Analyzing Web Tracking

Sharat Dhananjaya
Northwestern University
sharatdhananjaya2023@u.northwestern.edu

Emily Kohlberg
Northwestern University
emilykohlberg2024@u.northwestern.edu

Aining Li
Northwestern University
annieli2025@u.northwestern.edu

David Zhang
Northwestern University
davidzhang2023@u.northwestern.edu

ABSTRACT

For many companies, cookies and other trackers are vital for collecting data about their users. For instance, online websites might use Google Analytics to gather statistics and analytics for search engine optimization and other marketing purposes.

We combed through trackers used on popular websites using a free, online tools, Ghostery. By gathering this data, we hope to educate and inform users about how their personal data is being gathered, outside of what is mentioned in companies' privacy policy.

1 INTRODUCTION

Our main goal was to determine how many trackers and what kinds of trackers were being used across many of the most popular websites. This included websites related to food, social media, entertainment, and shopping among other areas. Using the Ghostery tracking tool and AdBlocker, we compiled lists of trackers across 50 different websites. Then, we did an in-depth analysis in which we broke down the trackers into different types such as advertising, consent management, customer interaction, and email among others. Using that information, we compiled averages for the number of trackers used in each website category as well as the average numbers for each tracking type based on the website category. Furthermore, we gathered information on the top 10 trackers and the percentage of each tracker type by website category.

We also parsed through the privacy policy information of various websites to determine what information they claim to store. We then compared this to what we found using Ghostery and AdBlocker to determine any discrepancies. If the website uses tools or frameworks, such as Google Analytics, on the users' data without disclosing this in their privacy policy, we want to share this with the user.

2 BACKGROUND

Most people browse countless websites in a day, and we are becoming more and more aware of the ways that websites are tracking more and more of our data. This might be concerning to some who value their privacy or at least want to be more informed about what information these websites are actually tracking/collecting, but most people simply don't have the time or patience to look through every single website they visit to understand what data these websites are using. A further issue lies in the fact that not all privacy policies are useful to an average person who might not understand the technical jargon or specific privacy concepts. There are also problems with vague wording or companies not properly disclosing what and how information is being collected. Furthermore, as cloud computing rapidly increases in popularity, users expect to have access to their applications in cloud environments and access them from anywhere in the world [4]. As a result, ensuring that user data is secure and safely stored remotely is of high priority for any website or company that collects information from users [4].

With the emergence of the digital economy, users' privacy is threatened by the collection and leakage of personal information. Websites use various tracking components such as cookies, LSOs, Javascript, Iframes, cloud computing, and remote servers. Using an HTTP proxy, the website will modify HTML pages by adding Javascript code before delivering them to the client [1]. Using this tracking code, a company is able to collect information about mouse movements, keyboard input, and other user movements [1]. One approach to measuring the privacy of user data is to assign scores based on tracking components used on a certain website. For instance, there could be LSO and cookie scores based on cookies Time to Live (TTL) [2]. In addition, there could be points assigned based on the presence of iframes or Javascript which do not use a TTL since they do not persist on the client side [2].

After the introduction of Content Delivery Networks (CDNs), there was additional concern regarding the caching of user content in a network of servers [3]. Websites can use these CDNs along with other third-party services such as advertisements, trackers, analytics, and social integration widgets to increase the effectiveness of user tracking [3].

With how ubiquitous the idea of web tracking has become, more people are beginning to place a larger emphasis on protecting their privacy and want to better understand what ways websites might be tracking them beyond what they already know. Oftentimes, people do not know the extent to which their data is being stored and distributed. Without robust security measures in place, stored personal information can be susceptible to data leaks and privacy attacks.

3 DATA

To get a good idea of what data different websites collect and track, we collected and analyzed many popular websites.

