Lessons learned from studying third-party data leaks in web services

Sampsa Rauti
sjprau@utu.fi
University of Turku
Turku, Finland

ABSTRACT

The rise of digitalization has led to a surge in the utilization of web services. Online services offer a convenient way to carry out many everyday tasks. However, third-party web analytics used in many essential web services cause privacy issues as confidential personal data may inadvertently be transmitted to these third parties. In this paper, we examine the reasons for third-party data leaks in web-based services from a software engineering point of view based on our earlier studies as well as the existing literature. We also offer several recommendations and guidelines for developers to alleviate these privacy issues in the future.

CCS CONCEPTS

• **Security and privacy → Web application security**.

KEYWORDS

Data leaks, third parties, online privacy, web security

ACM Reference Format:
Sampsa Rauti. 2023. Lessons learned from studying third-party data leaks in web services. In *2023 8th International Conference on Information Systems Engineering (ICISE 2023), December 16–18, 2023, Bangkok, Thailand*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3641032.3641043

1 INTRODUCTION

The use of web services has become increasingly popular due to their accessibility, scalability, and ease of maintenance. Because of this development, there are various essential web services that process sensitive personal data. However, this raises concerns about the potential data leaks to third-party entities like analytics services. Ensuring that unnecessary data collection does not happen and users are properly informed about the collection of their personal data would be crucial, but is not always realized well in practice. Sadly, this crucial issue is often neglected or poorly executed in practice, and even many essential web services such as digital health care service and governmental websites leak sensitive data without users having any possibility to be aware of this.

The importance of exploring privacy challenges in web development that result in third-party data leaks and finding solutions to mitigate and prevent these leaks can hardly be overstated. As personal data is often highly sensitive, it must be handled with utmost care to protect website visitors' privacy. Data leaks erode users' trust in web services. Users entrusting their personal information to a website expect it to be securely processed. If privacy challenges are not addressed, the result is likely to be a decline in trust and adoption of web services. Additionally, respecting privacy of users is an ethical responsibility for companies, organizations and software developers. For all of these reasons, it is important to examine the reasons for failing to safeguard user privacy and exposing users to data leaks. The methods to mitigate and prevent third-party data leaks in the future must also be explored.

This study investigates the reasons for third-party data leaks in web services from a standpoint of software engineering. We base this discussion on the lessons learned from our earlier studies as well as the existing literature. Furthermore, we also present various recommendations and guidelines for web developers in order to alleviate these privacy issues in the future. The study contributes to the field of online privacy studies by shedding light on the nature of third-party data leaks in web services and factors leading to these design flaws. The research aims to foster a culture of privacy awareness and responsible data handling among organizations and web developers.

The remainder of the paper is structured as follows. Section 2 discusses the nature of third-party data leaks in the web environment and gives some examples of these leaks in different application areas. Section 3 addresses the reasons for data leaks and challenges in software development that lead to such privacy issues. Section 4 provides an examination of ways to prevent third-party data leaks and improve websites as well as the software development process in terms of privacy. Finally, Section 5 concludes the paper.

2 THIRD-PARTY DATA LEAKS

Third-party web analytics have become increasingly common on modern websites. These analytics tools collect and analyze user data, providing insights into website performance, user behavior, and marketing effectiveness. Despite these benefits to companies owning websites, there are legitimate concerns regarding online privacy risks associated with the use of analytics tools.

Potential data leaks are one such risk. We use the term data leak to stress the unnecessary and detrimental consequences of transferring data to third parties. We refer to the cases where personal data inadvertently leaks to third parties potentially causing harm to website visitors, rather than referring to the data breaches where unauthorized access to personal data takes place. Third-party analytics tools can collect personal and identifiable data without users' explicit consent or knowledge. These tools often track and record various identifying data points, including the visitor's IP addresses [5], device specific identifiers and other technical details

ICISE 2023, December 16–18, 2023, Bangkok, Thailand

ACM ISBN 979-8-4007-0917-3/23/12
https://doi.org/10.1145/3641032.3641043

that can be used to single out individual users. Detailed profiles can then be created for individuals using this data, enabling targeted advertising and personalized marketing approaches.

