# Privacy and web analytics in whistleblowing channels

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Software Engineering
April 2024
Esko Vuorinen

UNIVERSITY OF TURKU
Department of Computing

Esko Vuorinen: Privacy and web analytics in whistleblowing channels

Master of Science (Tech) Thesis, 53 p.
Software Engineering
April 2024

---

In today's digital landscape, where data is being collected in virtually every interaction between users and websites, safeguarding users' privacy has become exceedingly paramount. This becomes even more critical with applications that handle information of sensitive nature such as whistleblowing channels, where preserving the anonymity of reporters is extremely important. The EU whistleblower directive mandated for larger corporations to establish these channels to promote transparency and accountability within their organizations. This thesis discusses how well Finnish companies have established safe and trustworthy whistleblowing channels by going over what kind of information is being leaked during the whistleblowing process, and can it lead to the identification of a user.

The thesis starts off by introducing the legislation that is relevant to data management within the whistleblowing channels in Finland, mainly the GDPR and the whistleblowing directive, along with Finland's implementation of these regulations. Following this, the thesis goes over what kind of privacy risks and threats exist inside the whistleblowing channels. Subsequently, two case studies are carried out to show how well the 15 biggest companies in Finland protect the privacy of the whistleblower during the utilization of each company's whistleblowing channel. The thesis also looks into how dark patterns i.e. deceptive designs, inside the companies' cookie banners, are used to manipulate users into making unbeneficial decisions for themselves. Additionally, the thesis assesses the clarity and comprehensibility of the companies' privacy policies, focusing on how well they are written in terms of transparency and understandability. Based on the research done in this thesis, all of the 15 companies leaked data that could lead to the identification of the reporter. Furthermore, most of the companies fell short in terms of the both the effectiveness of the content in their privacy policies, and the harmful usage of dark patterns in their cookie banners.

Keywords: privacy, whistleblowing, whistleblowing channels, GDPR, personal data

# Contents

# 1 Introduction

As technological advancements continue to change the landscape of the internet, privacy and the concerns surrounding it have become a pivotal topic in today's world. The handling, storage, processing, and use of data collected from users is nowadays happening in almost every interaction between the Internet and the user, from social media platforms to online banking. In order to keep up with this data collection, users must be vigilant about their own online activities, and at the same time organizations and developers need to account for it to safeguard the privacy of users.

Whistleblowing channels, which are not exempt from concerns of user privacy, are crucial tools that let people (usually workers or insiders of a specific company) report any misconduct or misbehaviour by the company to the proper authorities, enabling them to take corrective action. As a whole, whistleblowing plays an essential role in society by helping in detecting corruption, fraud, bribery, or any other form of misconduct by different companies and organizations.

At the University of Turku, we conducted a scientific study on the impact of data collection on user privacy during the whistleblowing process [1], and published it in November 2023. Building on the gained insights from this study, my thesis aims to dive deeper into the privacy challenges that exist inside the whistleblowing process and explore the impact different consent levels (maximum and minimum amounts of cookies) have on data collection practices in that process.

## 1.1 Research Motivation

The European Union (EU) passed the Whistleblower Directive in December 2019. The directive's main goal was to create confidential and secure reporting channels and prohibit any kind of retaliation against individuals who use these channels. Finland implemented this directive through the Whistleblower Act, where it was made mandatory for private sector companies with over 50 employees and public sector companies with over 250 employees to create a reporting channel[1]. In addition, the fast implementation of the directive raised some concerns regarding whether the channels were established in a rush, potentially leading to inadequate implementation of safety and privacy methods.

What is more, in today's digitalized world, it has been demonstrated that individuals can be identified through data collection on websites [2]. This raises concerns regarding privacy inside the whistleblowing channels. It is imperative that these channels would be made in a way, that ensures the anonymity of its user. In other words, the channels should be made in a way that prevents the reporting individual from being identified by the company or any associated third-party.

Despite the prohibition of any form of retaliation against whistleblowers, it has been demonstrated that in the cases of where these individuals are identified, the person or corporation being accused may take some revengeful acts or use intimidation tactics against the whistleblower [3]. This has the possible effect of making potential whistleblowers opt out from reporting unethical behavior, leaving such actions unnoticed by appropriate authorities.

[1] https://www.twobirds.com/en/trending-topics/the-eu-whistleblowing-directive/implementation-status/finland

## 1.2 Research Questions and Structure of Thesis

In the context of privacy concerns in whistleblowing channels, my research aims to shed light on the effectiveness and challenges associated with whistleblowing channels in preserving anonymity and privacy of users, while also taking into consideration the potential impact of different cookie settings. Research questions of the thesis will be following:

- RQ1: Can the whistleblower be identified, and if so, by whom?
- RQ2: How well do the Finnish companies protect the anonymity of a user utilizing whistleblowing channels?
- RQ3: Does cookie consent have the expected effects and are there dark patterns?
  - RQ3.1: Is there a difference between using different cookie settings?
  - RQ3.2: Do the cookie banners use dark patterns?
- Q4: Is the user informed adequately about data processing practices?

Chapter 2 examines the legislation frameworks regarding privacy in whistleblowing channels, mainly the General Data Protection Regulation (GDPR) and the Whistleblower Directive. Chapter 3 defines privacy risks and threats in whistleblowing channels. These two chapters are made as a literature review to answer RQ1, and give necessary foundational information about privacy. In Chapter 4, two case studies are presented, to answer RQ2 and RQ3.1. In the case studies, we look into the 15 biggest companies in Finland and evaluate how well they safeguard the privacy of an user utilizing their whistleblowing services with the utilization of different cookie setting. Chapter 5 discusses the so-called Dark patterns i.e. deceptive designs and evaluates are they used by the 15 biggest companies of Finland to answer RQ3.2. In Chapter 6, we go through the privacy policies of the companies to investigate the amount of transparency companies employ when talking about their data collection practices. This helps us answer RQ4. In Chapter 7, we discuss the implications of our findings to the whistleblower, and some recommendations

for website developers. Finally, Chapter 8 goes through the contributions that this thesis has made to the field.

# 2 Legal Considerations Regarding Privacy

The concept of privacy itself is difficult to define and has evolved rapidly with the advancements in technology. In 1891, Samuel Warren and Louis Brandeis described privacy as the right to be left alone. Today the concept has evolved, and it includes a broad range of considerations, the complicated interactions between individuals, their personal information as well as the wide and ever-changing landscape of technology. In addition, the language describing privacy and the information considered being personal differs in various cultures and places. [4] However, in present-day, privacy as a concept can be broadly defined as the individual's right to choose what information about themselves is shared and who gets that information.

## 2.1 Legal Frameworks

In order to ensure the privacy protection in whistleblowing channels it is necessary to understand the relevant laws, regulations and directives governing data protection, privacy, whistleblowing and their connections to each other. In the following paragraphs, we will go through the most important ones in regard to the privacy of whistleblowing channels.

The General Data Protection Regulation (GDPR), one of the most crucial legal frame-

works, came into effect on May 25, 2018, with the aim of enhancing data protection for users. The GDPR clarified rules, established certain requirements, and provided clear instructions for processing, storing or transferring personal data in the EU. This regulation brought significant changes in the data protection strategies for companies. Overall, the legislation concerning data protection issues has rapidly evolved in response to the increasing amount of personal data collection practices in our daily lives. [5]

In addition to GDPR, In October 2019 the European Union enacted the so-called EU Whistleblower Directive where protection of individuals who report the misconduct or unethical behaviour of companies or organizations is highlighted. It emphasizes the importance of confidential and secure whistleblowing channels and shields the whistleblowers from any retaliation that may occur. [6]

In addition to these regulations and Directives, Finland has enacted national laws to ensure data protection, privacy and whistleblowing possibility of their citizens. After implementing the requirements of the General Data Protection Regulation (GDPR), Finland has introduced additional measures through the Data Protection Act to enhance the protection of individuals' privacy rights. The Data Protection Act points out the principles and requirements for handling personal data, ensuring the privacy and rights of individuals [7]. Furthermore, Finland has also enacted the Whistleblower Act, which mandates that the public sector companies with over 50 employees and private sector companies with over 250 employees would have to establish secure and confidential whistleblowing channels. [8]

In this chapter, we will discuss how these legal frameworks help in protecting privacy of the whistleblowing channels. They do so by establishing clear guidelines, requirements and standards for data protection and whistleblower practices. In other words, they create the overall framework that companies can and should follow in order to create confidential reporting channels.

## 2.2 EU Whistleblower Directive

The primary objective of the Whistleblower Directive is to create a secure and reliable environment for whistleblowers, enabling them to report any misconduct without the fear of retaliation. By encouraging and protecting whistleblowers, the EU aims to enhance the detection of misconduct by companies or organizations. [9]

The Directive gives legal protection to any whistleblower who reports information under certain conditions. These conditions include: the reporter has reasonable grounds to believe that the information being reported is accurate at the time of reporting, the reported information falls under the legislation, and the reported information is communicated to component authorities using the provided channels. Moreover, the directive encourages the use of internal reporting channels, when whistleblowers feel that there is no chance of retaliation, and the concern can be handled effectively, before utilizing external reporting channels. [9]

The directive presents several key provisions and requirements that member states must implement in their national legislation to protect whistleblowers. Here are some key provisions [8], [9]:

- **Reporting Channels:** The directive mandates the creation of secure and confidential reporting channels through which whistleblowers can report information about wrongdoings. These channels should be easily accessible, and users should have the ability to report anonymously if they choose to do so.
- **Protection from retaliation:** Whistleblowers are protected against any form of retaliation in response to their reports.
- **Confidentiality and anonymity:** The directive underlines the importance of maintaining confidentiality regarding the whistleblower's identity throughout the entire

reporting process. In addition, whistleblowers also have the right to remain anonymous, and any personal data collected during the process must be handled in accordance with data protection laws.

- **Feedback and follow-up:** Whistleblowers have the right to (if possible) get feedback on their report. In addition, be informed about the actions taken as a result of their reports. Furthermore, the directive encourages to create mechanisms for possible follow-up conversations with the whistleblower.

The EU whistleblower directive applies to a wide range of both public and private sector companies and organizations, including every private sector entity that has over 50 employees. It also applies to multiple areas, such as public procurement, financial services, consumer protection, environmental protection, and public health. The whistleblowing directive aims to create a systematic approach towards whistleblowing across the EU while promoting a high level of protection and standardization of reporting mechanisms. [6], [9]

In conclusion, the directive represents a significant step forward in promoting transparency and accountability within the EU. This is achieved by encouraging whistleblowers to come forward without the fear of retaliation, and thereby improving the detection of wrongdoings while simultaneously protecting the whistleblower.

## 2.3 General Data Protection Regulation

In the whistleblower Directive it is described that the personal data used in the reporting channels falls under the protection of General Data Protection Regulation (GDPR) [9]. This means that any information related to a person or information related to an identifiable person is protected by the GDPR. The primary objective of GDPR was to unify the

requirements and standards for data protection in the EU. Before the GDPR, data protection guidelines were fragmented due to the old Data Protection Directive being insufficient, and each member state having their own varying national laws [10]. What is also worth noting about GDPR is its scope. Although being EU regulation, in Article 3 of the GDPR it is discussed that it affects any and all organizations outside the EU that handle or process any personal data of an EU citizen [11]. In this section, we will look at the parts of the GDPR that are related to privacy of the whistleblowing channels.

