Universiteit Leiden
The Netherlands

Opleiding Informatica

Web Privacy in the Netherlands: A Study of Tracker Dynamics

Ashraf El Madkouki

Supervisors:
Dr. A. Saxena & F. Corriera

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/08/2025

Abstract

This thesis presents a comprehensive investigation of web tracking practices across the Dutch web. By analysing over 1,8 million ".nl" domains against a curated list of 8,069 known third-party trackers, we quantify both the prevalence and structural features of online tracking. The dataset was gathered by developing a custom, multithreaded Python scraper to collect HTML content and extract external resources from over 30 million requests. Our empirical findings demonstrate that Dutch websites load, on average, 3,04 trackers each, predominantly operated by global platforms (98,2% of all instances), with local trackers accounting for only 1,8%. Among the most frequently observed trackers are those operated by Google and Meta, highlighting their dominance in the online tracking ecosystem. We then create a network using this data, with webdomains as nodes and shared webtrackers as edges. Network analysis reveals a densely interconnected, small-world structure (average path length 2,01, diameter 5, density 0,12) dominated by a handful of hub domains. Centrality measures (degree, betweenness, closeness) highlight major content-delivery and analytics platforms as critical conduits for cross-site tracking. Community structure uncovers a pronounced modularity: five large communities encompass over 30% of nodes, while the majority of clusters remain small. The analysis shows that the Dutch web-tracking ecosystem is largely reliant on a small set of tech-giants, such as Google and Meta. We conclude that targeting these few trackers may yield the greatest reduction in the tracking of personal data. Our work contributes large-scale research on Dutch web-tracking and can serve as a foundation for regulatory interventions.

Contents

1 Introduction

When the internet was made available for public usage in 1993, it consisted of a simple network of static documents. This era of the web was referred to as "web 1.0". In 2004, the term "web 2.0" was popularised to describe a web of dynamic, user-generated content and new monetization tools, such as targeted advertisements.

In today's data-driven economy, personal information has become a valuable asset for companies seeking to take part in the lucrative businesses that comes with it, like targeted advertising. To collect as much online data as possible and monitor user behaviour, companies use what is known as web tracking, that is "The practice when some content ('trackers') embedded in a webpage recognizes the users visiting the page.[22]" For example, when a user is reading a blog online, the website may automatically load a script from an advertisement network. This script can record the user visiting the website and how long they spend reading. This information is used to build a profile based on a user's interests and to show them more personalised advertisements in the future. According to the GDPR [13], the collection of online data through web-tracking can put users at risk of identity theft, fraud or discrimination. To avoid harming users' rights, it is crucial for websites and advertisers to abide by the GDPR data protection laws.

Although HTTP cookies are the most common example, web tracking methods can be distinguished into two main categories: stateful and stateless tracking [26]. Stateful tracking techniques, commonly known as cookies, store small strings of text in the client browser to record session state, preferences, or identifiers. This data is sent back to the server on subsequent requests [7]. Stateless tracking, or browser fingerprinting, "is the process of collecting information through a web browser to build a fingerprint of a device" [19]. By combining attributes such as user-agent strings, screen resolution, and installed fonts, fingerprinting creates a unique identifier that can re-identify users without relying on stored cookies. Besides these two classes of web trackers, third-party trackers have become increasingly prevalent. Unlike first-party trackers, which are owned and managed by the domain, third-party trackers can monitor users across multiple domains [8]. This evolution has made it significantly more difficult for users to prevent their data from being collected.

The new data economy has benefited many businesses and spawned new ways to engage with the online world. A common practice in today's digital economy is to offer seemingly "free" applications. While requiring no upfront payment, users instead "pay" with their personal data [14]. While this may seem advantageous for users, disclosing personal data raises serious privacy concerns. If user data falls into the wrong hands, they could become a victim of identity theft, by having fraudulent purchases made in their name, become victim to reputational damage from exposed personal information, or even face physical harm [9].

1.1 Research Problem

Despite extensive global measurements of tracking on the most popular sites, there is a lack of large-scale, country-specific studies. Several studies have researched global and historical data on web-tracking. Su et al. [34] studied the increase in tracking on educational websites, while Lerner et

al. [20] conducted a historical analysis on web-tracking. Others examine more common web-tracking techniques and measure their efficacy: Sanchez-Rola et al. [26] provides an overview of web-tracking techniques and applications, and Cahn et al. [7] studies how cookies are injected into the client browser. However, little to no research gives insight into the inner workings of web-tracking in the Netherlands, which, due to regulation from the EU and Autoriteit Persoonsgegevens (AP), can exhibit unique online behaviour. This thesis will address this gap in measurement of the Dutch web-tracking ecosystem and inform policymakers and regulators with local insights to improve privacy interventions in the Netherlands. The research questions that we aim to answer are as follows:

- **RQ1:** Which web-trackers are most prevalent on Dutch websites?
- **RQ2:** What is the difference in the usage of global versus local trackers on Dutch websites?
- **RQ3:** How does web-tracking vary across different categories of web-trackers?

We analyse over 1,8 million domains under the .nl top-level domain using an extensive list of more than 8,000 known tracker domains. Furthermore, we identify the top local and global trackers and use a subset of categorised trackers to see the most common applications of web-tracking. This thesis will provide an overview of prior research on web-tracking, its applications and regulatory and privacy issues. Then we will discuss the method of gathering the data and processing it to obtain a clean dataset. Lastly, we analyse our data and in particular the large-scale network. We use our findings to answer the research questions and discuss the conclusions in the final chapter

2 Related Work

Web-tracking has been extensively covered in past works, especially cross-domain studies. However, research on a country level are rare. In this chapter, we will review prior works on web-tracking from 3 different angles. Firstly, we will discuss large-scale empirical studies on web-tracking. Next we discuss works that aim to provide an understanding of the mechanisms of web-tracking. Finally we discuss studies that focus on the risks of web-tracking from a privacy and economic standpoint.

