Personal Data Footprint (2019)

Gusani, Renato
School of Computing
National College of Ireland
Dublin, Ireland
x19411076@student.ncirl.ie

***Abstract*—In this paper I will be documenting sources of my own personal information by requesting my data from Google alone, since Google's services is where most, if not my biggest personal data is held in, this will be the only source of personal data I will be requesting from, it also has the lowest amount of wait time compared to other platforms of 30 days before having access to your data, I will be providing the data's characteristics and structure, and how the proposed data would be used to formulate a small-scale research project. I will note how the data would be collected and the high-level approach to analyzing it, I will then use the KDD methodology to analyze my data through the different steps then finally I will go over my expected findings and implications as well as ethical considerations, conclusion and future work. (*Abstract*)**

Keywords—personal, ethical, analyze, legal, data (*key words****)***

I. INTRODUCTION

One of the biggest companies that collect our data is none other than Google, it collects data on pretty much everything you do from the videos you watch, the ads you click, your location, even your Device Configuration Service, IP address and cookie data. Google collects more data than our government does, Google promised that they do not sell data that has been collected from its users but rather use information collected to "make ads relevant" while we browse the web, which is basically redistributing our data to third parties. Google has one of the biggest, if not the biggest source to be able to request your data, Google owns a majority of the trendiest platforms such as YouTube which will be one of the first data I will be requesting and implementing into this report as it is by far one of, if not the main platform I use that collects my personal data.

1) Google Data

By requesting my data from Google, First, I would like to find out how long back the data originates from, if it is from when I registered my account or much later, and if any data is stored which I have not yet had any knowledge of (offline data).

After I have thoroughly analyzed all my data, I will discuss how this data would have been collected, and then in the end discuss the legal and ethical concern (if any) this data would concern and the implications this data would cause. Before all this I will first provide a dataset that will describe the structure and its characteristics, after I have reviewed all my data, I will talk on how the data would ensue a small-scale research project. I will also be reading upon other papers in the same area such as [1] [2] and [3]

2) Methodology

The methodology I will be using is the KDD methodology [4] the reason I have chosen this one instead of the CRSIP-DM methodology is because I know it better than the latter and also because I have been thoroughly exposed with this methodology during lectures and labs at National College of Ireland as well as reading research papers that have used the KDD methodology..

3) Expected Findings and Considerations

After I have thoroughly presented and analyzed my personal data, I will be presenting my thoughts on my expected findings and considerations.

After that, I will note aspects of my report that would ensue ethical and legal implications.

TABLE I. YOUTUBE DATASET

Data Folder	***No. of Files***	***File Type***	***Total Size***	***No. of Instances***
chats	7	Chrome HTML Document	327 KB	3,277
history	2	Chrome HTML Document	22.3 MB	10,000+
my-comments	1	Chrome HTML Document	150 KB	1,092
my-live-chat-messages	1	Chrome HTML Document	2.90 KB	20
playlists	9	JSON Files	2.96 MB	7
subscriptions	1	JSON Files	1.27 MB	29,281
videos	23	MP4 & JSON Files	1.68 GB	23

II. DATA

1) YouTube Data structure and characteristics:

The first six files from the "chats" data folder are labeled as unstructured links, once clicked it opens a very detailed and structured HTML document in the Chrome browser, from this data I am able to see every single comment in my chats with timestamps, the first data folder is very different to the "my-comments" data folder since these "chats" are private chats that you have with other youtubers, for example creating a group chat on YouTube to share videos and chat with friends, this data structure was identical in structure to "history", "my-comments" and "my-live-chat".

The history file by far had the most amount of data available than from any other data in this paper, it has two separate files which an enormous amount of data referring to every single video I have clicked from the creation ofmy own channel.

The subscriptions and playlists data files which are JSON files, I was only able to view in Notepad++ since the majority of this data is semi-structured, it contained repeated attributes such as "channel id", "subscription" followed by URL's to channel avatars of the subscribed channels.