The GDPR defines personal data under Article 4(1) as: "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person" [11]. In the context of whistleblowing and reporting channels, this entails all information that can be connected to the person who is reporting. Furthermore, if the whistleblower wishes to stay anonymous throughout the reporting process, it is vital that the companies or organizations recognize this and do not collect any personal data.

The GDPR can be recognized as a data governing system that encourages companies to carefully think about data collection as well as its destruction [12]. The key concepts of GDPR include personal data, data subject, data processing, data controller and data processor. The main purpose of GDPR is to protect the personal information of EU citizens and make sure that the fundamental right to privacy is preserved.

**Data processing** means any kind of usage and manipulation of data ranging from collecting data to destroying it. **Data subject** as a concept refers to the person whose data is being processed, in other words, the users using applications and websites. The person who is responsible for choosing the purpose and method which are used on the data is

called a **data controller**. **Data processor** refers to third parties external of the company that processes personal data behalf of the controller. [11]

The GDPR outlines fundamental guidelines for data processing by organizations to protect the privacy rights of individuals. These principles include [11]:

- Data processing should be lawful, fair, and transparent, ensuring that data subjects understand what data is being collected.
- Purpose limitation, meaning that personal data should be collected only for a specific purpose, reducing the risk of data exposure.
- Data minimization, where only necessary data should be collected and processed, further reducing the risk of data exposure.
- Importance of accuracy, ensuring that personal data is correct and up-to-date.
- Accountability and compliance, where organizations should be accountable for their data processing activities and comply with the principles of GDPR.

The GDPR poses a lot of requirements for corporations that operate or do business inside the EU. In addition, the potential penalties for violating the GDPR are quite large as the max out at either 4% of the global revenue of the company or 20 million euros, depending on which one is higher. Compared to the Whistleblower Directive or the previous privacy legislation (the Data Protection Directive), these penalties are substantial and can have a real financial impact on companies. [11]

Another thing to consider when talking about GDPR in whistleblowing channels is that the data subject (whistleblower) has certain rights: the right to ask and receive the collected information about them, the right to know the purpose of the collected data, right to ask for the collected data to be deleted, the right to access the collected data, and rec-

tify any inaccuracies in the collected data. The whistleblower channels need to establish transparent and effective methods for data subjects to exercise these rights. [11]

One more thing that the GDPR takes into account, is its scope. In recital 22 in the GDPR it is said: ”Any processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union should be carried out in accordance with this Regulation, regardless of whether the processing itself takes place within the Union.” [11] This means that despite the regulation being European, its impact extends to any processing of personal data involving EU citizens in other parts of the world as well.

Overall, the GDPR has fulfilled its goals in unifying and updating the legislation across EU member states, and in boosting personal data protection by increasing the transparency and care in data processing. Now organizations have clear guidelines that they have to follow when handling personal data and the rights of individuals have been precisely defined, which further improves the transparency of data handling.

## 2.4 Finnish Legal Frameworks

As an EU member state, Finland has integrated the GDPR and the Whistleblower Directive into its national legislation. Finland did this through the Data Protection Act and the Whistleblower Act. These acts contain extra provisions that were made to further increase the protection and privacy of Finland’s citizens and encourage the reporting of unethical behaviour of companies.

The Data Protection Act (DPA) in Finland plays a crucial role in protecting the privacy of individuals. It implements the GDPR by specifying the guidelines and requirements for handling and processing personal data [13]. Furthermore, it emphasizes that in order to collect personal data, explicit consent is needed from the user, and points out the in-

dividual's right to access, request corrections or request the deletion of any personal data collected [7].

Like already mentioned, Finland enacted the Whistleblower Act, which mandates that private companies with over 50 employees and public entities with over 250 employees needed to establish confidential and secure reporting channels by April 2023. What is more, the channels must have the option to report anonymously and any retaliation by the company towards the whistleblower is prohibited with a threat of severe ramifications. Any personal data processed in the channels in Finland falls under the protection of DPA and GDPR [14].

In conclusion, the two discussed Finnish legal frameworks (the Data Protection Act and the Whistleblower Act) work together to encourage and make it possible for whistleblowers to act without the fear of them being identified, their personal data being processed without permission or being retaliated against in any way. In addition, the acts give clear guidelines and requirements that help companies build reliable privacy-respecting whistleblowing channels.

# 3 Privacy Risks and Threats in Whistleblowing Channels

Digital services have become more prevalent and accessible in today's digitalized society, which has led to more people utilizing them. At the same time, today's websites have become full of numerous embedded third-party services such as web analytics tools. This raises concerns have certain digital services, where user's privacy is indispensable such as whistleblowing channels, thoroughly thought about and implemented adequate measures to safeguard the privacy of its users. Furthermore, in the case of whistleblowing, the organization or the company that is being reported should not be able to identify the whistleblowers, since they might deem them as a threat [15]. In other words, no personal or identifying data should leak to third-parties or the company during the reporting process in order to protect the whistleblower.

In recital 30 of the GDPR personal data is defined: "Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them." [11]

In this chapter, we delve into different web tracking technologies, and data collection practices. We have already showed how personal data is defined in the GDPR, and how it adheres to the EU and Finnish legislation, but now our focus shifts to what kind of information applications collect from users, how it is collected, and how it may lead up to the identification of the user. By going over the landscape of data collection we aim to provide a comprehensive understanding about the subject.

## 3.1 Data Collection

The data collection methods can be classified into three distinct categories. The first method of collecting information is taking data from the user's input itself. In other words, users themselves give information to the service provider, for example, to be a registered user for a service, new user would give their name or email, which can be used to identify the user. The second method involves the application indirectly collecting information such as IP addresses and device information about the user. The third method adds information from an outside source to already existing information about the user creating a more comprehensive profile about them. For example, a company might collect data from user's interaction with their website and then obtain information about their social media habits from an outside source. By combining these two datasets a company can build a more detailed profile of the user. The first method usually involves the user giving out their names, banking details, address etc. This is information that can be considered personal data, since it can be linked to a natural person. However, in case of this thesis we will focus on the two latter methods of information collecting since we are interested in what kind of information is collected without the user direct input into the application and can it lead to the identification of the user.

While the general belief might be that the ”worst” thing that the information collected

through websites can be used for is target marketing the reality is much different. The collected data can be utilized for various questionable purposes, for example it can be used in price discrimination, where companies set the price of a product to each customer individually from their personal valuation of the item [16], assessments of financial credibility, where data on who you are friends with in Facebook can affect if you get a loan or not [17] or even in government surveillance [18]. Three common and regularly used tracking technologies involve cookies, fingerprinting, and tracking pixels[1].

Cookies were first introduced by Netscape in 1994. The creation of cookies enabled stateful browsing instead of using the stateless HTTP protocol. Cookies work by allowing small pieces of data that are created by the web application to be saved on the browser. These can be used to save information such as various different identifiers, user authentication, and keep track of session identifiers which can enable the work of complex functions (e.g. Shopping cart). [19] Cookies can be categorized into first-party cookies or third-party cookies. If a cookie has the same domain and scheme (http, https) as the current page, it is a first-party cookie. On the other hand, if the domain or scheme is different from the visited website then it is a third-party cookie.

Fingerprinters collect information about your device, browser, network and combines it to create a unique fingerprint to you as a user. What is more, this data can also be linked to existing profiles or information about the user making the user more identifiable. Tracking pixels are usually a 1x1 pixel sized graphics, that are loaded when a user visits a website or opens an email. The small size of the graphic makes it hard for the user to notice. In addition, they are usually designed to be hidden partly or fully in the background color, so it is almost impossible for the user to notice them. Tracking pixels mixed with JavaScript are used to collect information about user behavior meaning their interactions with websites, and device information such as the operating system, IP address, and timestamp.[2]

[1] https://www.cookiepro.com/knowledge/website-tracking-technologies/

[2] https://www.lrb.co.uk/the-paper/v43/n07/donald-mackenzie/cookies-pixels-and-fingerprints

Third-party services have gained popularity due to their tracking capabilities, them being cost-efficient and their well-made functionality. Website providers favor to utilize them instead of building their own analytic tools to avoid the resource-intensive and time-consuming workload. Furthermore, the ability to gather vast amounts of data from various websites creates a bigger advantage to the third-party services, since from bigger amounts of data, it is easier to build better and more comprehensive insights about users. [20]

The persuasive nature of these third-party tracking services lies in their ability to monitor users on multiple websites, collecting data from multiple sources helps them to create highly detailed profiles of individual users. This raises concerns about the user's privacy and data security. In other words, at the same time as more and more information is collected, more (personal) data is exposed to these services and the sense of user privacy starts to fade. [21]

Figure 3.1 below represents how analytical tools work on the surface level with three main components depicted. Firstly, the user requests a web page from the server. Secondly, the server returns and logs the request. Finally, as the web page is loaded in the user's browser, it connects to third parties sending them data. The third-party connections can provide other functions than analytical data collection, such as advertising, or social media integration. However, the type and amount of data transmitted remains depended on the configurations and purposes of all third-party components.

## 3.2 Identifying Data

To get a clearer understanding about the kind of data that is collected on websites, we will look at other studies, which were conducted in a similar fashion as the case studies done for this thesis. In one such study [2], Heino et al. analyze the network traffic of 34 web services provided and maintained by the Finnish public sector. In the study the web

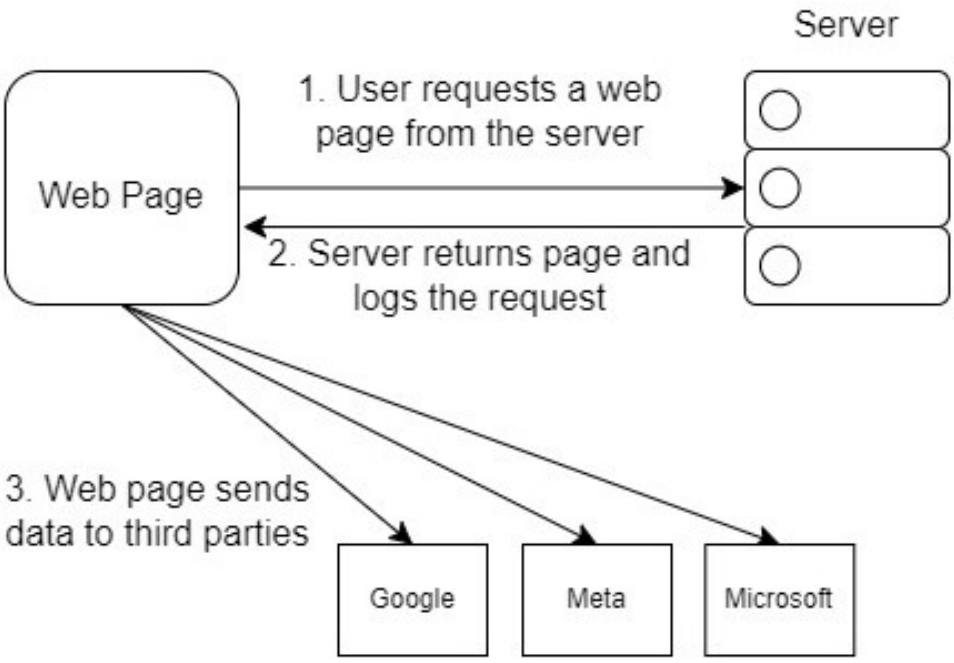


Figure 3.1: How third-party analytical tools work

services were tested by utilizing the most important functionalities, visiting crucial pages and logging into the services while all cookies were accepted. The traffic generated was recorded by using Google Chrome's Developer Tools (DevTools), which was then later analyzed to see what kind of data was collected. It was shown that almost every single web service collected some amount of data from the user, some more than others, using different kind of trackers. The personal data collected in the study will be shown in Table 3.1.