2.1 Measurements and Network Analyses

Without much notice from the general public, web tracking has increased significantly in our daily online usage. Many studies have sought to quantify the prevalence of web-tracking on different platforms and across different times. Su et al. [34] developed a framework to measure web-tracking activity on more than 17,000 educational websites and compared it to a matched control group. They found a significant increase in third-party tracker intensity over time. They suggest that this growth is attributed to the increase in interactive features and raises concerns about privacy risks in education. Lerner et al. [20] conducted a twenty-year "archaeological" study of web tracking. Analysing the data from their tool, TrackingExcavator, reveals a substantial rise in third-party trackers, which have become more complex in both data collection and behaviour. They argue that understanding the historical context of web-tracking is crucial in any policy discussion surrounding it. Yang et al. [37] compared web tracking on mobile devices and desktop browsers, showing that mobile tracking has reached levels comparable to desktop tracking and poses potentially more severe privacy risks due to continuous location monitoring and the collection of sensitive data. They suggest that users should be informed about the privacy risks of web-tracking, especially on mobile devices. Network analysis has been used to understand the dynamics of different complex systems, including offline social networks [4, 31], social media networks [12, 21, 2], banking transaction networks [30], criminal networks [32], collaboration networks [10, 28], dark networks [24, 6], and terrorist networks [15]. In this work, we use network analysis techniques to understand the dynamics of web-tracking networks.

2.2 Web Tracking Mechanisms and Techniques

Several studies have investigated the technologies and methods behind web tracking. Based on when these researches have taken place, analysing web tracking has become an increasingly important topic. Sanchez-Rola et al. [26] provide a comprehensive review of web-tracking techniques, applications, and countermeasures, distinguishing between stateful methods (cookies) and stateless approaches (fingerprinting). They find that it is crucial to understand these techniques, especially when creating policy around online privacy. Cahn et al.'s empirical study of web cookies [7] examines how cookies are injected into the client browser and how they monitor and record user behaviour. They conclude that third party trackers greatly outnumber first party trackers and are therefore a much greater risk. Laperdrix et al. [19] present an in-depth analysis of fingerprinting techniques, proving how combinations of device and browser attributes can generate unique identifiers for cross-session tracking. They suggest that protection mechanisms are in a constant arms race with new fingerprinting techniques. Castell-Uroz et al. [8] provide a comprehensive tutorial on web

tracking and how to detect and minimize it. They create ORM, an open source framework for a crawler that collects data and labels trackers, to measure web-tracking and give an overview of the top tracking domains.

2.3 Web-tracking from Privacy and Economic Perspectives

Mayer and Mitchell [22] did a combined study of the technology and the policy discussion surrounding web tracking. They reviewed the US and EU privacy frameworks (in 2012) and showed that mechanisms to protect users' privacy, like self-regulatory opt-out functions, are lacklustre. In *A Model of Data Economy* [14], Farboodi and Veldkamp highlight the growing value of user data, illustrating how transactions that appear to be "zero cost" actually involve users paying with their personal information. The paper also emphasizes the critical importance of privacy protection, as businesses increasingly seek to collect and exploit this data. *Privacy Harms* [9] by Citron and Solove define a legal framework for the different types of harms that can be inflicted on a person by means of privacy violations. This article provides a topology of harm to help courts tackle cases. Lastly, Peacock [3] suggests that current web tracking developments influences user agency and calls using the internet for private affairs entering an "unconscionable contract". This transaction is found to be unjust, as it "puts the burden of an economic transaction wholly on one side and in this case the online user."

3 Dataset

The current dataset that has been acquired consists of the scraped data of more than 1,8 million websites and more than 8,000 known tracking domains. The dataset was acquired in the following steps:

3.1 Gathering Websites

The primary objective of this study was to compile a comprehensive list of Dutch domains (i.e. domains ending in `.nl`). Determining the exact number of active `.nl` domains is challenging, since registrations and deletions happen every day.

Instead, this research leveraged the 2024 Common Crawl datasets. Common Crawl is a non-profit organisation that crawls the web and archives the data for the public to use. It captures HTML content from webpages, such as links and content, which are then stored in large datasets. To assemble as extensive a list as possible, we filter the 2024 crawls for hosts ending in `.nl`, and over 1,8 million unique Dutch domains were extracted. This was done with the help of a fellow student, Daniel Gelencser.

3.2 Finding Trackers

To identify third-party trackers on a given website, a comprehensive reference list of known trackers is required. Web trackers appear as external resource requests, among many non-tracking external links, so accuracy depends on both breadth and quality of the tracker list. For this study, we combined three authoritative sources:

1. **Su et al.** [34] provided 1,285 tracker domains identified in their empirical analysis .
2. **WhoTracksMe**, the open dataset published by Ghostery and Cliqz, contributed 5,288 third-party tracker domains [18].
3. **Disconnect**, via their publicly available tracking-protection database, added 6,379 tracker domains, each annotated with a tracker category [11].

After merging these lists and removing duplicate entries, the resulting master list contains 8,069 unique third-party tracker domains, of which 6,379 include category metadata.

3.3 Processing the Data

After collecting 1,8 million domains, we implemented a custom multi-threaded web-scraper, using Python, to retrieve, parse, and process each page. The scraper uses the `requests` library combined with `urllib3`. Retry to perform HTTP GET requests, and a `ThreadPoolExecutor` (from Python's `concurrent.futures` module; original implementation by Brian Quinlan, 2009) to parallelize

downloads while preventing race conditions via a SQLite-backed queue. This way, all HTML files of the domains are gathered.

Each HTML document is then parsed with BeautifulSoup4 to extract external links, from `<a>`, `<script>`, `<img>`, and `<iframe>` tags as well as URLs embedded in JavaScript. Each external domain is compared with our tracker database via simple SQL lookups. Discovered trackers are recorded through foreign-key relationships in the database, HTTP response codes are saved for each URL, and the raw HTML and script files are written to a structured directory for auditing and later analysis. This resulted in an SQL database of over 30 million rows, where each domain has one row for every external link that was scraped.

