Analyzing the Use of CNAME Cloaking in the Wild

by

Boris van Groeningen

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday July 10, 2023 at 14:00 PM.

Student number:	4719875	
Project duration:	November, 2022 – July, 2023	
Thesis committee:	Prof. dr. ir. G Smaragdakis,	TU Delft, supervisor
	Dr. G. M. Moura	TU Delft
	Dr. A. Katsifodimos,	TU Delft

Preface

During my first year of the master's program, I had the opportunity to select courses for the third quarter. One of the courses that caught our attention was the Hacking-Lab course, although we were initially unsure of what it entailed. As we explored the course, we discovered that we could choose from various projects. One particular project that intrigued us was the measurement of first- and third-party cookies on websites belonging to governmental and other official institutions. Given my interest in tinkering with web-related projects, I decided to pursue this project.

Throughout the course, our findings regarding the number of cookies on these websites proved to be quite interesting. During our discussions with our supervisor, Georgios Smaragdakis, the topic of CNAME cloaking was brought up. However, due to the limited duration of the course (only 10 weeks), we were unable to incorporate an investigation of CNAME cloaking into our project. Instead, we identified it as an area of potential future work that would require further exploration.

A few months later, when the time came to choose a topic for my thesis, I decided to continue exploring the realm of CNAME cloaking, picking up where I left off during the Hacking-Lab course. Given my positive experience working with Georgios (or George, if he likes you), I was excited to delve deeper into this subject under his guidance. Our bi-weekly meetings and discussions have been invaluable, and I would like to express my gratitude for his supervision throughout this project.

During my years at TU Delft, I have been fortunate to have a supportive group of friends who have accompanied me on this academic journey. Together, we motivated and pushed each other to excel, with the shared goal of graduating together. Bram, Ivo, Armin, Gerben, and Robin, thank you for the memorable times we've shared, both the highs and the lows. You probably all know what I mean by that. Your friendship has been a constant source of support and enjoyment. And of course, will continue to do so.

Last but not least, I would like to thank my family for (financially) supporting me throughout these years.

Boris van Groeningen, July 2023

Abstract

CNAME (Canonical Name) cloaking has emerged as a deceptive technique employed by website operators to obfuscate the true origin of their content. This master thesis aims to comprehensively examine the utilization and prevalence of CNAME cloaking across the web.

To achieve this, a custom program was developed to crawl websites and gather valuable insights such as cookies and embedded objects. DNS resolutions are performed to identify domains in the resolution chain that exhibit characteristics of cloaking, as per the defined parameters. The thesis leverages diverse datasets to analyze different segments of the web, providing a holistic view of the ecosystem.

This study focuses on several key aspects. Firstly, it investigates the most common types of cloakers encountered, shedding light on their prevalence and distribution within the web. Furthermore, the coexistence of Content Delivery Networks (CDNs), trackers, and cloakers is analyzed, providing a comprehensive understanding of their interplay and potential implications. Additionally, the Time-to-Live (TTL) values of cloakers are examined to gain insights into their temporal dynamics and potential strategies employed by operators.

By examining the prevalence and dynamics of CNAME cloaking, this research contributes to the broader understanding of this deceptive practice and its implications for privacy, security, and user experience. The findings of this thesis provide valuable insights for policymakers, web administrators, and security professionals to devise effective countermeasures against CNAME cloaking.

According to our findings, cloaking tends to occur more frequently on popular websites, indicating a correlation between website popularity and the likelihood of encountering cloaking behavior. Additionally, our analysis reveals that each cloaker tends to target specific types of websites, suggesting a degree of specialization or targeting within the cloaking ecosystem.

Moreover, we will delve into the origins and implications of both cookies and embedded objects in the context of cloaking. By examining the relationship between cloaking and these elements, we aim to gain a deeper understanding of the mechanisms and techniques employed by cloakers in their tracking practices.

Contents

List of Figures

List of Tables

Introduction 1

CNAME cloaking, also known as CNAME shadowing, is a deceptive technique employed by certain online platforms to mask the true origin of their content. It is a relatively recent phenomenon that has gained attention within the realm of web security and privacy.

The Domain Name System (DNS) serves as a fundamental infrastructure of the internet, translating human-readable domain names into IP addresses that computers can understand. In a typical DNS resolution process, a client sends a request for a domain name to a DNS server, which responds with the corresponding IP address [1]. However, CNAME cloaking subverts this process by introducing an additional layer of redirection [2].

Traditionally, the CNAME record in DNS allows a domain to be an alias for another domain [3]. For instance, when a user visits `subdomain.example.com`, the DNS resolution might reveal that it is an alias for `anotherdomain.com`. This redirection is transparent to the user, as the browser displays the content of `anotherdomain.com` while retaining the original URL in the address bar.

CNAME cloaking takes advantage of this DNS functionality to conceal the true source of content. By setting up CNAME records and leveraging various web technologies, malicious actors can make it appear as if the content is hosted on a trusted and reputable domain, when in reality, it is being served from a completely different and potentially untrustworthy source.

The motivation behind CNAME cloaking can vary. Some online platforms employ this technique for legitimate purposes, such as load balancing or content delivery network (CDN) management, to enhance performance and ensure availability [4]. However, it has also been adopted by malicious entities seeking to deceive users, evade security measures, or engage in unethical practices, including phishing, malware distribution, or illicit advertising [5].

In recent years, CNAME cloaking has become a significant concern due to its potential implications for online security, user privacy, and the overall integrity of the web. As a result, researchers, security professionals, and regulatory bodies have devoted increasing attention to understanding its mechanisms, identifying its prevalence, and developing countermeasures to mitigate its risks.

The subsequent sections of this thesis will explore the various facets of CNAME cloaking, including its techniques, impact on security and privacy, potential countermeasures, and recommendations for addressing this emerging threat.

1.1 Problem Statement

While CNAME cloaking presents various implications and challenges, it is primarily regarded as a problem due to the following reasons:

- **User Deception and Privacy Concerns**: CNAME cloaking can deceive users by making them believe that they are interacting with a trustworthy website or service when, in reality, they are unknowingly engaging with content served by an entirely different and potentially malicious source. This deceptive practice not only erodes user trust but also poses significant privacy risks, as user data can be collected, tracked, and potentially abused without their knowledge or consent .
- **Evading Security Measures**: By employing CNAME cloaking techniques, malicious actors can circumvent security measures such as domain reputation systems, blacklisting, and content filtering. This allows them to exploit the reputation of legitimate domains, making it difficult for security solutions to accurately identify and block malicious content. As a result, unsuspecting users may be exposed to various online threats, including phishing attacks, malware distribution, or unwanted advertising [6].

- **Impact on Web Ecosystem and Integrity**: CNAME cloaking disrupts the transparency and accountability of the web ecosystem. It undermines the integrity of domain ownership and obscures the true relationship between content providers and the websites or platforms hosting that content. This lack of transparency makes it challenging to hold responsible parties accountable for abusive or illegal activities, hampering efforts to enforce regulations and maintain a secure online environment [2].

The GDPR [7], enacted by the European Union (EU), is considered one of the most comprehensive data protection regulations to date. It empowers individuals with increased control over their personal data and imposes strict obligations on organizations handling such data. CNAME cloaking poses a significant challenge to the principles and objectives outlined in the GDPR, as it allows unauthorized data collection, profiling, and potentially compromises the confidentiality and integrity of personal information.

Furthermore, CNAME cloaking also raises concerns regarding cybersecurity. By disguising the true origin of a website or service, attackers can potentially create a perfect facade to launch various forms of cyberattacks, such as phishing, malware distribution, and identity theft. As a result, both individual users and organizations become vulnerable to significant financial, reputational, and legal repercussions.

1.2 Research Questions

In order to analyze the use and behavior or CNAME cloaking requires the following research questions to be answered.

Research Question 1

How prevalent is CNAME cloaking on the web?

Various datasets, containing domains from a certain sector, will be used to check for the presence of CNAME cloaking. For each of these, the most occurring cloakers will be analyzed. Furthermore, Time to Live (TTL) is also of interest for the encountered cloakers.

Research Question 2

What are the characteristics of websites that use CNAME cloaking?

This research seeks to identify the specific types of websites that are more prone to employing CNAME cloaking techniques. By examining a diverse range of online platforms, including e-commerce sites, social media platforms, financial institutions, governmental and news websites, the study aims to uncover patterns and trends that indicate a higher likelihood of CNAME cloaking usage.

Research Question 3

How is cloaking distributed amongst ranking intervals?

This research question seeks to investigate whether cloaking predominantly occurs on popular websites or if it is evenly distributed across the web. For this, we will only use our larger datasets to ensure an large enough sample size.

1.3 Outline

The remainder of this thesis encompasses essential background information, followed by a review of related work in the field. Subsequently, we provide a comprehensive explanation of our methodology, which can be found in Chapter 4 on page 13, along with a detailed overview of the datasets employed, as presented in Chapter 5 on page 17.

In the Results section, which can be found in Chapter 6 on page 20, we present the findings derived from the analysis of these datasets. This section serves as a comprehensive compilation of our observations, providing insights and conclusions drawn from our research.

Furthermore, we discuss our methodology and findings in Chapter 7 on page 33 as well as some of the limitations. Moreover, in this chapter we will conclude this research by summarizing our findings as well as discussing potential future work.

1.4 Contributions

Our research focuses on the analysis of CNAME cloaking using various datasets. We acknowledge that the web is a dynamic environment, constantly evolving. Therefore, examining how cloaking behavior changes over time is also a valuable aspect of our study. In addition to cookies, our methodology includes the assessment of embedded objects to identify instances of cloaking. This approach enables us to analyze the occurrence of cloaking across both cookies and embedded objects, providing a comprehensive understanding of its existence and distribution. Furthermore, the code used for the experiments, including all relevant datasets and outputs will be made available at `https://github.com/Boris304/cname-cloaking`. This can also be seen as a standalone tool to crawl websites and find tracking/cloaking.

Background 2

In this chapter, we will delve into the fundamental principles of web-related privacy, which serve as the cornerstone of this research. We will discuss several key concepts that form the basis of our investigation, namely DNS (Domain Name System), cookies, tracking, and cloaking.

2.1 DNS

The Domain Name System (DNS) is a fundamental technology that underpins the functioning of the internet. This section provides an overview of DNS, its purpose, and its role in translating human-readable domain names into machine-readable IP addresses. Understanding DNS is crucial for comprehending how websites are accessed and how information flows across the internet [1].

2.1.1 Definition and Purpose

The Domain Name System (DNS) is a hierarchical decentralized naming system that associates domain names with their corresponding IP addresses. It serves as a distributed database containing records that map domain names, such as `www.example.com`, to their respective IP addresses, such as 192.0.2.1. The primary purpose of DNS is to enable users to access websites and other internet resources using memorable domain names, rather than having to remember complex numerical IP addresses [8].

2.1.2 Functionality and Components

DNS operates through a distributed network of servers, each performing specific functions within the DNS resolution process. The key components of DNS include:

- **DNS Resolver**: The DNS resolver is a client-side software or service that initiates DNS queries on behalf of users or applications. It sends requests to DNS servers to resolve domain names into IP addresses [9].
- **DNS Server**: DNS servers store and distribute DNS records. They respond to queries from DNS resolvers by providing the requested information or referring the resolver to another DNS server that has the necessary data [8].
- **DNS Records**: DNS records contain information associated with domain names, such as IP addresses, mail server information (MX records), name server information (NS records), and other types of resource records (RR). These records are stored in DNS servers and retrieved during the resolution process [10].

2.1.3 DNS Resolution Process

When a user enters a domain name in a web browser or any other application, the DNS resolution process is initiated to translate the domain name into an IP address. The general steps involved in DNS resolution are as follows:

- **DNS Query**: The DNS resolver sends a query to the DNS server specified in its configuration or provided by the network.
- **Recursive Query**: If the DNS server being queried does not have the requested information, it will recursively query other DNS servers to resolve the domain name. This process continues until the IP address associated with the domain name is obtained [11].

- **Caching**: To improve efficiency and reduce the load on DNS servers, DNS resolvers typically cache the resolved DNS records. This caching allows subsequent queries for the same domain name to be resolved more quickly [12].