After searching for the most visited websites and reflecting on our personal usual Internet usage, we selected and listed 55 popular sites, such as Google, YouTube, and Twitter, and decided to analyze how they collect and gather data. In addition to wanting to understand how the most visited sites use trackers to collect information, we wanted to understand how different categories of sites differed in terms of the information they tracked. So we picked a wide range of website categories, covering social media, online shopping, food, search, news, government, entertainment, workplace/education, and booking. Although sites such as Target and CVS under the online shopping category were not the most viewed sites, we expected shopping sites to use the tracker extensively, so we selected many shopping sites for analysis.

The full list of 55 websites we analyzed is shown below.

- Social media: Twitter, Instagram, Facebook, Pinterest, whatsapp, Reddit, discord, TikTok
- Online Shopping: Macy's, Target, CVS, Walgreens, Walmart, urban outfitters, Zara, ASOS, Nordstrom, wholefoods market, kohls, trader joes, world market, eBay, riteaid
- Food: Mcdonalds, chipotle, Panera bread, Wendy's, Burger King, Taco Bell
- Search: Yelp, Google, Wikipedia
- News: Bloomberg, weather
- Government: illinois.gov, wa.gov, usa.gov, cia.gov
- Entertainment: YouTube, Netflix, Hulu
- Workplace/Education: northwestern, slack, quora, Coursera, khan academy, zoom, Gmail, Outlook, LinkedIn, piazza
- Booking: Airbnb, AMC theatres, united, hostel world

4 METHODOLOGY

We employed Ghostery, a powerful tracker and ad blocker, to collect information about trackers on websites. Specifically, after installing the Ghostery extension, we visited each website and learned from the Ghostery pop-up the trackers attached to the currently visited websites, the category they belong to, and the total number of trackers. For example, using ghostery, we can know that Macy's homepage uses more than 50 trackers without logging in: ad trackers such as DoubleClick Floodlight, Bing Ads, consent management trackers such as Tealium Consent, and website analytics such as Facebook Connect and Google Analytics. An example Ghostery pop-up is shown below.

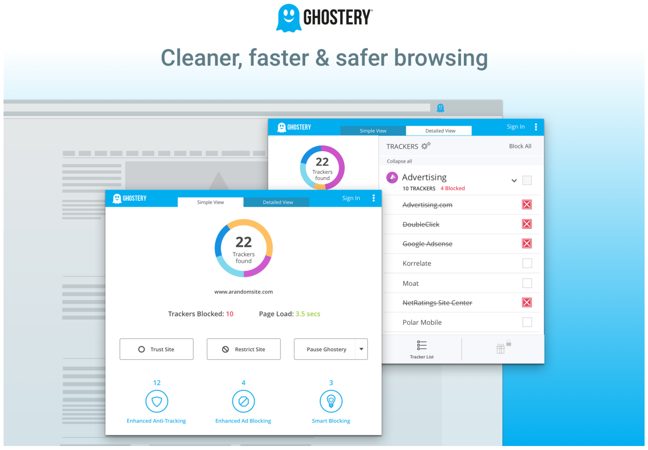
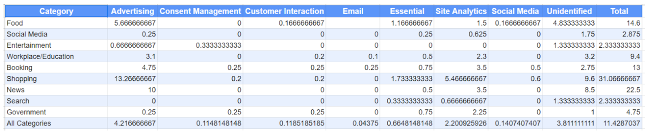


Figure 1: Ghostery screenshot (https://addons.mozilla.org/en-US/firefox/addon/ghostery/).

We then organized the collected trackers by site according to the following categories: Ads, Consent Management, Customer Interaction, Email, Essential, Site Analytics, Social Media, and Unidentified.

