Measuring the correlation between online behavioural tracking practices and offline privacy laws across different countries

Yashica Patodia
19CS10067

Ishan Goel
19CS10052

Aryan Mehta
19CS30006

Shrinivas Khiste
19CS30043

Satwik Chappidi
19CS30013

Mayank Kumar
19CS30029

Figure 1. Surveillience: You are being watched

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

Abstract

This study aims to measure the correlation between online behavioural tracking practices and offline privacy laws across different countries. We extracted the fingerprint and tracking features used by the top 400 websites from the Tranco list and analyzed and compared the differences in

the execution of offline privacy laws in 4 different countries namely, India, Japan, the USA and the Netherlands (Europe).

CCS Concepts: • **Behaviour Tracking Mechanism**; • **Finger-printing**; • **Cookies**;

Keywords: Behaviour Tracking Mechanism, Finger-printing, Cookie trackers

ACM Reference Format:
Yashica Patodia, Ishan Goel, Aryan Mehta, Shrinivas Khiste, Satwik Chappidi, and Mayank Kumar. 2022. Measuring the correlation between online behavioural tracking practices and offline privacy laws across different countries: . In *Proceedings of* . IIT KGP, 10 pages.

1 Motivation

Internet users are being tracked and profiled more frequently than ever, and their personal data is frequently used as payment for services. If we are to support and uphold the right to privacy, it is critical that all parties involved have a better understanding of this new reality. Let's try to understand what exactly is online tracking, what are the various ways in which one can be tracked and last but not least why this even matters to us. To answer the first question we know that Owners and operators of websites frequently gather data about how users interact with their websites online. The identity of the user's preferred language and/or display, websites visited, items bought, and transactions made may all be collected and recorded. The reasons for gathering this data vary and may include: • remembering a user's preferences (such as language, font size, and colour scheme) so that the website's look and feel are preserved for the user on subsequent visits; 1) analysing user navigation to improve website design; 2) establishing and maintaining a user's logged-on identity so that the user can move around the website without being required to log on again, or 3) tracking the user's location. We are worried about this last type of online behavioural tracking. Online behavioural advertisers frequently analyse the gathered data, create comprehensive profiles of website users, and classify them in accordance using sophisticated algorithms. The website owners/operators or a third party then target the user categories to present marketing content or advertisements that are thought to be pertinent to them.

The online behaviour of website visitors can be tracked and recorded by organisations using a variety of methods. While the tools for tracking in the examples below are some of the more typical ones, other tools may be created in the future as a result of the rapid advancements in online technology. 1) At the web server end, by using methods like inserting web beacons or bugs on web pages 2) At the web server end, by recording and retaining an authenticated user's dealings and behaviour on the website, such as information searched, transactions carried out, and his/her purchasing history. In this report we will mostly consider the following tracking tools 1) First and Third Party cookies 2) Browser Fingerprinting.

Let's try to understand why should these online tracking methods concern us. The following situations can occur: (a) Website users' information or browsing patterns are frequently collected by the website operator or owner without the users' knowledge or consent; (b) Website users' information or browsing patterns may even be collected by a third party without the users' knowledge or consent; (c) The collected information may be transferred to other parties by the website operators/owners or the third party without the users' knowledge or consent; (d) Information about a website user may be made available to third parties without the users' knowledge or consent.

Organizations must carefully evaluate whether the information collected lacks unique identifiers, and can be used to directly or indirectly identify a person. Organizations should be aware that they frequently compile a complex set of these identifiers that, when combined, may be sufficient in and of themselves to establish an individual's identity. This is where the correlation with the public laws across different countries comes into the picture.

2 Research Gap and Question

The research gap which we are trying to solve is that there has been research done regarding the behavioural tracking practices and research regarding offline privacy laws separately but we are trying to find the correlation between them across different countries which is the research gap which we are trying to solve. This is still identified as a problem but remains unsolved till this date.

Hence our research question is in quantitative terms is as follows:

What is the correlation between behaviour tracking practices specifically finger printing with the privacy laws drafted in several countries?

