User Data on the Internet

Amol Bhagavathi - amol-bhagavathi@uiowa.edu

The Internet allows users across the world to conveniently gain and share information. Additionally, the Internet has provided to be a rich environment for companies, such as Google, Microsoft, and Meta, to provide meaningful services for users. However, such services can be supported by advertising to a large user base. To make advertisements more effective, user information is often collected and used to match certain advertisements to certain individuals, raising valid concerns about data privacy. In this paper, I will provide a background on the topic, explore the technical aspect of how companies make targeted advertising feasible, and provide methods to prevent user data malpractices.

user data | targeted advertising | data privacy

Introduction

In an effort to shape technological governance, the European Union (EU) recently implemented the Digital Services Act (DSA). The act was primarily designed to shield consumers from malicious entities, such as scammers, operating within online marketplaces. Additionally, the DSA carries an even broader goal: to increase transparency between corporations and their users. Beyond its original goals, this act protects individuals from targeted advertising based on sensitive information like "age, ethnicity, political affiliation, and sexual orientation" (1).

From the DSA's goals, it is apparent that the DSA was aimed at Internet companies that connect buyers, like me, and sellers, such as Hoka, on the Internet. For instance, industry giants like Google and Amazon bridge the gap between buyers, searching for a desired product, and sellers, potentially selling said product. Given the large user base these companies have, they have the ability to access, retain, and use the wealth of user data they amass.

In the year 2010, the Internet had approximately 2 billion users. Barely a decade later, the Internet grew to 4.7 billion users (2). With the substantial growth in Internet users, as well as the centralization of services to a select few corporations, there have been concerns about safeguarding data integrity across the Internet.

Technical Details - From Acquiring Users to Using Gathered Data

A. Attracting Users. To gather user information, Internet companies have to convince users to use their product. While obvious, this foundational step allows the company to exchange their service for information about their user. To acquire users, Internet companies offer free and valuable services for consumers (3). For example, Google lets users search, publish, and store information through the multiple applications provided in the GSuite. Meta and LinkedIn provide social media platforms, where users can post updates and connect with friends. To use these services, users need to create an account, which ties their usage and other provided information to their identity. This strategy combines useful products with easy access, lowering the barrier for users.

B. Collecting User Data. Collecting, storing, and using user data is a technically complex task. Therefore, it is important that these companies have systems set up that can scale in proportion to the number of users gained. There are two primary ways companies gather customer information online: passive and active user data collection (3).

Passive user data collection leverages web technology to gain insights on its users and applications. For instance, Hypertext Transfer Protocol (HTTP) Logging captures each interaction between a web client and server, compiling a comprehensive record of their transactions (3). Stored as HyperText Markup Language (HTML) files, these transaction logs hold data that can be parsed and displayed by specialized software. The insights from this process aids in assessing the effectiveness of of business endeavors, such as advertising campaigns. It's important to note that this data is not attached to specific users, unless coupled with other technologies. Figure 1 depicts some of the information contained within the HTML files.

Other the other hand, cookies are identifying files dispatched from a web server to a computer. More specifically, a cookie is a tiny text file automatically delivered to a device and stored by the web browser. It holds a unique identifier, a string of random characters, effectively identifying an individual's browser to facilitate linkage to databases with user usage data and inferred interests (4). While smartphone apps differ from web browsers, they rely on similar technology to link devices with users (3). For example, when logged onto an app like Instagram, cookies ensure seamless session continuity between devices.

Unlike passive user data collection, active user data collection relies on users actively contributing information, often as a way to access the services offered. For example, GSuite applications, such as Google Drive, Photos, and Gmail. This kind of interaction requires user engagement with the platform to provide their information, consequently linking the information to their account. In contrast to passive data collection, the user is actively a part of the data collection process, potentially negatively affecting user engagement and sign-ups (3).

The data is usually stored in databases, places where large amounts of information can be organized and structured to the business's needs (5). I have personally used databases, such as MongoDB, Firebase, and surprisingly, Google Sheets, for data storage and retrieval. The information within the databases, when combined with complex algorithms to predict user behavior, can give a company a competitive advantage within business operations.

C. Using User Data. Depending on the chosen data collection method, companies can use their amassed user data in multiple ways. When relying only on HTTP Logging, the collected data can be parsed and presented in a easy to understand format that suits the company's needs. This immediate overview proves invaluable for quickly assessing application trends. For example, I have used this strategy for my personal website, as shown in Figure 2. This mechanism allows me to discern which sections of my application have the most traction, enabling me to prioritize those sections. Similarly, advertisers harness this technology to monitor the success of ad campaigns across applications, ensuring that resources are channeled towards the most responsive platforms.