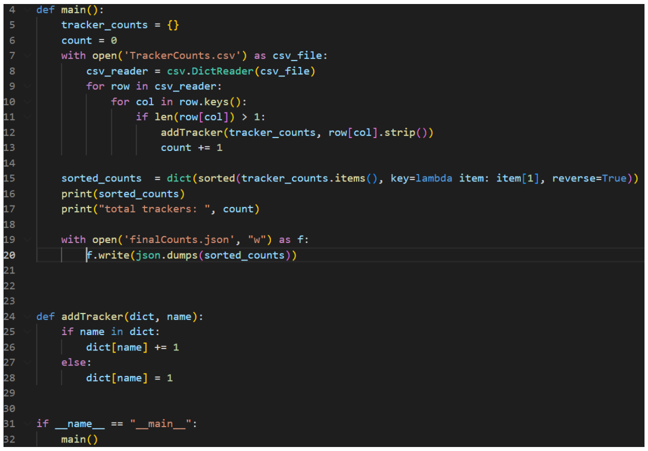
We further calculated and analyzed the numerical information on the number of trackers for each website. As shown in the figure below, we calculated the average number of different trackers for different website categories. The data were also presented in charts to make it easier to spot patterns and distribution. Understanding the number of trackers helps us know how trackers are used across sites, what types of data are collected and the potential impact on user privacy.



Category	Advertising	Consent Management	Customer Interaction	Email	Essential	Site Analytics	Social Media	Unidentified	Total
Food	5.666666667	0	0.1666666667		1.166666667	1.5	0.1666666667	4.833333333	14.6
Social Media	0.25	0	0	0	0.25	0.625	0	1.75	2.875
Entertainment	0.6666666667	0.3333333333	0	0	0	0	0	1.333333333	2.333333333
Workplace/Education	3.1	0	0.2	0.1	0.5	2.3	0	3.2	9.4
Booking	4.75	0.25	0.25	0.25	0.75	3.5	0.5	2.75	13
Shopping	13.26666667	0.2	0.2	0	1.733333333	5.466666667	0.6	9.6	31.06666667
News	10	0	0	0	0.5	3.5	0	8.5	22.5
Search	0	0	0	0	0.3333333333	0.6666666667	0	1.333333333	2.333333333
Government	0.25	0.25	0.25	0	0.75	2.25	0	1	4.75
All Categories	4.216666667	0.1148148148	0.1185185185	0.04375	0.6648148148	2.200925926	0.1407407407	3.811111111	11.4287037

Figure 2: Average tracker counts across categories

In addition, we conducted an in-depth analysis of specific trackers. We used an extension, Combine Sheets, to pull data from several spreadsheets into one and wrote python codes shown below to sort the trackers in terms of frequency of occurrence. We thus identified the top ten most used trackers from the trackers we collated. For example, we identified Google Analytics, Google, and Google Tag Manager as the top three trackers in terms of usage. Then, we did further research and analysis on these trackers. In particular, we looked for the extent of tracker usage: in terms of the percentage of web traffic tracked and the number of websites that loaded the tracker, as well as the functionality and purpose of the tracker.



```
def main():
    tracker_counts = {}
    count = 0
    with open('TrackerCounts.csv') as csv_file:
        csv_reader = csv.DictReader(csv_file)
        for row in csv_reader:
            for col in row.keys():
                if len(row[col]) > 1:
                    addTracker(tracker_counts, row[col].strip())
                    count += 1

    sorted_counts  = dict(sorted(tracker_counts.items(), key=lambda item: item[1], reverse=True))
    print(sorted_counts)
    print("total trackers: ", count)

    with open('finalCounts.json', "w") as f:
        f.write(json.dumps(sorted_counts))



def addTracker(dict, name):
    if name in dict:
        dict[name] += 1
    else:
        dict[name] = 1


if __name__ == "__main__":
    main()
```

Figure 3: Python code to sort the trackers in terms of usage

5 RESULTS

To recap we collected information from our dataset of websites by recording all of the trackers detected by Ghostery, and these are the results aggregated The above figure showcases the differences in average tracker count across website categories. Social Media, Entertainment, Search and Government all had less than 5 trackers