4 Research Method

In this chapter, we describe the methodology for processing the 1,8 million domains, constructing the network from an undirected weighted graph and extracting key metrics from it.

4.1 Building the Network

We model the tracker–domain ecosystem as an undirected, weighted graph $G = (V, E)$ using Python's networkX library, where each node represents a domain and each edge (u, v) is weighted by the number of distinct third-party trackers shared between domains u and v. Node weights and sizes are proportional to the weighted degree,

$$\deg_w(v) = \sum_{u \in V} w_{uv}.$$

To construct the tracker-domain network, we implement 2 stages of data processing in Python, with the first stage for acquiring a mapping of domains to trackers and the second stage for building the network and performing analyses. Firstly, the raw dataset consists of domain ID's and tracker ID's. The url list and tracker list are merged to create a mapping of domain to trackers. Then the list of domain-tracker mappings are filtered to only contain domains mapped to known trackers. Furthermore, the trackers list is enriched with a column `times_seen` that shows the amount of times each tracker appears in the mapping. Finally, we take a 6% random sub-sample of the full dataset (7,299 domains, 3,183,635 edges) for computational tractability. From this sub-sample we iteratively create the undirected, weighted graph by adding every domain as nodes and creating an edge for each shared tracker with other nodes.

4.2 Analysing Graph Metrics

In order to extract useful information from our data, we perform a number of analyses:

- Firstly, it is important to list the top 20 most prevalent trackers. Every tracker is sorted on `times_seen`. Then the top 20 trackers are put into a horizontal bar plot with each tracker's frequency.
- Next, we compute, for each domain, the number of unique known trackers. We plot the distribution as the count of domains on the vertical axis against the number of unique trackers in a bar chart.
- Finally, to analyse the prevalence of the tracker category, we take the subset of trackers that are categorised and count the occurrence of categories per tracker. We plot the categories and their frequency in a horizontal bar chart.

Next, we calculate key metrics from our weighted network:

1. **Average degree** is the mean number of connections per node. This shows whether nodes have few or many neighbours on average and implies a certain level of connectivity.

2. **Average weighted degree** is similar to average degree, but instead of averaging over the number of distinct edges per node, sums the weight of all edges per node and takes their mean. This captures the strength of connections in the network

3. **Graph density** is calculated by dividing the total number of distinct connections in the graph by the total number of possible edges

4. **Average path length** is the mean shortest path distance between all pairs of nodes and measures how fast information travels through the network. A lower value implies more direct connectivity between nodes

5. **Network diameter** is the longest shortest path in the network. This metric reveals the network's maximal length

6. **Degree distribution** is the distribution of the frequency of node degrees (k) as the amount nodes that have exactly (k) connections, **Weighted degree distribution**, similar to its unweighted peer, shows the frequency of nodes by their weighted degree. Plotting these in a log-log scale reveals whether the network exhibits a scale-free structure. We also perform the Kolmogorov-Smirnov test to see whether the network fits a hypothesised power-law tail. We compute the maximum distance D between the empirical and fitted distribution functions. Lastly, we evaluate the p-value to determine whether the power-law hypothesis can be rejected.

To identify structurally important domains in the network, we compute three centrality metrics:

- **Degree centrality** measures the proportion of direct connections a node has [27]. This helps to identify domains that might share trackers with many other domains and populate much of the web-tracking activity

- **Betweenness centrality** measures how often a node lies on the shortest path between other node pairs. This metric identifies possible domains that connect larger community of nodes

- **Closeness centrality** of a node is calculated by taking the inverse of the average shortest path distance from one node to all the other nodes. This indicates domains that can be easily reached through other nodes.

Lastly, we run the Louvain community detection algorithm [5] on our network, which maximises modularity in large graphs. Modularity is a metric that measures how strong a network is divided into communities. The Louvain algorithms works by assigning each node to their own community. Then repeatedly moves a node to a neighbouring community which results in the highest modularity. Based on the communities found, it turns communities into "super-nodes" for a condensed graph and repeats the process of optimising modularity until it can improve no longer. We apply the Louvain algorithm to find clusters of domains that share many common trackers and highlight the modular structure of the network.

5 Empirical Analysis

In this chapter we present the results from our analysis. We begin by discussing the prevalence of the top tracking domains in the Netherlands, then examine their categories, the division between global and local trackers and the overall distribution of unique trackers per domain. Next, we analyse the metrics obtained from our weighted common-tracker network, explaining the different characteristics it exhibits and what that implies for the Dutch web-tracking ecosystem. Finally, we explore the results of the Louvain community detection algorithm and discuss the implications of the domain clusters.

5.1 Top Trackers

Figure 1 shows the 20 most commonly observed third-party tracker domains in our dataset. Five of these domains are operated by Google (including YouTube), occupying ranks 1, 3, 5, 7, and 12. The most frequent tracker, `fonts.googleapis.com`, exceeds the runner-up, `facebook.com`, by 405.009 observations. According to Google's privacy FAQ, requests to `fonts.googleapis.com` collect the client's IP address, the requested URL, and HTTP headers (including user-agent and referer), but the Google Fonts API does not set cookies or build user profiles, its sole purpose is to serve web fonts and leverage HTTP caching to improve performance [17].