2.2 Cookies

In the context of the World Wide Web, cookies play a crucial role in facilitating various aspects of user interaction and enhancing the overall browsing experience. This section provides an overview of cookies, their functionality, and their significance in the web environment. Understanding cookies is essential for comprehending the mechanisms behind user tracking, session management, and personalization techniques employed by websites. This knowledge forms the foundation for further exploration of privacy concerns and regulatory frameworks surrounding cookies [13].

2.2.1 Definition and Purpose

Cookies, also known as HTTP cookies or web cookies, are small text files that websites store on a user's device (typically a web browser) to store information or preferences. These files consist of name-value pairs and are primarily used to enable stateful interactions between the user and the website. By retaining certain information, such as login credentials, session identifiers, and user preferences, cookies can enhance user experience and provide personalized content.

2.2.2 Functionality and Types

Cookies serve various functions, depending on their purpose and lifespan. They can be classified into two broad categories: session cookies and persistent cookies.

- **Session Cookies**: Session cookies are temporary files that are erased when the user closes the web browser. They facilitate session management by allowing websites to remember user actions or preferences during a single browsing session. For example, session cookies enable the retention of items in an online shopping cart as users navigate through different pages on an e-commerce website [14].
- **Persistent Cookies**: Unlike session cookies, persistent cookies remain stored on the user's device even after the browser is closed. They have a specific expiration date or remain valid until manually deleted by the user. Persistent cookies are commonly utilized for purposes such as remembering login credentials, customizing website settings, and delivering targeted advertisements based on user preferences [15].

2.2.3 Regulation and Privacy Implications

While cookies offer significant benefits in terms of personalization and user experience, they also raise concerns regarding privacy and data protection. Since cookies can track user behavior and collect personal information, they have been subject to scrutiny by privacy advocates and regulatory bodies. As a result, several countries and regions have introduced legislation and regulations to protect user privacy and regulate the use of cookies, such as the European Union's General Data Protection Regulation (GDPR) [7] and ePrivacy Directive [16].

2.3 Embedded Objects

Embedded objects play a crucial role in enhancing the functionality and interactivity of web pages. This section provides an overview of embedded objects, their purpose, and their impact on the user experience. Understanding embedded objects is essential for comprehending how multimedia content, interactive elements, and external resources are seamlessly integrated into web pages [17].

2.3.1 Definition and Purpose

Embedded objects refer to multimedia elements, such as images, videos, audio files, and interactive applications, that are seamlessly integrated within a web page. These objects are typically displayed directly within the web page's content, enhancing its visual appeal, interactivity, and information presentation. The primary purpose of embedded objects is to enrich the user experience, provide supplementary information, and facilitate the effective communication of ideas.

2.3.2 Functionality and Types

Embedded objects offer various functionalities, depending on their nature and purpose. They can be categorized into different types, including:

- **Images**: Images are one of the most common types of embedded objects. They can be static or dynamic, providing visual representations, illustrations, or infographics to support the textual content on a web page. Images contribute to the aesthetic appeal of the page and can help convey information more effectively.
- **Videos**: Embedded videos allow the seamless playback of multimedia content directly within a web page. They enable the presentation of dynamic visual information, such as tutorials, demonstrations, promotional videos, or entertainment content. Videos can be hosted on external platforms (e.g., YouTube, Vimeo) or self-hosted on the website's server.
- **Audio Files**: Embedded audio files provide a way to include sound elements within a web page. They are commonly used for background music, podcasts, interviews, or other audio-based content. Audio files can be played directly on the web page or streamed from external sources.
- **Interactive Applications**: Embedded interactive applications, such as maps, surveys, games, or forms, enable user engagement and interaction within the web page. These objects allow users to input data, make selections, or manipulate elements, enhancing the overall user experience and facilitating data collection or user feedback.

2.3.3 Significance and Impact

Embedded objects significantly contribute to the overall user experience and engagement on web pages. They help convey information more effectively by combining visual, auditory, and interactive elements. They, however, do not need to originate from the same domain as the website they are found on. This is important to note because trackers can use this to invade the website unknowingly to the user.

2.4 CDNs

Content Delivery Networks (CDNs) have become an integral part of the modern web infrastructure, significantly impacting the delivery of web content and improving user experiences. This section provides an overview of CDNs, their purpose, and their impact on website performance, scalability, and global content distribution. Understanding CDNs is essential for comprehending the mechanisms behind efficient content delivery and optimizing web performance [18].

2.4.1 Definition and Purpose

A Content Delivery Network is a globally distributed network of servers strategically located in various geographic regions. CDNs are designed to deliver web content, such as HTML pages, images, videos, and other static or dynamic files, to end-users with improved speed, reliability, and scalability. The primary purpose of CDNs is to reduce latency, minimize server load, and enhance the overall performance of websites, especially for users located far from the origin server.

2.4.2 Functionality and Benefits

CDNs offer several functionalities and benefits that contribute to enhanced web performance and user experience:

- **Content Caching and Edge Servers**: CDNs store cached copies of website content across multiple edge servers located in geographically diverse locations. When a user requests content, the CDN delivers it from the nearest edge server, reducing latency and minimizing network congestion [19].
- **Load Balancing**: CDNs distribute incoming traffic across multiple servers, optimizing resource utilization and preventing individual servers from becoming overwhelmed. This load balancing technique ensures consistent performance and reduces the risk of server downtime during peak usage periods [4].
- **Scalability and Capacity**: CDNs can handle significant spikes in traffic by dynamically scaling resources and leveraging their distributed infrastructure. This scalability enables websites to accommodate increased demand without compromising performance or user experience.
- **Global Content Distribution**: CDNs replicate and distribute content across multiple servers worldwide, allowing websites to serve users in different geographic regions more efficiently. This global distribution reduces the distance between users and content, resulting in faster load times and improved user satisfaction.
- **Security and DDoS Mitigation**: Many CDNs offer built-in security features, such as protection against Distributed Denial of Service (DDoS) attacks, intrusion detection, and web application firewalls. These security measures help safeguard websites and mitigate potential threats [20].

2.4.3 Impact on Web Performance and User Experience

CDNs have a significant impact on web performance and user experience by improving page load times, reducing latency, and ensuring reliable content delivery. By caching content closer to end-users, CDNs minimize the distance data must travel, resulting in faster and more responsive websites. This improved performance leads to enhanced user satisfaction, longer visit durations, and increased conversion rates.

Additionally, CDNs contribute to improved website availability and resilience. By distributing content across multiple servers, CDNs reduce the risk of single points of failure and enhance the overall reliability of websites, even during periods of high traffic or server outages.

2.5 Trackers

In the realm of web browsing, trackers play a significant role in monitoring user behavior, collecting data, and enabling various functionalities [21]. This section provides an overview of trackers, their purpose, and their impact on user privacy. Understanding trackers is crucial for comprehending the mechanisms behind targeted advertising, data collection practices, and the implications for user consent and control over personal information.

2.5.1 Definition and Purpose

Trackers, also known as web tracking technologies, are tools used by websites and online services to monitor user activity and gather information about their browsing behavior. These tracking technologies enable the collection of data related to a user's interactions with websites, including visited pages, clicks, searches, and other online activities. The primary purpose of trackers is to gather insights into user behavior, personalize content, and deliver targeted advertisements.

2.5.2 Tracker Types

Trackers can be classified into two main categories based on their origin and relationship to the website being visited: first-party trackers and third-party trackers.

- **First-Party Trackers**: First-party trackers are embedded on a website by its own operators. They are designed to collect data specifically related to the user's interactions with that particular website. First-party trackers are commonly used for essential functions such as session management, website analytics, and personalization based on user preferences [22].
- **Third-Party Trackers**: Third-party trackers are embedded on a website by entities other than the website operators. These trackers can monitor user activity across multiple websites, often without the user's explicit knowledge or consent. Third-party trackers are commonly associated with online advertising networks, data brokers, and analytics companies. They collect data from various websites to create user profiles and deliver targeted advertisements [23].

2.5.3 Privacy Implications and Regulation

The widespread use of trackers raises concerns about user privacy, data protection, and the potential misuse of personal information. The extensive tracking of user behavior across multiple websites can create detailed profiles that may infringe upon individual privacy. Furthermore, the lack of transparency and user control over data collection practices by third-party trackers has prompted regulatory scrutiny and efforts to establish safeguards.

Various regulatory frameworks, such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), have been implemented to address these concerns. These regulations emphasize the importance of obtaining user consent for tracking activities, providing transparent information about data collection practices, and giving users control over their personal information.

2.6 CNAME Cloaking

CNAME (Canonical Name) cloaking is a technique employed by some online services to conceal the true identity of the domain or subdomain they are using. This section provides an overview of CNAME cloaking, its purpose, and its implications. Understanding CNAME cloaking is important for recognizing the potential challenges it poses to transparency, security, and trust within the web ecosystem [2].

2.6.1 Definition and Purpose

CNAME cloaking involves the use of DNS (Domain Name System) records to redirect traffic from a domain or subdomain to a different domain while maintaining the original domain name in the URL. This technique is often used by online services, including content delivery networks (CDNs) and privacy-focused tools, to mask their actual domain and make it appear as if the service is being provided directly from the original domain.

The primary purpose of CNAME cloaking is to enhance privacy, security, and branding for the online service by concealing the underlying infrastructure and providing a seamless user experience. It allows the service provider to leverage the reputation, trust, and established security measures associated with the original domain.

2.6.2 Functionality and Implications

CNAME cloaking works by configuring the DNS records of the original domain to point to the domain or subdomain of the service provider. This makes it appear as if the content or service is being served directly from the original domain, despite the actual infrastructure being hosted by the service provider.

While CNAME cloaking can provide benefits such as improved performance, increased security, and easier maintenance for the service provider, it raises several implications:

- **Transparency and Trust**: CNAME cloaking can obscure the true origin of the content or service, making it challenging for users to determine which entity is actually providing the service. This lack of transparency can erode trust and make it difficult for users to assess the credibility and security of the service.
- **Security Risks**: Since the actual infrastructure is hidden behind CNAME cloaking, it can be challenging to verify the security measures implemented by the service provider. Users may unknowingly trust a service that does not meet their security expectations, potentially exposing them to data breaches or other vulnerabilities.
- **DNS Resolution Challenges**: CNAME cloaking can introduce complexities in DNS resolution, leading to potential issues with caching, DNS-based security measures, and accurate logging of traffic. It can also impact the ability to implement granular access controls or enforce geolocation restrictions effectively.

Related Work 3

3.1 CNAME Cloaking

The paper titled "Characterizing CNAME Cloaking-Based Tracking on the Web" [6] by Dao et al. holds significant relevance in the exploration of CNAME cloaking and its impact on online tracking. This comprehensive study provides valuable insights into the prevalence, behavior, and privacy implications associated with CNAME cloaking-based tracking techniques. By examining the use of CNAME cloaking by popular tracking providers and analyzing their behavior, the authors contribute to our understanding of the extent to which this tracking method is employed in the wild.

Prior to this study, limited research had been conducted specifically focusing on CNAME cloaking and its implications. The paper bridges this gap by conducting a detailed measurement study, shedding light on the scale and characteristics of CNAME cloaking-based tracking. By identifying various tracking domains that utilize CNAME cloaking and analyzing their behavior, the researchers uncover the types of tracking techniques employed and the information collected, thereby providing a deeper understanding of the privacy risks associated with this technique.

The authors also propose countermeasures to mitigate CNAME cloaking-based tracking, suggesting potential interventions at both the policy and browser-level. These proposed countermeasures contribute to the ongoing efforts in developing effective strategies to protect user privacy against CNAME cloaking techniques.

Furthermore, this paper serves as a valuable reference for researchers, policymakers, and industry professionals interested in the field of online tracking and user privacy. It establishes a foundation for future research endeavors in understanding the impact of CNAME cloaking and exploring further mitigation strategies. The insights and findings presented in this paper can guide the development of privacy-preserving technologies and inform the formulation of privacy policies that aim to address the challenges posed by CNAME cloaking-based tracking.

In summary, "Characterizing CNAME Cloaking-Based Tracking on the Web" is a highly relevant and significant contribution to the existing body of knowledge on online tracking and user privacy. Its comprehensive measurement study, identification of tracking domains, analysis of tracking techniques, and proposed countermeasures provide valuable insights for researchers, policymakers, and industry practitioners seeking to understand and address the implications of CNAME cloaking-based tracking techniques.