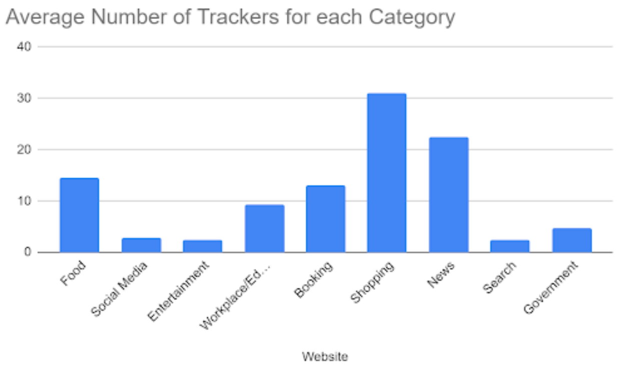


Figure 4: Average tracker count across categories

on average across websites and had the lowest tracker counts. On the other end, Shopping and News had the highest tracker counts with over 20 trackers on average across these websites.

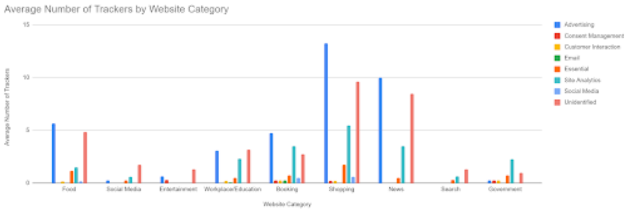


Figure 5: Average tracker count across categories, broken down by tracker type

This figure is similar to the previous figure but also shows the breakdown of tracker types within each category. When looking at the discrepancies between categories there are a few explanations to explain these differences. We used the Ghostery web browser extension to collect all of our data for this project. Ghostery looks at all of the API calls/HTTP requests being made on the web page which means that it can really only detect 3rd party or external forms of tracking. The tracking that is done by the website is often more difficult to track as this could be recorded through information that you willingly provide to the website, such as the posts you make on social media.

This difference can be shown when comparing Social Media to Shopping. Social Media has comparatively few trackers as the information can already be gathered from the posts you interact with or what you choose to post on the website. These types of websites usually also require a user to be logged in to use them, and we collected our data after signing into an account. Because these websites can already associate your interactions with your account directly, it doesn't need to add cookies to the browser or use any 3rd party trackers. In contrast, shopping websites sometimes require accounts but not always, so it makes sense that they use comparatively more trackers and cookies to have better targeted content the next time a user visits the website again.

Among the high tracker categories(Shopping, News, and Food), the largest percentage of trackers consisted of advertising trackers. This makes sense because these are all generally for-profit businesses that can benefit from gathering this information from site visitors. In the case of Food and Shopping, this can be used to generate targeted ads for specific products when users visit other websites or search engines like Google. For News, this information can be sold to advertisers to show ads specific to a user. The most common way that this is done is through cookies stored in the browser that are shared between websites all using the same trackers.

