0	4	The Hague
https://www.xvideos.com	Windows 10	Desktop	0	4	Dallas
https://www.xvideos.com	Windows 10	Desktop	0	3	The Hague
https://www.yahoo.co.jp	Android 10	Samsung Mobile	6	18	Dallas
https://www.yahoo.co.jp	iOS 15.6	Apple iPad	6	14	Dallas
https://www.yahoo.co.jp	iOS 15.6	Apple iPhone	5	18	Dallas

https://www.yahoo.co.jp	Linux	Desktop	6	14	Dallas
https://www.yahoo.co.jp	macOS (Monterey)	Desktop	6	14	Dallas
https://www.yahoo.co.jp	Windows 10	Desktop	6	14	Dallas
https://www.yahoo.com	Android 10	Samsung Mobile	5	6	Dallas
https://www.yahoo.com	Android 10	Samsung Mobile	0	1	The Hague
https://www.yahoo.com	iOS 15.6	Apple iPad	14	6	Dallas
https://www.yahoo.com	iOS 15.6	Apple iPad	0	1	The Hague
https://www.yahoo.com	iOS 15.6	Apple iPhone	5	6	Dallas
https://www.yahoo.com	iOS 15.6	Apple iPhone	0	1	The Hague
https://www.yahoo.com	Linux	Desktop	52	7	Dallas
https://www.yahoo.com	Linux	Desktop	0	1	The Hague
https://www.yahoo.com	macOS (Monterey)	Desktop	30	6	Dallas
https://www.yahoo.com	macOS (Monterey)	Desktop	0	1	The Hague
https://www.yahoo.com	Windows 10	Desktop	30	6	Dallas
https://www.yahoo.com	Windows 10	Desktop	0	1	The Hague
https://www.yandex.ru	Android 10	Samsung Mobile	11	20	Dallas
https://www.yandex.ru	Android 10	Samsung Mobile	11	11	The Hague
https://www.yandex.ru	iOS 15.6	Apple iPad	14	20	Dallas
https://www.yandex.ru	iOS 15.6	Apple iPad	14	11	The Hague
https://www.yandex.ru	iOS 15.6	Apple iPhone	11	16	Dallas
https://www.yandex.ru	iOS 15.6	Apple iPhone	11	11	The Hague
https://www.yandex.ru	Linux	Desktop	49	20	Dallas
https://www.yandex.ru	Linux	Desktop	48	11	The Hague
https://www.yandex.ru	macOS (Monterey)	Desktop	49	20	Dallas
https://www.yandex.ru	macOS (Monterey)	Desktop	48	11	The Hague
https://www.yandex.ru	Windows 10	Desktop	49	20	Dallas
https://www.yandex.ru	Windows 10	Desktop	48	11	The Hague
https://www.youtube.com	Android 10	Samsung Mobile	1	3	Dallas
https://www.youtube.com	Android 10	Samsung Mobile	1	3	The Hague
https://www.youtube.com	iOS 15.6	Apple iPad	0	3	Dallas
https://www.youtube.com	iOS 15.6	Apple iPad	0	1	The Hague
https://www.youtube.com	iOS 15.6	Apple iPhone	0	3	Dallas
https://www.youtube.com	Linux	Desktop	1	5	Dallas
https://www.youtube.com	macOS (Monterey)	Desktop	1	5	Dallas
https://www.youtube.com	Windows 10	Desktop	1	5	Dallas
https://www.zhihu.com	Android 10	Samsung Mobile	1	10	Dallas
https://www.zhihu.com	Android 10	Samsung Mobile	1	10	The Hague
https://www.zhihu.com	iOS 15.6	Apple iPad	1	10	Dallas
https://www.zhihu.com	iOS 15.6	Apple iPad	1	10	The Hague
https://www.zhihu.com	iOS 15.6	Apple iPhone	1	10	Dallas
https://www.zhihu.com	iOS 15.6	Apple iPhone	1	10	The Hague
https://www.zhihu.com	Linux	Desktop	1	16	Dallas
https://www.zhihu.com	Linux	Desktop	1	16	The Hague

https://www.zhihu.com	macOS (Monterey)	Desktop	1	16	Dallas
https://www.zhihu.com	macOS (Monterey)	Desktop	1	16	The Hague
https://www.zhihu.com	Windows 10	Desktop	1	16	Dallas
https://www.zhihu.com	Windows 10	Desktop	1	16	The Hague
https://www.zoom.us	Android 10	Samsung Mobile	15	25	Dallas
https://www.zoom.us	Android 10	Samsung Mobile	0	14	The Hague
https://www.zoom.us	iOS 15.6	Apple iPad	48	30	Dallas
https://www.zoom.us	iOS 15.6	Apple iPad	0	14	The Hague
https://www.zoom.us	iOS 15.6	Apple iPhone	47	30	Dallas
https://www.zoom.us	iOS 15.6	Apple iPhone	0	14	The Hague
https://www.zoom.us	Linux	Desktop	52	30	Dallas
https://www.zoom.us	Linux	Desktop	0	14	The Hague
https://www.zoom.us	macOS (Monterey)	Desktop	51	30	Dallas
https://www.zoom.us	macOS (Monterey)	Desktop	0	14	The Hague
https://www.zoom.us	Windows 10	Desktop	52	30	Dallas
https://www.zoom.us	Windows 10	Desktop	0	14	The Hague

Table C.1: List of Alexa Top 50 websites, snapshot from 16th July 2022, number of cookies per website and user agent for location Dallas and The Hague.