Hakuna Metadata (1)

Exploring the browsing history

Since I joined European Digital Rights (EDRi) in September 2016, one of most hottest topics that is being discussed in the Brussels bubble is the review of ePrivacy rules (ePR). As a complementary instrument to the General Data Protection Regulation (GDPR), ePR mainly deals with data protection and privacy in the electronic communications sector, such as the tracking of users when they browse the internet. Since the GDPR has been already finalized, advocacy around the ePR is probably the last chance to defend European citizens digital rights. One of my key responsibilities as the Ford-Mozilla Open Web Fellow is to bring practical understanding to policy/political debate, and I agreed with Joe that I will work on the issues that needs more technical clarifications. One such blur area in the ePR happens to be "metadata and the impact on privacy". So, this article is an explainer about the power of metadata and the reason why we need stronger policies in that context.

What is metadata?

Without getting too much into details about the technical or EU definitions of metadata, let us simply understand it as the **data about the data**. The table below illustrates the difference between the data and the metadata.

Activity	Data	Metadata
Alice **calls** Bob	(Voice) Conversations happened over the call between Alice and Bob	Time when Alice called Bob, duration of the call, Alice and Bob's physical location, any network related information such as routing path. Everything apart from the voice conversation can be treated as metadata here.
Alice sends an **email** to Bob	Content of the email	To, From, Date Time, Subject, network information, and anything related that you can imagine which is not the content of the email.
Alice takes a **photo**	Photo itself	Time of the photo taken, Geolocation information embedded to the photo, camera model, etc.
Alice **browses** the internet.	The content of the websites that he visits	URLs of the websites, description of the website which is usually included within the 'meta' tag of a web page, time of visit, visit count, network information.

Table 1: Data vs Metadata

Often we see that the data is considered to be sensitive and as a personal property it has to be protected. It is possible to protect your data using encryption technologies, for example GNU Privacy Gaurd (GPG) for emails. On the other hand, metadata is not treated to be very sensitive and for the same reason there are not many methods to encrypt it. It is due to the technical shortcoming of the basic building blocks of Internet Protocol (IP) stack. It does make sense to not encrypt the metadata right? Because if we encrypt the sender information on an email, your email client wouldn't know whom to send it to.

When the internet protocols were built, the intention was merely to establish a communication channel to connect the world. At that point of time, there were no much threats from government spying agencies, mass surveillance programs or from the advertisers. However, today we live in a world where everything we do on the Internet is being tracked and thus putting our privacy for sale on the data market. Even though the metadata has been a gold mine for Internet Service Providers (ISPs), Telecommunication providers from past two decades, the privacy risks of the metadata started to be a debatable topic since the Snowden revelations. Here are some of the quotes about the power of metadata from former big shots of government spying programs.

"*Metadata absolutely tells you everything about somebody's life. If you have enough metadata, you don't really need content.*"

- Stewart Baker, Ex- NSA General Counsel

"*We kill people based on metadata.*"

- Michael Hayden, former director of the NSA and ex- CIA

Metadata by its virtue is not invented to help privacy invaders; instead it was intended to fasten the process of classification and indexing of any kind of bulk data, without looking at the data itself. So, by definition, metadata enforces data protection by letting someone process the data, without even looking at the content inside. However, that is also the fastest way to profile the whole internet users, right? Earlier in October 2015, Share Lab presented this piece of investigative journalism which articulates the hidden power of email metadata. Indeed, it is scary to see what one can understand about personal behavior just from the "To", "From", "Subject" and "Timestamp" fields. Other than the scary use-cases, there are a handful of projects such as Proofmode (earlier known as Informacam - CameraV) which harness the power of metadata for combating against fake news. However, the number of projects which exploits that power for advertisement tracking and surveillance outbeats the genuine use cases of metadata.

Browsing history and the potential threat actors

Modern browsers such as Firefox, Google Chrome, Opera and Internet Explorer stores the browsing history to provide a user-friendly browsing experience. By default, these browsers store the history of all the previously visited websites, cached copy of the websites, form filling history, cookie information and also bookmarks. Depending on the operating system and the browser, these information will be stored in a specific location on the hard disk of your computer in a lightweight database. Some of us rant about this default nature of the browsers, as it compels users to manually opt-out of browsing history storing mechanisms and the privacy concerns associated with it. Browser history - specifically the website information and cached copy has its own advantage in terms of usability:

1. Automatic completion/suggestion of previously visited URLs.
2. Locally cached copies of the previously visited websites to boost up the browsing speed, which is very helpful when the Internet connection is very slow.