One thing to be noted is that the device IP address is always sent out with each web request, which is evident in the results shown in Table 3.1 as well (two services did not use any analytic tools, which means they did not leak the IP addresses) [2]. Some analytical tools might try to tackle this problem by anonymizing the IP address. For instance, Google Analytics 4, which replaced the earlier version in July 2023, has the capabilities to do this

| Sent data item | Number of services (Percentage) |
|---|---|
| IP Address | 32 (94.1%) |
| Browser | 32 (94.1%) |
| Operating System | 32 (94.1%) |
| Screen Size | 32 (94.1%) |
| Window Size(viewport) | 15 (44.1%) |
| Color depth | 25 (73.5%) |
| User Identifier | 2 (5.9%) |
| Device/Browser Identifier | 15 (44.1%) |
| Other Unknown Identifier | 31 (91.2%) |
| Timestamp | 26 (76.5%) |
| Language | 25 (73.5%) |
| List of Purchased Products/Services | 1 (2.9%) |
| Viewed Contents/Performed Actions | 32 (94.1%) |
| Connection Type | 6 (17.6%) |

Table 3.1: Personal data items sent to analyzed web services [2]

by removing the IP address of all users operating inside the EU. However, it should be mentioned that the service still uses the IP address to deprive some geographical location data before deleting it[3]. However, the omission of the IP address does not mean that it would be now impossible to identify the user. In most cases, with the location data gathered from the IP address, a lot of technical details and timestamps of the interactions between the user and the website, it is still plausible to identify the user. Furthermore, the anonymization is conducted by Google on their servers, which means the process is under their control. This means that the website host has no other choice than to trust that the third-party (Google) carries out proper anonymization and data handling.

Other studies using the same methodology have shown mirroring results. A study by Carlsson et al. analyzing network traffic of 15 Finnish web services often used by vulnerable groups (such as elderly, people with medical issues, people living in remote locations) showed that despite this personal data being very delicate, it was sent to various third parties [22]. Another study by Heino et al. showed that 95 websites of universities from

[3]https://support.google.com/analytics/answer/12017362?hl=fi

19 different EU countries have concerning number of analytic services [23]. These studies highlight that data collection exists in many different web applications and services, including in the ones where personal data should be protected at all costs.

As we have discussed there are many different kind of data that can be classified as personal, from person's name, bank details and real life address to their Operating system, IP address and Purchase history. This kind of information can be collected through all three aforementioned methods from a user. Privacy regulations like the GDPR have made it mandatory for websites, services and web applications to inform users how they are being tracked and provide an option for the user not to get tracked. In essence, collecting data and tracking users is not mandatory, rather it is a discretionary choice made by the service provider, excluding the extremely rare cases where a website or service has been compromised by hacking to conceal malicious data collectors or trackers.

In order to assess the potential risk each data item holds for identifying a person, we can classify them based on their high-level characteristics. Yabing Liu categorizes them in his article into three distinct dimensions [24]:

- Static vs. Dynamic
  - **Static** refers to data that remains the same over time, or at least does not change frequently, such as person's gender, phone number, name, email address etc.
  - **Dynamic** in contrast means data that may change over shorter or medium time intervals, for example user's session ID, geographical location and set of personal interests.
- Unique vs. Non-unique
  - **Unique** data serves to unmistakably identify an individual user from other users. Phone number and user's email are great examples of unique data.

  - **Non-unique** refers to data that many user may have in common, this includes things like, user's gender, name or date of birth.

- Shared vs. Distinct
  - **Shared** data is data that most likely stays the same across services and websites, this refers to data such as mailing address.
  - Conversely, **Distinct** data may vary across different services or websites. For instance, the timestamp of user logging in or session identifier.

| Sent data item | Static | Unique | Shared | Rating |
|---|---|---|---|---|
| IP Address | x | | x | Critical |
| Name | x | | x | Critical |
| Gender | x | | x | Minimal |
| Mailing Address | x | x | x | Critical |
| Email address | x | x | x | Critical |
| Browser | | | x | Minimal |
| Operating System | x | | x | Moderate |
| Screen Size | x | | x | Minimal |
| Window Size (viewport) | | | x | Moderate |
| Color Depth | x | | x | Minimal |
| User Identifier | x | x | | Critical |
| Device/Browser Identifier | x | x | x | Critical |
| Timestamp | x | | | Minimal |
| Language | x | | x | Moderate |
| Viewed Contents/Performed Actions | | x | | Moderate |

Table 3.2: Classification of data items according to how critical they are in the identification of a user

Table 3.1 shows different personal data items and number of services they were sent to from using 32 different web applications. Since the case study done in this thesis is done in similar fashion, we can use the data items from the list, with a small modifications to give a clear picture on what kind of data item are collected on the applications, and evaluate their potential in identifying an user utilizing the three dimensions discussed in the previous paragraph. This will not be a perfect list on which attribute contributes most

to the identification of the individuals but may guide on what data items possess potential risks in identifying a user. We will rank the data items into three categories: 'Minimal' for items that are less likely, on their own, to directly identify the individual, but can lead to the user identification when combined with other data items, 'Moderate' for items that usually contribute more significantly to the user identification than the 'Minimal' items, although most of the time they still require combination with other items for identification, and 'Critical' for items that, by themselves, can be used to identify users.

We gave data items: 'Browser', 'Gender', 'Screen Size', 'Color Depth', and 'Timestamp' the ranking of 'Minimal'. While these data items can change while the application is used on a different device or time, it should be noted that exceptionally distinct values within these items, either alone or combined, could directly reveal a user's identity. Moreover, it is important to highlight that typically users mostly use the same device for the same application, and these data items can be used to build more comprehensive profiles of a user. In addition, combining them or comparing them to existing profiles can lead to the identification of the user. As discussed before, personal data has a broad definition and these items can be considered part of that, since every data item collected by web application can be treated as personal information.

# 4 Case Studies and Analysis:

In this chapter, we will go over two case studies analyzing whistleblowing channels of 15 Finnish companies. In each case, we will navigate through the corporation's website to the point where it links to the whistleblowing service and the whistleblowing channel itself. The companies selected were chosen from a list of largest companies in Finland upkept by Asiakastieto[1], a Finnish organization supplying information about companies. A couple of corporations from the list needed to be excluded, since they seemed to have a whistleblowing channel accessible only to employees, and we could not access them. What is more, we have anonymized the companies to protect their identities and mitigate any possible biases. They will be referred to with labels such as Corporation 1.

To test how well user privacy was preserved we ran a test sequence where we acted as a ”regular” user moving from the home page of the company to the page that leads to the online-based whistleblowing service. Then the reporting channel was accessed, and information was entered to test any potential data leaks. However, we did not submit any information, and the process of reporting was aborted before doing so. The whole process was done twice, first while consenting to all cookies and data collection, and a second time where only the minimal number of cookies and data collection was consented to.

We recorded the network traffic with the help of Google Chrome's developer tools. The

[1]https://www.asiakastieto.fi/yritykset/top-listat

recording was done while cache was disabled to prevent the distortion in results due to previously cached data. In addition, we opened the whistleblowing channels services on a new tab to make sure that all data that leaked was recorded. The recorded network traffic was then saved in the log files for a closer look. As we analyzed the log files, two types of personal data stood out:

- Contextual data: Data that contains contextual information, which is sensitive, such as data showing user visiting a page that lead to the whistleblowing service or data showing user clicking a link that leads to a whistleblowing service. Obviously, if the whistleblowing service happened to be leaking data about a user utilizing it, for example if the URL of the service was leaked, this would count as sensitive data as well.
- Identifying data: Personal data that could lead to identifying the user of the web service, such as IP address or distinct device identifiers.

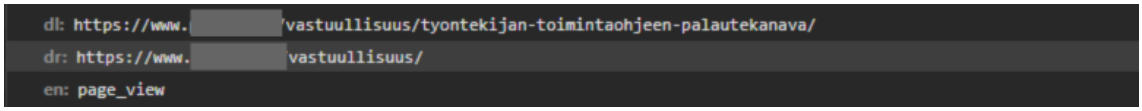


Figure 4.1: Part of Google Analytics payload: Highlighting the visited URL leak.

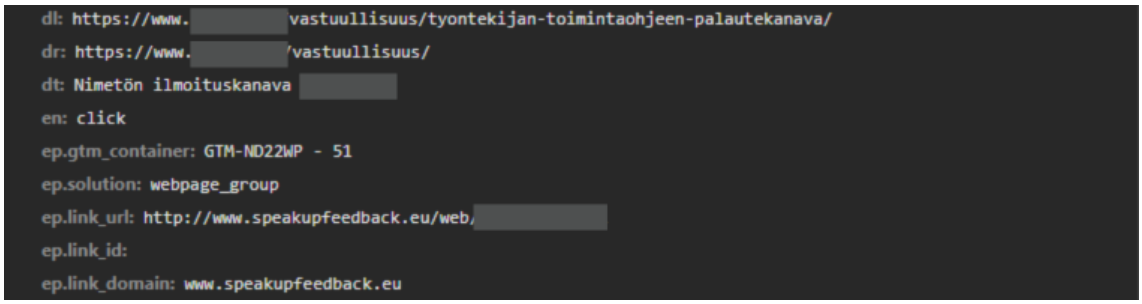


Figure 4.2: Part of Google Analytics payload: Highlighting link click and link address leak.

Figures 4.1 and 4.2 provide examples of how contextual data is depicted in payloads. The first figure illustrates how the visited URL leak looks like in Google Analytics payload, and the second figure represents how a link click and link address leak appear in

Google Analytics payload. In both figures, the corporation's name is greyed out to maintain anonymity.

In addition, we analyzed the dark patterns existing inside the cookie consent banners of the corporations and privacy policies of the studied websites, but they will be discussed in more detail in Chapters 5 and 6.

## 4.1 Results of the Study with Maximum Cookies

The results of this study show several privacy concerns inside the websites of the companies, especially in the case of sub-pages leading to the whistleblowing services. Majority of these pages were leaking data to third parties, in total, 14 out of 15 corporations (93.3%) leaked the data of a user clicking a link that leads to the whistleblowing channel, to a third-party actor. What is more, all 15 corporations leaked data indicating that the user was at least somewhat interested in whistleblowing. This was indicated by the user either visiting a page that had a whistleblowing-theme or clicking a link that leads to a whistleblowing service. Furthermore, a few of the corporations leaked this information to various third parties. In addition, we discovered a singular case where the whistleblowing channel itself was leaking sensitive information to a third party, and on 4 cases (26.6%) the link clicks were also found to be internally leaked within the respective company.

We found three primary type of contextual data leaks on the studied corporation websites:

- **The visited URL leaks**. In this data leak type, the address of the visited page is disclosed. By combining this data with identifying data on the user, it becomes possible to deduce user's interest in whistleblowing. It should be highlighted that the visited page needs to be related to whistleblowing, such as a corporation page that discusses whistleblowing and contains a link to the reporting channel, with a

website URL that reflects that focus.

- **The link click leaks**. The second leak type tells that the user has pressed a particular link that leads to the whistleblowing service. This leak type essentially confirms that the user has accessed the whistleblowing channel probably with the intention of filing a report.

