Data leaks to third parties in web services for vulnerable groups

Robin Carlsson*, Sampsa Rauti*, Timi Heino*,
* University of Turku, Department of Computing, Turku, Finland
crcarl,sjprau,tdhein@utu.fi

***Abstract*—Third-party analytics services are increasingly being used to improve sales and usability of websites. While these services often have great value for companies and organizations using them, they also raise privacy issues. When information is gathered using third-party analytics, third parties also receive lots of personal data about users. This is often especially true for many vulnerable groups who may be forced to use online services instead of on-site services, such as the elderly, people with medical issues, or people living in remote locations. We conduct a comprehensive study of 15 Finnish web services often used by vulnerable groups by analyzing their network traffic. Our findings show most of these services use third-party analytics and, despite their delicate nature, send highly sensitive personal data to third parties. The study also discusses the implications of the found data leakages and offers recommendations on how to improve privacy of online services from a software engineering point of view.**

Keywords*—privacy, web services, data leaks, vulnerable groups

I. Introduction

Digital technology provides a valuable means of delivering services to those who may have difficulties accessing them in person [1], [2]. As access to technology and internet connectivity has expanded, policymakers have recognized the importance of digital service delivery. For example, in 2019, Finland passed the Act on the Provision of Digital Services, which aims to improve the accessibility, quality, and security of digital services for all individuals. The COVID-19 pandemic has highlighted the need for online services, and it has further accelerated the shift towards digital service delivery.

In particular, many vulnerable groups can benefit from digital services. Vulnerable groups refer to individuals or communities that are at a greater risk of experiencing negative outcomes or facing challenges because of their socio-economic status, health, age, sexual orientation, or other factors [3], [4]. For example, elderly and older adults may become more vulnerable to health problems, isolation, and other challenges. Similarly, people with disabilities are often more susceptible to lack of access to services and opportunities, as are many immigrants and refugees.

Digital services, which can be easily accessed from anywhere with an internet connection, are particularly beneficial for individuals who may have difficulty leaving their homes or accessing transportation due to financial issues, remote location, disability, or health concerns [5]. Digital services also lower barriers to access, especially for people who may have language barriers, lack of documentation, or other challenges that may prevent them from accessing on-site services. It is also often argued that digital services can provide a higher level of confidentiality than on-site services, which can be beneficial for individuals who may be concerned about privacy or the potential for discrimination. However, online services may not be as confidential and private as expected, especially for many vulnerable groups [6].

This confidentiality aspect of digital services is the focus of the current study. We conduct an in-depth study of 15 Finnish web services often used by vulnerable groups by analyzing their network traffic while using the main functionality of the services. We study what third-party analytics these services use and explore what kind of personal data they transmit to third parties. Some previous studies have addressed third-party analytics services on websites used by vulnerable groups. For example, analytics in web-based governmental services [7], healthcare services [8] and pharmacies [9] have been studied. However, the current study discusses a wider selection of different online services frequented by vulnerable groups, while providing an in-depth analysis of found data leakages.

The rest of the paper is structured as follows. Section 2 outlines the study setting and methods used in the current study. Section 3 presents the results of our network traffic analysis and explores what kind of personal data the studied web services sent to third parties. Section 4 discusses the implications of our findings. Finally, Section 5 concludes the paper.

II. Research Setting and Method

The current study examined 15 web services targeted at or often used by vulnerable groups. Many of the selected services are offered by the Finnish government and cities, but also by different organizations and private companies. These services included among others immigration services, tax administration, public transport websites, mental health related web services, web portals for sexual minorities, as well as a private medical center and online pharmacy. The list of the services is shown in Table I. Aside from the websites of a couple of public

This research has been funded by Academy of Finland project 327397, IDA – Intimacy in Data-Driven Culture.

sector websites, we have opted not to identify the exact services. Instead, the type of service has been described.

The experiment involved testing web services by utilizing their main features and visiting key pages[1]. All cookies were accepted upon accessing the studied websites. Network traffic was recorded using Google Chrome's Developer Tools, which allows for examination of web page source code and network activity. When recording network traffic, caching was turned off and only the requests going to third parties were inspected. The cache was recorded traffic Figure 1 shows an example view of network traffic recorded with Chrome Developer Tools. The recorded web requests were saved as a log file and analyzed. Any data that could be used to identify the user and was sent to third-party analytics services was extracted from the log file. In particular, sensitive personal data was searched.