At this point, it is obvious that our browsing history is accessible to our browsers, which is why it is highly recommended to use open-source trustworthy browsers such as Mozilla Firefox ,which protects and respects your privacy. Whereas if you are using other browsers from the companies which are themselves the data brokers and advertisers, you end up giving away your browsing history to get tracked. So, assuming that we trust our browsers, let us exclude it from being a threat actor in our model.

Entity	Access to browsing history *	Comments
Malware in the computer	Full	Any program which has adequate privileges to start a browser process and browse the web potentially has the capacity to leak it. Such malwares have a high demand in the darknet. Other than that, there are browser hijacking malware which pollutes your history
Wifi Hotspot	Full	Using captive Wi-Fi is a common practice in many places, especially when using public hotspots.
Internet Service Providers (ISPs)	Almost full	ISPs can seek many insights, even when the traffic is encrypted. Have a look at "How Internet sees you" HTTP: The ISP knows which pages you're visiting and could see the data you send and receive.

		HTTPS: The ISP knows which domain you've visited but not the URL parameters, and not the contents of any data you send or receive.
Domain Name Service (DNS) Providers	Partial	Only the domain name queries and not complete URL.
Cookies (tracking, advertising and profiling companies)	Partial to almost full (depending on who's cookie it is)	Based on cookie origin policies, cookies from Website A can collect the history related to that.
Websites that you visit	Partial	Any websites that you visit would obviously know that you have visited them.

Table 2: Access to browsing history

In spite of the clear privacy implications, there is no clarity under the law about whether browsing history (more specifically the URLs) is to be protected as content or non-content metadata. Most of the lobbyists express their dissatisfaction about the changes between leaked and the official proposals of ePR. Out of many other concerns, the most questionable parts of the ePrivacy Regulations, at least for me is the permitted usage and exceptions of contents of communication. Things are a bit more complicated than that on two levels. Firstly, there are various cross-cutting issues (consent, tracking, ISPs, "value-added services", etc...) where metadata analysis comes up. Exceptions for web analytics could imply serious privacy concerns without stronger guarantees of statistical privacy.

Without getting much into that debate, let us explore browsing history as it provides a rich source of metadata of our daily interactions with the internet world. For the sake of simplicity and for understanding the power of this chunk of metadata, let us assume a malicious ISP (who can completely or partially see our browsing metadata) who does not respect the privacy policies to be the threat actor.

1. Analytics about the web history

Based on the browsing history contained in my computer, below is a simple analytics of the website domain names that I have visited the most. Like any other "normal" internet user I have used Google as the search engine; spent ample amount of time on social media sites such as Twitter, Facebook and LinkedIn; watched videos over Youtube; used Wikipedia as the primary source of information; shopped on Amazon; sought programming help over Github and Stackoverflow; and so on.

