State of Online Tracking – an Analysis

Kaustubh Lohani, Nayanika Ranjan

Instructor: Edward Stohr

Introduction

Have you ever seen yourself surfing Amazon for headphones, and now every page you visit offers advertisements for headphones

Nowadays, trackers are incorporated on websites. These trackers collect data under the pretense of improving the user experience. They are, however, used to profile a user's attributes and behaviors by analyzing location history, browsing history, and other data to deduce your gender, ethnicity, interests, and habits.

This online profile is then used for targeted advertising.

In this poster, we have tried to decode these trackers' usage, the websites that use them, and the organization that operates them.

Policymakers can use this study to design laws for online tracking. Another application could be to develop sophisticated tracker blockers.

Data

The data for this study is self collected using the DuckDuckGo Privacy Essentials Chrome Extension, and DuckDuckGo Tracking Radar.

Data Collection

Web Scraping:

- A list of popular websites was compiled using Kaggle.
- Further a Python program was built using the pyautogui package, to automate the collection of html for the blocked trackers sections from websites using the DuckDuckGo extension.
- Using the beautiful soup library in python the collected html files were parsed for the tracker data and saved in a csv file.
- Finally, a script was written for preliminary cleaning of the collected data.

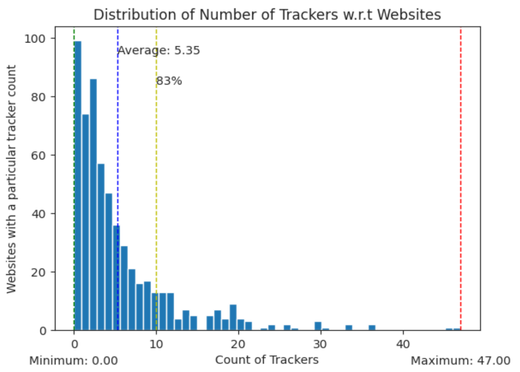
Data Preprocessing:

Using pandas library in python the scraped data was cleaned, and additional columns were added.

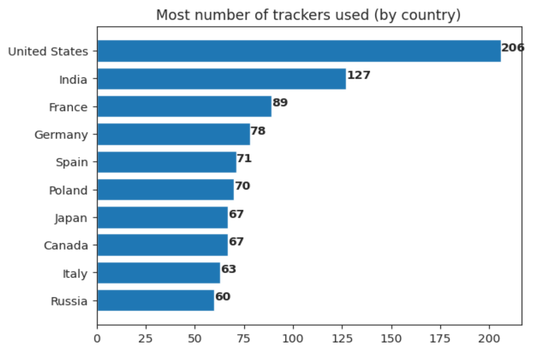
Collecting data related to organizations that owns & operates the trackers:

- First, the preprocessed data was grouped on tracker URL.
- A Python script was then used to search for the tracker URLs in the DuckDuckGo Tracker Radar data. The script extracted information about the owners of each tracker, was then combined with the existing data.

Results

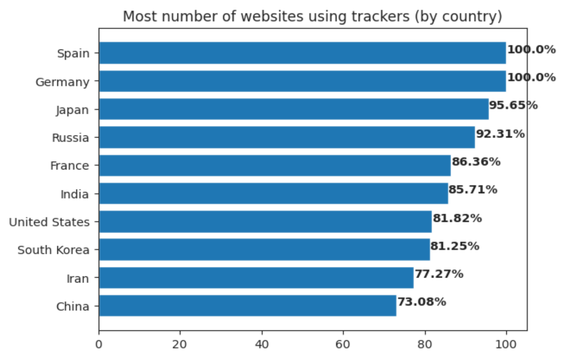


- On average, each website uses 5 trackers
- 83% of the analyzed websites use 10 or fewer trackers

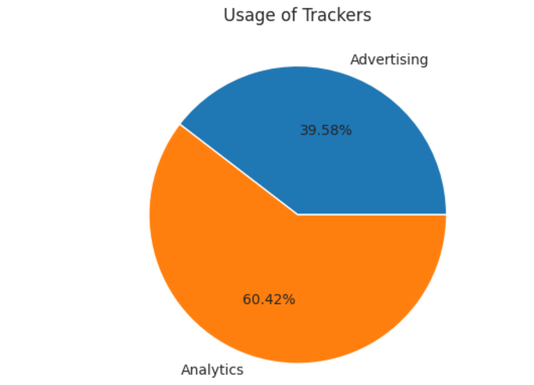


Out of the top 10 analyzed countries:

- United States-based websites use the greatest number of unique trackers.
- Russian websites use the least number of unique trackers.