For the last file "videos" this was the only file that had MP4 files that I was able to open and play in any video player, the videos were of what I had uploaded to YouTube, each of MP4 files also had an adjoining JSON files which, again, were semi-structured with only parts

being understandable such as the attributes “caption”, “definition”, “custom thumbnail” and “licensed content”.

a) How the Data would be collected

YouTube already have software in place for when a user registers an account with their platform, they automatically collect data from the user, mainly to provide relevant ads to the user. The data presented above in the dataset labeled as “videos” would be the easiest to collect data from as these are the current videos the user has uploaded to the YouTube platform, which in this case is me, the data labeled as subscriptions under the data folder attribute would be collected by when the user clicks the button “subscribe” under a channel, this make sure the user receives this specific channel’s videos as they’re uploaded to their subscriptions page.

TABLE II. *GOOGLE PLAY STORE DATASET*

Data Files	***No. of Instances***	***File Type***	***Total Size***
Devices.json	22	JSON File	769 bytes
Installs.json	2,314	JSON File	54.3 KB
Library.json	1,827	JSON File	35.3 KB
Order History.json	348	JSON File	8.87 KB
Purchase History.json	78	JSON File	1.96 KB
Reviews.json	47	JSON File	1.39 KB

2) Google Play Store Data structure and characeristics

Once the first file Devices.json is opened it gives the most recent data on the user’s device, Data such as “carrierName”, “playstoreClientVersion”, “manufacturer”, “modelName”, “deviceName” would be in this file, these are just a few examples of what is contained in this file, it has the most recent data of my device.

The second file named Installs.json, which has the greatest number of instances shows every application I have installed through google play store, device data is also available in this file as well as timestamps and the title of the app installed.

Library.json contains my current library of installed applications on my device and time stamps of when the applications acquired.

The order history file contains information on transactions that needed some type of payment, whether it was in-purchases or buying paid applications on the play store, the data was all combined into this one file and contained data such as the name, address code, city and the type of card used to make the transactions, but did not have information of what was purchased.

The second last file Purchase history seems to be an extension of the previous file since it has some repeated data elements except now more data is available on the name of the item(s) purchased,

Reviews.json which was the last file available from the play store data contains information on every review I have left on an application in the app store including timestamps and ratings.

a) How Google Play Store data would collected

Google Play Store collects data from when the users registers an account and agrees to the terms and conditions, google uses many different techniques, one of which is data mining, to collect data from users as well as developers from the apps in play store also have the capability to collect specific data on its users.

TABLE III. LOCATION HISTORY DATASET

Data File	***No. of Instances***	***File Type***	***Total Size***	***No. of Attributes***
Location History.json	32,217	JSON File	757 KB	10

3) Location Data

a) Structure & Characteristics: One the Location History,json file is opened in Notepad++, a total of 32,217 of instances is available, from all his data there are 4 repeated atrributes which are “timestamp”, “latitude, “longtitude” and “accuracy”.

The timestamp attribute shows the exact time the data was pulled at and the latitude and longitude attributes pinpoint the exact location in the world where the device was at, accuracy seems to tell how accurate this data is, as on many cases the accuracy went from 100% down to 0%, there is also an attribute which only repeats a few times which is “type”, “activity” and “confidence” which always seems to have empty values.

b) How location data would be collected: Location data is gathered through your device having location services turned on, although it has been said on numerous occaisions that companies have the capability to view your location without you having location services turned on, unless the battery is physically taken out of the device there is no way of knowing if you are being tracked.

TABLE IV. ANDROID DEVICE CONFIGURATION SERVICE

Data Folder	***No. of Instances***	***No. of Attributes***	***File Type***	***Total Size***
Device-	324	324	Chrome HTML Document	38 KB

4) Android Device Configuration Service

a) Data Structure: Once the HTML file is opened in Chrome a very detailed and long data file is available to view, it starts off with device and account identifiers such as Android ID, MEID(s), IMEI(s), ESN(s), Serial Number(s) and MAC addresses.