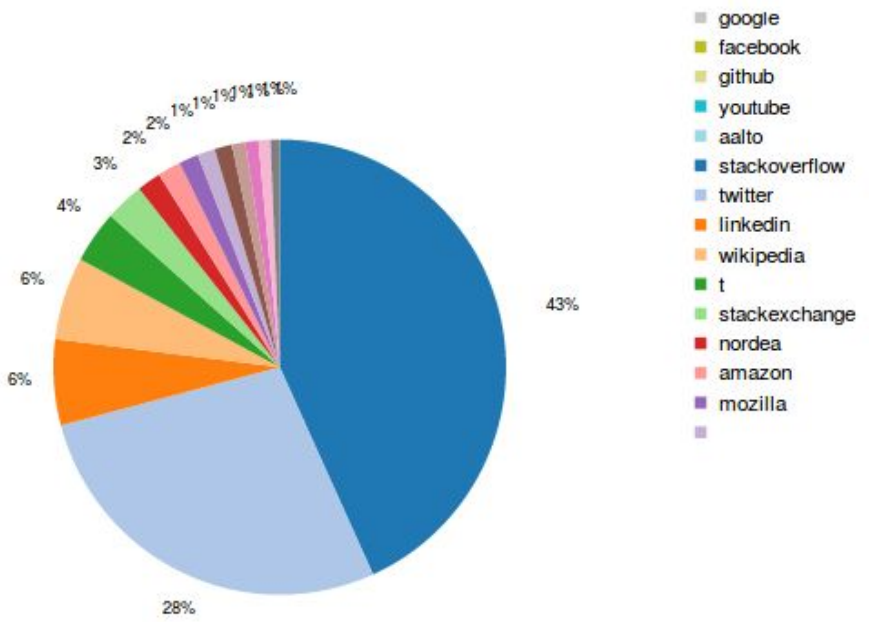

Figure 1: Most visited domain names

According to ePR and as per the global norms of deducing useful insight of the users, ISPs can use such analytics for their survey purposes. Under genuine use cases, these kind of statistics are helpful for fine-tuning the bandwidth for specific websites that are used more by the users. Even though the top 20 websites remain the same across all parts of the world, depending on demographics and social structure of a region, the websites that will appear after the top 20 are not always the same. Your contribution to big data analytics, starts right from here - just by contributing the domain names of the websites that you have visited. The same chunk can be used for profiling you as well. May be these websites in Figure 1, is most common to all and does not really profile as you different. But, imagine some of the porn sites or your favourite political parties web page! Well, that makes you little different than others right?

Figure 2 belows shows the suffixes or more technically the top-level domains (TLD) of the websites that I have visited the most. In many cases TLDs represent the countries that the websites are affiliated with. Also, websites like Google change the TLDs depending on the country from which you are browsing their website. For instance, even if you typed www.google.com from Belgium, it will be redirected to www.google.be automatically. Based on Figure 2, one can easily tell that I have connections with Finland (.fi), Belgium (.be), India (.in and .co.in) and some academic affiliation (.edu). While your ISP will obviously know these information, imagine the case when you are travelling!

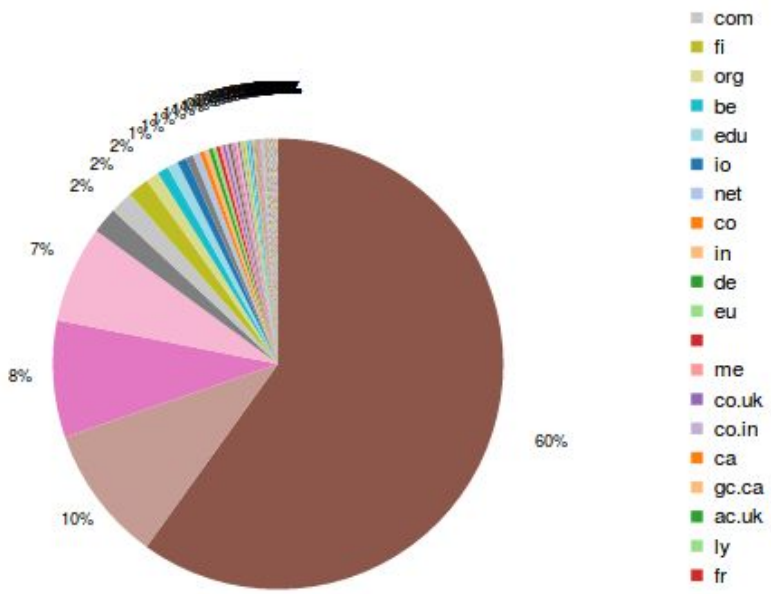

Figure 2: Suffix (TLDs) of most visited websites

Even you are in a foreign country, you still visit websites related to your home country. So, along with the ISP of the foreign country, your geographic affiliation or affinity is now evident to the DNS providers as well. At this point, you have contributed second chunk of information to the big data and profiling to two of the entities which can collect data about you.

2. Browsing patterns

If the internet traffic is HTTP, everything will be transmitted in plain text. So, ISPs can see full path of the URL (http://www.facebook.com/**zuck**). Whereas, when it is HTTPs only partial path is visible (https://www.facebook.com/) to the ISPs. To know more about how Internet works, refer to EDRi's paper on the same topic.