3 Study Design

All of our team members are undergraduate students from the department of Computer Science and Engineering, IIT Kharagpur. We are currently pursuing a course on Usable Security and Privacy under the guidance of Prof. Mainack Mondal where we have been taught about the correct procedures to conduct research in this field. Our study is a large-scale measurement. Our study is a within-subjects study because, for a single participant, we run multiple tests across different countries and check how they comply with a nation's privacy policy. The study design consists of work in the following broad categories

- Literature survey of system change in the effect of privacy laws

- Collecting data regarding different types of tracking mechanisms (subclasses of First, Third-party cookies and Fingerprinting)
- Building a robust system to gather tracking information.
- Performing EDA analysis to gather correlation metrics between the behaviour-tracking practices and online privacy-centric laws.

During our research, we will be scraping a lot of websites that might potentially have the private data of individuals. We will not collect any private data of any individual. The only information that we study is the kind of trackers being used and we do not intend on exploring the scraped data further.

We will only access public websites and public accounts so that we do not expose any private information in our research. Additionally, we maintain the anonymity of the websites by using a hash-based coding system with a private key.

- The study involves collecting users' public data in an anonymized fashion which is further stored and hashed. It imposes no risk to participants.
- We need the requested waiver to carry out research. Our approach will only use publicly available data.
- The research does not involve using identifiable private information or identifiable biospecimens, hence this question does not arise.
- Here the participants are the websites, being a non-human public entity we do not need to provide participants with additional pertinent information.

Debriefing does not apply to this study. We are using Selenium for automating our tests and extracting tracker information from websites. Browser fingerprint features statistics are obtained using FPMON browser extraction and cookie tracker information is obtained from the Ghostery browser extension.

4 Analysis Plan

We will be doing a qualitative and quantitative analysis to find the correlation between online behaviour tracking and offline privacy laws across different countries. The recruitment of potential websites will be done from the TRANCO list of the most popular websites across the globe that was created for research purposes. This website will contain any potential website that is likely to be important for our survey. We will select the top 400 websites from this list and then perform further screening to remove irrelevant or potentially vulnerable websites from our list.

The process of screening will be done as follows. We will make sure that the website that we are scraping is a public website and it does not host any private information that can raise privacy concerns. We will also not collect data from private accounts on public websites so that we do not expose any private information about any individual. Apart from this, we will only be surfing the websites following the desired protocols and observing behaviour tracking practises and not collecting any data that is being tracked.

To determine whether a website is public, we can check whether it can be accessed by the public without any requirement of authentication.

5 Procedure

5.1 Literature Survey

Literature Survey serves the two primary purposes

- Scoping down the project analysis for understanding which behavioural tracking practices to consider and which country's offline privacy laws are relevant to our project analysis.
- Understanding the importance of the correlation between online tracking practices and offline privacy laws.

Our data collection process involves collecting publicly available information from websites. We observe the offline privacy laws of the 4 selected countries. The laws in them are mentioned as follows:

- **European Union: General Data Protection Regulation (GDPR)** The General Data Protection Regulation (GDPR) is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA). It is an important component of EU privacy law and of human rights law. It also addresses the transfer of personal data outside the EU and EEA areas. The GDPR's primary aim is to enhance individuals' control and rights over their personal data and to simplify the regulatory environment for international business.
- **United States of America: California Consumer Privacy Act (CCPA)** The California Consumer Privacy Act (CCPA) is a state statute intended to enhance privacy rights and consumer protection. The act intends to provide the residents with the right to know what personal data is being collected about them, know whether their personal data is disclosed and to whom, and request a business to delete any personal information about a consumer collected from that consumer.
- **India: Personal Data Protection Bill 2019 (PDP)** The Personal Data Protection Bill 2019 (PDP) aims to provide for the protection of the privacy of individuals relating to their personal data, specify the flow and usage of personal data, create a relationship of trust between persons and entities processing the personal data, protect the fundamental rights of individuals whose personal data are processed, to create a framework for organisational and technical measures in processing of data, laying down norms for social

media intermediary, cross-border transfer, accountability of entities processing personal data, remedies for unauthorised and harmful processing.

- **Japan: Act on the Protection of Personal Information (APPI)** The purpose of this Act is to protect the rights and interests of individuals while taking into consideration the usefulness of personal information, in view of a remarkable increase in the utilization of personal information due to the development of the advanced information and communications society, by clarifying the responsibilities of the State and local governments, etc. for measures on the protection of personal information, and by prescribing the duties to be observed by entities handling personal information.