Afterwards a long list of Device Attributes is shown, some examples are Locale, Time zone, Hardware, Model, Brand and so on, next every feature on my device is shown along with it’s version in nominal data, after that seven attributes are shown which are, Connection Time, Build Fingerprint, Product, Brand, Radio Firmware Version, Bootloader Firmware version and finally Google play Services Version, all this data is shown in both categorial and nominal data in a very structured way, at the end Errors are shown under two Attributes listed under Connection Time and HTTP Response Code.

b) How the data would be collected: Android Device Configuration Data would be collected the same way Google collects every other one of it's data, users are able to constrict some of the data sent to Google but not much, Android OS is owned by Google, the information from your device would automatically be sent to Google as we have agreed to the terms and conditions when we first got our devices.

TABLE V. *CHROME DATASET*

Data Files	***No of Instances***	***File Type***	***Size***
Autofill.json	3	JSON File	1 KB
Bookmarks	68	Chrome HTML Document	11 KB
BrowserHistory.json	3	JSON File	1 KB
Dictionary	0	Microsoft Excel	0 KB
Extentions.json	4	JSON File	1 KB
SearchEngines,json	3	JSON File	1 KB
SyncSettings.json	60	JSON File	2 KB

5) Chrome Data Structure

The first file which is Autofill.json is an empty file once opened, after doing research I have found out that it is only empty as I had just previously cleared my history in the chrome browser which is why the autofill file is empty, in the case I had not cleared my history, in this file there would be words that I have searched into Chrome which auto filled on their own.

The second file which was bookmarks, was the only HTML file in the Chrome Data Folder, once I opened the HTML file in chrome I was able to view a total of four different kind of bookmark folders, "Bookmarks bar", "Other bookmarks", "Synced Bookmarks", and the last one being "Bookmarks v.2".

The third file BrowserHistory.json in my case was also empty as the first file as I had previously cleared my history before requesting this data, in the case that I hadn't, here would lay my entire chrome history data, from what I searched and what webpages were opened.

The fourth file Dictionary which was the only Execl type document in the Data folder was also empty and after much research I am still not sure what type of data would be stored in here.

The fifth file Extensions.json had every chrome extension I currently had installed onto my chrome browsr at the time of requesting my data.

SearchEngines.json in this case was an empty JSON file with nothing inside.

The last file however, SyncSettings.json had 3 attributes to begin with which were, "App", "App Settings", and "Preferences", after that a looped sequences of 2 attributes which were "name" and "value" repeated, in the end two attributes displayed which were "themes" and "managed Users". All this data was categorical.

a) How the Data would be collected

The way Google Chrome would collected personal user data would be through the information stored on our local machines such as browsing history information like the URL's of website visited, the cache of texts as well as other resources from those webpages, personal information and passwords, list of permissions that have been granted to the website(s), cookies from the website(s), data saved by add-ons and lastly a record of what was downloaded, this data would be specifically collected by chrome's basic browser mode.

III. SMALL-SCALE RESEARCH PROJECT.

1) Putting all the Google Data together

After gathering all my Google Data, I would like to figure out how safe my data is being kept is safe or whether its insecure. As well as whether my data is being shared with any third-parties such as medical institutions or governments.

2) Putting it to use

After I have collected and figured out the details of how my data is stored I would then like to figure out how much time of everyday I spent on the internet such as view time on YouTube, by acquiring and reviewing this data I would know if it is causing health concerns and whether I need to limit my usage.

IV. KDD METHODOLOGY.

1) Selection / Data

The KDD process I will be following will be in regard to [4]

This is the initial step, it will develop the scene for understanding what should be done with various choice such as transformation, algorithms, representation and so on. The data I have selected for this process is my entire Personal Google Data as used in this project which is my YouTube Data, Google Play Store Data, Location History Data, Chrome Data, and lastly Android Device Configuration Data. I will be implementing all this Data into one Dataset; this step is important because Data Mining learns and discovers from the accessible data.

2) Preprocessing / Target Data

In this step, data reliability is improved, it supports data clearing to handle the missing quantities and removal of noise. This will handle any missing values in my data, the statistical technique I will be using is called Clustering-based anomaly detection, using this technique I can analyze any cluster which has noise, data instances falling outside of the cluster will be marked as anomalies.

3) Transforming / Preprocessed Data

At this stage, the creation of usable data for Data Mining is made available and used. Techniques here involve dimension reduction for example feature selection and record sampling but what I will be using is discretization of numerical attributes and functional transformation, this step is a must for the success of the entire KDD process, it will distinguish the data with its counterparts.