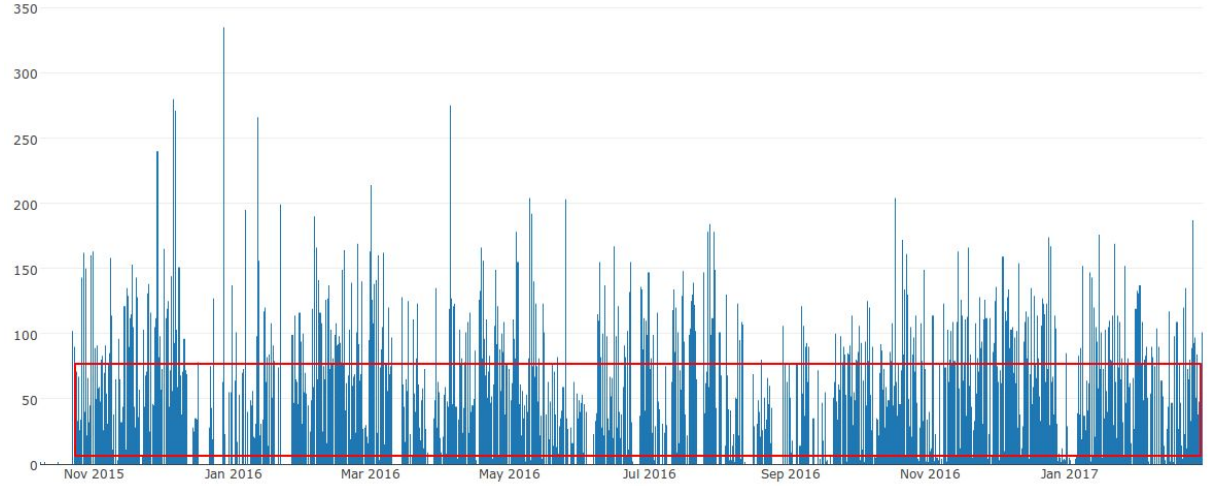

Figure 3: Number of unique URLs visited over time

Since the full path of URL is visible to the ISPs when your traffic is not encrypted, they can start analysing your behavior online. Figure 3 represents a graph of the total number of unique websites that have visited over time based on my browsing history. As one can see, I visit 10-150 unique URLs on an average over the period of November 2015 to January 2017. Some peaks in the graph beyond this range shows a lot of anomalies in my browsing pattern. These anomalies could potentially indicate certain specific events of my life. It could be increased workload, planning my travel, searching for a job or anything that you can imagine.

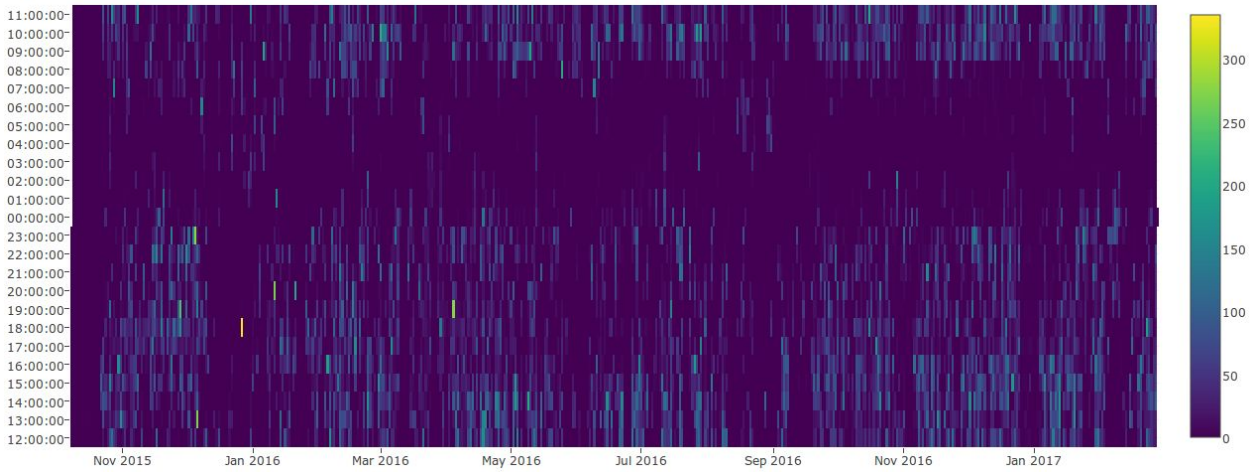

Figure 4: Heatmap of browsing pattern - unique URLs visited over time

Another way of looking at the browsing patterns is by plotting a heatmap of the same data *i.e.* the number of unique URLs visited over time as shown in Figure 4. While Figure 3 shows the anomalies in the browsing pattern, the heatmap gives a snapshot of the lifestyle in an easily understandable manner.