Due to privacy restrictions for the usage of VPN in Canada, we were not able to observe the following privacy law prevalent there.

- **Canada: Personal Information Protection and Electronic Documents Act (PIPEDA)** The Personal Information Protection and Electronic Documents Act (PIPEDA) is a Canadian law relating to data privacy. It governs how private sector organisations collect, use and disclose personal information in the course of commercial business. In addition, the Act contains various provisions to facilitate the use of electronic documents. PIPEDA incorporates and makes mandatory provisions of the Canadian Standards Association's Model Code for the Protection of Personal Information.

Coming to the second part which involves understanding the behaviour-tracking practices to consider we will be analysing the following :

- 1st party cookies: The website (or domain) you visit is directly responsible for storing first-party cookies. These cookies enable website owners to gather analytics information, remember language preferences, and carry out other helpful tasks that enhance user experience.
- 3rd party cookies: A third-party (non-owner) cookie is one that is installed on a website and is used to gather user information for the third party. Similar to regular cookies, third-party cookies are set up so that a website can recall information about the user later.
- Browser Fingerprinting: Websites can gather information about your operating system, active plugins, time zone, language, screen resolution, and other active settings using the powerful technique known as browser fingerprinting.

5.2 Data Collection

The data has been collected from TRANCO from the top 400 websites. For each website we have extracted fingerprinting features using FPMON and from Ghostery extracted cookies features. The type of features of fingerprinting are as follows

- Window features: devicePixelRatio, innerWidth, colorDepth, outerWidth, etc.
- Audio features: createAnalyser, createOscillator, createGain, createScriptProc, etc.
- Flagged features: doNotTrack, msDoNotTrack etc.
- Storage features: sessionStorage, localStorage, indexedDB, openDatabase, etc.
- Battery features: getBattery, charging etc.

We extract various types of cookie trackers using the Ghostery extension which are divided into the following categories:

- Advertisement trackers: Advertising cookie trackers are responsible for the targeted ads to users which they identify by collecting data that catalogues the behaviour of the user on a website.
- Site analytics trackers: These trackers collect data such as a number of clicks, time visited and other analytical features of a user visiting a website.
- Essential trackers: These include cookies which are necessary for a website to work such as website login, logout etc.
- Social media trackers: Social media cookies are cookies used to connect a website to a third-party social media platform. They remember a user's details after the user signs in to a social media account from a website.
- Miscellaneous trackers: These include other types of cookies such as session cookies, zombie cookies etc.

	url	date	JS Attributes Tracked	Fingerprinting Features	Aggressive Features	Aggresive Categories	Sensitive	loadtime
0	https://www.google.com/	1669380406460	4	2	0	0	User_agent;Cookies_enabled	2892
1	https://www.youtube.com/	1669380457050	0	0	0	0	NaN	46879
2	https://www.linkedin.com/	1669380474401	0	0	0	0	NaN	855
3	https://www.amazon.com/	1669380511384	35	17	7	7	Online;Geolocation;Storage;User_agent;Platform...	9280
4	https://www.warnerbros.com/	1669380538162	11	5	1	1	User_agent;App_version;Storage;Product;Platform	18298
5	https://www.bing.com/?toWww=1&redig=5BE1241F75...	1669380544709	4	1	0	0	Timezone	701
6	https://www.trendyol.com/	1669380553679	5	1	0	0	Storage	5626

Figure 2. Data Extracted across TRANCO list websites

5.3 Building the Application and Writing Code

We needed data regarding the various tracking practices in various websites. To collect this data we have used the FPMON and Ghostery extension.

Let us first look at how we used the FPMON extension. This extension collects data about various JS Attributes that are tracked in a website for fingerprinting and displays this info in its extension. To get this info from the extension we needed to modify the code of the extension to send the relevant information to the browser via console logs so that this information can be captured by a web crawler.

The code for FPMON was available on GitHub. We looked into the code to analyse where it outputs the details of

browser fingerprinting and modified the code to log the data in the console. This updated code was then packed in the form of a CRX file so that it can be enabled on the browser.