Finally, it is important to briefly define the term "personal data". In this context, we adopt the definition used by GDPR and the Finnish Office of the Data Protection Ombudsman: personal data is "all data related to an identified or identifiable person"[2] [3]. Examples include IP addresses, location data, and ID numbers used by analytics services to track users. It is important to note that data items that can be combined together to identify a person are also considered personal data. For example, screen size, while not directly an identifying piece of data, can be used in profiling a user. GDPR also defines certain "special categories of personal data" such as data on ethnic origin, health-related data and data concerning a person's sexual orientation[4]. The processing of this type of personal data is generally prohibited under GDPR, unless certain conditions are met, such as obtaining an explicit consent from the individual. Additionally, organizations must implement appropriate technical and organizational measures to protect this type of personal data from unauthorized access and processing.

III. Results

Figure 2 shows the third parties found in our network traffic analysis. Google and Facebook (Meta) are the most frequent third-party services, followed by LinkedIn and React&Share. Despite the legal concerns for the use of Google Analytics in EU [10], Google's services are still widespread even on websites frequently used by vulnerable groups, many of which are essential services. On the other hand, many public sector bodies such as Tax Administration and Digital and Population Data Services Agency have been careful to choose third-party companies based in Finland. On average, there were 2.6 third-party analytics services per website. This number can be considered quite high, as we are talking about web services handling sensitive data of vulnerable groups.

The data shared with third parties includes details like IP addresses, device and user identifiers, User-Agent headers with details about the OS and browser, and other technical data like screen resolution. The device's IP address, which is included in every web request, is a key piece of information used to identify the user [11], [12]. When contextual data, such as information about the page the user is visiting or the action they are taking, is connected with the previously discussed identifying pieces of information, privacy is compromised.

Table I lists the studied websites and presents the related main findings. The studied public sector websites did not contain a large number of third-party analytics services, but it is clear that they still cause some privacy issues. A *Finland information website* had 3 different third-party services such as Google Analytics. As the website consists of a large number of information pages about various topics and offers these pages in several languages (e.g. for refugees and migrants), Google can easily track the topics a user reads and the language they choose.

On the website of *Finnish Immigration Services*, there are 2 third-party analytics services that tracked the pages the user visits. This may be considered highly problematic, as delicate information such as the user accessing information about seeking asylum in Finland is revealed to third parties. Similarly, *Tax Administration* as well as *Digital and Population Data Services Agency* made use of third party analytics, and page visits are tracked. For example, the fact the user is accessing information about a specific tax is revealed to a third party.

Public transit service websites of two Finnish cities were also studied. One of these services contained 7 third party analytics services, and sent search terms input by the user to 5 different third parties. Search terms can include for example destinations the user is intending to reach. The second public transit service website only tracked page visits.

We also analyzed two websites frequently used by sexual minorities. The first one of these is a website of a *sexual minority rights organization*. The subpages a user visits were tracked by three third-party analytics services, including for example pages where the visitor can choose to join different associations related to sexual minorities. Also, the web form used to join organizations is located in an external service that has Google Analytics. This means the user's intention to join a specific association can be revealed. The second website was a *sexual minority web portal*. Visited pages are tracked on this website, including for example a page where users register to a discussion forum and threads on the discussion forum (a thread topic is a part of a URL which was sent to Google).

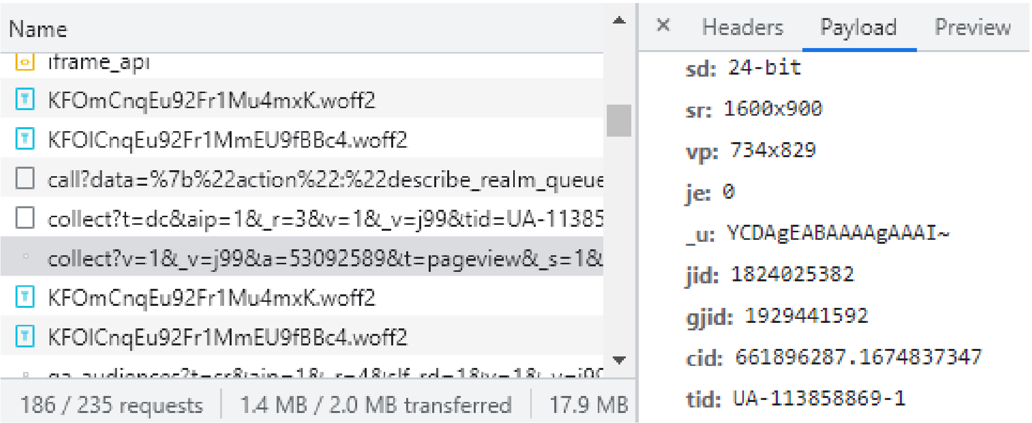
The most glaring privacy issues were found in web services provided by a private *medical center* and *online pharmacy*. When booking an appointment to a doctor

[1] We have thoroughly documented the testing sequence of each web service. This documentation, as well as a list of the tested web services, can be provided upon request.