3.2 Cookie Synchronization

The paper titled "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask" [24] by Papadopoulos et al. addresses an important aspect of online tracking, namely cookie synchronization. This research study provides a comprehensive examination of cookie synchronization techniques, their implications for user privacy, and the potential risks associated with this practice.

In recent years, cookies have played a crucial role in tracking users' online activities across websites. Cookie synchronization, also known as cookie syncing, is a technique employed by advertising networks and tracking companies to share user identifiers or data stored in cookies between different domains. By synchronizing cookies, these entities can build a more comprehensive profile of a user's browsing behavior, preferences, and interests.

Prior to the work conducted by Papadopoulos et al., there was limited research specifically focused on cookie synchronization and its implications. This paper fills this research gap by providing a detailed exploration of cookie synchronization techniques, shedding light on the underlying mechanisms, motivations, and potential risks involved.

The researchers examine the prevalence of cookie synchronization across various advertising ecosystems, including the identification of key players and the methods employed to facilitate cookie syncing. Through their analysis, Papadopoulos et al. offer insights into the potential privacy concerns raised by this practice, as it allows for the linking of user data across different websites and domains without the explicit knowledge or consent of the user.

Furthermore, the paper presents an in-depth discussion of the privacy implications and risks associated with cookie synchronization. It examines the potential for data leakage, tracking beyond user opt-out preferences, and the lack of transparency and control for users. By highlighting these risks, Papadopoulos et al. contribute to a better understanding of the privacy challenges posed by cookie synchronization and the need for appropriate safeguards.

The work by Papadopoulos et al. also proposes potential mitigations and countermeasures to address the privacy risks associated with cookie synchronization. By advocating for improved transparency, user control, and industry guidelines, the authors provide practical recommendations for minimizing the negative impact of cookie synchronization on user privacy.

This paper is highly relevant for researchers, policymakers, and industry professionals interested in the field of online tracking and user privacy. It serves as a valuable reference, offering a comprehensive overview of cookie synchronization techniques, their implications, and potential mitigations. The insights and findings presented in this study can inform the development of privacy-enhancing technologies, industry best practices, and regulatory frameworks aimed at safeguarding user privacy in the context of cookie synchronization.

In conclusion, "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask" by Papadopoulos et al. is a significant contribution to the existing body of knowledge on online tracking and user privacy. By providing a comprehensive exploration of cookie synchronization techniques, privacy implications, and potential mitigations, the authors shed light on an important aspect of online tracking and contribute to the ongoing efforts to protect user privacy in the digital ecosystem.

3.3 Security Risk of CNAME Cloaking

The paper "Oversharing Is Not Caring: How CNAME Cloaking Can Expose Your Session Cookies" [25] by Aliyeva and Egele provides a comprehensive exploration of the security risks posed by CNAME cloaking, with a specific focus on the exposure of session cookies. By conducting meticulous analysis, the authors demonstrate how CNAME cloaking can inadvertently lead to the leakage of session cookies, thereby jeopardizing user authentication, authorization, and overall account security.

One of the notable contributions of this paper is its investigation into the mechanisms through which CNAME cloaking can result in the exposure of session cookies. The authors shed light on the technical intricacies involved in this process, highlighting the interplay between CNAME chains, cross-origin requests, and cookie policies. Through their analysis, they establish a clear understanding of the vulnerabilities introduced by CNAME cloaking and the potential risks to user privacy and security.

Moreover, the paper highlights the real-world implications of cookie exposure resulting from CNAME cloaking. By demonstrating how an attacker can exploit this vulnerability to gain unauthorized access to user accounts or perform session hijacking, Aliyeva and Egele emphasize the urgency of addressing this issue. Their findings underscore the need for robust countermeasures and security measures to mitigate the risks associated with CNAME cloaking.

In addition to uncovering the security risks, the authors propose potential mitigation strategies and protective measures to address the cookie exposure resulting from CNAME cloaking. By advocating for improvements in web security practices, Aliyeva and Egele contribute to the ongoing efforts in developing safeguards to protect user data and privacy. Their recommendations encompass aspects such as secure cookie management, improved web policies, and browser-level interventions.

The paper by Aliyeva and Egele is highly relevant for researchers, practitioners, and policymakers interested in understanding the security implications of CNAME cloaking and its impact on user privacy. The findings presented in this study provide a deeper understanding of the risks associated with CNAME cloaking, specifically in terms of session cookie exposure. Furthermore, the proposed mitigation strategies can guide the development of improved security measures, informing the design and implementation of privacy-preserving technologies and policies.

3.4 Cloaking and Governmental Websites

The paper "Measuring Web Cookies in Governmental Websites" [26] by Gotze et al. addresses the issue of user tracking on governmental websites, providing valuable insights into the prevalence of third-party tracking and the associated privacy risks. Although this study focuses on web cookies rather than CNAME cloaking specifically, it is highly relevant to our research on CNAME cloaking due to its examination of tracking practices and their implications for user privacy.

Gotze et al.'s study investigates popular governmental websites across different countries, assessing the extent to which these websites engage in third-party tracking during user visits. The findings of the study reveal a significant concern: up to 90% of the analyzed governmental websites generate cookies from third-party trackers without obtaining user consent. Even in countries with strict user privacy laws, non-session cookies created by trackers persist on these websites, potentially compromising user privacy.

The relevance of the paper to our research on CNAME cloaking lies in the parallel objective of understanding user tracking. While CNAME cloaking primarily focuses on the obfuscation of tracking infrastructure through DNS manipulation, the insights provided by Gotze et al. shed light on the real-world implications of tracking practices on governmental websites.

The study's recommendations for responsible governmental website development are particularly noteworthy. The authors suggest avoiding the embedding of third-party resources, such as social media plugins and web advertising media, and minimizing the inclusion of references to external URLs that can potentially download additional content. These recommendations align with the objectives of our research on CNAME cloaking, as they emphasize the need to minimize the integration of potentially malicious or privacy-invasive external entities within a website's infrastructure. These entities can be seen as embedded objects, which we will be analyzing of CNAME cloaking.

Furthermore, the paper highlights the importance of regular audits of governmental websites to assess the state of third-party tracking and promptly remove any identified trackers. This emphasis on transparency and accountability resonates with our efforts to detect and mitigate CNAME cloaking techniques, as both areas of research aim to safeguard user privacy and ensure compliance with privacy regulations.

In summary, the paper "Measuring Web Cookies in Governmental Websites" by Gotze et al. provides valuable insights into the prevalence and implications of user tracking on governmental websites. By considering the findings and recommendations of this study, our research on CNAME cloaking can benefit from a broader understanding of privacy risks and potential countermeasures. The insights gained from this paper inform the development of detection and prevention mechanisms tailored specifically for CNAME cloaking, enhancing our ability to protect user privacy in the context of web tracking.

Methodology 4

This chapter provides a technical perspective on key concepts introduced in Chapter 2 on page 4 and lays the foundation for the experimental setup and data analysis. It explores technical definitions such as CDNs, tracking, and cloaking. The chapter explains the role of CDNs in the web ecosystem and highlights that they serve a beneficial purpose. It also discusses tracking, which is considered invasive, and introduces cloaking as a specific form of tracking. The chapter then delves into the methodology for data collection, outlining the specialized tool used to identify instances of cloaking-based tracking. It explains the process of extracting cookies and embedded objects, performing DNS resolutions, and flagging CDNs, trackers, and cloakers. The chapter concludes by outlining the analysis that will be conducted on the collected data, focusing on the distributions of CDNs, trackers, and cloakers, as well as deeper examinations of frequent cloakers and their origins.

4.1 Technical Definitions

In this section, we will delve into a more technical perspective on some of the concepts previously discussed in Chapter 2 on page 4. These concepts play a pivotal role in shaping our experimental setup and are subsequently utilized for conducting in-depth analysis on the collected data.

4.1.1 CDNs

As mentioned earlier, Content Delivery Networks (CDNs) play a crucial role in the web ecosystem, as they are responsible for delivering a significant portion of content to users [3]. Notably, Google and Facebook are prominent examples of CDNs. In our experimental setup, we will specifically examine the presence of CDNs and their impact on the domains we analyze.

It is important to note that CDNs are not regarded as harmful entities when encountered while traversing the web. On the contrary, they serve a beneficial purpose by efficiently delivering content to users.

For a comprehensive list of the CDNs we consider in our experiments, please refer to Chapter 5 on page 17. This chapter provides detailed information about the datasets we utilize, including the complete list of CDNs included in our analysis.

4.1.2 Tracking

Trackers are widely recognized as being more invasive than CDNs when it comes to privacy concerns [21]. In our experimental research, we leverage a consolidated dataset that combines information from various sources to ascertain whether a given domain should be classified as a tracker. To achieve this classification, we cross-reference the top-level domain of each encountered domain with our meticulously curated trackers list, as elaborated in Chapter 5 on page 17. Domains that match the criteria in our trackers list are then flagged in our output data, enabling further analysis and investigation, as illustrated inFigure 4.1. This systematic approach allows us to gain comprehensive insights into the presence and behavior of trackers in our study.

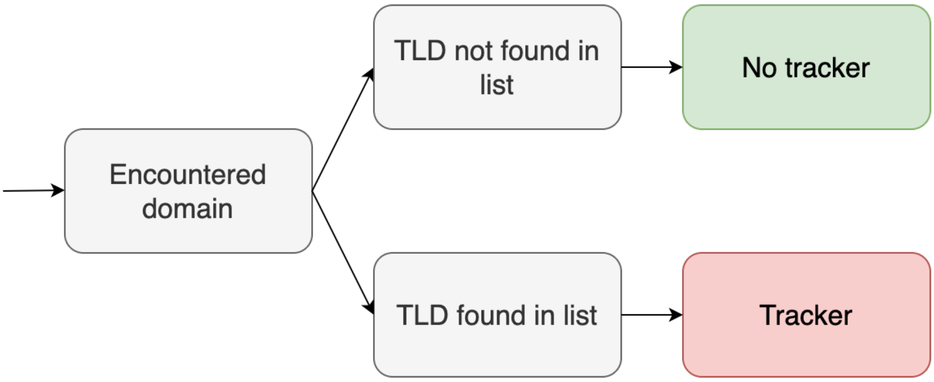

Figure 4.1: Overview of tracking definition.

4.1.3 Cloaking

In this report, we define *cloaking* as a form of tracking known as cloaking-based tracking. This occurs when the initial domain, discovered through cookies or embedded objects, contains a tracker from a different top-level domain in the DNS resolution chain. We classify the tracking domain involved in this scenario as a cloaker. An overview can be found below in Figure 4.2.

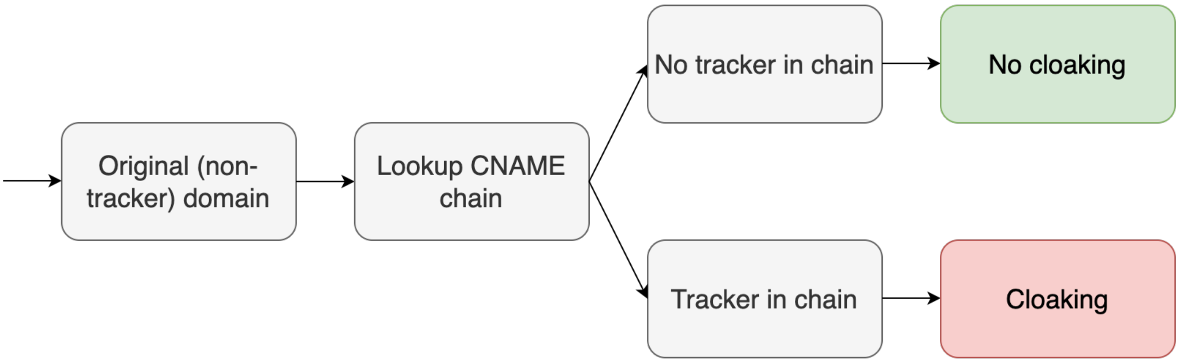

Figure 4.2: Overview of cloaking definition.

An illustrative example of the aforementioned scenario is depicted in Figure 4.3 below. In this case, an embedded object associated with the domain `voordeelnieuwtje.net` resolves to the domain `fjm0v.voluumtrk.com` before ultimately resolving to an IP address. According to our list of trackers, this domain qualifies as a tracker. Therefore, this specific instance qualifies as a case of cloaking.