- **The link address leaks**. In this third leak type, the address of the destination of the link click is disclosed. In all cases examined in this thesis, this destination address contained the name of the utilized whistleblowing service.

In majority of cases, these leaks are highly relevant to third party data collection. The leaks take place in the corporation web page that leads to the whistleblowing channel. In addition, the leaks also happened internally meaning that the information is also received by the corporation that runs the website. Since all companies have the ability to track the visited pages of users on their own websites even without using third-party analytic services, the first kind of data leak is not that relevant when the leak happens internally. What is more, third type of data leaks falls into the same kind of irrelevancy if the leak is internal, since the company already knows the name of the whistleblower service they use. However, the second type of data leak is highly relevant in the case of internal leaks, since if the company gets the information on who clicked the link to access a reporting channel, the whistleblower might get retaliated against. Additionally, there is a possibility that this second leak type can be observed by the company through the user interface of utilized third-party analytical tool, but we can not confirm this with the set-up used in this study. Lastly, apart from the corporation website, information such as visited URL from reporting channel alongside technical details about the user leaking is possible and addressed in this study. Figure 4.3 below shows these different types of data leaks and number of leaks to third parties per corporation.

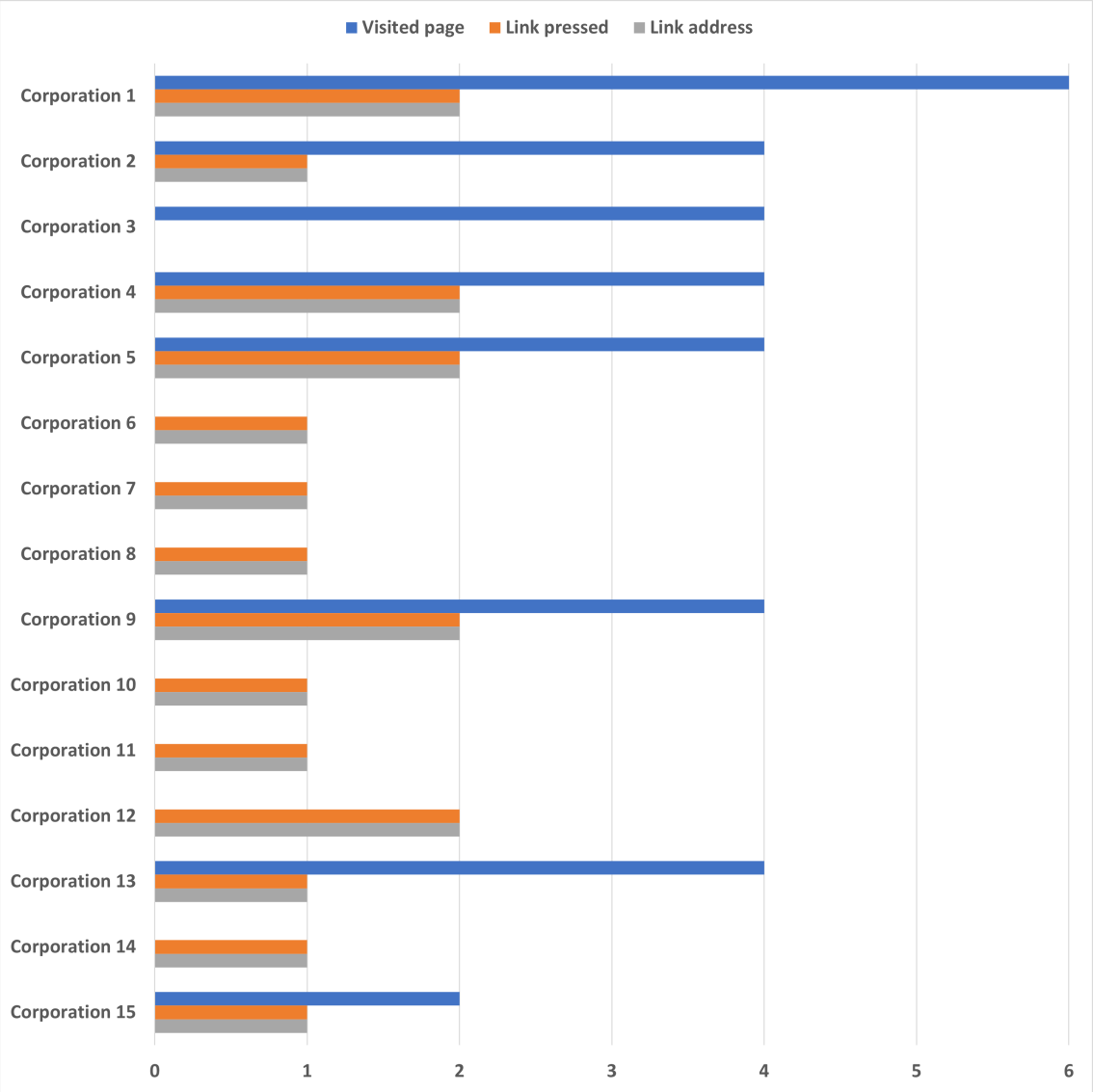


Figure 4.3: Numbers of different types of data leaks for each corporation while using maximum amount of cookies.

This study identified a total of 32 different third-party analytical services across the corporations analyzed. This means that the average of 2.1 data collection tools were used by a corporation. However, what should be noted as well is that majority of the corporations employ more than 3 third-party data collection services, while a singular company stands out utilizing as much as 11 such services on their website. Nonetheless, there was considerable amount of overlap between the applications used. For example, Google was found on the majority of the websites that were examined.

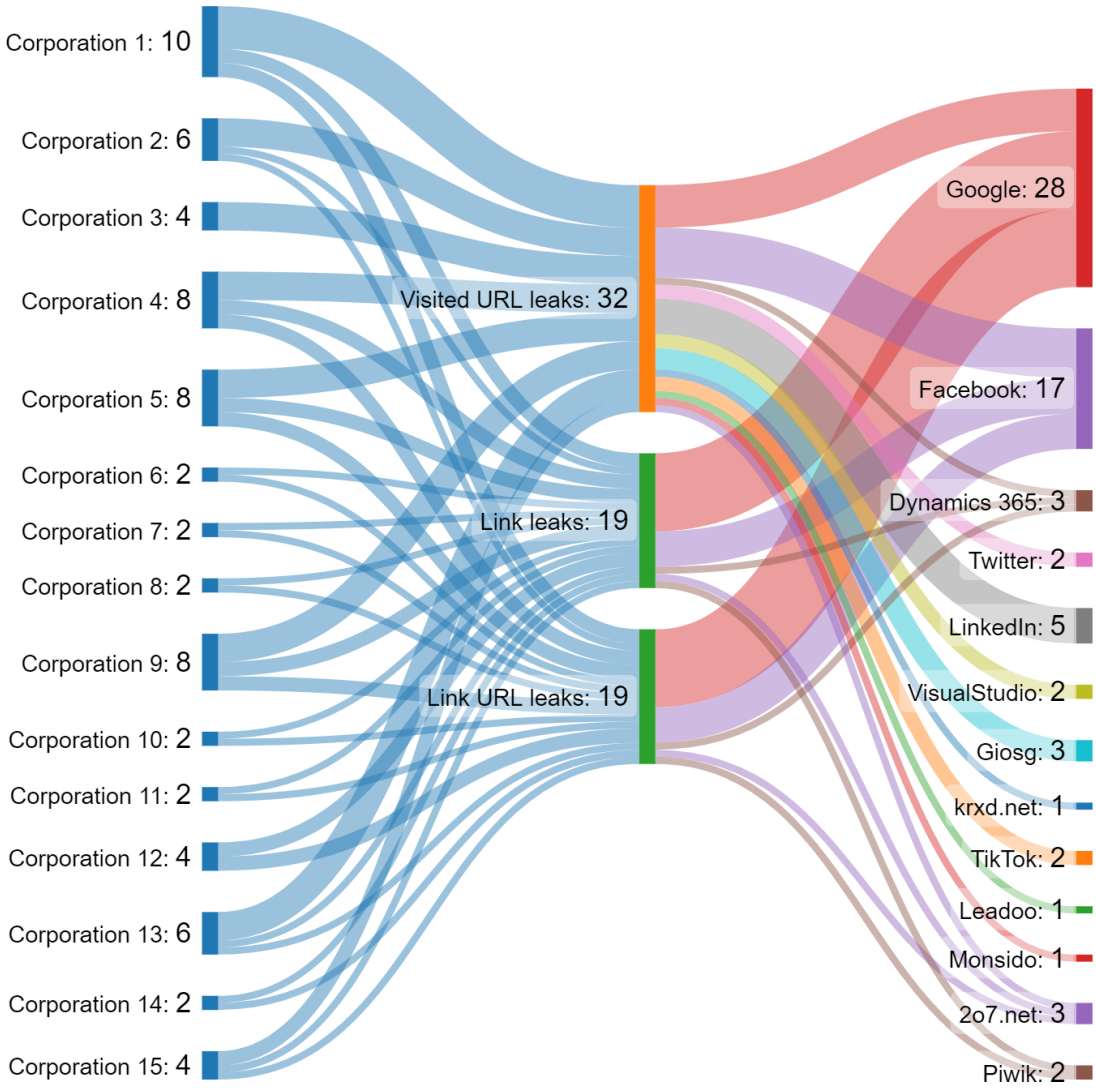


Figure 4.4: On the left: corporations and the amount of leaks per corporation, on the middle: data leak types, and the sum of each of leak type, on the right: third parties data leaked to, and the amount of data leaks going to each third party.

In the found third-party services, there were not only globally renowned analytics services like Google Analytics and Facebook (Meta), but also lesser-known and smaller-scale services, such as 2o7.net which is used by Adobe Systems for their web analytics and tracking services. All of the companies leaked at least one data item belonging to the three categories to a third-party data collection service. However, there was a single corporation (Corporation 3), that did not leak the link being pressed or the URL of the whistleblowing

service to any third parties.

It is noteworthy that, in this study, the real number of visited URL leakages was higher than what was accounted for here. The reason lies in the URL being too generic to indicate enough intention of using a whistleblowing service by the user. For example, some website addresses that lead to the reporting channels of the inspected corporations looked like this: 'www.company.fi/code-of-conduct', which does not indicate any intention of whistleblowing. On the contrary, this is how the address looked like on the pages that were counted as leaks: 'www.company.fi/whistleblowing'. This kind of address clearly implies the page's focus in whistleblowing. In other words, generic pages such as "contact information" were not considered as part of the visited URL leaks, as it would impossible to say from that piece of information alone that their intention is to use whistleblowing service.

Figure 4.4 shows that Google Analytics is the most common third-party analytics service that was utilized by the corporations. It was observed in 14 out of 15 (93.3%) studied corporations, and it was solely responsible for 28 out of 70 (40.0%) of the data leaks found in this study. Overall, this gives a picture how extensive Google is in the realm of data collection. Facebook (Meta) follows Google being the second largest third party data collector found in this study being accountable for 17 out of 70 (24.3%) data leaks, which does not come as a surprise either, since Facebook is also a major player in the data collection world. The remaining the third-party services that leaked data outside the domain of the corporations, had a ranging number of leaks from 1 (1.4%) to 5 (7.1%) out of 70. LinkedIn being the biggest of these having 5 leaks, followed by Dynamics 365 (Microsoft), Giosg and 2o7.net each having 3 leaks.