Half of the top-20 domains are operated by major technology and social-media companies (Google, Meta, Microsoft, Pinterest, and X, formerly Twitter). The remaining domains include website-building platforms such as Shopify and WordPress, indicating that many third-party requests support essential site functionality. Some noteworthy exceptions are:

- `statcounter.com`: provides visitor analytics by logging data such as timestamps, IP addresses, browser and OS versions, device information, and referer URLs to help site owners understand user behaviour [33].
- `addtoany.com`: a universal sharing-button platform that integrates with Google Analytics to report sharing events directly within the analytics dashboard, which may explain its prevalence [1]. Along with Statcounter, Addtoany provides a tracking service for commercial purposes.
- `unpkg.com`: a fast, global CDN for delivering npm packages—commonly used to load JavaScript assets in web templates—automatically mirroring every file published to npm with low latency via Cloudflare's edge network [35].
- `gmpg.org`: hosts the XHTML Friends Network (XFN) metadata profile, defining semantic values for the HTML `rel` attribute to describe social relationships, a specification originally developed by the Global Multimedia Protocols Group [16]. Meuser et al. [23] showed in a 2015 study that this domain was among the highest in Pagerank [25]. The domain being a reference to a basic HTML function possibly explains its high frequency.

When classifying the top 20 trackers by functionality, we see that 40% of top trackers, such as `fonts.googleapis.com` and `unpkg.com`, send resources to a website, boosting its performance.

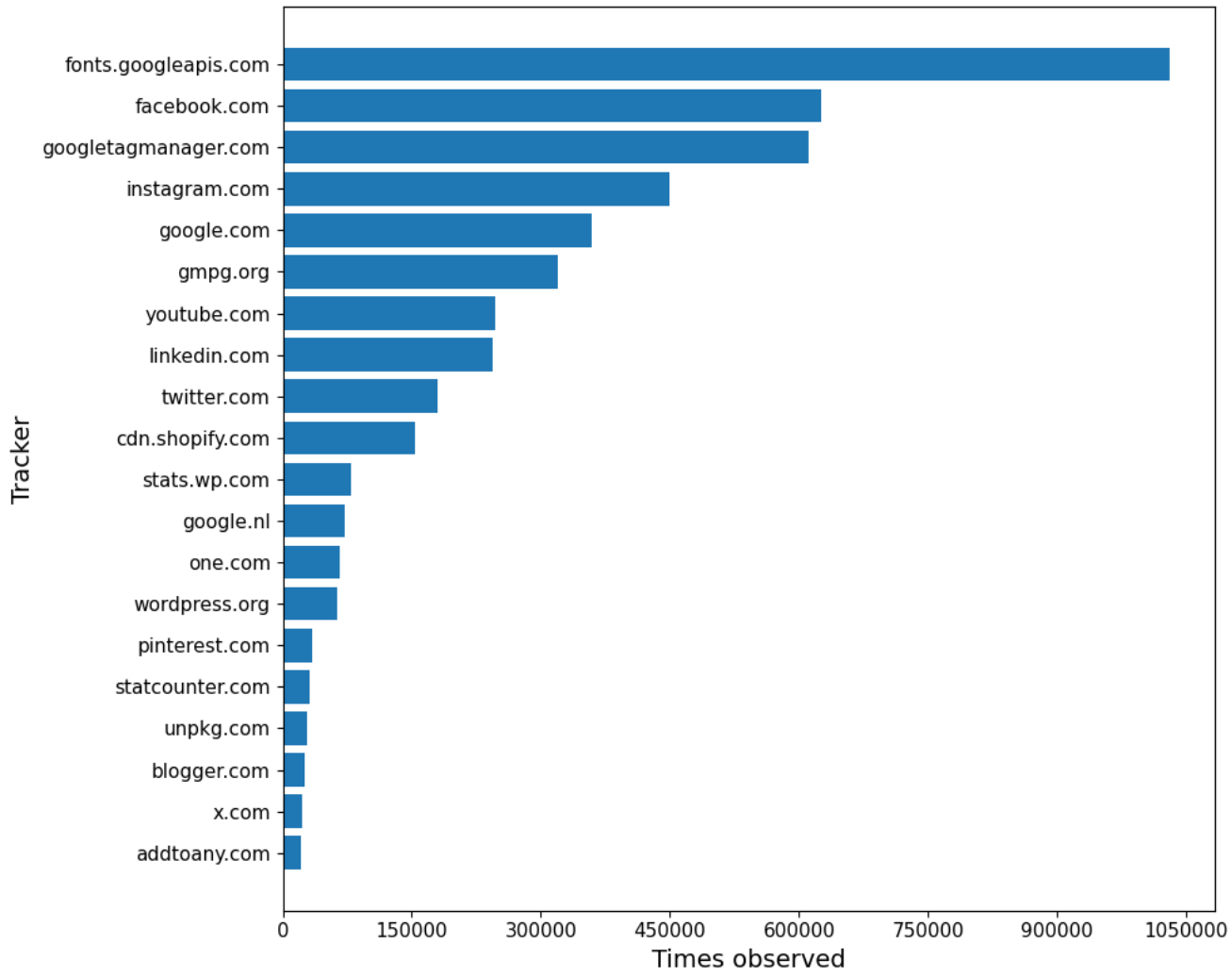

Figure 1: Top 20 trackers by amount of observations in dataset of 1,8 million web domains

30% of top tracker domains serve the purpose of driving analytics on a website (`statcounter.com`, `googletagmanager.com`). The remaining 30% support social media websites or e-commerce platforms (`facebook.com`, `shopify.com`). Among these top trackers, only one of them (`google.nl`) is a Dutch domain.

From our findings it becomes clear that large tech companies such as Google and Meta dominate the tracker activity on Dutch websites. The clear lack of `.nl` domains might imply that local online analytics business is not as strong in the Netherlands and instead relies on the global analytics market. Furthermore, it seems that website performance is a top priority on websites, with user profiling as a close second.

5.2 Distribution of Trackers

Figure 2 shows the distribution of unique trackers per domain, and Table 1 provides a concise summary of these results. For instance, in our dataset the domain *erfgoed20.nl* contains the maximum of 15 unique trackers, whereas *0202338046.nl* contains only one. The distribution is markedly right-skewed: the mode is 1, the mean is 3.04, and the median is 3—indicating that, on average, Dutch websites include up to three trackers.