4) Data Mining / Transformed Data

I am now able to decide on which kind of Data Mining to use for example classification, clustering or regression. This depends on the KDD goals and also on the previous steps. The technique I will use is clustering. Having the technique, I can now decide on the strategies. This stage involves choosing an individual technique to be used for searching patterns which in my case will be understandability which will be used for decision trees .Now that the Data Mining algorithm is reached,

at this stage I may need to utilize the algorithm several ways until a suitable outcome is made, an example is by turning the algorithms control parameters such as the minimum instances in a single leaf of a decision tree.

5) Evaluation / Patterns

At this step I assess and interpret the patterns and its reliability to my goal characterized in the first step, here I consider the preprocessing steps as their impact on the Data Mining algorithm results. This step focuses on the utility and understanding of the induced model. In this step, the identified knowledge is also acknowledged for further usage. The last step is the use and overall feedback and discovery by the results given by Data Mining.

6) Interpretation

Now, I am prepared to include the knowledge into another system for further research. This knowledge becomes effective in the sense that I can make changes to the system and measure the impacts (if any). The succession of this step decides whether the KDD process is effective. There is a lot of challenges in this step such as losing "lab conditions" under which I have worked under, for example the knowledge was discovered from a certain static depiction, it's usually only a set of data, but now the data has become dynamic. The Data structure may change certain quantities that become not available, and the data domain might be altered, such as an attribute that may have a value that was not expected before.

V. Expected Findings and Implications

1) Findings

This Data provides a tremendous amount of data on the Google Network which could be used more than just Google itself, the data gathered from example the last Dataset, Chrome, might cause harm to the data owner if there was private information which the data owner did not want to be exposed, or even the device configuration data, which would give a lot of private information about the users device.

2) Implications

If the data was to be exposed the user of the data might follow up with a lawsuit or if any case it was exposed by a private anonymous group the user might not be able to take any action which would cause a lot of harm, consent is one of the first implications, if this was not my personal data.

VI. Ethical Considerations

1) Research Studies

Looking back upon the Google data one of the main ethical considerations that would come into play would be Informed consent and Do no harm, what is meant by this is that the Data discussed about would have to be given consent by the owner before any of the data is talked about as well as that no harm is done to the person whose data is being discussed.

Also, Anonymity and Confidentiality are the two most important ethical considerations when referring to someone's personal information such as in this paper.

2) Additional Information

If the ethical considerations are not looked over in depth a lot of harm and repercussions might harm the issuer of the paper.

VII. Conclusion and Future work

1) Conclusion

After thoroughly analysing my personal data from google I am able to distinguish between how far back my data originates to and if any offline data is being kept, I now also know that a very small limited amount of data is available to the public and this is known as google does not let us download every single data that we personally produce by using their platforms and services .

2) Future Work

If I was to return to this topic one of the first things I would like to follow up with is how the data from different Google sources complements each other and what the data tells me about this specific user (In this case is me) how much time is spent using Google's YouTube and if interaction (comments) plays a big part.

References

[1] O. Gencoglu, H. Similä, H. Honko, and M. Isomursu, 'Collecting a citizen's digital footprint for health data mining', in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 7626–7629, doi: 10.1109/EMBC.2015.7320158.

[2] M. Harjumaa, S. Saraniemi, S. Pekkarinen, M. Lappi, H. Similä, and M. Isomursu, 'Feasibility of digital footprint data for health analytics and services: an explorative pilot study', BMC Medical Informatics and Decision Making, vol. 16, no. 1, p. 139, Nov. 2016, doi: 10.1186/s12911-016-0378-0.

[3] K. Nakajima, Y. Mizukami, K. Tanaka, and T. Tamura, 'Footprint-based personal recognition', IEEE Transactions on Biomedical Engineering, vol. 47, no. 11, pp. 1534–1537, Nov. 2000, doi: 10.1109/10.880106.

[4] Fayyad et al. 1996, 'KDD Process/Overview', Overview of the KDD Process.[Online].Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html. [Accessed: 21-Dec-2019].