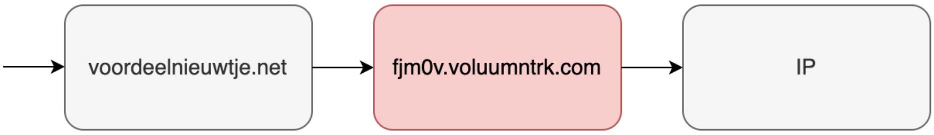

Figure 4.3: A simple case of cloaking.

4.2 Data Collection Overview

In our experimental methodology, we have developed a specialized tool designed to identify instances of cloaking-based tracking on the web. This tool utilizes Selenium [27] and Chrome DevTools [28] to access websites when provided with a specific domain. This, therefore, also used Google Chrome perform the experiments [29]. Once a webpage is successfully reached, we extract both cookies and embedded objects for subsequent cloaking analysis. In the case of cookies, we extract pertinent information such as their name, domain, and TTL (Time to Live), and subsequently group them based on domain to optimize DNS resolution later in the process.

Next, we perform DNS resolutions for both the discovered domains of embedded objects and cookies. If any domain within the chain of resolutions is identified as a tracker with a top-level domain different from that of the original website, we classify it as an instance of cloaking as discussed in Section 4.1 (Technical Definitions). Additionally, we identify and flag Content Delivery Networks (CDNs) and regular trackers, which are then recorded in a JSON object for further analysis purposes. This comprehensive approach allows us to effectively detect and examine instances of cloaking-based tracking, as well as identify other relevant entities for further investigation.

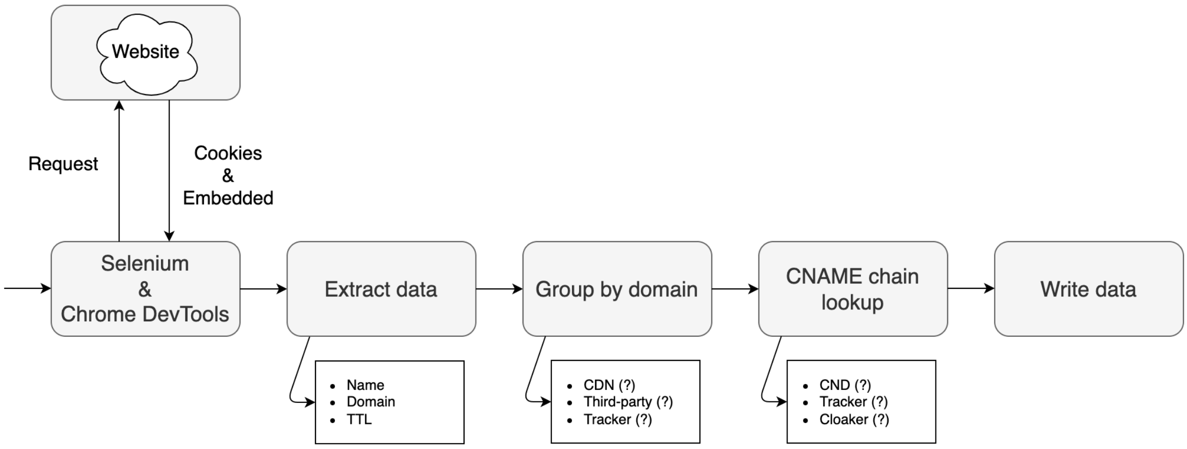

Figure 4.4: High-level overview of the experimental setup.

Upon successful execution, the resulting JSON object will contain information about all reachable websites. Each website is divided into two parts: cookies and embedded objects. As previously mentioned, cookies are grouped by domain to minimize the number of required DNS resolutions. For cookies, we store basic cookie data such as the name and expiration date represented as a UNIX timestamp. For each cookie domain, we conduct DNS resolutions and record the TTL and domain name. These domains are then checked for the presence of CDNs, tracking, and cloaking, and appropriately flagged in the data. Regarding embedded objects, only the base URL is considered. For each unique embedded object found on the website, we perform the same DNS resolutions as for the cookie domains. This process allows us to gather information about the TTL and domain name of the embedded objects. Overall, the JSON object provides comprehensive details about cookies, including DNS resolutions and categorization, as well as information about embedded objects obtained through DNS resolutions.

4.3 Analyis

Once all the data from the various datasets is collected, our analysis will focus on examining the distributions of CDNs, trackers, and cloakers. We will explore the prevalence of each category within the datasets and identify the most common cloakers specific to each dataset. Additionally, we will investigate whether these cloakers originate from cookies or embedded objects. In our analysis, we will also consider the Time-To-Live (TTL) values associated with the identified cloakers. This information will provide insights into the longevity of the cloaking activity. For larger datasets, we will further explore the distribution of cloakers within ranking intervals. This analysis will help identify any patterns or trends related to the prevalence of cloakers among websites in different ranking ranges.

Additionally, we will conduct a detailed examination of the more frequent cloakers. This analysis will involve studying their DNS resolutions and origins to gain deeper insights into their behavior and characteristics. Furthermore, we will investigate the websites from which these cloakers originate and determine the nature of these websites.

By exploring the origins of the cloakers and studying the types of websites they are associated with, we aim to understand the patterns and tendencies of cloaking practices.

Datasets 5

In this short chapter, we will discuss the datasets to be used for our experiments. This includes datasets for the domains we will be crawling and lists containing known trackers and CDNs as well.

5.1 Domains

In our experimental approach, we utilize a range of diverse datasets, each encompassing different types of domains. These datasets serve as valuable resources for our research. Firstly, we employ the Alexa Top 1M dataset [30], which contains a comprehensive compilation of the most frequently visited domains ranked in descending order. From this dataset, we extract a subset specifically comprising Dutch websites, allowing us to conduct targeted experiments within this context. Additionally, we incorporate a curated list of governmental domains provided by Rijksoverheid [31], which further enhances the comprehensiveness and relevance of our analysis. Moreover, we include domains associated with the G20 and Covid [32], with a particular focus on news-related domains. These extensive datasets contribute to the richness and depth of our experimental investigations, although they do require a significant investment of resources. Furthermore, we will compare the presence of cloaking to that of the Fukuda dataset [31], which contains domains on which cloaking-based tracking has been detected as of January 2020.

Table 5.1: The datasets used for the experiments.

Dataset	# Domains
Alexa	707k
Dutch	11k
Rijksoverheid	1.8k
G20	5.8k
Covid	198
Fakuda	1762

5.1.1 Alexa

The Alexa dataset has been extensively used in the creation of our algorithm. It offers a good balance of websites and a sufficient number of trackers for detection purposes. However, it is important to note that out of the 707k domains in the dataset, experiments were only performed on the first 10k domains to improve the speed of the experiments. As we move further down the ranking, the number of unreachable websites increases. Here are the first 10 domains from the Alexa dataset to provide an insight into the types of popular websites included:

1. google.com
2. youtube.com
3. baidu.com
4. facebook.com
5. bilibili.com
6. qq.com
7. zhihu.com
8. amazon.com
9. twitter.com
10. wikipedia.org

5.1.2 Dutch

The Alexa dataset is the foundation of the Dutch dataset. Namely, we have taken all domains ending in **.nl**. This means that some domains from the Alexa and Dutch datasets might overlap. However, this overlap is quite small, namely 20 websites. Doing experiments on all Dutch websites provides us with an understanding of cname-based cloaking on a more local level. The first 10 domains in the Dutch dataset are:

1. `nu.nl`
2. `ad.nl`
3. `google.nl`
4. `nos.nl`
5. `telegraaf.nl`
6. `funda.nl`
7. `digid.nl`
8. `dumpert.nl`
9. `marktplaats.nl`
10. `buienradar.nl`

5.1.3 Rijksoverheid

The Rijksoverheid dataset comprises all websites belonging to the Dutch government. Since it consists of government websites, it is not expected to encounter cloaking-based tracking. However, if such tracking were to be present, it would be an interesting and invasive finding, considering the sensitive nature of government websites and the potential privacy implications. Some prominent examples of this dataset are: `rijksoverheid.nl`, `rivm.nl` and `government.nl`.

5.1.4 G20

The G20 dataset used in this study consists of official websites from the 19 countries that are part of the G20 [33]. The presence of any cloaking within this dataset would indeed have serious privacy implications, considering that these websites are predominantly governmental in nature.

Governmental websites often deal with sensitive information and are trusted sources of information for the general public. If cloaking-based tracking were to be found on these websites, it would raise concerns about the privacy and security of user data. Given the authoritative nature of these websites, any privacy breaches or unauthorized tracking activities could have significant implications.

Therefore, analyzing the presence of cloaking within the G20 dataset is crucial to understand the extent of privacy risks associated with such practices on official governmental websites.

5.1.5 Covid

The dataset comprising official COVID-19-related webpages is relatively small compared to other datasets we used. However, it holds significant importance due to the nature of the websites being official and governmental, specifically focused on COVID-19 information. The presence of CNAME cloaking in these websites would have serious privacy implications, considering the substantial amount of traffic these websites generated during the pandemic.

5.1.6 Fakuda

The dataset used in this study consists of domains where cloaking has been previously detected. In our report, it is named after the GitHub page on which it can be accessed [34]. It serves as a means of observing how the online landscape evolves over time in terms of cloaking behavior. One of the questions we aim to address is whether cloaking disappears once it is exposed.

While it is possible that some domains may discontinue their use of cloaking techniques after being exposed, it is important to note that the online landscape is constantly changing. New cloaking methods may emerge, and existing cloakers may adapt their techniques to evade detection. Therefore, the disappearance of cloaking cannot be guaranteed solely based on exposure. Continuous monitoring and analysis are necessary to stay updated on the evolving nature of cloaking practices. Some prominent examples in this list are:

- `nike.com`
- `nintendo.com`
- `discovery.com`

5.2 Trackers

To determine whether a domain encountered during our analysis is a tracker, we have compiled multiple well-established tracker lists, including Adguard DNS, Easylist, Nocoin, and Easyprivacy [35]. These lists serve as valuable resources for identifying known trackers on the web.

To enhance the comprehensiveness of our tracker detection, we have created a consolidated list by taking the union of these four individual lists. This unified list provides us with the broadest possible coverage when it comes to detecting and recognizing trackers across various domains. By leveraging this comprehensive compilation of tracker information, we can effectively identify and classify domains involved in tracking activities during our experiments.

5.3 CDNs

In our experimental investigations, we are also interested in detecting the presence of Content Delivery Networks (CDNs). To accomplish this, we have curated a concise list of popular CDNs for comparison with the domains encountered during our experiments. We have sourced most of these CDNs from the list provided in the paper titled "Seven Years in the Life of Hypergiants' Off-Nets" [36] by Gigis et al.

By cross-referencing the domains we encounter with this curated list of popular CDNs, we can determine whether a domain matches the top-level domain associated with a CDN. This analysis allows us to identify the involvement of CDNs and assess their impact on the web content we examine during our experiments.

The list of popular CDNs we have compiled for our experiments include the following domains: Google, Facebook, Instagram, Netflix, Akamai, Alibaba, Cloudflare, Amazon, CDN Networks, Limelight, Apple, Twitter, Msegde, and Fastly. These CDNs represent well-known entities in the realm of content delivery and are included in our analysis to determine the presence and influence of CDNs in the domains we encounter during our experiments.

Results 6

In this chapter, we will explore various datasets to analyze the occurrence of cloaking and other related factors. Our investigation begins by examining the distribution of CDNs, trackers, and cloakers within each dataset. It is important to note that, according to our definition, cloakers are considered trackers, but we specifically account for them as cloakers in our analysis.

Additionally, we will investigate the prevalent cloakers within each dataset, examine Time-To-Live (TTL) values, and explore the categorization of websites and cloakers. Moreover, we will showcase examples of prominent websites to provide a clearer understanding of the subject matter.

As discussed in Chapter 5 on page 17, we will be using various datasets to conduct our experiments. In Table 6.1 below you can find the number of reachable pages per dataset during execution. A page being unreachable can be caused by a number of things, namely:

- The website is down.
- Connection timeout → our program continues on to the next domain when no connection has been established for a set amount of time. This is more often the case with international websites, as seen by the **G20** row in the table.
- The website enforces anti-automation techniques, rendering our program ineffective.

Table 6.1: Datasets and their reachable pages.