[2] https://gdpr.eu/eu-gdpr-personal-data/

[3] https://tietosuoja.fi/en/what-is-personal-data

[4] https://tietosuoja.fi/en/processing-of-special-categories-of-personal-data



```
iframe_api
KFOmCnqEu92Fr1Mu4mxK.woff2
KFOICnqEu92Fr1MmEU9fBBc4.woff2
call?data=%7b%22action%22:%22describe_realm_queue
collect?t=dc&aip=1&_r=3&v=1&_v=j99&tid=UA-11385
collect?v=1&_v=j99&a=53092589&t=pageview&_s=1&
KFOmCnqEu92Fr1Mu4mxK.woff2
KFOICnqEu92Fr1MmEU9fBBc4.woff2
186 / 235 requests | 1.4 MB / 2.0 MB transferred | 17.9 MB
```

```
sd: 24-bit
sr: 1600x900
vp: 734x829
je: 0
_u: YCDAgEABAAAAgAAAI~
jid: 1824025382
gjid: 1929441592
cid: 661896287.1674837347
tid: UA-113858869-1
```

Fig. 1: A sample view of Chrome Developer Tools. Network traffic with HTTP requests and a payload with some pieces of identifying data is shown.

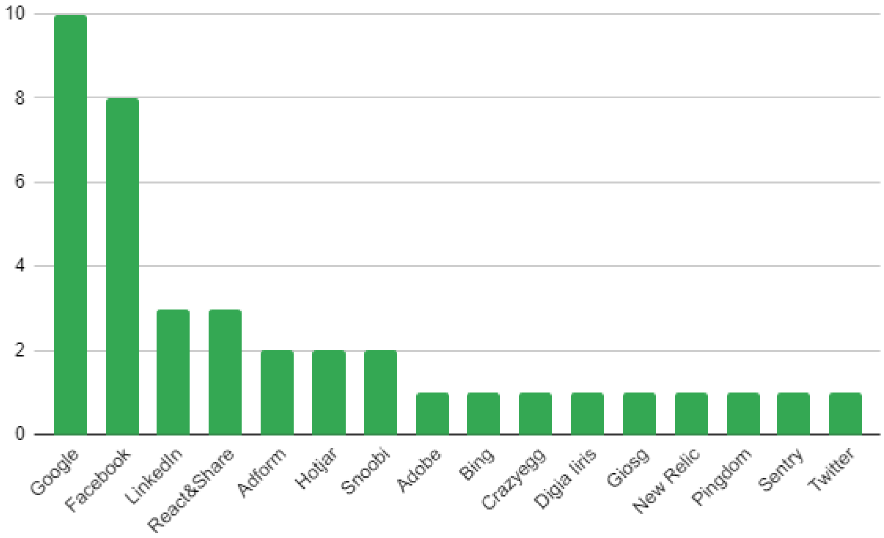


Fig. 2: The third parties found in the network traffic analysis.

on a website of the studied medical center, information about the intended appointment, the doctor's field of specialization (e.g. neurology), as well as the location and date of the appointment were leaked to 4 different third-party analytics services. There were also various sensitive information pages (such as a page about depression), where 5 different analytics services tracked user visits. Altogether, 8 third-party services were present on this medical center website. Possibly even worse, the online pharmacy we studied leaked the information about the prescription medicine the customer intended to order to 3 different third-party services! In addition to this, the medicines the user views were leaked as well. The pharmacy website made use of 4 third parties in total. In both of these cases, sensitive information about a patient's medical conditions can be deduced. It is very likely these privacy leaks are not intended by website developers. At the same time, the users of the websites also remain oblivious to their sensitive health data being leaked.

A *student health services* website had one analytics service tracking page visits. As in other cases in this study, this can be a problem with pages dealing with specific health related topics, such as a peer support group for depressed students. Similarly, a website of a *therapy center* had 2 third-party analytics services recording visited pages – for example, a page for a depression test. A *mental health chat* we studied used Google Analytics. The service was implemented as a single-page application, so it did not track separate page views, but a visit on the website is still reported. On the other hand, a *public sector social and healthcare service* portal was the only website with no third-party analytics in the current study.