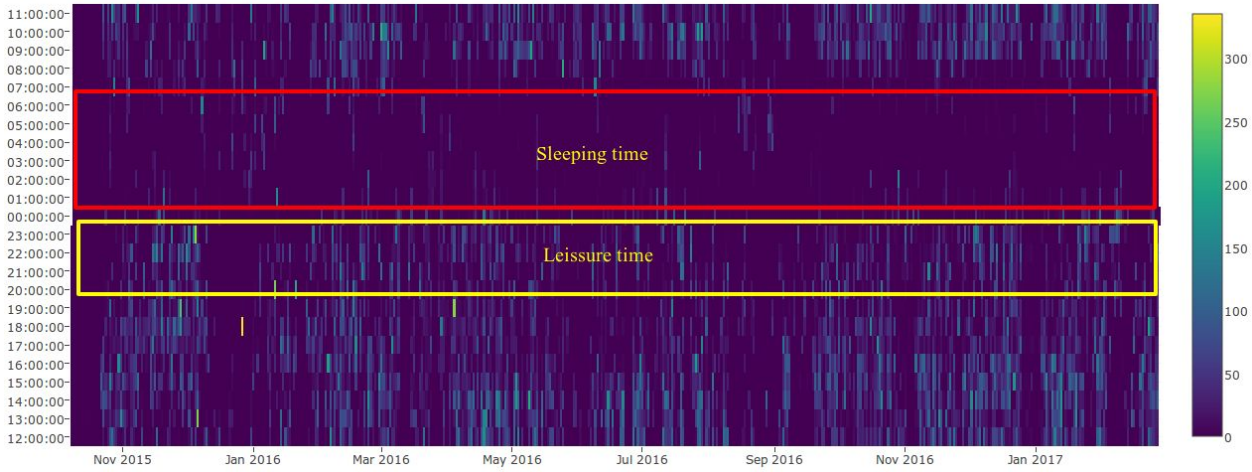

Figure 5: Heatmap of browsing pattern - sleeping (idle) and leisure time

There are consistent patterns in the lower half and the upper quarter of the graph. Even within those patterns, we can see two different sets, which depicts my work time browsing and after-work leisurely activities as it fades out from 20:00 hour onwards. In the figure 5, from 12:00 AM till 07:00 AM, there is a constant strip of dark patch which represents less activity over the internet, or in other words it is the time when I sleep.

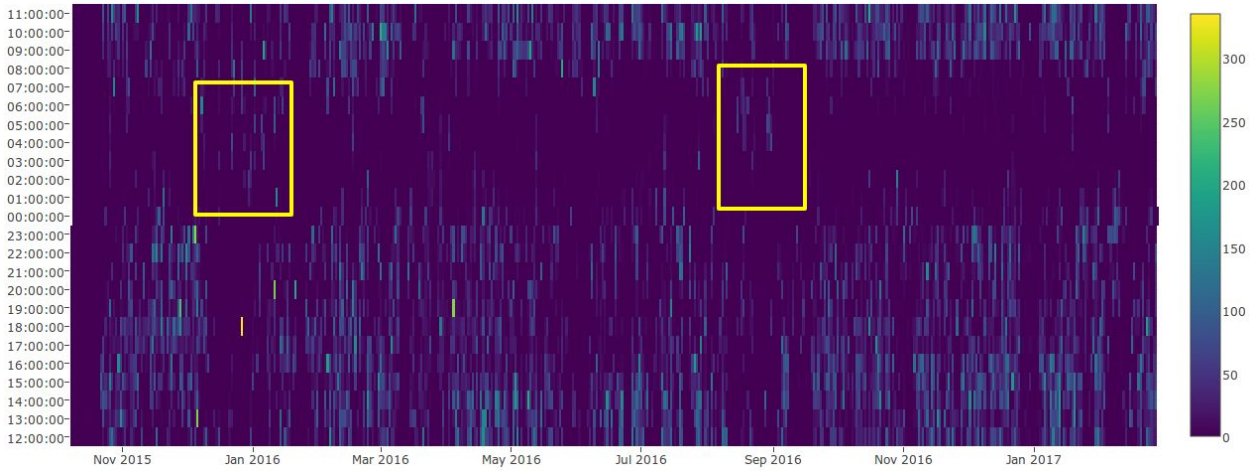

Figure 6: Heatmap of browsing pattern - travel

As highlighted in the figure 6, there are certain patches within the strip of my sleeping pattern. When correlated with the change in name suffixes (with reference to figure 2), it was found out to be work-related travels. In other words, I had travelled to a different timezone and continued to work from 9.00 AM to 7:00 PM as I have done on any other regular day.