	Total domains	**Reachable domains**	**Percentage (%)**
Alexa	10000	8858	88.58
Dutch	10709	9770	91.23
Rijksoverheid	1812	1332	73.51
G20	5813	3435	59.09
Covid	198	156	78.79
Fakuda	1762	1698	96.37

6.1 Alexa

As depicted in the Sankey diagram shown in Figure 6.1 and the distribution table provided in Table 6.2, the majority of CDN-related activity is observed in the embedded objects. Notably, what stands out is the significant number of trackers found in cookies compared to the other datasets. This observation highlights the noteworthy presence of trackers within the cookie domain, indicating their potential impact on user privacy. Furthermore, cloaking-based tracking is almost evenly split amongst cookies and embedded objects.

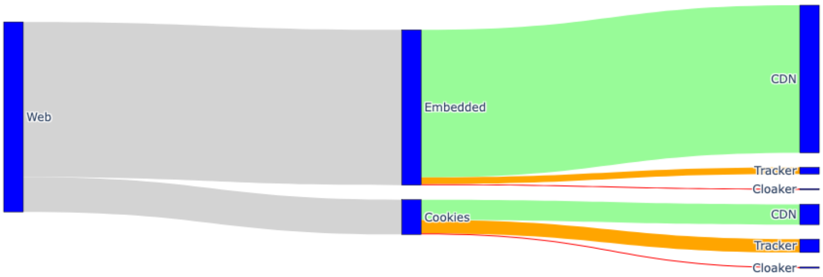

Figure 6.1: Sankey diagram of the Alexa dataset.

Table 6.2: Activity distribution for the Alexa dataset.

	Cookies	**Embedded**
CDN	1840	13398
Tracker	1207	604
Cloaker	102	95

6.1.1 Cloaking encounters

By far the most common cloaker in the Alexa dataset is `wl.hpyrdr.com` as seen in Figure 6.2. We will have a closer look at `wl.hpyrdr.com` later on in Section 6.7. Besides `wl.hpyrdr.com`, most cloaking happens with embedded objects

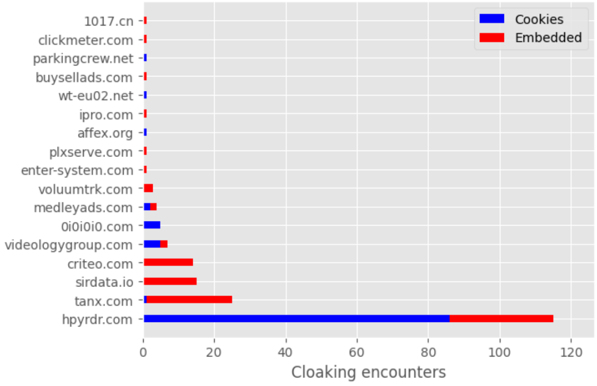

Figure 6.2: Bar graph containing the cloakers and their frequencies in the Alexa dataset.

6.1.2 Ranking

We have analyzed the cloaking percentage between ranking intervals. As we can clearly see from Figure 6.3, most of the cloaking happens on the more popular websites. With 4.5% cloaking happening in the first 1% of reachable websites. This translates to the first 88 websites. This is double of what happens at the less popular websites, where cloaking occurs at frequencies of 1.3% and 1.6% respectively.

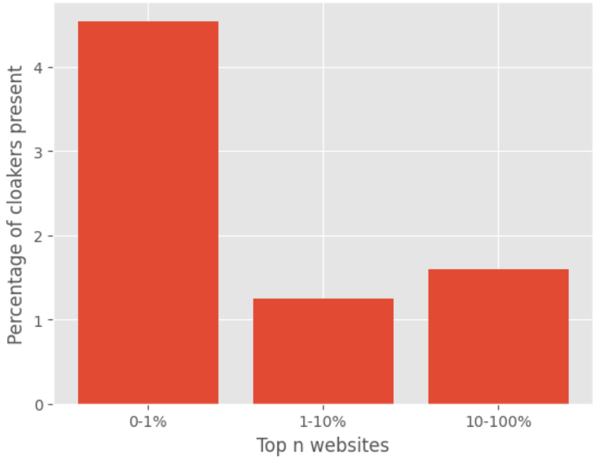

Figure 6.3: Cloakers per ranking in the Alexa dataset.

6.1.3 TTL

Upon examining the TTL values as seen in Figure 6.4, several observations can be made. Firstly, the TTL values vary significantly across the listed domains, ranging from as low as 10 seconds to as high as 1800 seconds (30 minutes). This variation suggests that different cloaking domains employ different caching strategies and have different preferences for the duration of DNS resolution caching.

Domains like `squarespace.com`, `plxserve.com`, and `affex.org` have relatively low TTL values, indicating that their DNS resolutions need to be refreshed more frequently. On the other hand, domains such as `criteo.com`, `wt-eu02.net`, and `buysellads.com` have higher TTL values, indicating that their DNS resolutions are cached for a longer period.

The presence of both short and long TTL values among cloaking domains could indicate different strategies employed by cloakers. Some domains might opt for shorter TTLs to ensure frequent updates and adaptability, while others might choose longer TTLs for improved efficiency and reduced DNS resolution overhead.

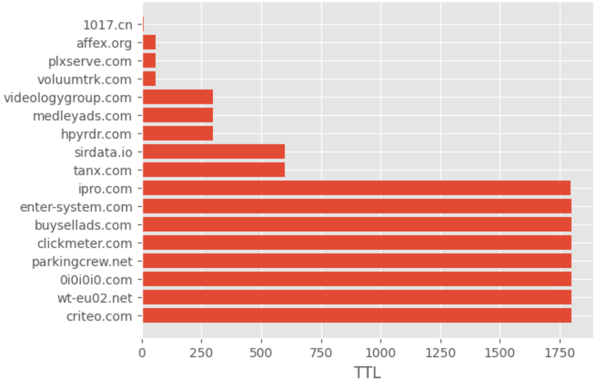

Figure 6.4: TTLs of cloakers in the Alexa dataset.

6.2 Dutch

In the Dutch dataset, we can observe a similar pattern to the Alexa dataset with respect to the presence of cookies and embedded objects. Figure 6.5 illustrates a Sankey diagram to visualize the flow of these elements, while Table 6.3 presents a distribution table for a more detailed analysis. The findings are as follows:

According to the Sankey diagram and the distribution table, the Dutch dataset reveals a substantial presence of CDN-related activity primarily in the embedded objects. This suggests that a significant portion of the content delivery network is utilized for serving embedded objects.

Cloaking is, however, less evenly split amongst cookies and embedded objects. In this case, most cloaking originates from embedded objects, which might more difficult to differentiate in Figure 6.5, therefore have a look at Table 6.3.

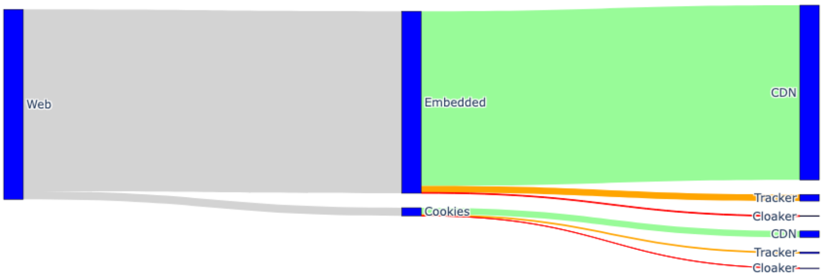

Figure 6.5: Sankey diagram of the Dutch dataset.

Table 6.3: Activity distribution for the Dutch dataset.

	Cookies	**Embedded**
CDN	542	14772
Tracker	129	556
Cloaker	5	59

6.2.1 Cloaking encounters

Unlike in the Alexa dataset, we don't have a massive outlier such as `wl.hpyrdr.com` this time. This time, the most popular cloakers are `polarmobile.com`, `criteo.com` and `sirdata.io`. All having originated from embedded objects.

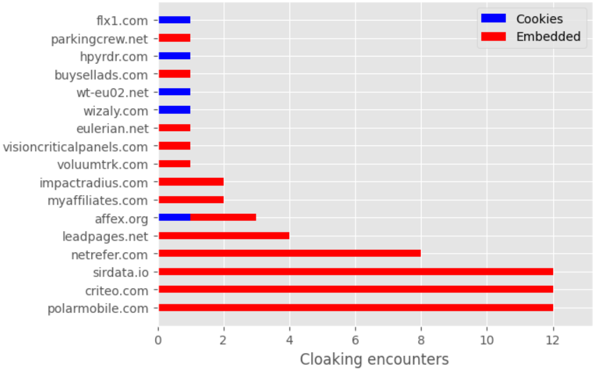

Figure 6.6: Bar graph containing the cloakers and their frequencies in the Dutch dataset.

6.2.2 Ranking

As with the Alexa dataset, more cloaking happens on more popular websites as seen in Figure 6.7. With 2.1% of cloaking happening in the top 1% of websites, which equals to 97 websites. This number is higher than the top 1% of websites in the Alexa dataset due to the higher number of reachable pages in the Dutch dataset.

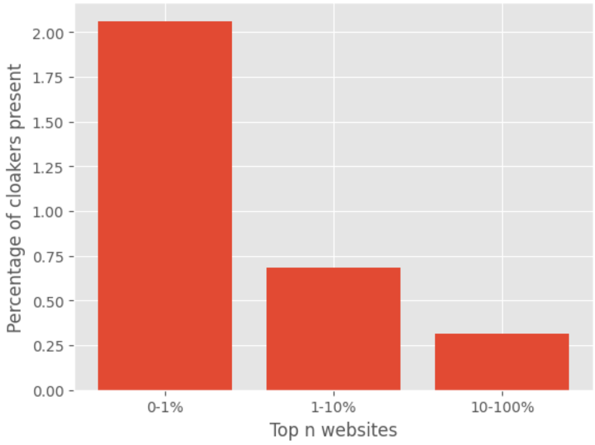

Figure 6.7: Cloakers per ranking in the Dutch dataset.

6.2.3 TTL

Similar to the Alexa dataset, the TTL values in for this dataset also exhibit variation among the domains as seen in Figure 6.8. Some domains, such as `voluumtrk.com`, `impactradius.com`, and `myaffiliates.com`, have very short TTL values, indicating that their DNS resolutions are cached for a very brief period, possibly for faster updates and improved responsiveness.

On the other hand, domains like `criteo.com`, `buysellads.com`, and `leadpages.net` have longer TTL values, suggesting that their DNS resolutions are cached for a longer duration. This can reduce the frequency of DNS resolution requests and potentially improve performance.

When comparing the TTL values between the two lists, it is interesting to note that domains like `wt-eu02.net`, `criteo.com`, `buysellads.com`, `affex.org`, and `parkingcrew.net` appear in both lists with similar or slightly different TTL values. This suggests that these domains consistently maintain their TTL settings across different datasets or experiments, indicating a deliberate choice in their caching strategies.

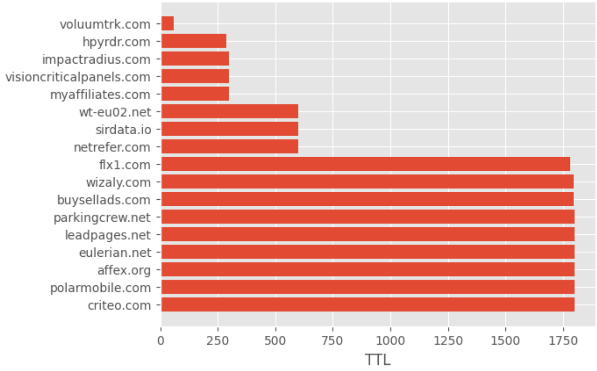

Figure 6.8: TTLs of cloakers in the Dutch dataset.

6.3 Rijksoverheid

Fortunately, no instances of cloaking have been detected on any of the websites within the Rijksoverheid dataset. However, it is worth noting that trackers are still present, which can be considered potentially harmful to user privacy. These trackers primarily originate from cookies, as illustrated in both Figure 6.9 andTable 6.4. This analysis provides valuable insights into the distribution and sources of trackers within mgovernmental websites, highlighting the importance of monitoring and addressing potential privacy concerns.