Along with identifying personal data, third-party web analytics tools track contextual data, that is, different user actions. This is done in order to observe how well the user engages with the company website and whether marketing goals are met [7]. The website owner can configure analytics tools to track different user actions [1]. At the same time, the information of this user behavior is also sent to the third party providing the web analytics service. The frequently tracked actions and corresponding data leak types include:

- *Accessing a specific page.* The URL address of the page the user is accessing is transmitted to a third party. The URL of the page alone often conveys sensitive information about the user's interests, intentions and actions that should not be shared.
- *Performing a search.* The search terms the user types in a search field on a website and uses to perform a search are regularly leaked to third-party analytics services. Like visited pages, these terms can also reveal delicate contextual information. Because the user is one who freely decides the contents of the search terms, this data is often even more sensitive than the visited URL addresses.
- *Submitting a transaction.* The different actions and transactions carried out on websites, such as purchasing a product or booking an appointment, may be leaked to third parties. This can often provide a third party with sensitive details on users and intimate actions they undertake in their private lives.

An example of the first type of data leak would be a user accessing a page on a mental health website that discusses and provides instructions on seeking help when having suicidal thoughts. This contextual data leaking along with identifying information can potentially be very harmful to the user. A scenario involving the second type of data leak is, for instance, searching for a specific prescription medicine on an online pharmacy website. An individual's health data is one of the most sensitive types of personal data and also very valuable, and should never be shared with third parties without an explicit consent or other valid legal basis. Finally, an example of the third data leak type is booking an appointment for a doctor on a medical center's website. If the web developers have not paid necessary attention to third parties and their data collection practices, confidential details on the booking such as the name of the doctor and location can be leaked to third parties. Unfortunately, all of these are real-life examples of the data leaks that have been found and addressed several times in our research and the previous studies by other researchers [8, 16].

To summarize, third-party analytics services record identifying information on the user, combined with several types of contextual information. Visits on sensitive web pages, delicate searches, and several transactions and actions that should be kept private can potentially leak to third parties. Web developers and the organizations managing websites often seem to lack a comprehensive understanding of the sensitive data transmissions occurring within their services. This is why it is crucial to both explore the reasons for these software privacy issues and design flaws, and suggest effective methods to mitigate or prevent them from happening.

3 CHALLENGES AND REASONS FOR DATA LEAKS

Popularity of web analytics

Third-party web analytics have become widespread because they allow analyzing website visitors' behavior, measuring website performance, optimizing user experience and tracking conversions. Almost with any modern website, including analytics seems to be a de facto standard which is not questioned all that much. Third-party analytics are included on many websites as a part of a package deal without the customer really expressing any need for this feature. Oftentimes, web developers also do not seem to acknowledge the potential risks the inclusion of analytics may introduce, such as data collection without explicit user consent or the sharing of sensitive personal data with third-party entities [2, 6].

The widespread use of analytics services may often also be the result of the organizations' lack of awareness. When the website development is outsourced, developers are eager to include additional features in the website. The clients are often uncertain about their actual needs, but may be keen to accept third-party analytics with the package because it is a standard practice that supposedly aids in achieving business goals. Potential privacy issues are not considered.

Unfamiliarity with used services

Website maintainers and developers may be unaware what exact personal data items are shared with third parties by web analytics tools. Because of this, the destinations of personal data may also be unknown – specifically the fact that many third-party services may transmit data outside the EU. The fact that third-party services often capture sensitive data on different delicate pages if not configured properly or omitted from those pages may not be apparent to developers or data protection officers [15]. There are also other third-party services, such as tools associated with website performance monitoring and chat services, that may be incorporated in a website without fully realizing the consequences for user privacy.

Many organizations and companies use several web analytics tools, with the goal of getting the most comprehensive information about their web site visitors. As the third-party analytics services are often not researched thoroughly, there may be several overlapping services with the same purpose collecting similar data and leaking it to third parties. The developers may not take the necessary time to understand the features of the used third-party analytics services and peruse their documentations and privacy policies, although this would be crucial to understand their data collection and processing practices.

Use of platforms and content management systems

Modern websites are usually built on top of off-the-shelf software platforms or content management systems (CMSs). These solutions allow the developer to enable third-party analytics very effortlessly, and in many cases, web analytics may already be a built-in feature.

For example, the popular CMS Wordpress offers an easy way to install the Google Analytics plugin, whereas the websites made with the cloud-based website builder Wix include built-in Wix Analytics . While the use of platforms in general is not the problem, developers have to be familiar with the platform they are using and aware of all the third-party connections their website initiates and the personal data it potentially sends out as a result of using ready-made code.

Unclear requirements and poor stakeholder communication

When requirements are defined inadequately and poorly, misunderstandings often follow. In terms of web privacy, this will most likely lead to a website that does not meet the needs of its users. This is often quite apparent as the client has failed to effectively communicate to the developers that the user's data should be safeguarded in a specific situation or website function. This can result in the presence of third-party analytics in places such as violence shelter websites or whistleblowing services where they are completely unnecessary and pose a threat to confidentiality. The issue may become particularly challenging for small companies with limited resources and a limited understanding of data protection. Developers can make their own incorrect assumptions about the scope of data collection. They may also inadequately inform clients about how personal data is collected, and which parties it is shared to. Lack of stakeholder involvement easily leads to misalignment and unsatisfactory project outcomes [9, 13].