But how does the technical parsing of HTTP Logging files work? Web analytics platforms such as Google Analytics, Adobe Analytics, and Vercel Analytics are commonly used tools I have encountered to track, dissect, and display website traffic patterns.

While cookies are slightly different than HTTP Logging files, cookies have a similar underlying technical processes to harvest the data. Cookies fashion user profiles by associating a series of user interactions over time. Some cookies are tailored for cross-device tracking, giving companies a holistic view of user behaviors across multiple devices and platforms. This cohesive perspective allows for consistent user experiences by loading cookies from prior states from different devices onto the current device being used (3).

These methods are combined with historical insights to create real-time tailored user experiences that respond quickly to user interactions. Some examples include InfoSeek, who leveraged cookies to refine search results, and Lycos, who added banner advertising influenced by past searches. Doubleclick, on the other hand, employs cookies to gauge user exposure to specific ads, thereby optimizing ad delivery (3). This strategic approach prevents advertisers from wasteful expenditure on unproductive sites or users.

D. How Data is Misused. The surge in data collection from users has revealed a pressing concern about the potential malpractice with such information. Companies offering multiple free services often rely on advertising as their primary revenue source. Therefore, more identifiable data is seen as a competitive edge within the industry. Advertisers are willing to invest more in precise, targeted advertising, incentivizing companies to prioritize data collection over user privacy.

For example, Google has come under scrutiny from European data regulators on multiple occasions. In the past, it incurred a €50 million fine from France for unclear data consent policies. Furthermore, Ireland's Data Protection Commission investigated allegations that Google violated GDPR regulations by sharing personal data with advertisers (6).

Similarly, Uber came under similar scrutiny over data privacy. Despite a privacy policy that explicitly stated that employees should not access customer ride histories, Uber staff exploited a tool known as the "God View" to track journalists, politicians, and celebrities (6).

These data-related malpractices extend beyond surveillance. The U.K.'s Information Commissioner's Office issued fines to two affiliated entities—Leave.EU, a pro-Brexit campaign group, and Eldon Insurance—for indiscriminately employing personal data in marketing campaigns without acquiring proper consent (6). These examples underscore the complex and conflicting interests surrounding data usage and user privacy.

How to Prevent Data Misuse/Malpractices

Because the interests of service providers and users often clash, it's important to find ways to make sure user data is collected, stored, and used correctly according to the regulations created by the government. One way to do this is to make it more costly for companies if they don't follow the privacy laws (7). However, there needs to be reliable methods to catch companies with user data malpractices.

For example, in California, companies have to put a "Do Not Sell My Personal Information" (DNSMPI) button on their websites. It can be easily checked if a website has this button by parsing the webpage's structure, or Document Object Model (DOM), for a DNSMPI button. But some actions, like secretly moving personal data, happen in hidden places like private servers, making it harder to track (7).

One way to catch these transactions is through closed book audits. In these audits, fake data is sent to websites. This can change how parts of the website look or work. By measuring how significant these changes are, it can be inferred whether the company in question has valid data privacy practices. For example, if the changes go over a certain limit, can be reasonably inferred that the company might be engaging in hidden transactions.

By using these methods, the cost of breaking data privacy regulations increases, incentivizing companies to can make sure they follow regulations to keep user data safe, even though their interests don't align with user interests.

Implications to Users

Most data collected from the users are not inherently dangerous. For example, a name, birthdate, email, etc. will not cause significant damage to the user, unlike collecting credit card or financial information. Therefore, most user data collected is inherently non-dangerous data. However, what does it mean when the data is wrongly handled?

For very serious causes, data handling malpractice can lead to users ingesting content they did not intend to ingest. This can lead to manipulation, such as political manipulation like that that occurred in the Facebook-Cambridge Scandal (8).

However, the mishandling of user data, even when seemingly harmless, can have far-reaching impacts beyond immediate damage. One potential concern is identity theft or fraud. While a user's name, birthdate, and email might not seem harmful on their own, when combined with other pieces of information available online, they can be used to create a detailed profile that malicious actors can exploit for various illegal activities.

Furthermore, data mishandling can erode user trust and privacy. When users provide their information to a service, it is expected that the data is treated with care. Any breach of this trust can result in reputational damage to the company responsible and might drive users away from the platform.