Like mentioned above the study found a total of 32 analytical tools in total among the 15 corporations studied, and what it is worth noting that only 13 out of 32 (40.6%) of the analytical services leaked any of the contextual information that was of interested in this

study. This implies that there are ways to utilize these tools while not compromising the privacy of a user. In addition, it shows that some data collection tools are safer than others regarding the privacy of a user. Furthermore, in the study we found that the same third-party data collection tool might leak the studied contextual information in one company, but not the other, which implies that it is possible to deploy these services in a way that preserves user privacy.

## 4.2 Results of the Study with Minimum Cookies

To give a picture on how much cookie consent matters to the privacy of a user we did the study again with the same methodology, only changing the user consent to only allow the minimum amount of cookies on the cookie consent banner. The results found in this study imply that cookie consent matters and only allowing the minimum amount of them helps reduce the amount of tracking that happens on a website. However, it does not totally stop the user tracking as still 5 out of 15 (33.3%) leaked the data of user clicking the link leading to a reporting channel on the corporation website. In addition, 8 out of 15 (53.3%) companies leaked information suggesting user's interest in whistleblowing, either through leaking the information about user visits to a whistleblowing-themed page or user link clicks that lead to the whistleblowing channel. A few of these companies leaked this information to various third parties. On the positive side 7 out of 15 (46.7%) companies did not leak information to third parties, and we did not detect any internal leaks in this study.

We still focus on the same three categories (the visited URL leaks, the link click leaks and the link address leaks), and as already mentioned they are all relevant in third-party data collection, and in the cases of internal leaks the second type (the link clicks) are especially important. Figure 4.5 below shows these types of data leaks: displaying the number of

data leaks to third parties per corporation and the data leak type.

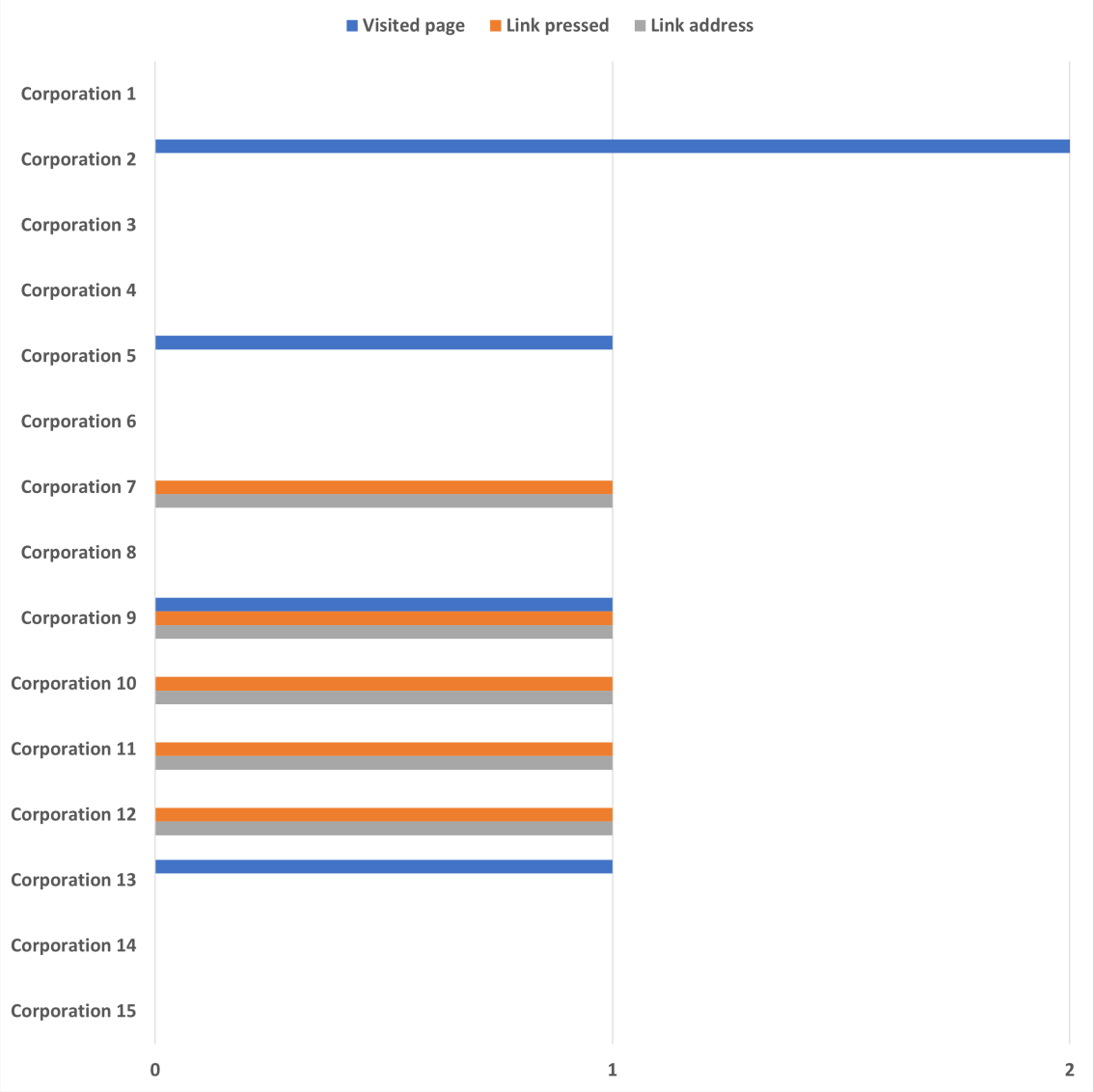


Figure 4.5: Numbers of different types of data leaks for each corporation while using minimum amount of cookies.

The total amount of third-party analytic tools we found in this study was 9. An average of 0.6 tools per corporation. However, it should be noted that 5 of the corporations did not seem to have any single third-party analytical tool activated during this study, and among the remaining companies, the maximum amount of these tools utilized was 2, observed in Corporations 2, 7, 11, 12, 14, and 15.

As we see in Figure 4.6, Google is the largest third-party data collector found in this study

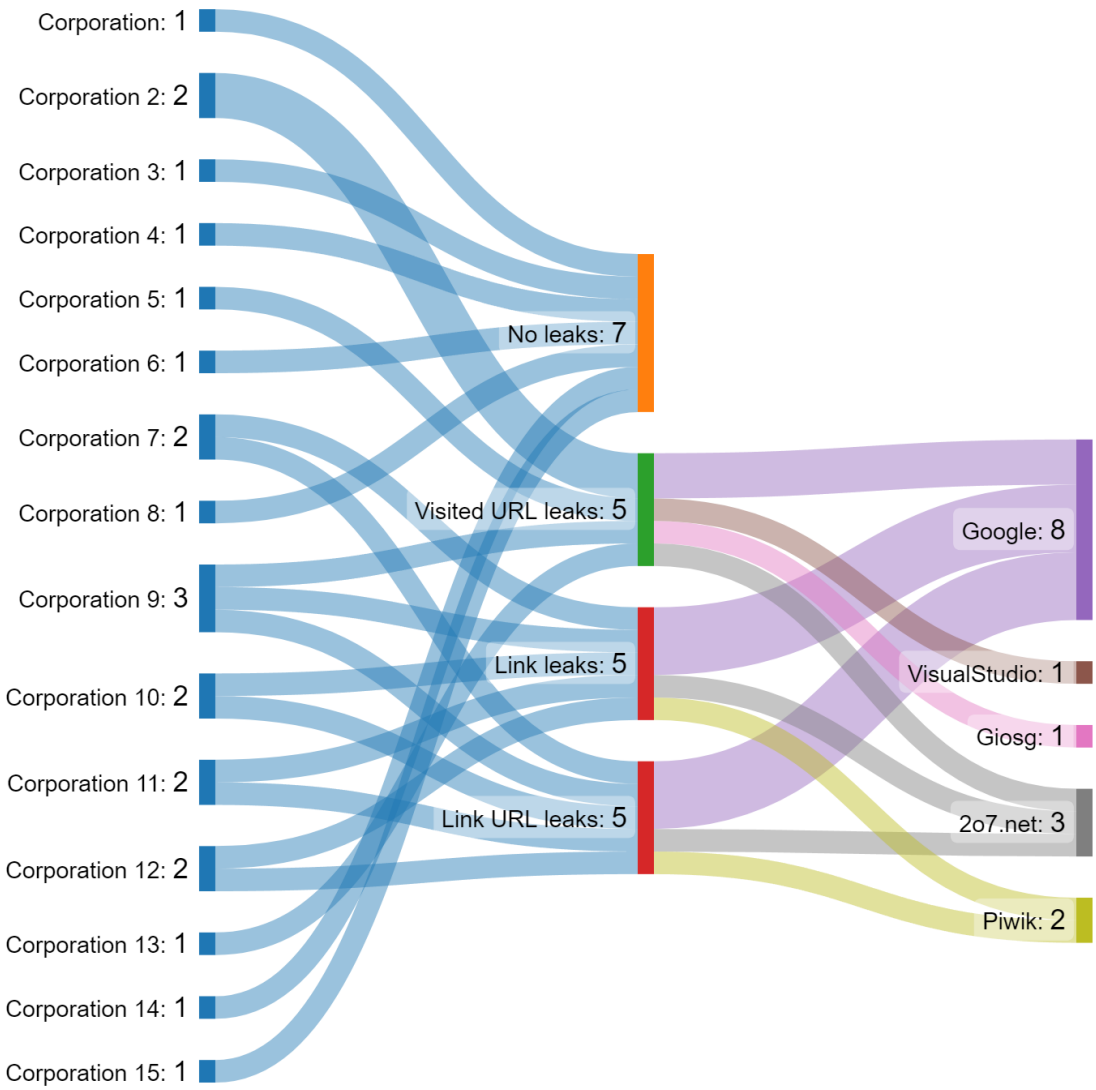


Figure 4.6: On the left: corporations and the amount of leaks per corporation, on the middle: data leak types, and the sum of each of leak type, on the right: third parties data leaked to, and the amount of data leaks going to each third party.

being present in 5 of 15 companies (33.3%). It was also responsible for 8 out of the 15 (53.3%) leaks of the study. The rest of the found analytical third-party tools were each utilized by only one company, and were responsible between 1 (6.7%) and 3 (20.0%) leaks with 2o7.net being the biggest with 3 leaks followed by Piwik with 2 leaks, followed by VisualStudio and Giosg each having only 1 leak.

In the case of this study as well as the previous one the real number of visited URL leaks

was larger than presented but as the visited address did not indicate user interest in whistleblowing we did not count them. Only 5 out of 9 (55.6%) of the found data collection services gathered contextual data that was of interest in this study regarding whistleblowing. This emphasizes the point of the possibility of deploying these services in a way that safeguards user privacy.

## 4.3 Comparison

The studies showed clear differences in results. In other words, it is evident that there is a difference in data collection practices based on the amount of cookies used. In the maximum cookie study it was shown that all 15 corporations leaked contextual data that was of interested, while in the minimum cookies study only 8 out of the 15 corporations had contextual leaks. This shows that using minimum cookies lessens the tracking that goes on the websites of the companies, however what is alarming is, it does not totally disable the third-party analytical tools. What is more, in the first study it was shown that 14 of the corporations leaked data of pressing the link that leads to the whistleblowing service. On the second study, only 5 of the 15 companies leaked this information.