Blind spot in privacy-by-design

While privacy is usually taken very seriously when building software systems processing sensitive data, many websites seem to be an exception when it comes to third-party data leaks. Websites are a crossing point where tech giants get the possibility to spy on confidential communication and sensitive data transfers between website users and organizations, companies or other service providers. It can be seen as an interface where at the same time, users are inputting sensitive personal data and analytics companies are harvesting data under the guise of providing business insights to companies and improving user experience.

Unfortunately, these two functionalities coincide quite frequently. One can simply think about a user booking a doctor's appointment on a website of a medical center or ordering a prescription medicine from an online pharmacy. These transactions should be highly confidential but they are also very interesting from a conversion tracking point of view. Too often this means that technology giants such as Google or Meta are present to record the transaction and gather the data.

Unfamiliarity with the application area

If developers are unfamiliar with the application area the developed website falls into, this can pose challenges for the development process when it comes to online privacy considerations. When developers lack an understanding of the specific field or industry in which a web service operates, they may unintentionally overlook or underestimate the privacy implications of data collection by third parties.

Due to this lack of knowledge, developers may not be aware of specific data items that are collected (e.g. a name of a prescription medicine a user has ordered) and the potential consequences leaking these details has for user privacy. This can lead to unintended disclosure of sensitive user information without appropriate consent. When privacy requirements of a specific application area are not known to developers, it is difficult for them to assess how necessary different data collection practices and third-party services are for the web service. Personal data is then collected and shared with third parties simply because this is usually done or perceived as necessary.

Ignoring varying requirements of subsites

Many larger websites have different subsites or subsections, that is, websites subordinate to the main website and hosted under the main site's domain name. The challenge here is that the whole website often uses the same third-party analytics services. The fact that subsites and separate web applications may have varying privacy requirements is regularly forgotten. A good example of this are various voting advice applications provided on websites of Finnish media companies. As the media company's website with its news articles and advertisements is quite a different environment as the voting advice applications where the user's political opinions may be leaked to third parties, this difference should also be reflected in the third parties present in the voting advice application. Unfortunately, this is not always the case.

Many organizations and companies may also have individual sensitive pages where visits and user behavior should not be tracked. While web analytics and tracking on a university website may not sound that serious at first, there may be pages related to seeking help from mental health issues or reporting harassment cases, for instance. Web developers should exercise caution with pages containing this kind of content and always make sure information about the user displaying them is not leaked.

Dark patterns and poor transparency

Consent management platforms are typically used to display cookie banners on websites and help in complying with data protection regulations. This enables the website operator to inform users properly and provide them with choices regarding cookies and data collection. Ironically, however, cookie banners are often designed in a way that persuades users to accept tracking cookies and data collection [12]. These choices often benefit the website owner, and ultimately, the third parties, rather than the users' privacy.

Dark patterns come in various forms, such as making the accept button in a cookie banner more visible and enticing than the reject button [14]. The user may easily press the button routinely without thinking, unwittingly accepting data collection practices. In such cases, it becomes questionable whether the action qualifies as making an informed decision and providing conscious consent. Also, even if the user is vigilant and chooses minimum cookies and data collection, the website developer is still the one who decides what "necessary cookies" are on a specific website, as the term is subject to interpretation. Even when the user reads the privacy policy provided on the website, all too often they still cannot be sure of what

personal data is shared with third parties and who these parties are, as the provided information is regularly incomplete.

4 SOLUTIONS AGAINST THIRD-PARTY DATA LEAKS

Using self-hosted analytics

For most data collection purposes, choosing a self-hosted analytics solution would be wise and sufficient. This approach ensures that the personal data collected by a company or organization remains within its own infrastructure and data is not unnecessarily shared with any third parties. In other words, the gathered data is processed and stored locally and the owner of the website maintains complete control over it. Instead of sharing data to several external parties, data privacy is prioritized in this approach and the risk of data being accessed or used by unauthorized parties is minimized.

In-house analytics give the organization freedom to fully define its data handling practices. Compliance with privacy regulations and data confidentiality can be ensured firsthand. One example of a free and open-source web analytics platform that can easily be deployed locally is Matomo [3, 4]. In addition to giving the website owner full control over data, Matomo also strives to provide privacy-focused data collection by offering features such as not storing visitors IP addresses and honoring Do Not Track requests.