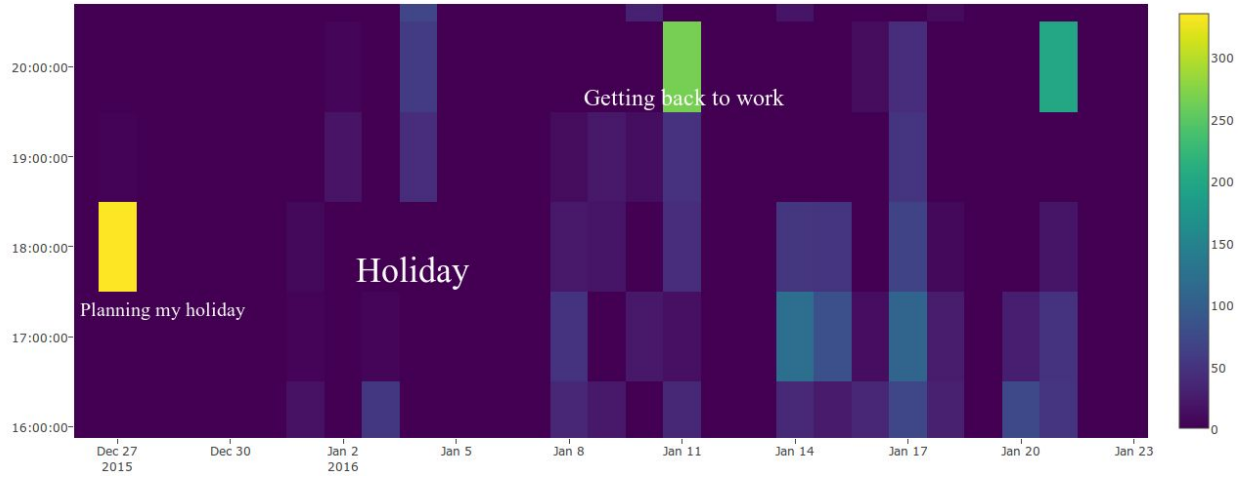

Figure 7: Heatmap of browsing pattern - Holiday season

If we zoom in the graph more (As represented in figure 7), there are patterns which show high number of browsing, a patch of almost no activities even during the regular working hours, then a sudden increase in browsing activities and finally resuming to normal working hour pattern.

This depicts that I planned for my holiday (checking into flights, confirming hotel booking, etc.), took a break from work, returned from the holiday (sudden increase is possibly due to following up on emails and activities that I might have missed during my trip) and finally resuming my work.

So, at this point, one can know about my working hours, sleep time, work-related travel and my holiday schedules just using my browsing metadata. That is quite a lot of information about me retrieved just from the metadata right?

3. Potential adwords

As mentioned earlier, browsing history falls into the grey area of whether to be treated and protected as content data or as the non-content metadata. Unlike many other metadata, where it is not possible to retrieve the complete content data just by using the metadata associated with it, it is possible to retrieve all the contents of the websites that you have visited by crawling over the list of URLs from your browsing history. Whether or not it happens in reality, to avoid giving the list of URLs directly to advertisers, the ISPs can automate their analytics system to crawl over the list to seek insight on what you might have seen while browsing. By giving away just the keywords deduced from the websites that you have visited to the advertisers, the ISPs can potentially bypass the privacy laws by claiming it as anonymized.

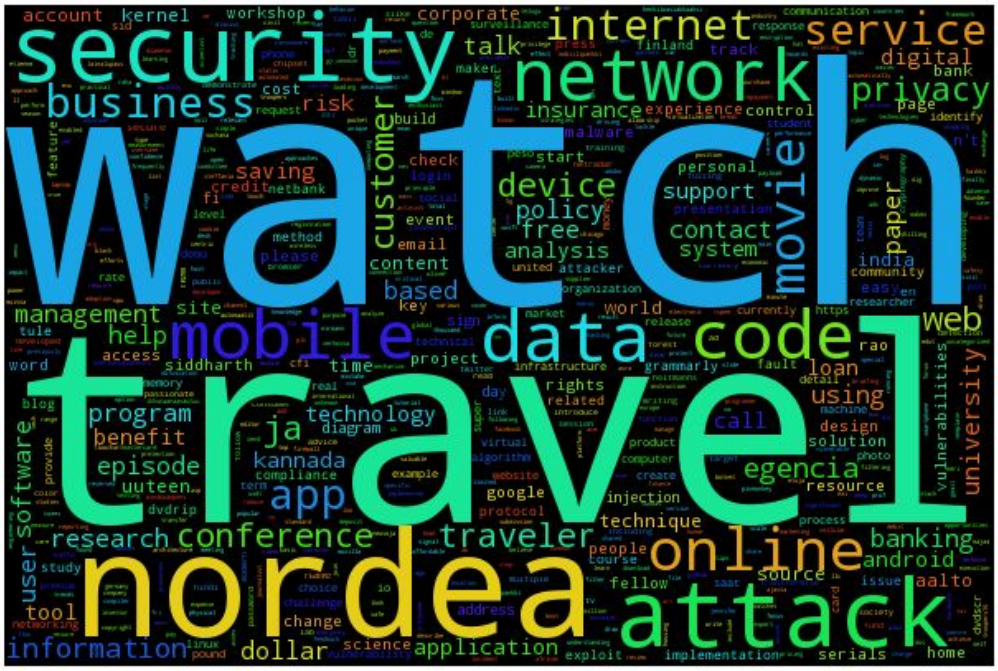

Figure 8: Wordcloud generated by crawling over the list of URLs from the browsing history