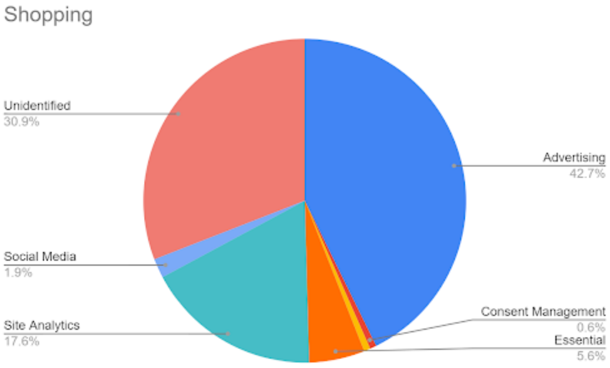


Figure 6: Tracker Breakdown for Shopping Websites

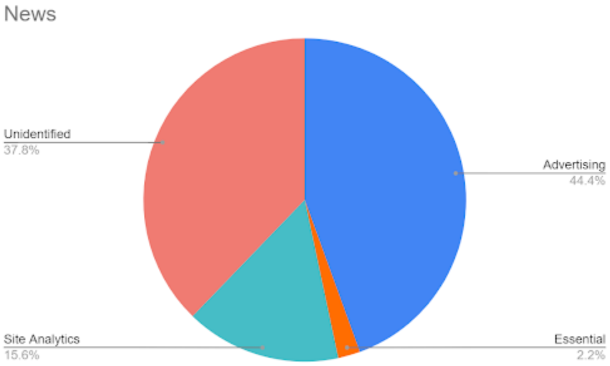
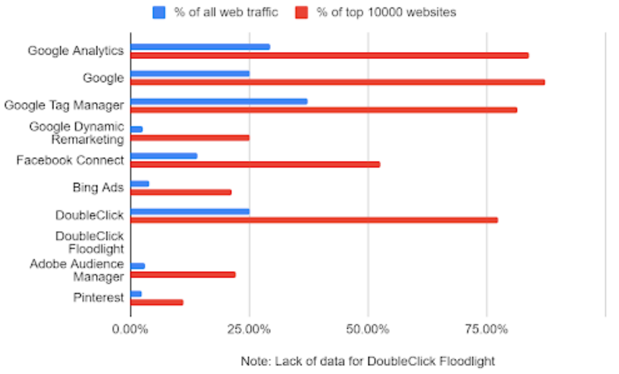


Figure 7: Tracker Breakdown for News Websites

5.1 Analyzing Trackers

Figure 6, shows the occurrences of the top ten trackers within our dataset. We then used the companion to Ghostery, WhoTracksMe which is a website that provides information about common trackers present on websites. According to their statistics, all of the associate Google trackers, including DoubleClick and excluding Google Dynamic Remarketing, were being used by more than 75% of the top 10,000 websites, and by more than 25% of all websites. This shows that a lot of the websites currently operating on the



Note: Lack of data for DoubleClick Floodlight

Figure 8: Top 10 Trackers by Occurence Across Dataset

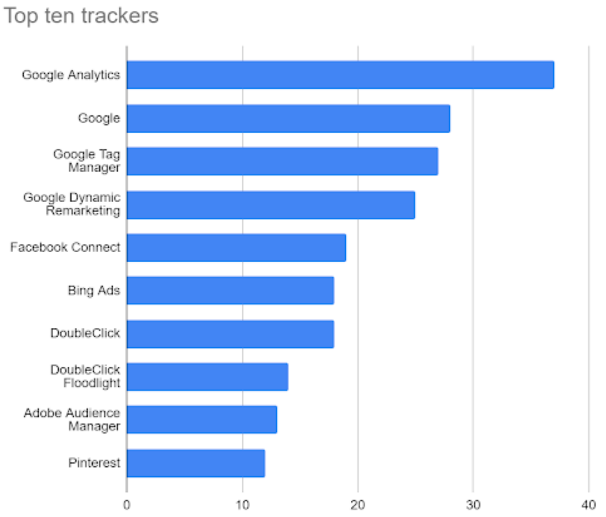
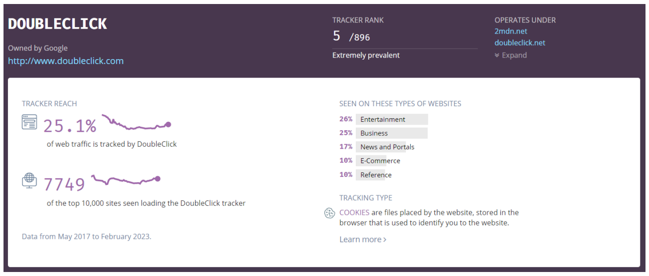


Figure 9: Usage of trackers across the entire web (WhoTracksMe)

web depend on these same trackers, which are all also managed by Google. These trackers from Google are also used for a variety of purposes indicated by the tracker type, Google Analytics is for Site Analytics, Google Tag Manager is Essential, and DoubleClick/Google are for Advertising. Although it isn't disclosed, it is expected that Google has some degree of access to the information that these trackers are collecting. These results also indicate that many of these tech companies might have more access to your data than you previously thought as the tools/trackers that they provide are being used across the majority of websites. Even if you never visited Google, Facebook, or Bing, they will still have methods of getting your data.