In terms of manipulation, improper data handling can enable more than just content ingestion. It can facilitate the creation of personalized, misleading narratives or advertisements that exploit individuals' vulnerabilities and biases. This type of manipulation can influence opinions, behaviors, and even decisions users make, as shown in the Facebook-Cambridge Scandal.

On the other hand, the responsible collection and utilization of data, can offer substantial benefits to users and society. Google, even with its past issues with data privacy, uses data to improve search results, enhance user experience, and offer personalized recommendations. This can save users time and provide them with content that aligns with their interests.

Additionally, aggregated and anonymized data can be leveraged to gain insights into broader trends that are insightful to the public. For example, Google's Flu Trends utilized search data to estimate flu activity in various regions, offering valuable information to public health officials (9). This illustrates the potential for data to serve the greater good when used responsibly and ethically.

Conclusion

The differences between company interests and user privacy is a central concern within the ever-changing landscape of user data collection. While the digital realm presents opportunities for personalized experiences and improved services, the misuse of data remains a critical challenge. This paper explored data handling, from passive and active collection methods to the potential consequences of mishandling seemingly harmless information. Additionally, this paper explored the implications on the user and potential was to ensure safe data practices within companies.

Bibliography

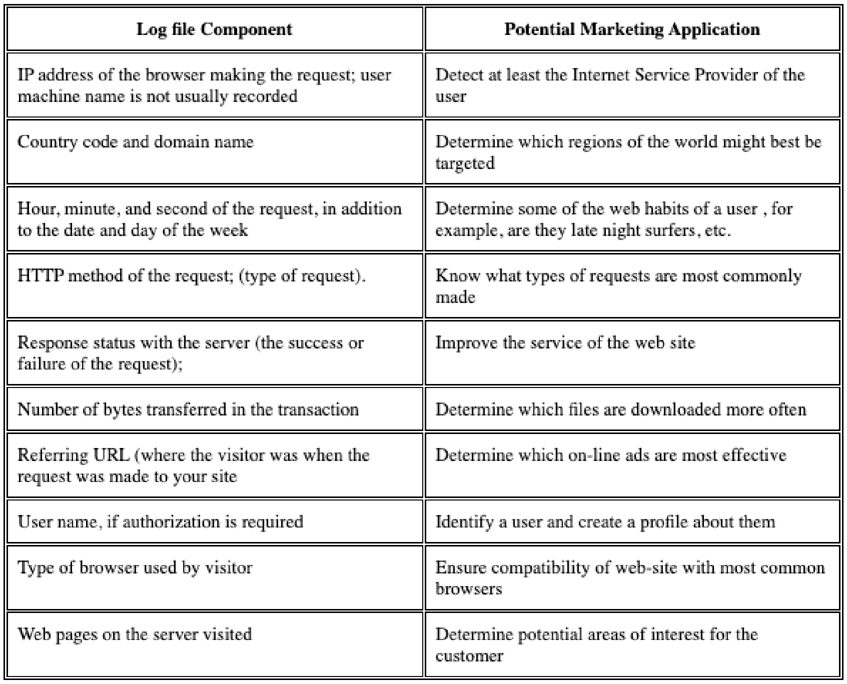
1. European Commission. Questions and Answers: Digital Services Act*. 4 2023.
2. Hannah Ritchie, Edouard Mathieu, Max Roser, and Esteban Ortiz-Ospina. Internet. *Our World in Data*, 2023. https://ourworldindata.org/internet.
3. Catherine Scovitch Koji Takamura Yoriko Matsuda, Paul Rosenstein. Direct marketing on the internet. 2023. https://web.mit.edu/ecom/www/Project98/G2/dmi.htm.
4. Wikipedia contributors. Web tracking — Wikipedia, the free encyclopedia, 2023. [Online; accessed 31-August-2023].
5. What is a data store? https://aws.amazon.com/what-is/data-store/.
6. Swaroop Sham. What is data misuse? 2020. https://www.okta.com/blog/2020/06/data-misuse/.
7. Mihailis Diamantis, Maaz Bin Musa, Lucas Ausberger, and Rishab Nithyanand. Forms of disclosure: The path to automated data privacy audits. *62 Harv. J. L. & Tech.*, 2023. doi: 10.2139/ssrn.4458285. Forthcoming.
8. IGA KOZLOWSKA. Facebook and data privacy in the age of cambridge analytica. 2018. https://jsis.washington.edu/news/facebook-data-privacy-age-cambridge-analytica/.
9. Wikipedia contributors. Google flu trends — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Google_Flu_Trends&oldid=1170206551, 2023. [Online; accessed 31-August-2023].