The amount of utilized third-party data collection services varies greatly between the studies as well. In the maximum cookie study the average amount of data collection tools used by a corporation was 2.1 whereas in the minimum cookies one the number was only 0.6, and subsequently the total amount of different third-party analytical tools used were 32 and 9. Google was found to be the biggest third-party data collector in both of the studies. This did not come as a surprise, given that it is one of the biggest data collectors on the planet. However, what came as a surprise, was that Facebook (Meta) totally disappeared during the second study, and while it had 17 out of 70 leaks in the first study it had 0 out of the 15 leaks in the second study.

While these studies mainly focused on the contextual data leakages, it is important to note that all of the surveyed tools leaked personal data capable of identifying the user. These data points included things such as client ID numbers (CID), which serve as unique identifiers for specific device-browser combinations, as well as IP addresses, operating systems, device screen sizes, and various other technical details. While, as mentioned in Chapter 3, some of the data points can not individually be used for the identification of a user such as operating system or screen size. However, other details such as CID numbers and IP addresses pose more of a threat, and generally they can be used to identify the user. What is more, combining the data points can and will create more comprehensive fingerprint of the user, making it then possible to identify the user that way. This goes in accordance with Court of Justice of the European Union where IP addresses were considered as personal data in the Breyer case [25].

Together, the identifying - and contextual information (such as link click) can create a clear picture about the user and their actions such as whistleblowing. Especially technology giants such as Google and Facebook (Meta), are highly likely to have an understanding which IP address and device identifiers belong to which user, enabling them to identify who the user is, and what their actions were.

Furthermore, the corporations studied may be able to figure out this connection as well, without relying on information obtained through a third-party service. This may be very well be the case specifically in instances where the user utilizes the whistleblowing service through the user of company devices (work laptops, smartphones or tablets). The likelihood of identifying the user within the company increases when reporter accesses the whistleblowing channel through the company network. For obvious reasons, it is imperative that corporations are not able to learn the identities of whistleblowers.

In both studies, we found one case where the actual whistleblowing service itself leaked data to a third party, all of the other remaining data leak instances were associated with

the company pages that lead to the whistleblowing services. This particular service was utilized by Corporation 5, and it leaked data to a third-party entity called New Relic, a company specializing in web analytics, based in the USA. It appears that the Corporations 5's whistleblowing service is developed by a software company that operates in Finland, and they incorporated the New Relic analytics tool as an integrated component. In this case, the leaking of contextual data (the URL of the whistleblowing channel) and identifying data (such as user's IP address) to New Relic represents a severe violation of user privacy, especially considering the sensitive nature of this kind reporting service. Given that the company promoting this whistleblowing channel as "safe and compliant with EU directives", we can assume they are not aware of the shortcomings of their reporting service.

The occurrence of such leaks (the URL and IP address) inside the whistleblowing service is cause for great concern. It is worth noting, that the company from which the whistleblowing service was obtained is known for incorporating the New Relic analytics tool in their products, which is likely the reason for the leakage. The other whistleblowing channels that underwent inspection did not exhibit any leakage and appeared to be appropriately implemented.

Naturally, it is a great result that 14 out of the 15 implemented whistleblowing solutions of the inspected corporations did not have third-party data leaks. However, the found leak openly demonstrates that mistakes can happen even in the development of highly confidential whistleblowing services, such as the inclusion of third-party analytical tool and inadequate testing of the product. Furthermore, the company using the reporting service has also neglected their part by not assessing the privacy of the product adequately.

In these studies, the internal leaks to corporations themselves are more difficult to track than leaks to third-party applications. Since it is possible for any website owner to track user visits as well as the IP addresses of the visitors, it means that on the server side it

is feasible to track the user visiting the page leading up to the whistleblowing service. However, this could not be studied with the set up we used for these studies. In spite of that, for the companies to learn about information like clicking of a link, they need to be specifically sent that data. We found that 4 out of the 15 companies (26.7%) leaked this information internally on the maximum amount of cookies study. The address where the link takes you was also leaked to the company in each case. The corporations that received this information can be certain that the user who has visited the page regarding whistleblowing, also entered the whistleblowing service probably with the intention of filing a report. While it is understood, that the whistleblowing service should not be implemented and upkept by the company, but rather trusted to external parties, the pages leading to the reporting channel that are hosted by the company remain a significant privacy concern. In the minimum amount of cookies study, we did not find any internal leaks.

# 5 Dark Patterns

We have discussed that as our lives have moved more and more to the internet, at the same time the collection of valuable information through different data analytical tools has increased. In recital 32 of the GDPR it says: ”Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject’s agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement.” [11] That means that in order to collect data from the user, the host of the website needs the user’s consent. This is one of the main reasons cookie banners exist.

## 5.1 Introduction to Dark Patterns

While the legislation tries to give user the choice rather than the website of what cookies to use, it is clear that many try to bypass this with the use of dark patterns, which are also known as deceptive designs. Dark patterns try to ”trick” the user to make a decision more favourable to the owner of the website by making one choice stand out over the other [26]. For example differentiating the contrast between the colors of the accept and decline buttons. A 2018 study done by Kulyk et al. suggests that most of the website users just accept the use of cookies as a ”necessary evil” in order to access websites [27].

The term dark patterns was coined by Brignull in 2010, and he described them as ”tricks used in websites and apps that make you do things that you didn’t mean to, like buying or signing up for something.” [28], in the case of cookie banners, they are used to accept variety of (tracking) cookies, that ordinary users have no interest in. Even in this very specific domain that is cookie banners, there are difficulties creating criteria that would be able to capture the essence of dark patterns and that could be turned into accountable variables such as the intentionality of the deception. What is more, the designers for cookie banners evolve and find new ways to try to push the user the make the choice they want. Despite all of that, when we studied the cookie banners of the 15 corporations we followed the European Data Protection Boards guidelines when evaluating the designs of the cookie banners.

The European Data Protection Board categorizes seven different practices from A to K that can be found in cookie banners [29]:

- Practice A: The reject button does not exist on the first layer. In these kind of cookie banners there might be an accept button and a button that lets user access more options but does not contain a reject button on the same layer as these buttons.
- Practice B: There are pre-ticked boxes in cookie banner. Here cookie banners show already accepted cookies as pre-ticked boxes on the first layer or the second layer. Usually each category of cookies the controller wants to save have their own box, and now that box would be pre-ticked before the user actually makes the choice. What is more, in recital 32 of the GDPR states: ”Silence, pre-ticked boxes or inactivity should not therefore constitute consent.”, so this means that pre-ticked boxes are in violation of the GDPR [11].
- Practice C: Cookie banners contain link, not a button, to reject the deposit of cookies, this is also called deceptive link design.

- Practice D & E: Cookie banner's configuration has different colors and contrast between the available options. Practice D refers to the deceptive colors, where for example the accept button can be a different color than the reject button. Practice E refers to cases where the contrasts between the buttons and their background differs, for example it can be that the accept button is on a white background which highlights it, while the reject button would be more of a darker background which makes it blend in.

- Practice H: Some cookie banners highlight the possibility to accept read/write option in the first layer but does not include the option to reject this in the same layer. This can lead the average user to believe he has no other option than to accept to the deposit of cookies.

- Practice I: As mentioned before there are various categories for different kind of cookies, this practice takes into account inaccurately classified ”essential” or ”strictly necessary” cookies. In other words cookies that for example process personal data should not be considered ”essential”, and them being classified to this and therefore accepted by the user thinking they had no other choice would be in violation of this practice.

- Practice K: Controllers should make it possible to withdraw the given consent. While not directly linked to the cookie banners, this practice is also discussed, as many controllers seem to not show user easily accessible way to access their privacy setting and withdraw their consent. It is said that controllers should make easily accessible solutions such as small icon or a link placed in a standardized place to access these settings.

## 5.2 Evaluation of Corporations' Cookie Banners

While evaluating the cookie banners of the 15 studied companies we decided to focus on Practices: A, B D and E. This choice was made, since practices C and H did not appear in the cookie banners we studied. What is more, evaluating Practices I and K are impossible to evaluate solely based on the cookie banners and would need further examining. The results of the I category can be seen in Chapter 4 of this thesis as it discusses the cases where only minimum amount (essential) of cookies were accepted and still some personal information was processed meaning that some analytical cookies were used as well. Practice K, on the other hand, is not completely related to only the cookie banner so we left it out as well. Furthermore, the four chosen practices have a certain simplicity and unambiguity, which makes them easier to assess, enhancing the trustworthiness of the results. The evaluation was done in two phases. Firstly, the author of this thesis analyzed the four selected practices (A, B, D, E), and then subsequently, a second researcher would independently review and verify these observations.

Table 5.2 displays the results of our evaluation of the inspected cookie banners. The red color indicates that a deceptive design was present, while the green implies that no such designs were found. As can be seen here, multiple insufficiencies were found, them being mainly psychologically misleading utilization of color and contrast between accept and reject buttons and their backgrounds (type D and E practices). Overall, 14 out of 15 (93.4%) of the studied corporations used deceptive designs in their cookie consent banners, with all of them using at least practices D and E. In other words, all but one of the companies tried to influence the user into choosing the accept button over other options. Corporations 4, 11 and 13 had more deficiencies than the rest, each of them lacking a reject button on the first layer of the cookie consent banner (type A practice), and Corporation 13 using pre-ticked boxes on the consent form (type B practice). Only one of the companies managed to avoid implementing design elements that could be considered dark patterns.

| Website | Type A No reject button on first layer | Type B Pre-ticked consent boxes | Type D Deceptive colors | Type E Deceptive contrast |
|---|---|---|---|---|
| Corporation 1 | | | | |
| Corporation 2 | | | | |
| Corporation 3 | | | | |
| Corporation 4 | | | | |
| Corporation 5 | | | | |
| Corporation 6 | | | | |
| Corporation 7 | | | | |
| Corporation 8 | | | | |
| Corporation 9 | | | | |
| Corporation 10 | | | | |
| Corporation 11 | | | | |
| Corporation 12 | | | | |
| Corporation 13 | | | | |
| Corporation 14 | | | | |
| Corporation 15 | | | | |

Table 5.1: The presence of dark patterns found on the studied cookie banners

# 6 Privacy policies

## 6.1 Introduction to Privacy Policies

A privacy policy is a legal document, which discloses how a company collects, processes and manages a users' data. This includes telling how personal data, such as names, emails, IP addresses, device information as well as other identifying information, is handled. In Article 12, the GDPR discusses how the controller collecting personal data needs to provide information to individuals about their data processing practices concisely, transparently and in a manner that is understandable and easily accessible, utilizing clear and simple language. Article 14 of the GDPR reviews the information that a controller needs to provide when personal data is not obtained directly from the data subject, in there it discusses how in the existence of automated decision-making, such as profiling, the controller needs to provide relevant information regarding the logic involved and the estimated repercussions of such processing for the data subject. [11] The use of cookies or similar tracking technologies to collect information about data subjects' actions on a website or online service fall under the category of profiling, and therefore, data controllers must inform data subjects about the use of such technologies and provide meaningful information about the logic utilized, such as the names of the services used or what kind of actions are being tracked.

Despite the GDPR requiring transparent and clear communication between the controller and users, privacy policies have been noticed to use vague and obscure language, undermining their purpose and value [30]. This is further noted with studies that show a significant amount of data collection activities happening in different web services and mobile applications which are then not accurately represented in their privacy policies [2][31].