Figure 8 represents the wordcloud generated by crawling over the most visited websites by me. Not so surprisingly, being a security and privacy researcher, I can see those words in this cloud,

along with other keywords related to my identity - both from professional and personal life. This cloud was derived by excluding all the social media and search engine related URLs.

Yet another buzzwords which we hear often these days is - “Data mining” and “machine learning”. Data Mining refers to seeking useful insights programmatically from the collected bulk data, whereas machine learning is to use that insight for data-driven decision making. One of the features of these methods known as [Named Entity Recognition (NER)] which allows to classify the text into categories such as organizations, persons and locations.

Figure 9: Wordcloud generated by Name-Entity Recognition - Organizational entities

If we run NER algorithms on the text retrieved by crawling over the list of URLs that you have visited, it provides more clarity to the keywords that can be potentially generated. Figure 9 shows the keywords related to organizational entities from the websites that I have visited. This narrows down my generic profile to target me on the keywords found in this cloud. For example, I could be a potential customer for insurance companies, University and management related jobs.

Further down the line, figure 10 represents the names of the people found in the websites that I have visited the most. Surprisingly, I turned out to be the self-obsessed person who visits websites of his own or the websites that talks about himself. In the Person names cloud, I can see some of my academic co-authors, role models or the people whom I follow. Imagine that there

are the names of Tim Cook or Steve Jobs! I am probably a potential customer for Apple! So, the list of adwords targeted towards me could include Apple products here onwards.

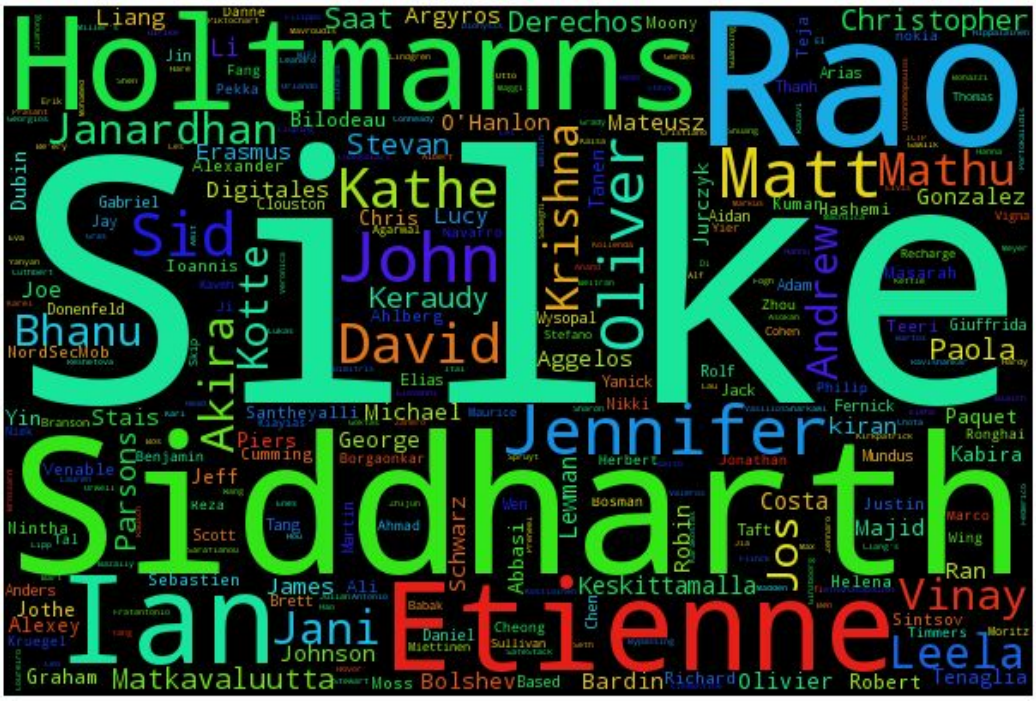

Figure 10: Wordcloud generated by Name-Entity Recognition - Person names

How about my next travel destination? Can it be predicted from my web history? Possible yes - it could be Brazil, China or Singapore!