A similar approach was used for the Ghostery extension. Ghostery extension gave us details about the first-party and third-party trackers active on the website and displayed this information on its UI. The source code for the extension was available on GitHub. On analysing the large code base we were able to figure out from where the UI component was getting its data and logged it into the console. This updated code was again packed into a CRX as before.

To automate the entire process of visiting the websites and getting the information about the trackers and fingerprinting we used the selenium library. Using this we automatically started a chrome browser enabled our extension on that browser using the CRX files and then iterated over the top websites from the TRANCO list and visited the websites one by one using selenium, waited for a few seconds for the website to load and got the information about the websites trackers and fingerprinting using the console logs from the updated codes of the extensions. All this data was saved into a CSV.

This code was run using a VPN service namely Proto VPN so that we could mimic visiting the website from different countries which were necessary for our experiment. Data about the top websites were obtained as described for all countries and saved into separate CSVs.

6 Results

We extracted the browser fingerprint features tracked by different websites using a browser extension known as FP-MON which measures and rates fingerprinting activity on any website in real-time. We automated the process and extracted the fingerprint features for the top 400 websites in the Tranco list for 4 different countries namely, India, Japan, USA and Netherlands. Using our code simulator written in selenium, we extract the total number of Javascript attributes tracked, the number of fingerprint features, the number of aggressive features, the aggressive categories and the sensitive categories.

We extracted all the trackers used by websites by using a browser extension known as Ghostery. Ghostery is an anti-tracking tool which shows all the trackers used by the website and divides them into categories such as Advertising trackers, Site analytics trackers, essential trackers, etc. We were unable to completely automate the process of extracting the information from the trackers identified by Ghostery. We tried various methods which are shown in the next section. To get some preliminary results, we manually extracted the trackers used by the top 50 websites in India from the Tranco list and plotted various analytical graphs which are shown in Section 5.4. The results for the top 10 websites are shown in the table below.

Website	Total trackers	Advertising/ Site Analytics	Essential
google.com	2	2	0
youtube.com	3	3	0
linkedin.com	3	3	0
amazon.in	2	2	0
warnerbros.com	7	6	1
bing.com	4	3	1
outlook.live.com	3	3	0
trendyol.com	2	1	1
azure.microsoft.com	11	11	0
github.com	1	1	0
Total	45	41	4

6.1 Additional Research Work: Attempts for Ghostery

The goal is to find the number of trackers on a particular website. An existing extension categorizes all the trackers present on a website called “Ghostery.”

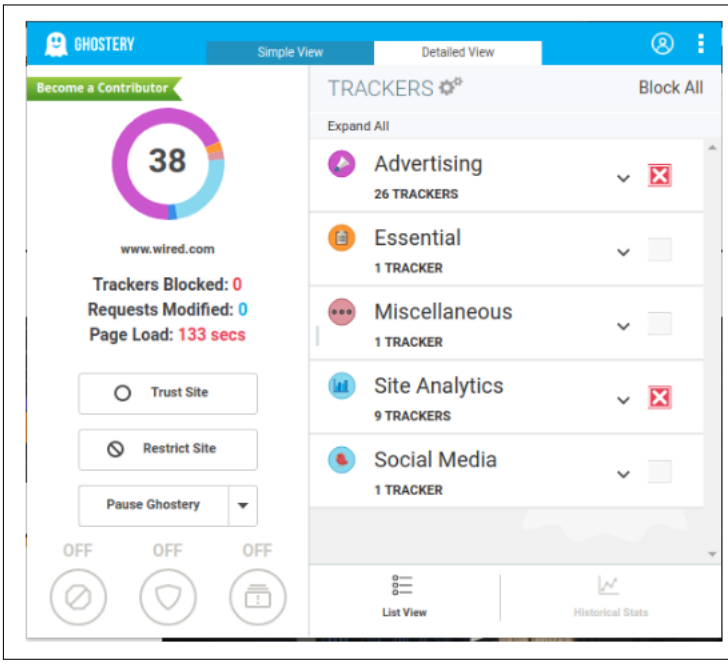

Figure 3. Ghostery

We wish to extract this information from the extension. We also want to automate this process using Selenium in Python.

Selenium has the tool to extract the console logs of a website, so we had to modify the extension such that it logs the categories to the console.