Reference Usage

Sources 1 and 2 were used in the introduction of the essay to provide background information. Sources 3 and 4 where mainly used when describing the technical details of user data on the Internet. Source 5 was used to explain how user data is stored. Source 6 was used to identify instances of data misuse. Source 7 was used to identify potential solutions to identify data malpractices. Source 8 was used to highlight a specific instance of data malpractice. Source 9 was used to highlight a specific instance of the usefulness of data. The 2 figures were used as a visual for data collection mechanisms.

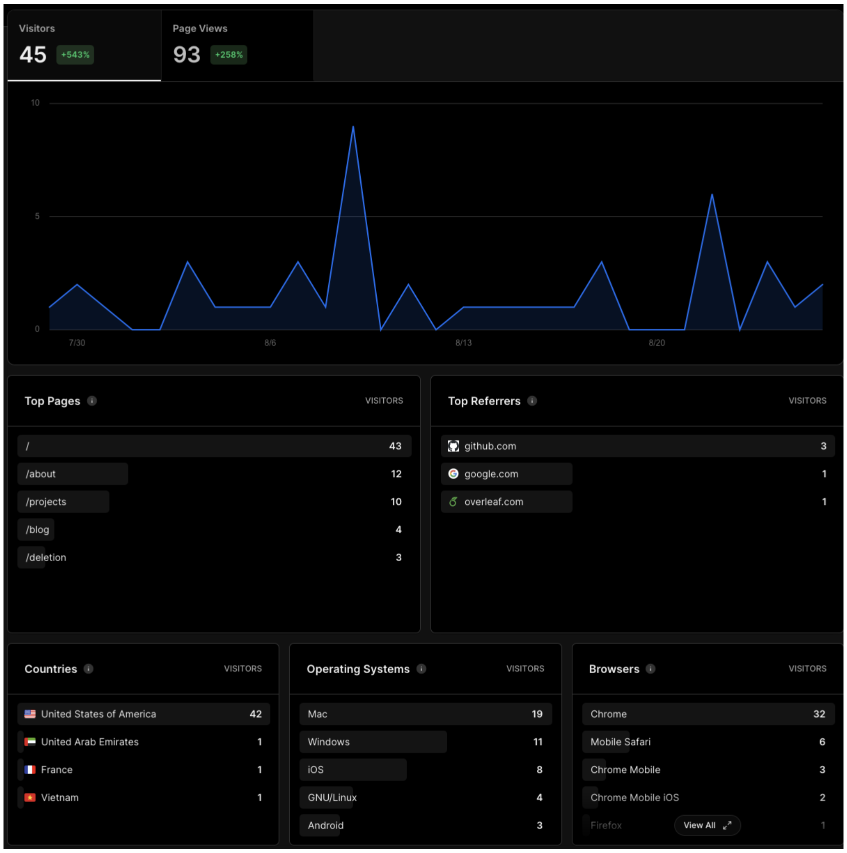
Appendix

See next page



Log file Component	Potential Marketing Application
IP address of the browser making the request; user machine name is not usually recorded	Detect at least the Internet Service Provider of the user
Country code and domain name	Determine which regions of the world might best be targeted
Hour, minute, and second of the request, in addition to the date and day of the week	Determine some of the web habits of a user , for example, are they late night surfers, etc.
HTTP method of the request; (type of request).	Know what types of requests are most commonly made
Response status with the server (the success or failure of the request);	Improve the service of the web site
Number of bytes transferred in the transaction	Determine which files are downloaded more often
Referring URL (where the visitor was when the request was made to your site	Determine which on-line ads are most effective
User name, if authorization is required	Identify a user and create a profile about them
Type of browser used by visitor	Ensure compatibility of web-site with most common browsers
Web pages on the server visited	Determine potential areas of interest for the customer

Fig. 1. Image taken from (3) showing data collected in HTTP Logging files.



Top Pages	VISITORS
/	43
/about	12
/projects	10
/blog	4
/deletion	3

Top Referrers	VISITORS
github.com	3
google.com	1
overleaf.com	1

Countries	VISITORS
United States of America	42
United Arab Emirates	1
France	1
Vietnam	1

Operating Systems	VISITORS
Mac	19
Windows	11
iOS	8
GNU/Linux	4
Android	3

Browsers	VISITORS
Chrome	32
Mobile Safari	6
Chrome Mobile	3
Chrome Mobile iOS	2
Firefox	1

Fig. 2. Image showing analytics from my personal website.