5.2 Analyzing Privacy Policy

Finally we did comparisons between the privacy policies and the detected trackers on a few websites and this is the summary of our results. Chipotle discloses that it collects using the following: 3rd Party tracking, analytics, and IP/related browser information when a user visits the site. There is also information that is stored directly through the website such as First Name,Last Name, phone number, payment info, that are stored whenever users make a transaction. Overall, Chipotle does disclose that they have 3rd party trackers on their site but it is kept intentionally vague what the information is being used for and who the information is being sent to. Macy's actually has a fairly detailed privacy policy and they disclose all of the types of Personal Information that they collect. They also state they use cookies, pixels, and SDK(Software Development Kits) to collect information along with other tracking information. This disclosure seems fairly thorough and Macy's does accurately disclose what types of trackers they are utilizing on their page. In general, this is a small sample of websites, but most companies do accurately disclose the presence of trackers and also some information on what the data is being used for. The only issue is that these websites don't accurately list out what the 3rd parties are, they just vaguely state: "for advertising purposes" or "our partners and affiliates".



DOUBLECLICK

Owned by Google
http://www.doubleclick.com

TRACKER RANK
5 /896
Extremely prevalent

OPERATES UNDER
2mdn.net
doubleclick.net
Expand

TRACKER REACH
25.1%
of web traffic is tracked by DoubleClick

7749
of the top 10,000 sites seen loading the DoubleClick tracker

Data from May 2017 to February 2023.

SEEN ON THESE TYPES OF WEBSITES

26%	Entertainment
25%	Business
17%	News and Portals
10%	E-Commerce
10%	Reference

TRACKING TYPE
COOKIES are files placed by the website, stored in the browser that is used to identify you to the website.
Learn more ›

Figure 10: WhoTracksMe statistics https://www.whotracks.me/trackers/doubleclick.html

6 LIMITATIONS AND CHALLENGES

The biggest limitation we faced was concluding what information about a user is being stored. This was the original goal of the project, and we had intended to create our own tool to do this. We were able to create a tool that could access the cookies on a computer's browser using Selenium. When we were getting the cookies stored from websites, we were able to get the full token value as well as the stored name and the source/origin of the token. These are somewhat helpful for detecting the presence of tracking cookies from specific websites, it was difficult to obtain any meaningful information of what these cookies were used for by the website.

This makes sense from a networking standpoint as tokens are primarily used as an identifier stored within a browser, and then sent the next time a user accesses the website again. Cookies can be used for a range of purposes from storing preferences and customization options, to customizing ads and products shown to a user. However, a lot of the logic is handled by the website on the server-side, so exactly how these cookies are used will be difficult to determine from just this cookie data. An approach that we tried with this is cross-referencing with a cookie database to see if there is an existing data on what the purpose of a specific cookie is, but there were also limitations to this approach as there was not a good way to automate it, there was a limited dataset of cookies present, and we weren't able to to verify the validity of the information present.

Another component of our project was to analyze the privacy policies on websites. We encountered a challenge of being able to effectively parse out the privacy policy information and interpret these results. As most websites are fairly different, the first challenge would be to use Selenium to navigate to the privacy policy section on a website. Then being able to interpret the information present on each website with an approach that generally works for most websites. It wouldn't be practical to perform natural language processing or some other AI analysis of the text in the file, so an approach that is more feasible would be extracting key points by using HTML tag.

Because of these challenges and limitations, we decided to shift our focus away from creating an automated tool that provides a user with information about their data to a manual analysis of a large number of popular websites to find trends in web privacy.