6.1.1 Approach 1. We studied the structure of the chrome extension and managed to log the required information to the

, ,

console of the UI component of the extension. An important

Figure 4. Approach 1

note is that three different consoles are at play here. These are

- **Website Console**: Normal console for any website

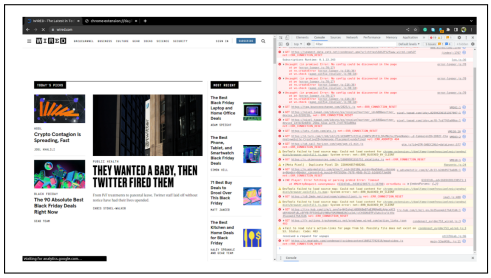

Figure 5. Website Console

- **Background Console**: Console for the background process for an extension (persistent)

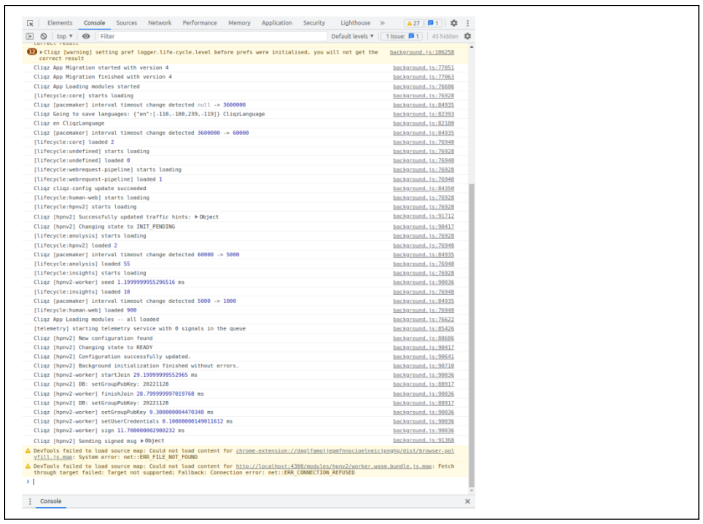

Figure 6. Background Console

- **Extension UI Console**: The UI for an extension is treated as a separate website, and this has its separate console. This is only active when UI for the extension is opened.

While Selenium has GUI control over the contents of a website, it only has API controls for the browser. There is no browser API to open the popup UI, and therefore can only be clicked manually. So we could not automate this process.

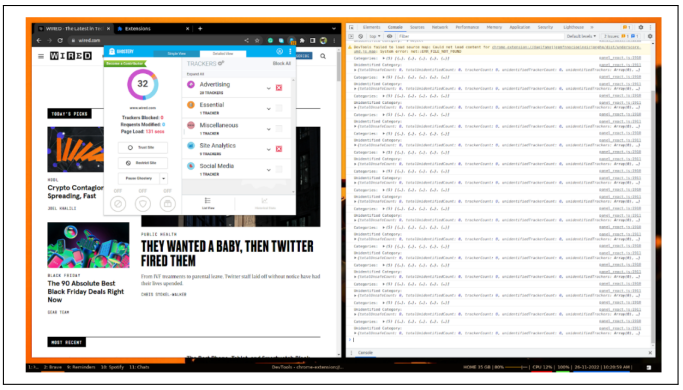

Figure 7. Extension UI Console

6.1.2 Approach 2. On further understanding of the react framework, the UI component built in React framework uses a Redux Store. This is a single immutable object that stores all the information that different UI sub-components use. We tried to create a dummy component to access this store and print our data. This was not a simple task as we had hoped and failed to do it.

6.1.3 Approach 3. An extension has many processes, of which two are 'background.js' and content-scripts. The background script is persistent and runs constantly. This file has about 16000 lines, and we could not find the code that categorizes the trackers.

6.1.4 Approach 4. On the other hand, a content script is the JavaScript code responsible for adding HTML elements to the current website. There are loaded the first time a website is loaded. These scripts are not written in trivial Javascript and must be compiled using webpack. Additionally, these content scripts have no direct way to access the Redux store, which we know has our data. So even this approach could have been more successful.

Despite trying four different approaches, we could not retrieve tracker information from Ghostery in an automated fashion (we figured out manual extraction but infeasible).

7 Exploratory Data Analysis

We have collected data regarding the fingerprinting features tracked across various websites in different countries.