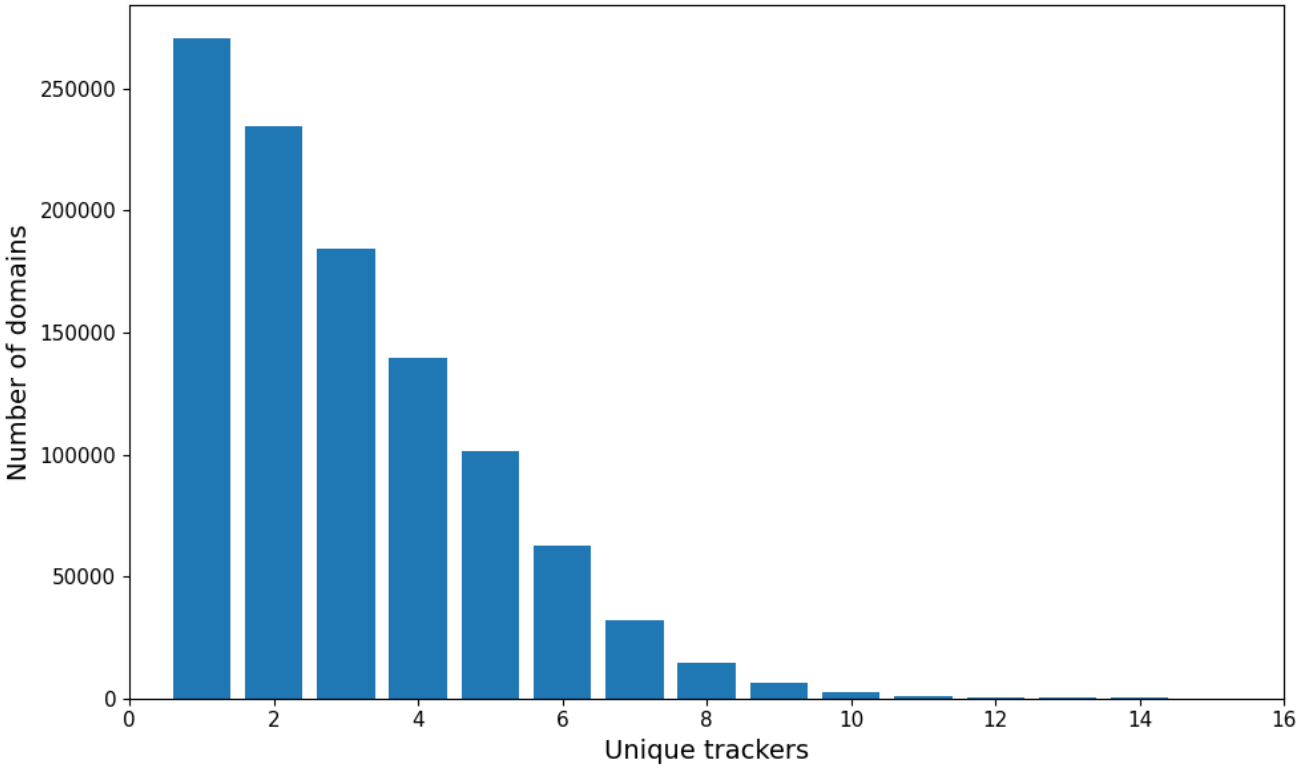

Figure 2: Distribution of unique trackers per individual domain

To distinguish tracker origin, we classify them as *local* (domains ending in `.nl`) or *global* (all other domains). Of the 8,069 distinct trackers, only 40 (0,5%) are local, while 8,029 (99,5%) are global. Considering all 5,114,947 tracker instances, 92,526 (1,8%) originate from local domains and 5,022,421 (98,2%) from global domains (see Table 1 for details).

Figure 3 summarizes the distribution of the 8,069 distinct trackers in the Disconnect dataset. Of these, 1,690 (20,9%) remain uncategorised, leaving 6,379 trackers distributed across 21 defined categories. Among the classified trackers:

- Advertising trackers are by far the most common, with 3,754 trackers (58,9% of categorized trackers).
- Analytics trackers (380) and site-analytics trackers (525) together account for 905 trackers (14,2%).

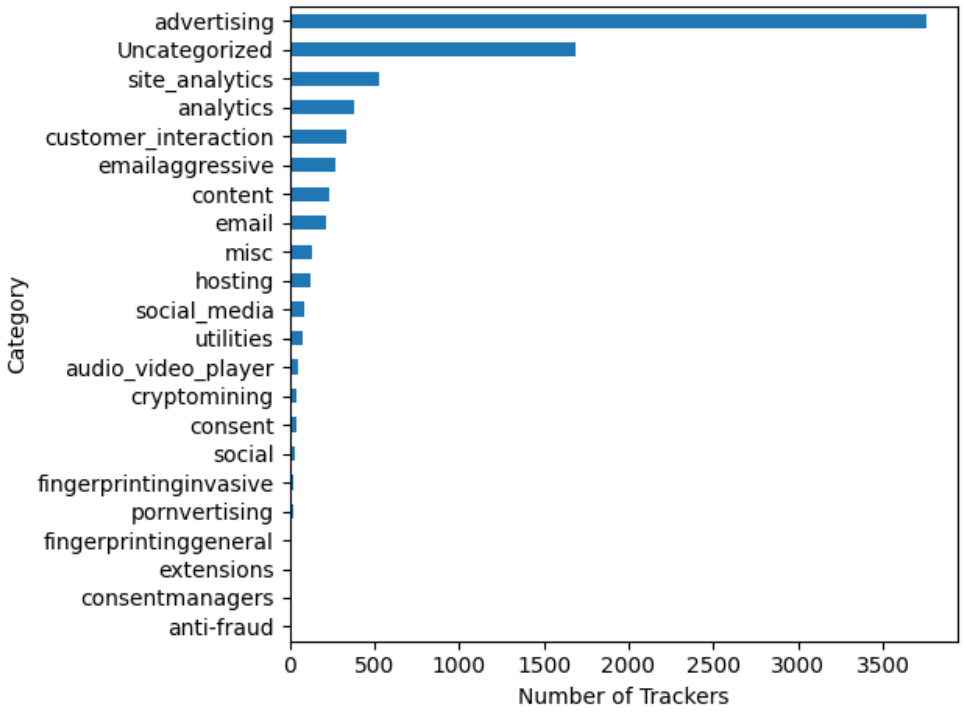

Figure 3: Categories of trackers, ranked by amount of distinct trackers

- Customer-interaction trackers make up 337 entries (5,3%), while email trackers (214) and aggressive-email trackers (268) together total 482 trackers (7,6%).
- Each of the remaining categories contributes less than 3% of the total.