Even though privacy policies are legal contracts between users and service providers, there are few problems in the users' behavior towards the privacy policies that need to be highlighted. Users have shown that when privacy policy is presented to them, it is read quite carefully, but on other hand when the policy is not presented to them by default, most users never click a link to read it. In addition, even the situation where users click the link it is noticed that users end up just skimming through the text. [32]

Furthermore, a study on perception and attitude towards privacy policies held by the users, has shown that users might not understand what is written in the policy, and even while understanding what it says, more than half show lack of knowledge about what the actual content is. For example, while users were shown to have an understanding about the clear impacts of identifying data collection, they lack the expertise to comprehend the consequences of collecting something like geographical data, and how this data could be further abused. [33] This emphasizes the disconnect between the controller and users regarding privacy policies and highlights the problem of privacy policies not being written in a clear, understandable, and transparent manner.

In summary, privacy policies are legal documents that have become widespread standard across websites to fulfill the GDPR-mandated need for informing users. They detail how a company collects, processes and manages user data including personal information. However, despite the GDPR requiring clear and transparent communication, studies have displayed significant gap between the actual data collection practices and how they are pre-

sented in privacy policies. This is further highlighted with the use of vague language in the policies, which undermines their value and purpose. Additionally, users' attitudes and behaviors toward privacy policies have shown a spectrum of engagement, from most not reading it if it not showed as default, to skimming through it, to not truly understanding the content of it. Overall, this highlights the compelling need for clearer and more user-friendly privacy policies to bridge the gap between controllers and users, while at the same time making sure that users are adequately informed about their data privacy.

## 6.2 Evaluation of Corporations' Privacy Policies

While evaluating the privacy policies of the 15 studied companies, we focused on the following three questions:

- Does the privacy policy tell the users about collecting identifying personal information?
- Are all the third-party data collectors mentioned in the privacy policy?
- Does the privacy policy inform users about collecting data from clicking links?

These categories were chosen, since they adequately display how well the privacy policies reflect the data collection practices, and they are each related with the process of identifying whistleblowers or their actions. It is crucial that these three categories are clearly presented in order to inform users about the potential privacy implications. The transparency of the inspected policies was examined by comparing their content to the actual network traffic.

Table 6.2 displays the results of the evaluation of the inspected privacy policies. The green color implies that the privacy policy informs the user appropriately, while the red indicates

| Website | Tells about collecting identi-fying personal data | All third parties mentioned | Informs that data about link clicks is collected |
|---|---|---|---|
| Corporation 1 | | | |
| Corporation 2 | | | |
| Corporation 3 | | | |
| Corporation 4 | | | |
| Corporation 5 | | | |
| Corporation 6 | | | |
| Corporation 7 | | | |
| Corporation 8 | | | |
| Corporation 9 | | | |
| Corporation 10 | | | |
| Corporation 11 | | | |
| Corporation 12 | | | |
| Corporation 13 | | | |
| Corporation 14 | | | |
| Corporation 15 | | | |

Table 6.1: The level of transparency in the analyzed privacy policies.

the lack of relevant information. Our study found that the privacy policies often do not reflect the actual data collection practices, most have some important information missing or written in ambiguous way, leading to the loss of transparency. Due to this 6 out 15 (40.0%) of the inspected corporations were found to fail informing users about collecting personal information. Furthermore, 14 out of 15 (93.3%) of the privacy policies did not mention anything about data from link clicking being collected. Additionally, 8 out of 15 (53.3%) companies failed to mention all of the third-party data collectors. In terms of the criteria, only one (6.6%) of the corporations (Corporation 3) managed to inform users the actual extent of the data collection activities that happen on their website, while rest (93.3%) failed to do so.

The results found here did not exactly come as a surprise, as various studies on privacy

policies have shown parallel results to what we have found here [2], [30], [34]. In general, the results display a concerning problem regarding privacy policies, that being discrepancies between what is said in privacy policies and the degree of the actual data collection. In addition, the lack of transparency, meaning the lack of information available to the user about the nature of collected data and details of its collectors, as well as the use of vague and unclear language in the privacy policies is the norm rather than the exception.

# 7 Discussion

One of the more perplexing aspects of third-party analytical tools is their continued considerable data collection practices, despite receiving massive fines [35]–[37] and public disapproval. This shows how valuable the collected data truly is, keeping these tools active despite the negatives implies that they rack up a considerable amount of profit. For corporations, data that helps personalize adds and optimize user experience has too much upside regardless of penalties, public backlash and malicious purposes one can use it for.

## 7.1 Implications for Whistleblowers

The identification of a whistleblower can lead to multiple different problems for the reporter. The possible retaliation by the company can greatly affect one's personal career progression for example [38]. In addition, potential whistleblowers seeing these consequences might be discouraged from using the whistleblowing service altogether. This undermines the goal of the whole whistleblowing process and obstructs the uncovering of crucial information. Like already mentioned throughout the thesis, whistleblowing channels need to be built in a way that protects user-privacy and makes them trustworthy. Any data leakage to third parties obviously breaks this principle.

What is more, even legal cases or ongoing investigations can rely on the information gath-

ered from the whistleblowing services. In this case, a data leak leading to the identification of the reporter can endanger those efforts, as upon learning the identity of the whistleblower the parties under investigation can cover up the evidence or retaliate against the reporter.

In addition, to the negative consequences already discussed, this kind of sensitive data leaking to different third-party actors is ethically questionable. The more data third parties gather, the more comprehensive profiles they can build about the users. What is more, the third parties need to store the data somewhere, and as more data is collected, it becomes exceedingly likely that a ”fourth party” acquires the same data.

## 7.2 Recommendations for Web Developers

While maintaining user privacy is something that is essential for every web developer to think about, developers of websites that handle sensitive data such as whistleblowing services, should take several crucial things into consideration. Firstly, one thing to think about is, whether third-party analytical tools are needed for something like the whistleblowing channels, as their potential added value to the service is at least questionable. These kind of services are ordinarily utilized by users who intent on exposing unethical behavior that can include things such a criminal – or at least highly questionable acts by those in position of power, and any other forms of workplace misconduct. These actions encompass things such as breaches of corporate social responsibility, instances of harassment or discrimination, violations of workplace safety, as well as financial fraud. Compromising the anonymity of the user, by including third-party analytical tools on the website that handle sensitive information seems unreasonable, and can lead to serious ramifications, impacting the user’s social life, career prospects, and even physical safety.

On the other hand, web analytic tools are intended to monitor user’s actions, often under

the disguise of ”improving the user experience” of the website. At the same time whistleblowing channels are services that users, at least ideally, do not frequently utilize. The best case scenario being that these channels would never be needed. Moreover, the individuals using these services are ultimately only interested in their functional utility and ensuring they can stay anonymous. Both of these requirements can be fulfilled without any continuous user surveillance. There has been multiple studies that show that the third-party analytical services, that collect sensitive data from other websites, are prone to leaking it. This data can be utilized, and the possibility of identifying of the user emerges, typically to the private enterprises responsible for the development of these tools [39]–[41]. With large corporations being involved, as is the case in the inspected whistleblowing websites in this thesis, the risk of accessing this leaked information to deduce the identity of a whistleblower among their employees becomes a legitimate possibility.

The already discussed factors speak strongly against any use of any third-party analytical services on these kinds of websites. Furthermore, even if there would be some kind of necessity for these services to be utilized, they should either be tailored individually for the website in question or be locally deployed. The potential privacy violations discussed in the thesis can not be permitted to occur in the context of whistleblowing process.

Apart from abandoning the use of analytics tools altogether, other precautions should be taken to ensure the privacy of users. One thing, the developers could do, is a thorough network traffic analysis, similarly to what was done in the case study of this thesis. This would show the developers are there any data leaks on their websites. It should be emphasized that this kind of analysis does not require specialized tools, specific expertise or many hours of work. This being the case, cost and time constraints should not pose as obstacles, and the lack of performing this kind of testing could be viewed as plain negligence by the developer team.

The case study in this thesis proved that cookie consent has a huge impact on the extent

of data collection that happens on websites. The dark patterns inside the design of cookie banners become increasingly significant, as they impact the amount of data being collected. Their role further grows, when sensitive data is involved, as this means that more of this confidential data is being collected. The use of these manipulative design techniques has become commonplace in the website design landscape, and previous research on the subject [42], [43] supports this claim, as they show that the most of website cookie banners feature design elements that can be categorized as dark patterns. The European Data Protection Board has compiled a report describing multiple common practices that intentionally and actively mislead the users, as discussed in Chapter 5. In order to address the issue of dark patterns, developers should try to adhere with the recommendations outlined in the report and avoid utilizing the practices it highlights.

Privacy policies share the same problem with dark patterns in the sense that it has become far too common to use vague, obstruct, or overtechnical language, internationally making the content difficult for the average user to understand. Other studies [2], [34] have concluded very similar results as this thesis, that in majority of the cases, privacy policies are misleading and hard to understand due to their use of unclear language. To rectify this kind of problem, would mean to change the prevailing norm that it is acceptable to purposefully withhold information from the user regarding the collection of data.

This thesis underscores the fact that the user privacy inside the whistleblowing process is not only depended on the whistleblowing service itself, and that sensitive contextual data can leak from other places as well. This highlights the importance to address the confidentiality of company websites, which can inadvertently leak this kind of sensitive data, such as pages that include a link to the reporting service. The case studies in this thesis indicate, that most of the time clicking these links leak sensitive contextual information to third-party analytical tools, which in itself already puts users' privacy and safety in danger.

All the recommendations discussed in this section have one fundamental issue at their

core: web analytics have become exceedingly prevalent, and modern websites seem to deploy as many of these tools as they can. In the majority of the websites, both in the case study of this thesis and in previous research [2], [22], multiple different third-party tools are utilized at the same time on the same website, collecting essentially the same data from the user. Although there might be an actual necessity to collect some analytics from the website, it is difficult to understand why multitude of almost identical services would be needed to accomplish this, as they basically collect the same data points. In addition, as more services collect sensitive data, more possibilities, for this data to leak, emerge. As the number of leakage points increases, they become exceedingly more difficult to notice, and might escape the notice of website developers and administrators, even if there had been proper precautions implemented. The current state of the internet and website development contributes to the overabundance of these analytical services, as their use is has become commonplace and is perceived as the norm, even while there is limited justification for their necessity. This factor becomes increasingly critical, when we talk about whistleblowing channels, as the need for these kinds of services diminishes. That is why, while designing these services, companies and developers should seriously think about opting out of the use of any third-party analytical services.

# 8 Conclusion

The thesis begins by reviewing the EU legislation regarding whistleblower services and the personal information they handle, along with Finland's implementation of this legislation. Subsequently, inspecting what kind of different privacy risks and threats exists for the whistleblowing channels. After that, two case-studies are presented, which include the examination and evaluation of the 15 biggest companies in Finland, on how well they preserve and safeguard their users' privacy, with one study focusing on the maximum level of cookie consent and the other on the minimum level of cookie consent. In addition, we looked through the cookie banners and privacy policies of the mentioned companies, to identify if any dark patterns or vague language were used in their designs or content. Lastly, we discussed some implications that the results could have for whistleblowers, and recommendations for website developers to avoid privacy problems.

**RQ1: Can the whistleblower be identified ,and if so, by whom?**

The whistleblower can be connected to a natural person with the data that is being collected on the websites. The personal information being collected on website include data items such as names, phone numbers, addresses, IP-addresses, device identifiers, screen - and window sizes, etc. A singular data item can lead directly to the identification of an individual, and if it is not possible solely on one item, the combination of multiple items can yield the same result. What is more, contextual data such as clicking a link that leads to the

whistleblowing channel or a website URL, which includes the word: ”whistleblowing” or ”reporting-channel”, implies that an user is interested in whistleblowing. The combination of personal data and contextual data can lead to the identification of whistleblowers.