Let us look at some key observations that we found during our analysis of the data.

First let's have a look at the histogram plots of the number of JS attributes that were tracked in a country in various websites. FPMON extension collects data about 140 JS attributes and checks whether these are tracked in the particular website. This gives an idea of the extent to which tracking is done in the particular website. We have added plots for each of the countries and also a combined plot to compare the tracking across countries.

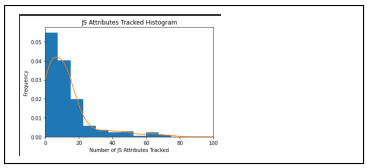

Figure 8. JS Attributes: USA

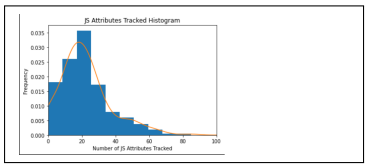

Figure 9. JS Attributes: India

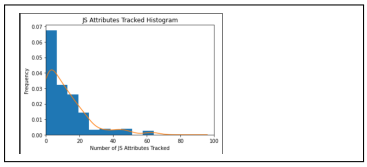

Figure 10. JS Attributes: Japan

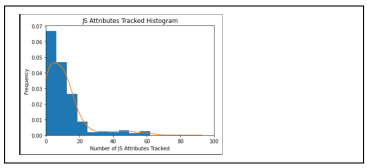

Figure 11. JS Attributes: Netherlands

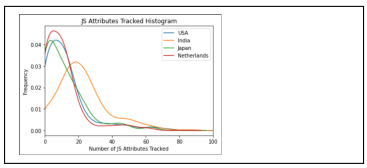

Figure 12. JS Attributes Comparison Graph

Observations 1: As the distance of the peak from the origin increases more the features are being tracked. As we can see that the maximum features are being tracked in India and the least for Japan which correlates the privacy policy being drated in both of these countries as well. India has least restrictive privacy policies hence the greatest number of JS attribute being tracked.

Next, we have histogram plots regarding the number of fingerprinting features that have been tracked. FPMON classifies the JS attributes into various fingerprinting features and gives us the number of features that were tracked. Here too we have added plots for each country and comparison plot for all the countries.

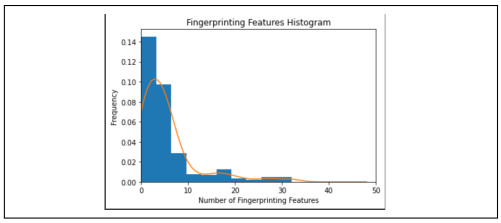

Figure 13. Finger Printing Attributes: USA

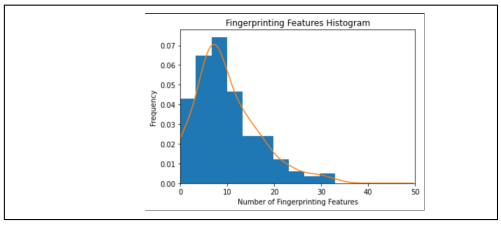

Figure 14. Finger Printing Attributes: India

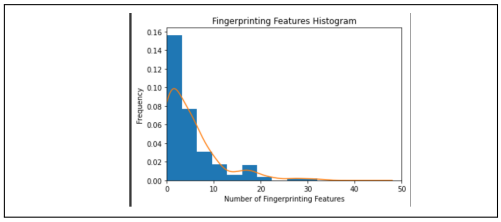

Figure 15. Finger Printing Attributes: Japan

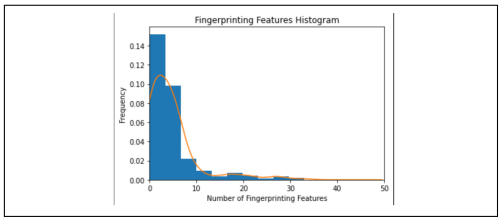

Figure 16. Finger Printing Attributes: Netherlands

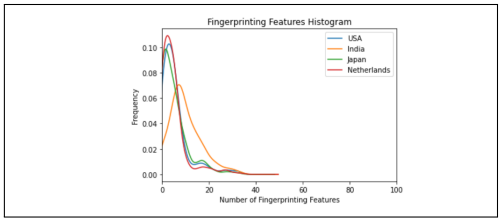

Figure 17. Finger Printing Attributes Comparison Graph

Observation 2: As the distance of the peak from the origin increases more the features are being tracked. As we can see that the maximum fingerpriting features are being tracked in India and the least for Japan which correlates the privacy

policy being drafed in both of these countries as well. India has least restrictive privacy policies hence the greatest number of Fingerprint attribute being tracked.