This distribution underscores the web's heavy reliance on advertising for revenue generation and on analytics for monitoring user behaviour. More advanced techniques, such as fingerprinting (both general and invasive) and cryptomining, collectively represent fewer than 2% of all trackers.

5.3 Network Analysis

In order to understand the structure of the Dutch web-tracking ecosystem, we model the domain-tracker network in an undirected, weighted graph. We will discuss the metrics found in this network and examine the degree distributions to describe characteristics of the network.

Table 2 reveals an average degree of 872 and an average weighted degree of 873. This implies a high level of interconnectivity. Furthermore, the similarity in values for the weighted and unweighted average degree's show that almost all nodes are connected to the same trackers, thus yielding the same weights. The interconnectedness of the network is further strengthened by the graph density of 0,12 and network diameter of 5. This means that out of all possible connections, 12% of them are present in this network and the longest path from one end of the graph to the other is 5. With the average path length of 2, this means the network exhibits small world properties; High

Table 1: Summary of Unique-Tracker Distribution per Domain and Origin of Trackers

Metric		**Value**	**%**
Distribution of Unique Trackers per Domain			
Mode	—	1	—
Mean	—	3,04	—
Median	—	3	—
Maximum	`erfgoed20.nl`	15	—
Minimum	`0202338046.nl`	1	—
Origin of Trackers			
Distinct trackers (total)	—	8,069	100%
Local trackers	`.nl` only	40	0,5%
Global trackers	Other domains	8,029	99,5%
Tracker instances (total)	—	5,114,947	100%
Local instances	From `.nl` domains	92,526	1,8%
Global instances	From other domains	5,022,421	98,2%

Statistic	**Value**
Number of nodes	7.299
Number of edges	3.183.635
Average degree	872
Average weighted degree	873
Network diameter	5
Graph density	0,12
Average path length	2,01

Table 2: Key network characteristics for the tracker–domain network.

interconnectivity, with relatively small paths within.

Figures 4 and 5 reveal a relatively right-skewed distribution. A few domains have a vast amount of higher degree and weighted degree nodes, while a large majority of nodes form the mid range. This middle range of moderate degree nodes explains why the average (weighted) degree remains relatively high. While the (weighted) degree distribution is relatively right-skewed, it deviates from a scale-free configuration. This is most likely due to the prominent middle range of moderate degree nodes. These diagnostics imply that targeting the highest degree nodes can mitigate most of the tracking activity, but more moderate (weighted) degree nodes should not be overlooked, as their prevalence cannot be underestimated.

Observing the centrality metrics shows a correlation between the nodes with the highest degree centrality and the amount of embedded social media links on a website [29]. Examining the number one and two domains on degree centrality shows a large amount of social media links. Comparing these domains to our dataset, it reveals that all trackers embedded in these domains appear in

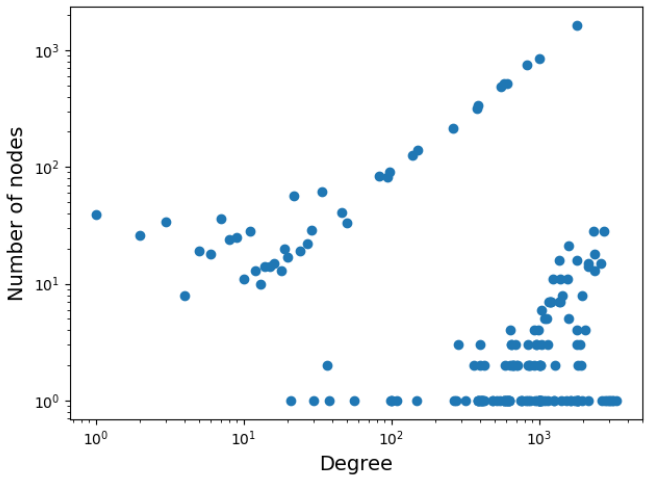

Figure 4: Degree distribution in log-log scale

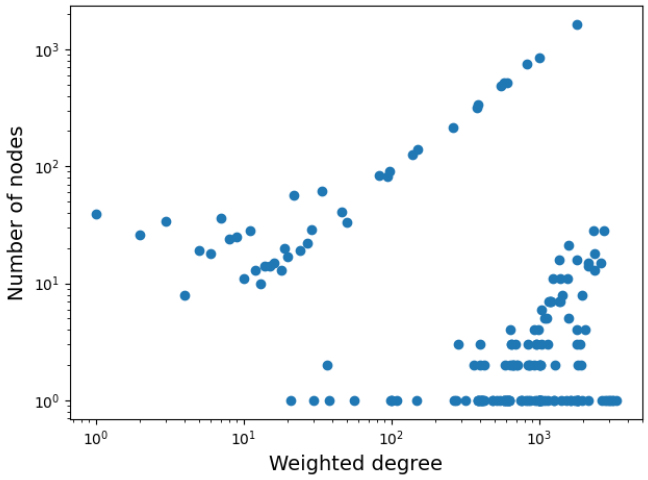

Figure 5: Weighted degree distribution in log-log scale

Figure 1. This implies that the top trackers cause the highest interconnectivity. The websites with the highest betweenness centrality share the same theme as those with the highest degree centrality. By having very prominent trackers embedded in your site, the connection to other webpages increases. Out of the 7,239 websites that were analysed, 260 of them have a closeness centrality of 1,0. This means these 260 websites are connected to every other site in the network. From a regulatory perspective, examining these 260 websites alone might capture the full extent of the Dutch cross-site tracking.