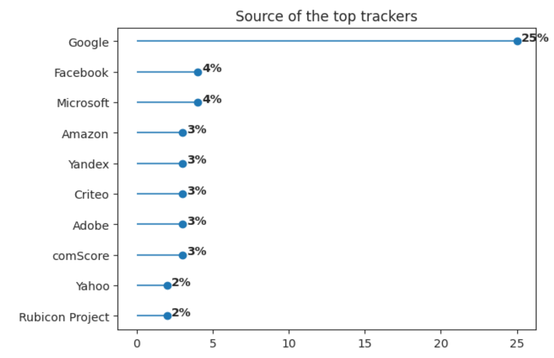


- All the analyzed websites in Spain (13) & Germany (15) used embedded trackers.
- 82% of the analyzed websites in the United States (44) used trackers.

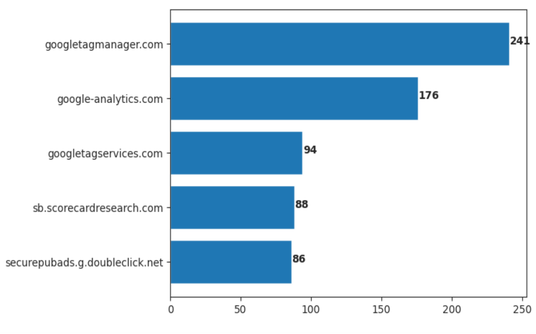


Among the trackers identified:

- 40% were used for advertising.
- 60% were used for Analytics.

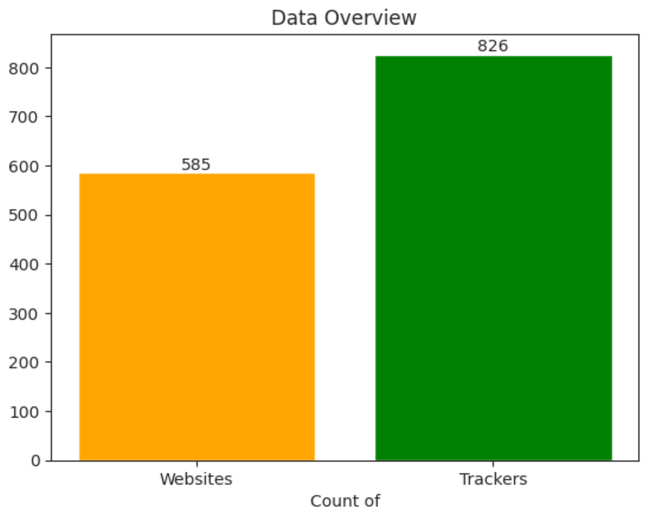


- 25% of all the trackers analyzed were owned by Google.
- 8% of all the trackers analyzed were owned by Facebook (4%) & Microsoft (4%).



- The most used tracker was the Google Tag Manager.
- Out of the top 5 most used trackers, 4 were owned by Google.

Data Overview



Country (Top 10)	Count of Websites Analyzed
United States	44
India	28
Russia	26
China	26
Japan	23
France	22
Iran	22
South Korea	16
Argentina	15
Germany	15

Conclusion and Future Work

Among the 585 websites from 105 countries, only 104 websites – 17.8% had no embedded trackers.

Legislations such as GDPR & CCPA (California Consumer Privacy Act) require websites & associated third parties to get consent before collecting and processing personal data. However, a study found that user data is collected and processed even after users opt out.

Therefore, using a tracker blocker such as DuckDuckGo or browsers such as Mozilla and Safari, which have tracker blockers baked into them, is the user's best bet to be safeguard their data against the trackers.

This study is currently done on a limited number of popular websites (585) for each country (105 – countries) and can be expanded further by increasing the number of websites analyzed.

Data collected by the trackers should also be looked at to analyze its impact on influencing users' choices.