Another important question is that from these features that were tracked which of these contained sensitive information. The plots on Aggressive Features convey this information. We have again provided histogram plots of the number of aggressive features that were tracked to observe the severity of the tracking in various countries.

FPMON also gives us information regarding which fingerprinting feature was tracked. We have noted the most frequently observed features and compared them across the countries.

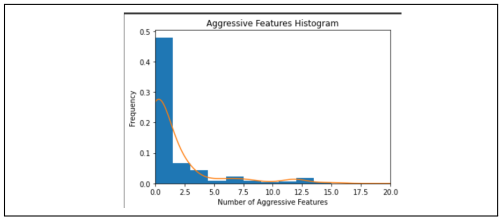

Figure 18. Aggressive Feature Attributes: USA

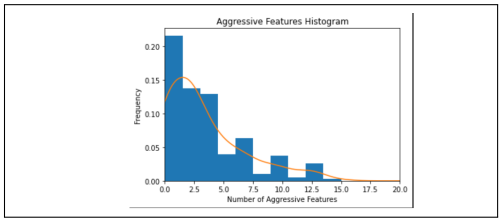

Figure 19. Aggressive Feature Attributes: India

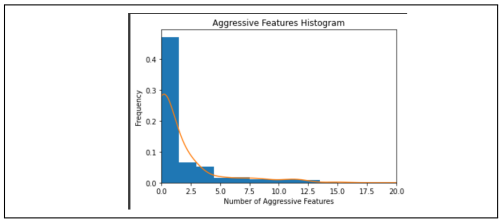

Figure 20. Aggressive Feature Attributes: Japan

Observation 3: As the distance of the peak from the origin increases more the features are being tracked. As we can see that the maximum aggressive features are being tracked in India and the least for Japan which correlates the privacy policy being drated in both of these countries as well. India has least restrictive privacy policies hence the greatest number of attributes being tracked.

To further find the correlation we have charted out the bar plots across top 4 features of 'Senstivity' namely

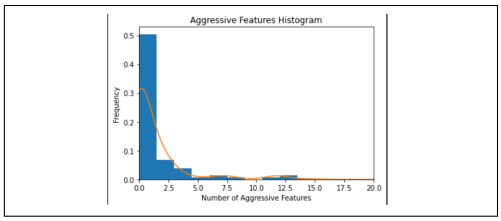

Figure 21. Aggressive Feature Attributes: Netherlands

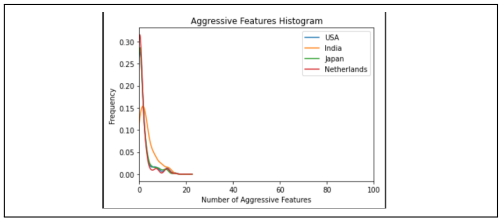

Figure 22. Aggressive Feature Attributes Comparison Graph

- User - Agent
- Storage
- Content Language
- Timezone

Sensitive categories are the fingerprint features which in the browser fingerprinting of various different websites. We have also drafted a word cloud which denotes the senstive categories for finger printing which occur most frequently.

Figure 23. Sensitivity Category Word Cloud

In all the 4 graph plots we can see that India has the maximum frequency of sensitive hence the least privacy offline laws with the least being in Japan. This is also support our previous hypothesis.