Network traffic analysis

As third-party tools and libraries become more complex, developers are not always able to keep themselves in the know about all the outgoing connections their website might have. The URLs the user has visited and the data they have input on the website may be transmitted to third parties [11]. This is why it is very important to analyze third-party requests, their payloads as well as their destinations. This is the only way to detect the actual leaks of personal data, as privacy policies and documentations of third-party analytics tools may also fail to report some aspects of their functionality. Analytics tools can also be inadvertently configured to capture events and data that should not be sent to third parties.

There are many different tools to monitor and capture third-party requests on a website. The Google Chrome browser provides developer tools that can be used to inspect network traffic and allows the user to use a filter to only see third-party requests. The payloads of these requests can also be analyzed. Developers should take time to make sure that identifying personal data is not unnecessarily sent to third parties. The same also goes to many contextual details such as URL addresses of sensitive pages and information about performing sensitive actions (say, for instance, seeking crisis assistance for mental health related problems). This kind of careful examination of network traffic should always be part of performing security testing for a website. The process should be repeated regularly as the functionalities on the website are updated, pages are added, and analytics services may change.

Careful assessment of third parties

The third-party services and tools used on a website should be carefully evaluated. The inclusion of each third-party service that collects user data is an architectural decision that should be justified and documented. Every third party should be necessary and the functionalities of these services should not overlap. There are too many developers (and platforms) who decide to employ third-party services and trackers simply because that is a prevalent trend among many websites and has become a common practice in the digital landscape.

The used third-party services may also be evaluated as a part of data protection impact assessment. This is an important process for organizations and involves making sure that they comply with data protection regulations like the GDPR. This process involves assessing the potential risks and negative impacts on individuals' privacy and identifying measures to minimize or eliminate the found risks.

One important method of strengthening privacy is data minimization [10]. It refers to the principle of only collecting, processing, and storing the minimum amount of personal data needed to satisfy a specific purpose. When gauging the effects a third-party service has on the user's privacy, it is important to consider what kind of data they process and whether this is really necessary. For instance, Google Analytics, which usually collects quite a lot of identifying and contextual personal data which may not be kept in Europe, may be evaluated to pose a substantial risk to user privacy. In this case, developers have to consider using some other service collecting less data, using self-hosted analytics, or getting rid of web analytics altogether.

Privacy audits

When a website includes sensitive data processing, a privacy audit should be carried out to evaluate third party services and detect data leaks. It is also important to ensure compliance with privacy regulations. For highly sensitive contexts such as medical center websites and online pharmacies, an external privacy audit would be beneficial. External and independent auditors bring an objective and unbiased perspective to the assessment process. The auditors can offer insights and recommendations based on their knowledge of industry best practices and privacy regulations. Aside from the technical implementation and privacy measures, audits also assess the organization's privacy policies and notices in order to ensure the user is adequately informed about collection, use, sharing and protection of personal data. This also encompasses evaluating the transparency of information given to websites users and verifying that policies align with legal requirements and actual processing of personal data carried out by the organization and potential third parties. Additionally, cookie notices and potential dark patterns they contain should be evaluated.

Understanding the application area

Developers should strive to gain a solid understanding of the application area or industry the website is being built for. By immersing themselves in the intricacies of the specific domain, they can better comprehend the unique data protection requirements that arise in the particular field or sector. Seeking guidance from application area experts when needed and frequently communicating with the key stakeholders is also crucial. Domain-specific knowledge is needed to make informed decisions on the website's privacy aspects and whether third-party services can and should be included.

Seeking guidance from application area experts can significantly contribute to the software development process and makes applying the privacy-by-design approach easier in a specific domain. Experts can offer knowledge and insights that help developers successfully address the challenges of data protection within the given industry. Developers should maintain frequent communication and actively engage with key stakeholders such as clients, users, and data protection officers. This gives diverse perspectives and makes it possible to align the website's privacy design with the requirements of all parties involved in the development process. The acquired domain-specific knowledge is then used developers to make the architectural decisions on the potential third-party services in the website.

Privacy culture

Cultivating a culture of privacy awareness in web development involves fostering a mindset that prioritizes protecting users' personal data. It is important to instill an understanding of privacy principles and regulations, such as GDPR, among team members. This includes educating developers and testers about data protection, potential risks of mishandling data, dangers associated with third parties, and the consequences of non-compliance. This way, the development team can make more informed design decisions throughout the web development process and integrate privacy considerations into every stage of development. Promoting a culture of privacy awareness requires continuous education and adaptation to evolving privacy regulations. This helps in avoiding data leaks and building and maintaining trust with users.