Lastly, we run the Louvain community detection algorithm [5] on the network and discuss the results. In this network consisting of domains and shared trackers, communities represent clusters of domains that are interconnected with each other through a set of third party trackers, but less connected to domains in other communities. By identifying these communities, we can gain insight on groups of domains that share similar tracking behaviour and helps to locate areas where

Table 3: Top 5 domains sorted by centrality

Domain	**Value**
Degree Centrality	
https://www.geldersefietsvierdaagse.nl	0,461
https://dutchbiz.nl	0,437
https://patrijs40ursem.nl	0,430
https://ndz.nl	0,416
https://alsea.nl	0,410
Betweenness Centrality	
https://dutchbiz.nl	226 660
https://www.jelrik.nl	225 544
https://www.roges.nl	191 357
https://www.lazytiger.nl	190 023
http://www.hanlitzgroup.nl	174 852

intervention is most effective. reveals 61 modules with hub-driven connectivity:

- The five largest communities contain 1,745, 1,009, 811, 594, and 561 nodes, collectively accounting for over 30% of all domains. This could mean that a large portion of Dutch websites have similar tracking behaviour. These large communities might represent sectors or categories of websites that rely heavily on shared third-party trackers, such as webshops or social media websites.
- Over half of the communities have ≤ 4 nodes, representing niche clusters. This indicates that the majority of domains in the network share smaller, more specialised trackers that would be used by specific types of websites. Identifying these tracking communities and intervening would therefore be more difficult as their tracking activity is more spread out.

Focusing on the core communities for intervention will have the greatest impact on reducing the overall network connectivity, without disrupting the functionality of smaller, less connected sites.

6 Conclusions

In this chapter we discuss the conclusions of the research and add some final words, as well as explore potential limitations of the study and present future research that draws from this study.

Web tracking is defined as embedding third-party content, such as cookies or fingerprinting scripts, into a webpage to gather information about users visiting the site. This information records the user behaviour on the website and is commonly used by advertisers to build unique identifier of a user to show them targeted advertisements.

The motivation for this research is the lack of country-specific studies on web-tracking. As outlined in chapter 1, we wanted to explore which trackers were most prevalent on Dutch web-domains, find the balance in local (`.nl`) trackers and global (`.com`, `.net`, etc) trackers and the distribution of different categories of web-trackers across Dutch websites.

To answer these research questions, we analysed over 1,8 million ".nl" domains and more than 8,000 third-party trackers. Studying this data revealed that Dutch domains load 3 unique trackers on average. Next we find that only 1,8% of all trackers on Dutch domains are local (.nl) trackers, while the vast majority (98,2%) are global trackers (e.g. .com, .net, etc). Furthermore, of the 6,379 categorised trackers, 58,9% comprise advertising trackers, 14,2% comprise analytics, and customer-interaction tools 5,3%. We see that the top 20 trackers are dominated by trackers from tech giants, such as Google and Meta.

Next we construct a weighted, undirected network of common trackers, where each node represents a domain and their edges are weighted by the number of shared trackers. This graph is created from a subset of 7,299 nodes. Analysing this network reveals small-world properties [36] by metrics such as an average path length of 2,01, diameter of 5 and a density of 0,12. This suggests a few major hubs influence most of the tracking. An average degree of 872 and weighted average degree of 873 reveals that most domains are highly interconnected. Finally, the Louvain community detection algorithm [5] finds 61 communities, of which the 5 biggest communities contain 30% of all domains in the network.

From our findings we can conclude that blocking or disrupting highly connected trackers and tracker hubs from our network, such as `fonts.googleapis.com`, may lead to a significant reduction of cross-site tracking. This gives insight to users on how to mitigate potential tracking of their online data and regulators on prioritising hub communities of web-trackers.

Because we constructed and analysed the network on a 6% sub-sample, we omitted a large section of our dataset that may have included rare domain-tracker links, thus potentially introducing a sampling bias. Another limitation of the research is the non-exhaustive list of trackers. We have only seemed to find 8,069 tracker domains, which were not all present in the complete raw dataset of 30 million rows. Finally, static HTML scraping, as done in this study, does not capture every type of tracker loaded onto a website, such as dynamically loaded scripts. These unexplored trackers may have revealed unique insight.

6.1 Future work

Future studies will focus on measuring the changes in Dutch (".nl") webpages before and after the introduction of the GDPR in 25 may 2018. This study will crawl historical data from websites (via the Wayback Machine) and research the change in tracking activity after the GDPR. Through this research we hope to quantify the impact that the GDPR has had on the Dutch web-tracking ecosystem and identify potential privacy vulnerabilities that the GDPR does not address.

In addition, we plan to conduct a sectoral analysis of Dutch (".nl") domains, exploring different categories in Dutch web-domains. By gathering data from a variety of sources (list of webshops, news, government, etc), we can group websites by category and crawl their data to reveal tracking activity within each group. This research may reveal what type of websites have the highest tracking activity, informing users to be more careful of certain websites and helping policymakers to target higher-risk groups.

References

[1] AddToAny. Addtoany share buttons api. https://www.addtoany.com/buttons/api/, 2025. Accessed 20 May 2025.

[2] Ivan Bermudez, Daniel Cleven, Ralucca Gera, Erik T Kiser, Timothy Newlin, and Akrati Saxena. Twitter response to munich july 2016 attack: Network analysis of influence. *Frontiers in big Data*, 2:17, 2019.

[3] Nataliia Bielova. Web tracking technologies and protection mechanisms. In *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2607–2609, 2017.