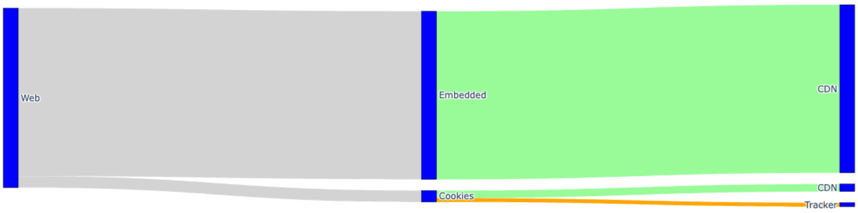

Figure 6.9: Sankey diagram of the governmental dataset.

Table 6.4: Activity distribution for the Governmental dataset.

	Cookies	**Embedded**
CDN	28	633
Tracker	15	0
Cloaker	0	0

As no instances of cloaking-based tracking have been identified within the Rijksoverheid dataset, there is no need to perform ranking and TTL analysis for this specific dataset. The absence of cloaking reinforces the notion that these governmental websites prioritize transparency and user privacy, providing a safe browsing experience for visitors.

6.4 G20

According to the Sankey diagram in Figure 6.10 and the distribution table in Table 6.5, the occurrence of cloaking within the G20 dataset is relatively low. Out of the total websites analyzed, only three websites were identified as having instances of cloaking. Interestingly, two of these websites originate from Germany (`ale.ombudsrat.de` and `berlin.de`), while the third website is located in Michigan, USA (`michiganlottery.com`).

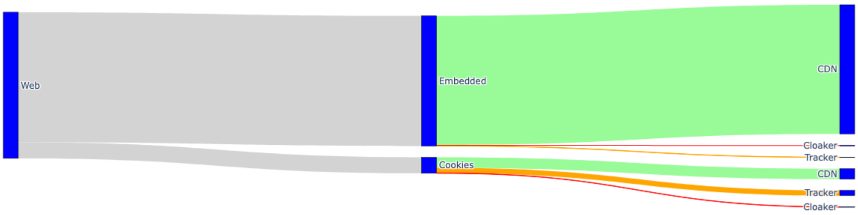

Figure 6.10: Sankey diagram of the G20 dataset.

Table 6.5: Activity distribution for the G20 dataset.

	Cookies	Embedded
CDN	269	6625
Tracker	135	15
Cloaker	2	1

6.4.1 Cloaking encounters

Figure 6.11 does not provide a lot of insight into the cloakers precence in the G20 dataset as there is little to go off of. We will, however, look at the cloakers in this dataset later on in Section 6.7.

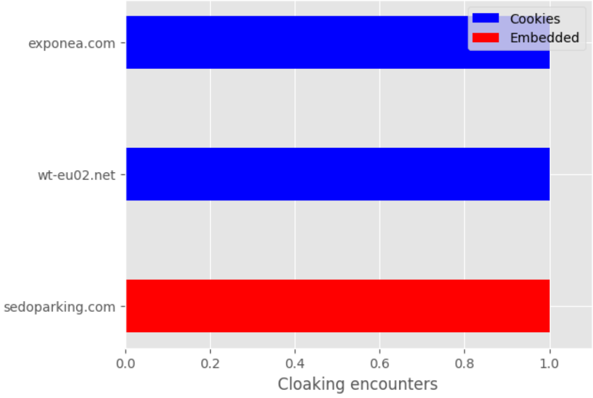

Figure 6.11: Bar graph containing the cloakers and their frequencies in the G20 dataset.

6.4.2 TTL

Based on the TTL values presented in Table 6.6, it is evident that the majority of the values cluster around 1800. Given the consistency of these values, it is possible that the actual TTL values for these domains are indeed 1800. However, it is important to consider the potential impact of caching during the process of accessing the domains. The slight variation in the retrieved TTL values could be attributed to the caching mechanism, which may have slightly reduced the TTL values at the time of retrieval. Therefore, a bar graph may not provide any additional insights, as the TTL values appear to be relatively stable and consistent within the observed range.

Table 6.6: TTL table for the cloakers in the G20 dataset.

Cloaker	TTL
`sedoparking.com`	1790
`wt.eu02.net`	1797
`exponea.com`	1762

6.5 Covid

Only 1 website has been subject to cloaking in the Covid dataset. Which is the German website `welt.de/themen/coronavirus-epidemie`. We will look at the DNS resolution in Section 6.7. Furthermore, some regular tracking still takes place, mostly through cookies present on the websites.

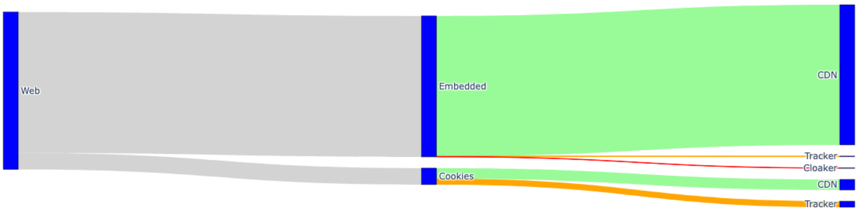

Figure 6.12: Sankey diagram of the Covid dataset.

Table 6.7: Activity distribution for the Covid dataset.

	Cookies	**Embedded**
CDN	37	493
Tracker	21	2
Cloaker	0	1

6.5.1 Cloaking encounters

Since we have only encountered one case of cloaking-based tracking in the Covid dataset, we will leave out the bar plot. The only cloaker we have come across is `transmatico.com` on the domain `welt.de/themen/coronavirus-epidemie`, which is a news website. We will briefly discuss `transmatico.com` in Section 6.7 later on.

6.5.2 TTL

In the Covid dataset, we have observed only one instance of cloaking, which is a relatively low occurrence. As a result, there is insufficient data to present in the form of figures and tables. However, we can still provide some insights regarding the cloaker found in this dataset, namely `transmatico.com`.

The TTL (Time to Live) value associated with the domain `transmatico.com` is considerably longer than what we have observed in previous cases. With a TTL of $21,600$ (equivalent to 6 hours), it indicates that the domain is cached for a significant period of time. This prolonged TTL suggests that the DNS resolution for this domain is stored in caches for an extended duration, reducing the need for frequent DNS lookups.

While the reason for such a long TTL is not immediately apparent, it is possible that the website administrators have intentionally set it to optimize performance or reduce server load. Alternatively, it could be a configuration specific to the DNS infrastructure of the domain.

Given the limited presence of cloaking in this dataset, it is important to exercise caution when drawing conclusions about cloaking behavior in general. However, the extended TTL value of `transmatico.com` provides an interesting observation that may warrant further investigation in future studies.

Furthermore, we will briefly discuss this case within the prominent example Section 6.7.

6.6 Fakuda

The Fakuda dataset, in comparison to the previous study conducted in [6], exhibited a significant decrease in the prevalence of cloaking-based tracking as seen in both Figure 6.13 and Table 6.8. While the previous study reported that all domains in the dataset contained instances of cloaking, our analysis of the Fakuda dataset revealed that only 11 out of the total websites showed the presence of a cloaker. This reduction in cloaking instances suggests a potential shift or change in the cloaking behavior over time.

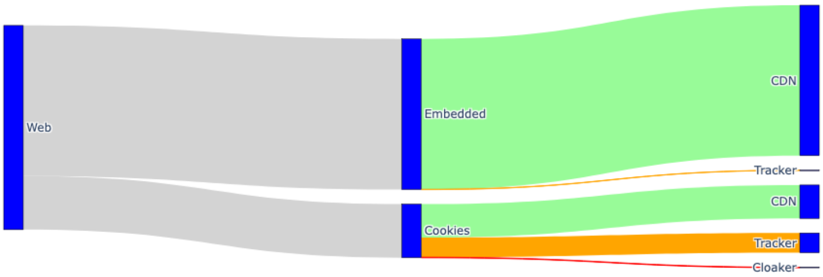

Figure 6.13: Sankey diagram of the Fakuda dataset.

Table 6.8: Activity distribution for the Fakuda dataset.

	Cookies	**Embedded**
CDN	886	3984
Tracker	552	5
Cloaker	11	0

6.6.1 Cloaking encounters

As illustrated in Figure 6.14, it is evident that all instances of cloaking within the Fakuda dataset originate from cookies. There are a total of three distinct cloakers identified: `enter-system.com`, `affex.org`, and `videologygroup.com`. These cloakers are responsible for the cloaking behavior observed within the dataset.

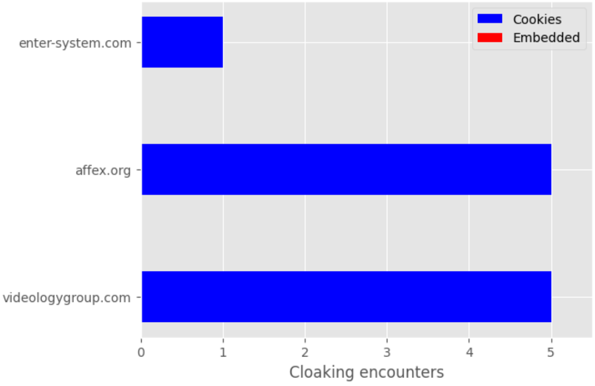

Figure 6.14: Bar graph containing the cloakers and their frequencies in the Fakuda dataset.

6.6.2 TTL

The TTLs differ quite a lot in this dataset. The cloaker `enter-system.com` has by far the longest TTL as seen in Figure 6.15.

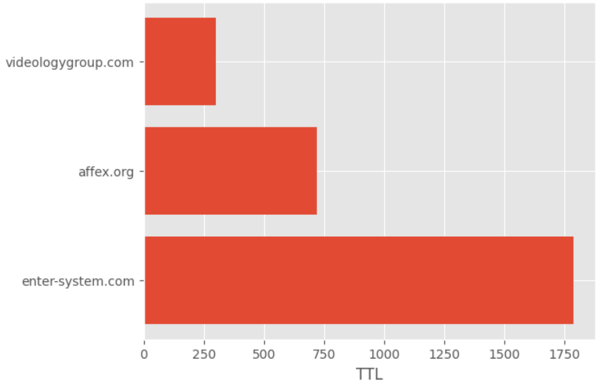

Figure 6.15: TTLs of cloakers in the Fakuda dataset.

6.7 Prominent examples

In this section, we will analyze some prominent cases of cloaking we have encountered during our experiments.

6.7.1 Domain: `aliexpress.com`

Aliexpress is a highly popular website that generates significant traffic. During our exploration of the Alexa dataset, we discovered embedded objects associated with the domain `www.alimama.com` on this website. The resolving process of this domain involves several redirects before reaching an IP address. Notably, one of these intermediate domains, with the top-level domain `tanx.com`, is identified as a tracker as seen in Figure 6.16. Consequently, this particular instance is classified as a case of cloaking.

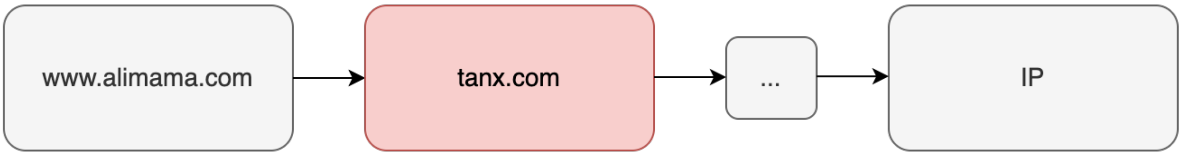

Figure 6.16: DNS resolution of `www.alimama.com`.

Alieexpress is not the only website on which this particular cloaking chain occurs. The embedded object by the `www.alimama.com` domain is also present on other Ali-related websites.

6.7.2 Cloaker: `wl.hpyrdr.com`

The most prevalent cloaker encountered in the Alexa dataset is `wl.hpyrdr.com`. This cloaker appears on websites a total of 86 times within embedded objects and 29 times within cookies. Interestingly, on 14 websites, it is present both as an embedded object and a cookie, resulting in a combined total of 97 websites where cloaking by `wl.hpyrdr.com` is detected.

A noteworthy observation is that the majority of these websites fall into the category of adult content. Additionally, several domains are resolved to the cloaker domain `wl.hpyrdr.com`, as illustrated in Figure 6.17. This finding further supports the prevalence and significance of the `wl.hpyrdr.com` cloaker within the Alexa dataset.