We also studied an online service of a *violence victim shelter*, which contains information for violence victims. The website leaked visited pages to 3 different analytics services. Curiously, even when the user interacts with an information page by clicking it to open small sections on different topics, Google receives the information. This gives Google very fine-grained information on different issues and questions a potential victim of violence is

TABLE I: The studied websites and main findings

Description	# of 3rd parties	Main findings
Digital and Population Data Services Agency	2	Tracking of visited pages
Finland information website	3	Tracking of visited pages
Finnish Immigration Service	2	Tracking of visited pages (e.g. instructions on seeking asylum)
Medical center	8	Information about the appointment (date, location, doctor's specialization) leaks
Mental health chat	1	Tracking of visited pages
Online pharmacy	4	Information about the prescription medicine leaks
Public transport service 1	7	Information about searches (e.g. bus destinations) leaks
Public transport service 2	1	Tracking of visited pages
Sexual minority rights organization	3	Tracking of visited pages (e.g. registration pages and forms of different related associations)
Sexual minority web portal	2	Tracking of visited pages (e.g. forum registration page)
Social and healthcare service portal	0	(no third party tracking)
Student health services	1	Tracking of visited pages (e.g. depression peer support)
Tax Administration	1	Tracking of visited pages
Therapy center	2	Tracking of visited pages
Violence victim shelter website	3	Tracking of visited pages and opened of page sections

thinking. The website takes the user's safety into account e.g. by offering an "emergency exit" button for quickly leaving the page and providing instructions on how to clear cache and browsing history. This is good, but the privacy issues seem to be forgotten at least when it comes to third-party analytics. As violence victims may be agitated and in a hurry when visiting the website, it is worth asking whether the use of analytics services is appropriate on such a website. Although visitors may accept the cookies upon arrival, they probably are not in the right state of mind to think about analytics services or assume information on visited pages and opened text sections may leak.

IV. Discussion

We have seen that several websites used by vulnerable groups leak data to third parties, and in many cases this data is highly sensitive. The key findings of the current study are the presence of third-party analytics even on web services provided by the government and leaks of highly sensitive health data in web-based private healthcare services.

Several vulnerable groups may have insufficient digital skills and knowledge to make use of web services in an effective and secure way and understand the possible privacy risks involved [13]–[15]. Problems such as low literacy skills, cognitive decline or sensory impairments can further make using web services and understanding texts such as consent banners and privacy policies more difficult. The vulnerable groups that often have the greatest need for the digital services are also the ones who are likely to suffer most harm when their privacy is compromised [16]. While traditional face-to-face services will continue to be important to vulnerable groups, it is also essential to improve the existing digital services so that they are accessible and secure to use for everyone. Following the privacy-by-design approach when developing the systems and informing the users about third parties in the web service in a clear and transparent manner plays a big role in this development.

Web application developers should be aware of analytics services used in their websites, but different platforms and frameworks used in web development make it very easy to turn on web analytics without much consideration. Therefore, it is important for developers to better analyze what kind of data flows out of their application to third parties, and the use of each analytics service should be well justified. Network traffic analysis similar to this study can be used to observe the outgoing data transmissions.

Web developers may also not fully understand the types of personal data and implications of using analytics services. While the collected data can be used to improve the website's user experience, it is also used for profiling by large analytics companies. As more information is collected over time, analytics services are able to create comprehensive profiles on users and their behavior. In recent years, the use of browser and device information for user profiling on the web has become increasingly sophisticated [17].

It is unclear if the analytics providers actually store and utilize the health data they receive, but it is unacceptable for this information to be shared in the first place. However, the leaked sensitive data is only sent to analytics service providers who may not have a reason to use it, and the data is unlikely to enter the open data market. Utilizing the data effectively may also require manual effort and understanding of the specific website's implementation.

Our results, and the findings concerning healthcare related web services in particular, give reason for further study of similar services in Finland and elsewhere in the world. Aside from the scientific aspirations, we also think that it is very important to quickly improve the privacy of the studied websites. To this end, we have informed the maintainers of the web services with the worst information leaks.

V. Conclusion

In this study, we investigated data breaches on Finnish websites used by vulnerable groups. Out of 15 websites studied, 14 leaked personal data to third parties. The most alarming cases were found in the healthcare sector, where the web services leaked details on intended appointment bookings and prescription medicine orders. Our results call for further investigation using a larger sample of websites used by vulnerable groups. In particular, our findings suggest the confidentiality of web-based services in the healthcare sector requires further study. Furthermore, examining similar web services in other countries would be beneficial. Additionally, future research could explore data breaches in more depth by experimenting and comparing different consent options on the studied websites.