Using Ghostery, we were able to find the full list of trackers on a website much easier than by creating our own tool. While Ghostery does provide a full list of trackers and data about most of the trackers, it does not tell the user what information each tracker is storing, which is a similar issue to what we faced before. However, with the new direction of our project, this was less important because we were still able to get information about what the tracker was for and how it works.

7 DISCUSSION

This project only focused on the third party trackers that are found on a website, which does not include any internal collection and storage of a user's data. It also does not include what aspects of a user's information are being tracked, or even what these third party trackers are doing with the data.

The average number of trackers from our list of 55 websites was 11.4. This number is pulled down by categories such as Entertainment and Social Media; however, we predict that the number of trackers on these websites was lower due to the vast internal databases that they have to store information. Additionally, we were checking the trackers from the home page, and there may be more throughout the website. The average is brought up significantly by the Shopping category. This makes sense because targeted advertisements are a valuable marketing strategy. Depending on how a specific user interacts with the internet, they may be more susceptible to privacy concerns than others who use lower risk websites. With some websites having over 100 third party trackers, there can be no guarantees where a user's data could end up. These companies could also be selling/distributing information or there could be data breaches.

Additionally, we do not think that most people have the proper data privacy education to understand the implications of these trackers or even know that a lot of these trackers are present. While some information about distributing data is available in the privacy policies of a website, a lot of people could just skip through those. Users should be able to access this information more easily without having to download an external tool to check. This could help them make more informed decisions about how they want to interact with websites in terms of the data they share.

There is a lot of future work that could be done for this project. While not something we had the time or resources to do, it would be interesting to create a tool that acts as an extended version of Ghostery and shows the partners of each of the trackers on a website to map out how far a user's data can travel. Another potential extension of this project would be to do a more in-depth analysis of what information each of these trackers store and compare it to the privacy policy that a website provides. Something we noticed about the privacy policies that we looked at was that a lot of them claimed to distribute information to third parties for various purposes, but they did not always list those third parties. It would be interesting to see what discrepancies or information they leave out of their privacy policy with regard to what information they send.

8 CONCLUSION

This paper analyzes the external trackers found on a variety of different types of websites using Ghostery. We determined that as a whole, advertising trackers and unidentified trackers make up the majority of external trackers websites use. Shopping, news, and food websites generally use more trackers than other categories of website.

Many websites use the same trackers, and we determined that Google owns many of the most prevalent trackers across all tracker purposes. This indicates that Google, and probably other large tech companies, have access to much more information about users than they are probably aware. This could potentially be a large privacy concern for people if they knew about the extent that their information was tracked.

Web privacy is still a relatively new field, and the internet continues to grow and become more prevalent in day to day life. Many users lack the education in data privacy that they would need to make safe and secure decisions about their data on the internet.

REFERENCES

[1] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. 2006. Knowing the User's Every Move: User Activity Tracking for Website Usability Evaluation and Implicit Interaction. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland) *(WWW '06)*. Association for Computing Machinery, New York, NY, USA, 203–212. https://doi.org/10.1145/1135777.1135811

[2] Asma Hamed, Hella Kaffel Ben Ayed, Dali Kaafar, and Ahmed Kharraz. 2013. Evaluation of third party tracking on the web. 471–477. https://doi.org/10.1109/ICITST.2013.6750244

[3] Georg Merzdovnik, Markus Donko-Huber, Damjan Buhov, Nick Nikiforakis, Sebastian Neuner, Martin Schmiedecker, and Edgar Weippl. 2017. Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools. https://doi.org/10.1109/EuroSP.2017.26

[4] R. Velumadhava Rao and K. Selvamani. 2015. Data Security Challenges and Its Solutions in Cloud Computing. *Procedia Computer Science* 48 (2015), 204–209. https://doi.org/10.1016/j.procs.2015.04.171 International Conference on Computer, Communication and Convergence (ICCC 2015).