8 Division of Work

8.1 Yashica Patodia : 19CS1007

I was responsible for the following tasks. Firstly, I was involved in drafting the questionnaire for our meeting the

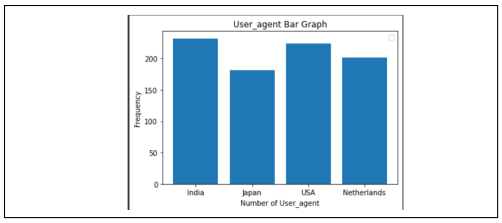

Figure 24. User Agent Bar Graph

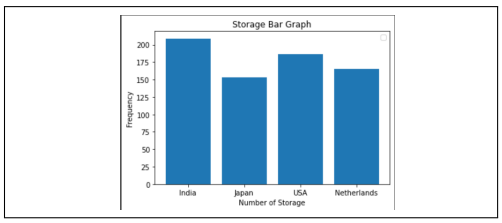

Figure 25. Storage Bar Graph

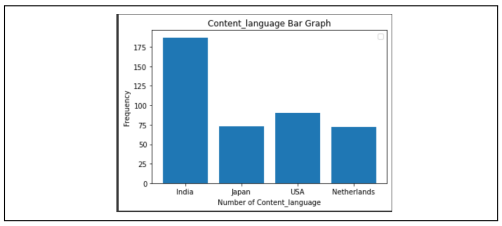

Figure 26. Content Bar Graph

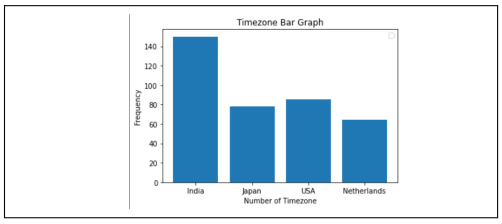

Figure 27. TimeZone Bar Graph

professor and IRB proposal in the initial task.In the second stage of the project I was involved in doing the literature survey of all the privacy laws prevalent across different countries. Thridly I was involved in the data analysis part along with Ishan specifically for the sensitive categories of finger printing.One of my major contributions in the last stage of the project was to write the entire report with the help of my team members.

8.2 Ishan Goel : 19CS30052

I helped with generating the selenium script along with my team from which we extracted the sensitive categories. I was also responsible for the plots and data analysis that is done on the results of tracking on various websites and from different countries for the sensitive categories of fingerprinting.

8.3 Aryan Mehta: 19CS30006

I read various methods used to extract tracking information from websites and we decided to use Ghostery for identifying the cookie trackers used by websites. I wrote the code of automating the fingerprint features extraction process from FPMON with my team members and then ran multiple experiments for different countries namely, Japan and Netherlands (Europe) using VPN to save the fingerprint features found by FPMON in a csv file for the top 400 websites from the TRANCO list. I also contributed to major parts of the report along with my team members.

8.4 Satwik Chappidi: 19CS30013

I was responsible for the automation extraction process of cookie trackers using Ghostery. I tried several different methods but was unable to completely automate the process. I have listed the different methods in the report. I also collected some data manually.

8.5 Shrinivas Khiste: 19CS30043

I was responsible for doing the exploratory data analysis part from the data obtained about the trackers. I plotted various plots explaining the correlation of online tracking with offline privacy laws. I also helped in making the changes in the crx file of the FPMON extension to get the fingerprint features from the extension directly.

8.6 Mayank Kumar: 19CS30029

I was responsible for preparing the IRB report along with my teammates. At start, I analyzed how different ad-blockers and cookie trackers work, and visited their open GitHub repositories to understand how easy or difficult is it to extract their details. We decided to use FPMON and Ghostery as they suited our requirements the best. I observed in detail the offline privacy laws of India, Canada and USA. I was also responsible for optimizing the selenium script to extract tracking information, and ran the code for top 400 websites from the TRANCO list for India and USA.

9 Future Work Plan

The potential benefits of the research is as follows:

- We will learn offline privacy laws of different countries and how strictly they are followed by the websites.
- There has been research regarding behavioural tracking practices and research regarding offline privacy laws separately but no research exists for measuring the correlation between them across different countries.

- The study would help us to find which is the best offline law practice across different countries in the world.

For future work, we want to develop robust tools to extract cookies features through an automated script. We can extend our work to more countries with privacy laws to identify the ideal offline privacy law which can be used to make tracking practices around the world much better. Various tests like the chi-square test can be conducted to find the proper correlation between online tracking practices and offline privacy laws.

10 GitHub Repository

The code can be found in the following repository
Link

11 Presentation

The presentation video and the slides can be found here
Link

Received 26 Nov 2022; revised 26 Nov 2022; accepted 26 Nov 2022