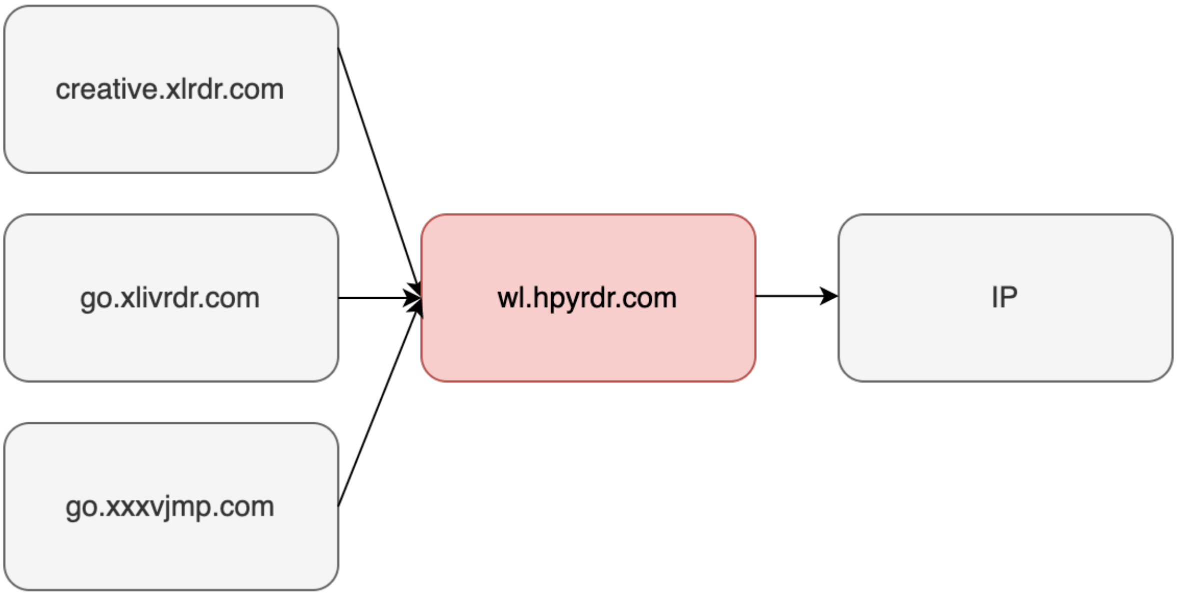

Figure 6.17: DNS resolutions to `wl.hpyrdr.com`.

6.7.3 Cloaker: `criteo.com`

Criteo is another frequent tracker/cloaker that we have encountered on numerous occasions during exploration of the Alexa and Dutch datasets. It appeared 14 in the Alexa dataset and 12 times in the Dutch dataset. It originates from an embedded object associated with the domain `www.manage.com`. In Figure 6.18, you can examine the DNS resolution process for this domain. Within the resolution chain, multiple top-level domains of `criteo.com` are present, indicating the involvement of Criteo in this particular case of tracking/cloaking.

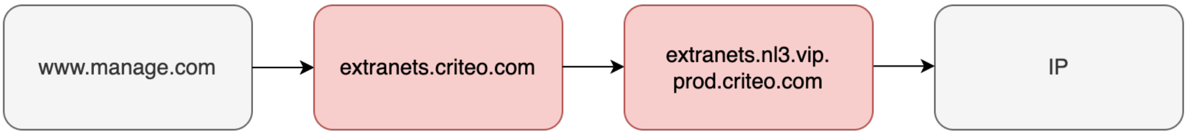

Figure 6.18: DNS resolution of `www.manage.com`.

This case is particularly interesting, especially in light of recent events where Criteo, one of the identified cloakers in the dataset, has been fined €40 million for GDPR infringements [37]. This highlights the importance of studying and addressing the issue of cloaking-based tracking, as it can have legal and regulatory implications related to privacy and data protection.

6.7.4 Cloaker: `web.polarmobile.com`

The cloaker `web.polarmobile.com` always originates from an embedded object by the domain of `privacy.polar.me` on websites in the Dutch dataset. Most of these websites are magazine website. For example: `weekend-online.com`, `zin.nl` and `hpdetijd.com`. The DNS resolution can be seen in Figure 6.19 below.

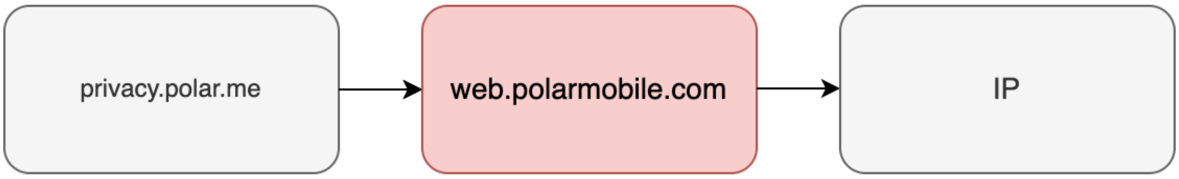

Figure 6.19: DNS resolution of `privacy.polar.me`.

6.7.5 Cloaker: `sirdata.io`

What is also interesting is the cloaker `sirdata.io`. This cloaker appears on mostly the same webpages as `polarmobile.com` and `criteo.com` through embedded objects. It is a little less subtle than other cloakers since only the change is `www.sirdata.com` to `www.sirdata.io`. This is technically a case of cloaking under our definition since the domain differs and the resolved domain exists in our tracking list. For completeness purposes the DNS resolution can be seen in Figure 6.20.

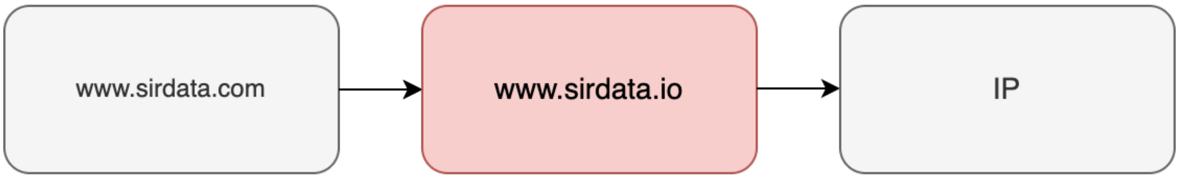

Figure 6.20: DNS resolution of `www.sirdata.com`.

6.7.6 Cloaker: `transmatico.com`

The only cloaker present in the Covid dataset is `transmatico.com`, more specifically `cname.transmatico.com`. The originating website is `welt.de/themen/coronavirus-epidemie`, which contains an embedded object by the domain of `sonderthemen.welt.de`. This resolves to the cloaking domain as seen in Figure 6.21.

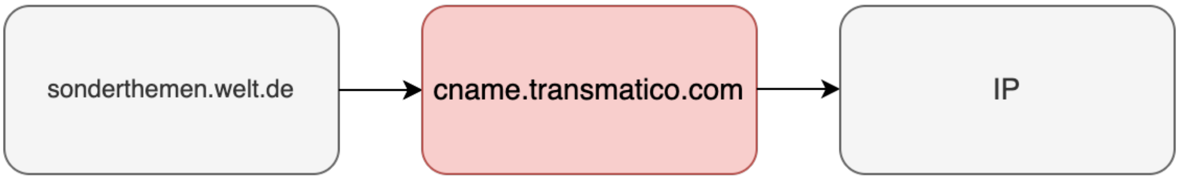

Figure 6.21: DNS resolution of `sonderthemen.welt.de`.

6.7.7 Domain: `www.berlin.de`

The official website/portal of Germany's capital, Berlin, contains a cookie by the domain of `w7.berlin.de`. This cookie redirects to a domain that is considered to be a tracker. This domain is different than the website's and is thus considered to be cloaking. The DNS resolution process for this domain is as follows, as shown in Figure 6.22:

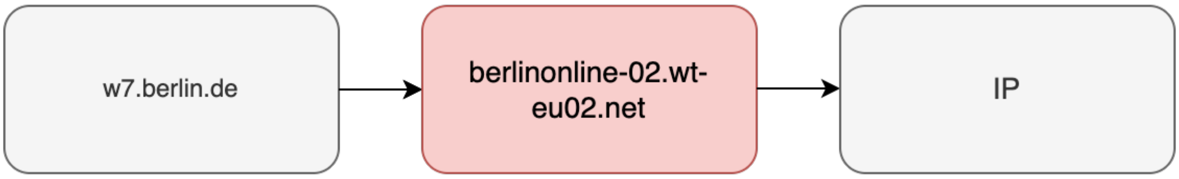

Figure 6.22: DNS resolution of `w7.berlin.de`.

This is a clear example of CNAME cloaking occurring on official websites. It is important to note that this may not necessarily be done with malicious intent but could be a result of negligence or lack of attention by the website administrators.

Discussion 7

This chapter aims to provide a comprehensive summary of the findings obtained in relation to the research questions posed in Chapter 1 on page 1. Additionally, we will address the limitations of our research and propose areas for future improvement.

7.1 Research Questions

In this section, we will comprehensively answer three research questions that delve into the use of CNAME cloaking in the wild. These research questions form the core of our analysis and provide valuable insights into the prevalence and characteristics of CNAME cloaking. Let us explore each research question in detail:

Research Question 1
How prevalent is CNAME cloaking on the web?

The results of our analysis reveal significant variations in the prevalence of cloaking across different datasets. The Alexa dataset exhibits the highest percentage of cloaking, with a rate of 1.59%. In comparison, the Dutch dataset shows a lower incidence of cloaking at 0.37%. As expected, the governmental and COVID-related datasets have considerably lower levels of cloaking, aligning with their nature and purpose.

Furthermore, our analysis highlights the dynamic nature of cloaking-based tracking behavior over time. A comparison of our results with the findings of [6] in the Fakuda dataset reveals a significant decrease in the number of cloakers over the past three years. This demonstrates that the exposure and scrutiny of cloaking practices have played a role in reducing their prevalence.

These findings underscore the importance of monitoring and addressing CNAME cloaking, as well as the potential impact of research and awareness in mitigating its prevalence and protecting user privacy.

Additionally, we conducted an analysis of the TTL values associated with the cloakers found in the datasets. The TTL represents the duration for which a DNS record is cached before it needs to be refreshed. We observed that different cloakers exhibit varying TTL values.

Some cloakers had longer TTLs, ranging up to half an hour, indicating that their DNS records are cached for a significant period of time. On the other hand, certain cloakers had shorter TTLs, suggesting that their DNS records are refreshed more frequently.

The discrepancy in TTL values among different cloakers may be attributed to various factors, such as the specific configuration choices made by the cloaker operators or the strategies they employ to evade detection or mitigate potential disruptions. Understanding these TTL patterns can provide insights into the operational behavior and resilience of cloakers, aiding in the development of effective countermeasures and detection mechanisms.

Research Question 2
What are the characteristics of websites that use CNAME cloaking?

In Section 6.7, we examined several cases of cloaking and observed a pattern where specific cloakers tend to target specific types of websites. For example, the cloaker `wl.hpyrdr.com` is predominantly found on adult websites, indicating a specific focus on this type of content. Similarly, cloakers like `polarmobile.com` appear to target magazine-like webpages.

These findings suggest that cloakers may have specific preferences or strategies in selecting their target websites. Understanding these patterns can provide valuable insights into the motivations and behaviors of cloakers. It also emphasizes the need for tailored countermeasures and detection techniques that account for the varying characteristics and preferences of different cloakers.

Research Question 3

How is cloaking distributed amongst ranking intervals?

In our analysis of the Alexa and Dutch datasets, we examined the distribution of cloaking percentages across different popularity intervals. The popularity intervals were based on the ranking of websites in terms of their traffic or popularity.

Interestingly, we found that the highest percentage of cloaking occurred within the most popular interval, specifically the 0-1% interval. This means that a significant proportion of cloaking activities is concentrated on the most popular websites.

This finding raises concerns about the prevalence of cloaking on highly visited websites and highlights the potential risks to user privacy and security. It suggests that even popular and widely recognized websites are not immune to the practice of cloaking and underscores the need for robust detection and prevention measures.

7.2 Limitations

The ever-changing nature of the web poses challenges in identifying trackers solely based on their domain. New trackers may emerge over time, unbeknownst to existing tracker lists at the time of their creation. Additionally, changes in domain names can lead to false negatives, where previously identified trackers may go undetected.

Moreover, the experimental setup incorporated a 60-second waiting period for webpage responses to ensure efficiency. However, this timeframe may have resulted in a lower number of reachable pages, particularly for international websites such as those in the G20 dataset. This effect was particularly noticeable when encountering Chinese domains within the Alexa dataset. The limitations imposed by the waiting period were necessary to strike a balance between thorough analysis and overall efficiency in data collection.