Our results serve as a cautionary tale to software developers and data protection officers in charge of web services handling sensitive data. It is crucial for service providers to be aware of their responsibility in safeguarding customer privacy in areas where they are most susceptible, including increased awareness and control over the chosen third-party services. The use of multiple third-party analytics services on websites used by vulnerable groups is difficult to justify and undermines users' trust. Vulnerable groups should be able to trust the privacy of online services as much as the corresponding on-site services.

References

[1] A. L. Culén and M. Van Der Velden, "The digital life of vulnerable users: designing with children, patients, and elderly," in *Nordic Contributions in IS Research: 4th Scandinavian Conference on Information Systems, SCIS 2013, Oslo, Norway, August 11-14, 2013. Proceedings 4*. Springer, 2013, pp. 53–71.

[2] N. W. Eyrich, J. J. Andino, and D. P. Fessell, "Bridging the digital divide to avoid leaving the most vulnerable behind," *JAMA surgery*, vol. 156, no. 8, pp. 703–704, 2021.

[3] M.-A. Choukou, D. C. Sanchez-Ramirez, M. Pol, M. Uddin, C. Monnin, and S. Syed-Abdul, "Covid-19 infodemic and digital health literacy in vulnerable populations: A scoping review," *Digital Health*, vol. 8, p. 20552076221076927, 2022.

[4] B. L. Chang, S. Bakken, S. S. Brown, T. K. Houston, G. L. Kreps, R. Kukafka, C. Safran, and P. Z. Stavri, "Bridging the digital divide: reaching vulnerable populations," *Journal of the American Medical Informatics Association*, vol. 11, no. 6, pp. 448–457, 2004.

[5] S. O'Sullivan and C. Walker, "From the interpersonal to the internet: social service digitisation and the implications for vulnerable individuals and communities," *Australian Journal of Political Science*, vol. 53, no. 4, pp. 490–507, 2018.

[6] Y. M. Baek, E.-m. Kim, and Y. Bae, "My privacy is okay, but theirs is endangered: Why comparative optimism matters in online privacy concerns," *Computers in Human Behavior*, vol. 31, pp. 48–56, 2014.

[7] A. R. Zheutlin, J. D. Niforatos, and J. B. Sussman, "Data-tracking on government, non-profit, and commercial health-related websites," *Journal of general internal medicine*, pp. 1–3, 2021.

[8] 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*, 2022, pp. 197–211.

[9] A. R. Zheutlin, J. D. Niforatos, and J. B. Sussman, "Data-tracking among digital pharmacies," *Annals of Pharmacotherapy*, vol. 56, no. 8, pp. 958–962, 2022.

[10] S. Winklbauer and R. Horner, "Austrian DPA Decides EU-US Data Transfer through the use of Google Analytics to Be Unlawful," *European Data Protection Law Review*, vol. 8, p. 78, 2022.

[11] V. Mishra, P. Laperdrix, A. Vastel, W. Rudametkin, R. Rouvoy, and M. Lopatka, "Don't count me out: On the relevance of ip address in the tracking ecosystem," in *Proceedings of The Web Conference 2020*, 2020, pp. 808–815.

[12] 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*, 2022, pp. 1–6.

[13] A.-M. Kaihlanen, L. Virtanen, U. Buchert, N. Safarov, P. Valkonen, L. Hietapakka, I. Hörhammer, S. Kujala, A. Kouvonen, and T. Heponiemi, "Towards digital health equity-a qualitative study of the challenges experienced by vulnerable groups in using digital health services in the COVID-19 era," *BMC health services research*, vol. 22, no. 1, p. 188, 2022.

[14] S. Ranchordás, "Connected but still excluded? Digital exclusion beyond internet access," *The Cambridge Handbook of Life Sciences, Informative Technology and Human Rights (Cambridge University Press, 2021, Forthcoming), University of Groningen Faculty of Law Research Paper*, no. 40, 2020.

[15] T. Heponiemi, K. Gluschkoff, L. Leemann, K. Manderbacka, A.-M. Aalto, and H. Hyppönen, "Digital inequality in finland: access, skills and attitudes as social impact mediators," *New Media & Society*, p. 14614448211023007, 2021.

[16] N. McDonald and A. Forte, "Privacy and vulnerable populations," in *Modern Socio-Technical Perspectives on Privacy*. Springer International Publishing Cham, 2022, pp. 337–363.

[17] N. Kaur, S. Azam, K. Kannoorpatti, K. C. Yeo, and B. Shanmugam, "Browser fingerprinting as user tracking technology," in *2017 11th International Conference on Intelligent Systems and Control (ISCO)*. IEEE, 2017, pp. 103–111.