[4] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, pages 137–143, 2006.

[5] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[6] Hanjo D Boekhout, Arjan AJ Blokland, and Frank W Takes. Early warning signals for predicting cryptomarket vendor success using dark net forum networks. *Scientific Reports*, 14(1):16336, 2024.

[7] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*, pages 891–901, 2016.

[8] Ismael Castell-Uroz, Josep Solé-Pareta, and Pere Barlet-Ros. Network measurements for web tracking analysis and detection: A tutorial. *IEEE Instrumentation & Measurement Magazine*, 23(9):50–57, December 2020.

[9] Danielle Keats Citron and Daniel J. Solove. Privacy harms. Technical Report Legal Studies Research Paper No. 2021-11
Public Law Research Paper No. 2021-11, GWU Legal Studies, February 2021. Published in *Boston University Law Review*, vol. 102, p. 793 (2022). Available at https://ssrn.com/abstract=3782222 or http://dx.doi.org/10.2139/ssrn.3782222.

[10] Giuditta De Prato and Daniel Nepelski. Global technological collaboration network: Network analysis of international co-inventions. *The Journal of Technology Transfer*, 39(3):358–375, 2014.

[11] Disconnect, Inc. License: Creative commons attribution-noncommercial-sharealike 4.0 international. Licensed under CC BY-NC-SA 4.0; summary at https://creativecommons.org/licenses/by-nc-sa/4.0/.

[12] David Ediger, Karl Jiang, Jason Riedy, David A Bader, Courtney Corley, Rob Farber, and William N Reynolds. Massive social network analysis: Mining twitter for social good. In *2010 39th international conference on parallel processing*, pages 583–593. IEEE, 2010.

[13] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union (OJ L 119, 4 May 2016, pp. 1–88), May 2016.

[14] Maryam Farboodi and Laura Veldkamp. A model of the data economy. Technical Report Working Paper No. 28427, National Bureau of Economic Research, February 2021.

[15] Ralucca Gera, Ryan Miller, Akrati Saxena, Miguel MirandaLopez, and Scott Warnke. Three is the answer: Combining relationships to analyze multilayered terrorist networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 868–875, 2017.

[16] Global Multimedia Protocols Group. Gmpg — global multimedia protocols group. https://gmpg.org/, n.d. Accessed 20 May 2025.

[17] Google. Privacy and data collection. https://developers.google.com/fonts/faq/privacy, 2024. Last updated 23 July 2024; accessed 20 May 2025.

[18] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. Whotracks.me: Shedding light on the opaque world of online tracking, 2018.

[19] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser fingerprinting: A survey. *ACM Transactions on the Web (TWEB)*, 14(2):1–33, 2020.

[20] Ada Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *Proceedings of the 25th USENIX Security Symposium*, pages 997–1010, Austin, TX, Aug 2016. USENIX Association.

[21] Mariana Macedo and Akrati Saxena. Gender differences in online communication: A case study of soccer. *arXiv preprint arXiv:2403.11051*, 2024.

[22] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE Symposium on Security and Privacy*, pages 413–427, 2012.

[23] Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, and Christian Bizer. The graph structure in the web – analyzed on different aggregation levels. *Journal of Web Science*, 1:33–47, 2015.

[24] Ryan Miller, Ralucca Gera, Akrati Saxena, and Tanmoy Chakraborty. Discovering and leveraging communities in dark multi-layered networks for network disruption. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1152–1159. IEEE, 2018.

[25] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, April 1998.

[26] Iskander Sanchez-Rola, Xabier Ugarte-Pedrero, Igor Santos, and Pablo G. Bringas. The web is watching you: A comprehensive review of web-tracking techniques and countermeasures. *Logic Journal of the IGPL*, 25(1):18–29, 2017.

[27] Akrati Saxena, Ralucca Gera, and SRS Iyengar. Estimating degree rank in complex networks. *Social Network Analysis and Mining*, 8(1):42, 2018.

[28] Akrati Saxena and SRS Iyengar. Evolving models for meso-scale structures. In *2016 8th international conference on communication systems and networks (COMSNETS)*, pages 1–8. IEEE, 2016.

[29] Akrati Saxena and Sudarshan Iyengar. Centrality measures in complex networks: A survey. *arXiv preprint arXiv:2011.07190*, 2020.

[30] Akrati Saxena, Yulong Pei, Jan Veldsink, Werner van Ipenburg, George Fletcher, and Mykola Pechenizkiy. The banking transactions dataset and its comparative analysis with scale-free networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 283–296, 2021.

[31] Akrati Saxena, Pratishtha Saxena, Harita Reddy, and Ralucca Gera. A survey on studying the social networks of students. *arXiv preprint arXiv:1909.05079*, 2019.

[32] Daniel M Schwartz and Tony Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009.

[33] StatCounter. Statcounter support faq. https://statcounter.com/support/faq/, n.d. Accessed 20 May 2025.

[34] Zhan Su, Rasmus Helles, Ali Al-Laith, Antti Veilahti, Akrati Saxena, and Jakob Grue Simonsen. Privacy lost in online education: Analysis of web tracking evolution. In Xiaochun Yang, Heru Suhartanto, Guoren Wang, Bin Wang, Jing Jiang, Bing Li, Huaijie Zhu, and Ningning Cui, editors, *Advanced Data Mining and Applications*, pages 440–455, Cham, 2023. Springer Nature Switzerland.

[35] unpkg. documentation@14.0.3. https://app.unpkg.com/documentation@14.0.3, n.d. Accessed 20 May 2025.

[36] Duncan J. Watts and Steven H. Strogatz. Collective Dynamics of "Small-World" Networks. *Nature*, 393(6684):440–442, 1998.

[37] Zhiju Yang and Chuan Yue. A comparative measurement study of web tracking on mobile and desktop environments. *Proceedings on Privacy Enhancing Technologies*, 2020(2):24–44, April 2020.