Additionally, it should be noted that multiple runs of the same dataset can yield different outputs. Various factors, such as the time of day, can contribute to these variations. To ensure a comprehensive analysis, we performed multiple runs of each dataset and combined the results, thus encompassing a broader range of observations. This approach accounts for the inherent variability in data collection and enables a more robust analysis of the findings.

7.3 Reflections

During the development of our program, we faced certain challenges, one of which involved encountering an infinite loop in some of our runs. This issue was caused by inadvertently creating a DNS loop within our program. For instance, this occurred when performing DNS resolution for the domain `www.twitter.com`. The correct resolution should have resulted in an IP address, as depicted in Figure 7.1. Identifying and addressing such issues was an important part of refining our program and ensuring accurate results.

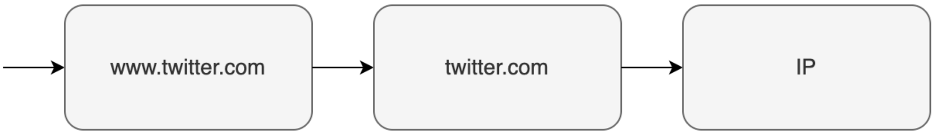

Figure 7.1: Correct DNS resolution for `www.twitter.com`.

Instead, we were always prepending **www.** in front of the domains, causing an infinite loop as seen in Figure 7.2.

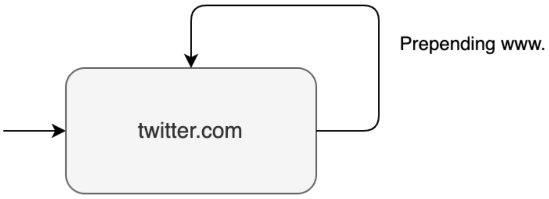

Figure 7.2: Incorrect DNS resolution for `www.twitter.com`.

After solving this issue, the program took around 2 seconds on average per website for the larger datasets. The G20 dataset took slightly longer on average, mainly due to the higher number of connection timeouts as seen in Table 6.1.

7.4 Reproducibility

As previously mentioned, the internet is a dynamic environment, constantly evolving and changing over time. Therefore, if we were to replicate the same experiments in the future, it is expected that the outcomes would differ, albeit perhaps only slightly. This highlights the importance of regularly monitoring the prevalence of CNAME cloaking to stay informed about any emerging trends or changes in behavior.

Furthermore, webpages themselves can undergo modifications, including the embedded objects they contain. These changes can also impact the results of experiments, as observed in the Fakuda dataset. In January 2020, this dataset exhibited 100% cloaking behavior. Such instances emphasize the need for ongoing research and analysis to capture and understand the dynamic nature of CNAME cloaking and its potential variations over time.

7.5 Future Work

CNAME cloaking continues to be a significant privacy concern, and further research is needed to address this ongoing issue. There are several avenues for future work that can contribute to a better understanding of CNAME cloaking and the development of effective countermeasures.

Increasing the sample size: Expanding the scope of the study by crawling a larger number of websites would be beneficial in identifying more instances of cloaking. A larger sample size would provide a more comprehensive dataset and enable the identification of potential patterns and trends related to cloaking behavior.

Pattern recognition and analysis: Analyzing the collected data for patterns and commonalities among cloakers can help in understanding their strategies and motivations. This could involve analyzing the characteristics of cloaked domains, their hosting infrastructure, or their relationship with tracking networks. Identifying such patterns could contribute to the development of more effective detection techniques.

Developing countermeasures: Developing robust countermeasures to detect and prevent CNAME cloaking is a complex task. It requires careful analysis of DNS resolutions, cookies, and embedded objects to identify potential cloakers. Future work could focus on developing machine learning algorithms or heuristics that can accurately identify suspicious DNS resolutions and link them to potential cloakers.

Evaluating the effectiveness of countermeasures: Once countermeasures are developed, it is crucial to evaluate their effectiveness in real-world scenarios. Conducting experiments and benchmarking against known cloakers can provide insights into the performance and limitations of the proposed solutions.

Collaboration and information sharing: Collaboration among researchers, industry experts, and regulatory bodies is essential to address the challenges posed by CNAME cloaking effectively. Sharing information and findings can help in creating a collective understanding of the issue and promote the development of standardized approaches and best practices.

Overall, future work should aim to expand our knowledge of CNAME cloaking, improve detection techniques, and develop effective countermeasures to safeguard user privacy and mitigate the risks associated with cloaking-based tracking.

7.6 Concluding Remarks

In conclusion, our research on CNAME cloaking and its implications for web privacy has shed light on several key aspects. We have analyzed multiple datasets, including the Alexa, Dutch, Rijksoverheid, Covid, Fakuda, and G20 datasets, to investigate the prevalence and behavior of cloakers across different types of websites.

Our findings indicate that the occurrence of CNAME cloaking varies among datasets, with the Alexa dataset showing the highest percentage of cloaking at 1.59%, followed by the Dutch dataset at 0.37%. This suggests that the extent of cloaking is influenced by factors such as website popularity and context. Additionally, our analysis reveals that cloakers tend to focus on specific types of websites, indicating a targeted approach in their tracking practices.

We have also explored the role of cookies and embedded objects in relation to cloaking. Cookies have been identified as a common medium for cloaking-based tracking, with certain domains associated with cloaking activity. Furthermore, the analysis of embedded objects has allowed us to investigate the presence of cloaking between cookies and embedded content, providing insights into the interconnectedness of these elements in the tracking ecosystem.

Throughout our research, we have encountered challenges and limitations, such as the dynamic nature of the web and the potential for false negatives in detecting cloakers. These challenges highlight the need for continuous monitoring and adaptation to uncover new and evolving cloaking techniques. Additionally, our study has identified the importance of considering privacy implications, especially when cloaking occurs on official and high-traffic websites, as observed in the Rijksoverheid and Covid datasets.

Bibliography

Here are the references in citation order Please note that discussed domains are not part of this list, since they have not provided information beside their DNS resolutions.

[1] WPBeginner. *What is DNS? and How Does DNS Work? (Explained for Beginners)*. Mar. 2020. URL: https://www.wpbeginner.com/glossary/dns/ (cited on pages 1, 4).

[2] Romain Cointepas. 'CNAME Cloaking, the dangerous disguise of third-party trackers'. In: (Dec. 2021) (cited on pages 1, 2, 8).

[3] Rebekah Houser and Daiping Liu. *CNAME Cloaking: Disguising Third Parties Through the DNS*. Oct. 2022. URL: https://unit42.paloaltonetworks.com/cname-cloaking/ (cited on pages 1, 13).

[4] *What is Load Balancing? - Load Balancing Algorithm Explained - AWS*. URL: https://aws.amazon.com/what-is/load-balancing/#:~:text=Load%20balancing%20is%20the%20method,a%20fast%20and%20reliable%20manner. (cited on pages 1, 7).

[5] Zeljka Zorz. *CNAME-based tracking increasingly used to bypass browsers' anti-tracking defenses - Help Net Security*. Mar. 2021. URL: https://www.helpnetsecurity.com/2021/02/24/browsers-anti-tracking/ (cited on page 1).

[6] Ha Dao and Kensuke Fukuda. 'Characterizing CNAME Cloaking-based Tracking on the Web.' In: *TMA*. 2020 (cited on pages 1, 10, 27, 33).

[7] *General Data Protection Regulation (GDPR) – Official Legal Text*. Sept. 2022. URL: https://gdpr-info.eu/ (cited on pages 2, 5).

[8] Infoblox. *DNS - What is DNS? Learn How Domain Name System Works | Infoblox*. Oct. 2022. URL: https://www.infoblox.com/glossary/domain-name-system-dns/ (cited on page 4).

[9] *DNS resolver*. May 2023. URL: https://www.computerhope.com/jargon/d/dns-resolver.htm (cited on page 4).

[10] *DNS Record Types: Defined and Explained - Site24x7*. URL: https://www.site24x7.com/learn/dns-record-types.html (cited on page 4).

[11] Geeky Much. 'DNS Queries — Recursive and Iterative - Networks Security - Medium'. In: (Jan. 2022) (cited on page 4).

[12] Bradley Mitchell. 'DNS Caching and How It Makes Your Internet Better'. In: (Mar. 2021) (cited on page 5).

[13] Joon S Park and Ravi Sandhu. 'Secure cookies on the Web'. In: *IEEE internet computing* 4.4 (2000), pp. 36–44 (cited on page 5).

[14] Kavya. 'What are session cookies? Do they need consent?' In: *CookieYes* (Mar. 2023) (cited on page 5).

[15] All About Cookies Editors. 'What are Session Cookies?' In: *All About Cookies* (Mar. 2023) (cited on page 5).

[16] *Proposal for an ePrivacy Regulation*. June 2023. URL: https://digital-strategy.ec.europa.eu/en/policies/eprivacy-regulation (cited on page 5).

[17] Christine Afandi. 'What are embedded objects?' In: *Wiredelta* (May 2022) (cited on page 6).

[18] Cheng Huang et al. 'Measuring and evaluating large-scale CDNs'. In: *ACM IMC*. Vol. 8. 2008, pp. 15–29 (cited on page 6).

[19] *CDN Vs Edge Server | How Does Edge Caching Work?* URL: https://www.belugacdn.com/cdn-vs-edge/ (cited on page 7).

[20] *What is a DDoS Attack? DDoS Meaning, Definition Types | Fortinet*. URL: https://www.fortinet.com/resources/cyberglossary/ddos-attack#:~:text=DDoS%20Attack%20Meaning,connected%20online%20services%20and%20sites. (cited on page 7).

[21] *Internet tracking: How and why we're followed online | Norton*. URL: https://us.norton.com/blog/privacy/internet-tracking (cited on pages 7, 13).

[22] Michal Wlosik. *First-Party Third-Party Cookies: What's the Difference? - Clearcode Blog*. Nov. 2022. URL: https://clearcode.cc/blog/difference-between-first-party-third-party-cookies/ (cited on page 8).

[23] *Third-party trackers | Firefox Help*. URL: https://support.mozilla.org/en-US/kb/third-party-trackers (cited on page 8).

[24] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 'Cookie synchronization: Everything you always wanted to know but were afraid to ask'. In: *The World Wide Web Conference*. 2019, pp. 1432–1442 (cited on page 10).

[25] Assel Aliyeva and Manuel Egele. 'Oversharing is not caring: How cname cloaking can expose your session cookies'. In: *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*. 2021, pp. 123–134 (cited on page 11).

[26] Matthias Gotze et al. 'Measuring Web Cookies in Governmental Websites'. In: *Proceedings of the 14th ACM Web Science Conference 2022*. WebSci '22. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 44–54. DOI: 10.1145/3501247.3531545 (cited on page 12).

[27] *Selenium WebDriver*. URL: https://www.selenium.dev/documentation/webdriver/ (cited on page 15).

[28] *Chrome DevTools Protocol*. URL: https://chromedevtools.github.io/devtools-protocol/1-3/Network/ (cited on page 15).

[29] Google Chrome. *Google Chrome Browser*. URL: https://www.google.com/chrome/ (cited on page 15).

[30] *Index of /archive/alexa/*. URL: https://toplists.net.in.tum.de/archive/alexa/ (cited on page 17).

[31] *Comunicatierijk*. URL: https://www.communicatierijk.nl/vakkennis/rijkswebsites/verplichte-richtlijnen/websiteregister-rijksoverheid (cited on page 17).

[32] Matthias Gotze et al. *"Measuring Web Cookies in Governmental Websites"*. URL: https://govcookies.github.io/ (cited on page 17).

[33] *About G20*. URL: https://www.g20.org/en/about-g20/ (cited on page 18).

[34] Ha Dao and Kensuke Fukuda. *GitHub - fukuda-lab/cname_cloaking*. URL: https://github.com/fukuda-lab/cname%5C_cloaking (cited on page 19).

[35] Justdomains. *GitHub - justdomains/blocklists: Domain-ONLY Filter Lists (for use with DNS / Domain blocking tools)*. URL: https://github.com/justdomains/blocklists (cited on page 19).

[36] Petros Gigis et al. 'Seven years in the life of Hypergiants' off-nets'. In: *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 2021, pp. 516–533 (cited on page 19).

[37] 'Advertising Company CRITEO fined €40 Mio'. In: (June 2023) (cited on page 30).