5 CONCLUSIONS

Privacy concerns and data leaks can erode user trust in the web services and the organizations behind them, causing reputational damage. If users learn that their privacy has been compromised, they may be hesitant to use the website and input their sensitive personal information in the future. This mistrust is likely to affect customer satisfaction and overall business success. Unfortunately, websites still seem to be a blind spot in terms of adequately protecting sensitive user data, as third-party analytics and confidential information submitted by the user meet with unexpected but grave consequences.

This paper has discussed the factors leading to insufficient privacy and data leaks in many essential web-based services. We hope that privacy practices such as using self-hosted analytics, comprehensive network traffic analysis, careful review of third parties, thorough privacy audits, gaining an improved understanding of the application area, and solid privacy culture can help to make difference and improve online privacy in the future.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Paschalis Bekos, Panagiotis Papadopoulos, Evangelos P Markatos, and Nicolas Kourtellis. 2023. The Hitchhiker's Guide to Facebook Web Tracking with Invisible Pixels and Click IDs. In *Proceedings of the ACM Web Conference 2023*. 2132–2143.

[2] Ari B Friedman, Lujo Bauer, Rachel Gonzales, and Matthew S McCoy. 2022. Prevalence of third-party tracking on abortion clinic web pages. *JAMA Internal Medicine* 182, 11 (2022), 1221–1222.

[3] Jonas Gamalielsson, Björn Lundell, Simon Butler, Christoffer Brax, Tomas Persson, Anders Mattsson, Tomas Gustavsson, Jonas Feist, and Erik Lönroth. 2021. Towards open government through open source software for web analytics: The case of Matomo. *JeDEM-eJournal of eDemocracy and Open Government* 13, 2 (2021), 133–153.

[4] Timi Heino, Robin Carlsson, Sampsa Rauti, and Ville Leppänen. 2022. Assessing Discrepancies between Network Traffic and Privacy Policies of Public Sector Web Services *(ARES '22)*. Association for Computing Machinery, New York, NY, USA, Article 11, 6 pages. https://doi.org/10.1145/3538969.3539003

[5] Peter Hense. 2020. Google Analytics: Injunctive relief, information requests and damages. In *Turning Point in Data Protection Law*. Nomos Verlagsgesellschaft mbH & Co. KG, 151–156.

[6] Xuehui Hu and Nishanth Sastry. 2020. What a tangled web we weave: Understanding the interconnectedness of the third party cookie ecosystem. In *Proceedings of the 12th ACM Conference on Web Science*. 76–85.

[7] Alicia Huidobro, Raúl Monroy, Manuel A Godoy, and Bárbara Cervantes. 2022. A contrast-pattern characterization of web site visitors in terms of conversions. In *Technology-Enabled Innovations in Education: Select Proceedings of CIIE 2020*. Springer, 31–51.

[8] Mingjia Huo, Maxwell Bland, and Kirill Levchenko. 2022. 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*. 197–211.

[9] James E Post, Lee E Preston, and Sybille Sauter-Sachs. 2002. *Redefining the corporation: Stakeholder management and organizational wealth*. Stanford University Press.

[10] Awanthika Senarath and Nalin Asanka Gamagedara Arachchilage. 2019. A data minimization model for embedding privacy into software systems. *Computers & Security* 87 (2019), 101605.

[11] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. 2022. Leaky Forms: a study of email and password exfiltration before form submission. In *31st USENIX Security Symposium (USENIX Security 22)*. 1813–1830.

[12] Alina Stöver, Nina Gerber, Henning Pridöhl, Max Maass, Sebastian Bretthauer, I Spiecker, M Hollick, and D Herrmann. 2023. a How Website Owners Face Privacy Issues: Thematic Analysis of Responses from a Covert Notification Study Reveals Diverse Circumstances and Challenges. *Proc Priv Enhanc Technol* (2023).

[13] Smiju Sudevan, M Bhasi, and K Pramod. 2014. A typology of stakeholder identification methods for projects in software industry. *MESM'2014* (2014), 1–5.

[14] Christian Voigt, Stephan Schlögl, and Aleksander Groth. 2021. Dark patterns in online shopping: Of sneaky tricks, perceived annoyance and respective brand trust. In *International conference on human-computer interaction*. Springer, 143–155.

[15] Xiufen Yu, Nayanamana Samarasinghe, Mohammad Mannan, and Amr Youssef. 2022. Got sick and tracked: Privacy analysis of hospital websites. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 278–286.

[16] Alexander R Zheutlin, Joshua D Niforatos, and Jeremy B Sussman. 2022. Data-tracking among digital pharmacies. *Annals of Pharmacotherapy* 56, 8 (2022), 958–962.