These data items are collected by various different entities, including multiple third parties and hosts of the visited websites. This means that multiple third parties such as Google and Facebook can identify the user, as well as the company hosting the websites. This becomes an obvious problem, since whistleblowers are usually company employees and their identification can lead to retaliation by the company.

**RQ2: How well do the Finnish companies protect the anonymity of an user utilizing whistleblowing channels?**

The case-study in this thesis displayed that even the biggest companies in Finland fail to protect the anonymity of the users utilizing the whistleblowing channels. It was shown that all 15 (out of 15) corporations inspected leaked contextual data, while simultaneously leaking personal data leading into the possibility of identifying individuals. In addition, it can be speculated that other companies, especially those with less resources would have a harder time to establish trustworthy and secure reporting channels. For the same reason, smaller corporations are less likely to take into account and address the possible contextual leaks on their own websites.

**RQ3: Does cookie consent have the expected effects and are there dark patterns?**

The results showed that cookie consent between the maximum and minimum amount of cookies affects the amount of data being leaked. It was shown that minimum amount cookies led to less amount of data being leaked than the maximum amount of cookies. With the maximum amount of cookies 15 out of 15 corporations leaked contextual data to third parties, while only 8 out of 15 corporations did the same with minimum amount of cookies. What is more, the number of third-party analytical tools were 32 and 9, showing

a huge difference on the amount of data collection between the consent choices. Despite internal leaks to the corporations being hard to track with our set up, we found that 4 out of 15 companies specifically leaked contextual data inside the companies with maximum amount of cookies accepted, and this internal leakage did not exist at all while only consenting to the minimum amount of cookies.

Dark patterns are deceptive design created in order to make users make decisions that benefit companies rather than themselves. The existence of these designs inside the cookie banners can make the user to consent to more amount of cookies that would be ideal for the user. We found that 14 out of 15 inspected cookie banners had dark patterns in their designs, with varying degrees of severity.

These deceptive designs try to make the user choose the option that is unbeneficial for themselves. This strategy effectively results in a broader scope of data collection, as indicated by the results from our case studies.

**RQ4: Is the user informed adequately about data processing practices?**

14 out of 15 of the evaluated privacy policies exhibited problematic tendencies, such as vague and obscure language, leaving out information such as all third-party actors, and failing to inform users clearly what kind of data is being collected such as personal data or data of a link click. In other words, for users to understand the extent of data collection that happens on the visited websites, even with thorough examination of the the privacy policy, would be either impossible or exceedingly challenging.

# References


[1] E. Vuorinen, P. Puhtila, S. Rauti, and V. Leppänen, "From whistle to echo: Data leaks in web-based whistleblowing channels," in *Secure IT Systems*, L. Fritsch, I. Hassan, and E. Paintsil, Eds., Cham: Springer Nature Switzerland, 2024, pp. 37–53.

[2] T. Heino, R. Carlsson, S. Rauti, and V. Leppänen, "Assessing discrepancies between network traffic and privacy policies of public sector web services," in *Proceedings of the 17th International Conference on Availability, Reliability and Security*, ser. ARES '22, Vienna, Austria: Association for Computing Machinery, 2022.

[3] M. Van Portfliet, M. Irfan, and K. Kenny, "When employees speak up: Human resource management aspects of whistleblowing," in *The Emerald Handbook of Work, Workplaces and Disruptive Issues in HRM*, Emerald Publishing Limited, 2022, pp. 533–547.

[4] K. Nissim and A. Wood, "Is privacy privacy?," 2018.

[5] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "Eu general data protection regulation: Changes and implications for personal data collecting companies," *Computer Law & Security Review*, vol. 34, no. 1, pp. 134–153, 2018.

[6] V. Abazi, "The European Union Whistleblower Directive: A 'Game Changer' for Whistleblowing Protection?" *Industrial Law Journal*, vol. 49, no. 4, pp. 640–656, Oct. 2020.

[7] Ministry of justice, Finland, "Data protection act," 2018. [Online]. Available: https://www.finlex.fi/en/laki/kaannokset/2018/en20181050.pdf.

[8] J. T. Stappers, *Eu whistleblower protection directive: Europe on whistleblowing*, 2021.

[9] European Council, "Directive (eu) 2019/1937 of the european parliament and of the council of 23 october 2019 on the protection of persons who report breaches of union law," 2019. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1937&from=EN.

[10] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius, "The european union general data protection regulation: What it is and what it means," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019.

[11] European Council, "General data protection regulation (eu) 2016/679," 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016R0679&qid=1689330735695.

[12] B. v. d. S. Chris Jay Hoofnagle and F. Z. Borgesius, "The european union general data protection regulation: What it is and what it means*," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019.

[13] P. Korpisaari, "Finland: A brief overview of the gdpr implementation," *Eur. Data Prot. L. Rev.*, vol. 5, pp. 232–237, 2019.

[14] Ministry of justice, Finland, "Ilmoittajansuojelulaki," 2022. [Online]. Available: https://www.finlex.fi/fi/laki/alkup/2022/20221171?search%5C%5Btype%5C%5D=pika&search%5C%5Bpika%5C%5D=1171%5C%2F2022.

[15] P. B. Jubb, "Whistleblowing: A restrictive definition and interpretation," *Journal of Business Ethics*, vol. 21, pp. 77–94, 1999.

[16] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, "Detecting price and search discrimination on the internet," ser. HotNets-XI, Redmond, Washington: Association for Computing Machinery, 2012, pp. 79–84.

[17] CNN, "Facebook friends could change your credit score," 2013. [Online]. Available: https://money.cnn.com/2013/08/26/technology/social/facebook-credit-score/index.html?hpt=hp_t2.

[18] CNN, "How the nsa piggy-backs on third-party trackers," 2013. [Online]. Available: https://cyberlaw.stanford.edu/publications/how-nsa-piggy-backs-third-party-trackers.

[19] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, *et al.*, "Can i opt out yet? gdpr and the global illusion of cookie control," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, ser. Asia CCS '19, Auckland, New Zealand: Association for Computing Machinery, 2019, pp. 340–351.

[20] N. Wehkamp, "Internalization of privacy externalities through negotiation: Social costs of third-party web-analytic tools and the limits of the legal data protection framework," ser. WWW '22, Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 525–533.

[21] T. Bujlow, V. Carela-Español, J. Solé-Pareta, and P. Barlet-Ros, "A survey on web tracking: Mechanisms, implications, and defenses," *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476–1510, 2017.

[22] R. Carlsson, S. Rauti, and T. Heino, "Data leaks to third parties in web services for vulnerable groups," in *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 2023, pp. 1208–1212.

[23] T. Heino, S. Rauti, R. Carlsson, and V. Leppänen, "Third-party services as a privacy threat on university websites," in *Proceedings of the 24th International Conference*

*on Computer Systems and Technologies*, ser. CompSysTech '23, New York, NY, USA: Association for Computing Machinery, 2023, pp. 134–138.

[24] Y. Liu, H. H. Song, I. Bermudez, A. Mislove, M. Baldi, and A. Tongaonkar, "Identifying personal information in internet traffic," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, ser. COSN '15, Palo Alto, California, USA: Association for Computing Machinery, 2015, pp. 59–70.

[25] *Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland [2016] ECLI:EU:C:2016:779, paragraph 49.*

[26] T. Mejtoft, E. Frängsmyr, U. Söderström, and O. Norberg, "Deceptive design: Cookie consent and manipulative patterns," 2021.

[27] O. Kulyk, A. Hilt, N. Gerber, and M. Volkamer, "This website uses cookies": Users' perceptions and reactions to the cookie disclaimer," in *European Workshop on Usable Security (EuroUSEC)*, vol. 4, 2018.

[28] H. Brignull, M. Miquel, J. Rosenberg, and J. Offer, "User interfacesdesigned to trick people," 2015. [Online]. Available: `https://www.deceptive.design/`.

[29] T. E. D. P. Board, "Report of the work undertaken by the cookie banner taskforce," [Online]. Available: `https://www.edpb.europa.eu/system/files/2023-01/edpb_20230118_report_cookie_banner_taskforce_en.pdf`.

[30] J. Reidenberg, J. Bhatia, T. Breaux, and T. Norton, "Ambiguity in privacy policies and the impact of regulation," *The Journal of Legal Studies*, vol. 45, S163–S190, Jun. 2016.

[31] R. Carlsson, T. Heino, L. Koivunen, S. Rauti, and V. Leppänen, "Where does your data go? comparing network traffic and privacy policies of public sector mobile applications," in *Information Systems and Technologies*, A. Rocha, H. Adeli, G. Dzemyda, and F. Moreira, Eds., Cham: Springer International Publishing, 2022, pp. 214–225.

[32] N. Steinfeld, ""i agree to the terms and conditions": (how) do users read privacy policies online? an eye-tracking experiment," *Computers in Human Behavior*, vol. 55, pp. 992–1000, 2016.

[33] D. Ibdah, N. Lachtar, S. M. Raparthi, and A. Bacha, ""why should i read the privacy policy, i just need the service": A study on attitudes and perceptions toward privacy policies," *IEEE Access*, vol. 9, pp. 166 465–166 487, 2021.

[34] L. Iwaya, A. Babar, A. Rashid, and C. Wijayarathna, "On the privacy of mental health apps," *Empirical Software Engineering*, vol. 28, 2023.

[35] F. Brian, "Facebook will pay an unprecedented $5 billion penalty over privacy breaches," 2019. [Online]. Available: `https://edition.cnn.com/2019/07/24/tech/facebook-ftc-settlement/index.html`.

[36] L. Seung, "Linkedin to pay $13 million in suit settlement for excessively spamming users," 2015. [Online]. Available: `https://www.newsweek.com/linkedin-13-million-class-action-lawsuit-emails-379975`.

[37] D. Clare, "Google agrees to pay $13 million in street view privacy case," 2019. [Online]. Available: `https://edition.cnn.com/2019/07/22/tech/google-street-view-privacy-lawsuit-settlement/index.html`.

[38] Ø. Kvalnes, "Whistleblowing," in *Communication Climate at Work: Fostering Friendly Friction in Organisations*, Springer, 2023, pp. 119–126.

[39] A. B. Friedman, L. Bauer, R. Gonzales, and M. S. McCoy, "Prevalence of third-party tracking on abortion clinic web pages," *JAMA Internal Medicine*, vol. 182, 11 2022.

[40] M. Huo, M. Bland, and K. Levchenko, "All eyes on me: Inside third party trackers' exfiltration of phi from healthcare providers' online systems," in *Proceedings of the 21st Workshop on Privacy in the Electronic Society*, ser. WPES'22, Los Angeles, CA, USA: Association for Computing Machinery, 2022, pp. 197–211.

[41] K. Schnell and R. Kaushik, “Hunting for the privacy policy – hospital website design,” 2022.

[42] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, and C. Wilson, “A comparative study of dark patterns across web and mobile modalities,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, 2021.

[43] T. H. Soe, O. E. Nordberg, F. Guribye, and M. Slavkovik, “Circumvention by design - dark patterns in cookie consent for online news outlets,” in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, ser. NordiCHI ’20, Tallinn, Estonia: Association for Computing Machinery, 2020.