Figure 11: Wordcloud generated by Name-Entity Recognition - Location entities

As the Figure 11 represents, I might have visited the websites which contained those locations which could probably be my next travel destination. Even without doing any fancy machine learning processing, I could attest that these were actually some of the places that I am planning to visit!

As mentioned before, if you are using HTTPs, the ISPs can see the full URL path in clear text. Along with them, the websites that you visit will obviously have to know that full path to deliver you exactly what you are looking for.

If you have searched for "vegetarian restaurants in Brussels"in Google , your Google query URL will be https://www.google.be/search?q=vegetarian+restaurant+brussel. Assuming that the ISPs will use the keywords you are searching to profile you again, it makes their job of deriving the adwords for your future targeted advertisements much more easier.

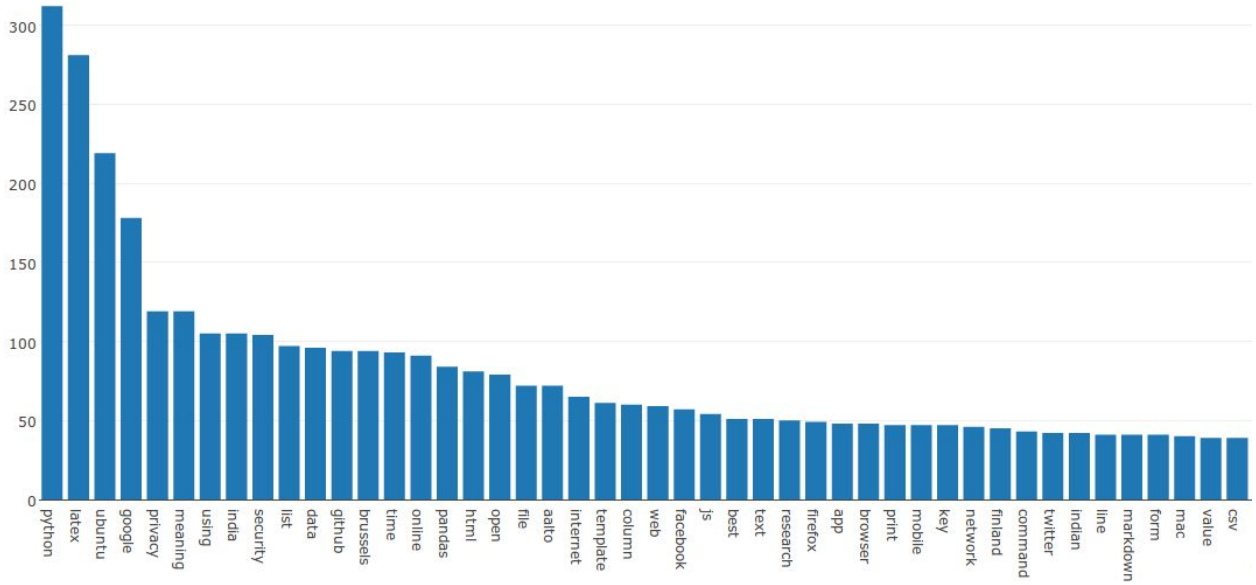

Figure 12: Word frequency graph of Google search keywords

Figure 12 represents the most searched words by me on Google. From this graph, it is evident that I use Python programming language, Latex for writing reports, use a computer with Ubuntu as the operating system, research on security/privacy, and so on. This itself along with the previous world cloud would be enough to profile me.

So, at this point, the ISPs know what makes you highly likely to click an advertisement link!

4. Pseudo social sphere

Unlike the metadata related to emails and phone call logs, the browsing history can be treated as one-dimensional metadata. Because, it is just the metadata about what you have browsed and it does not contain the influence of other people's interaction with you. On the other hand, email and phone call metadata contains the interaction you have done with others, along with the interactions done by others with you.

However, it is possible to seek insight on your affinity towards the people within your social circle using the one-dimensional browsing history metadata. For example, you will visit your close friends social media profile more frequently than you visit your ex-colleague's profile whom you know from first job. You might have visited the profile of your friend from the university more recently and frequently, than you visit your friend from high school. By capturing the number of visit counts and frecency (frequency + recency) from your browsing history, it is possible to reconstruct a pseudo social sphere (figure 13) , and thereby converting the browsing history to a two-dimensional data source.

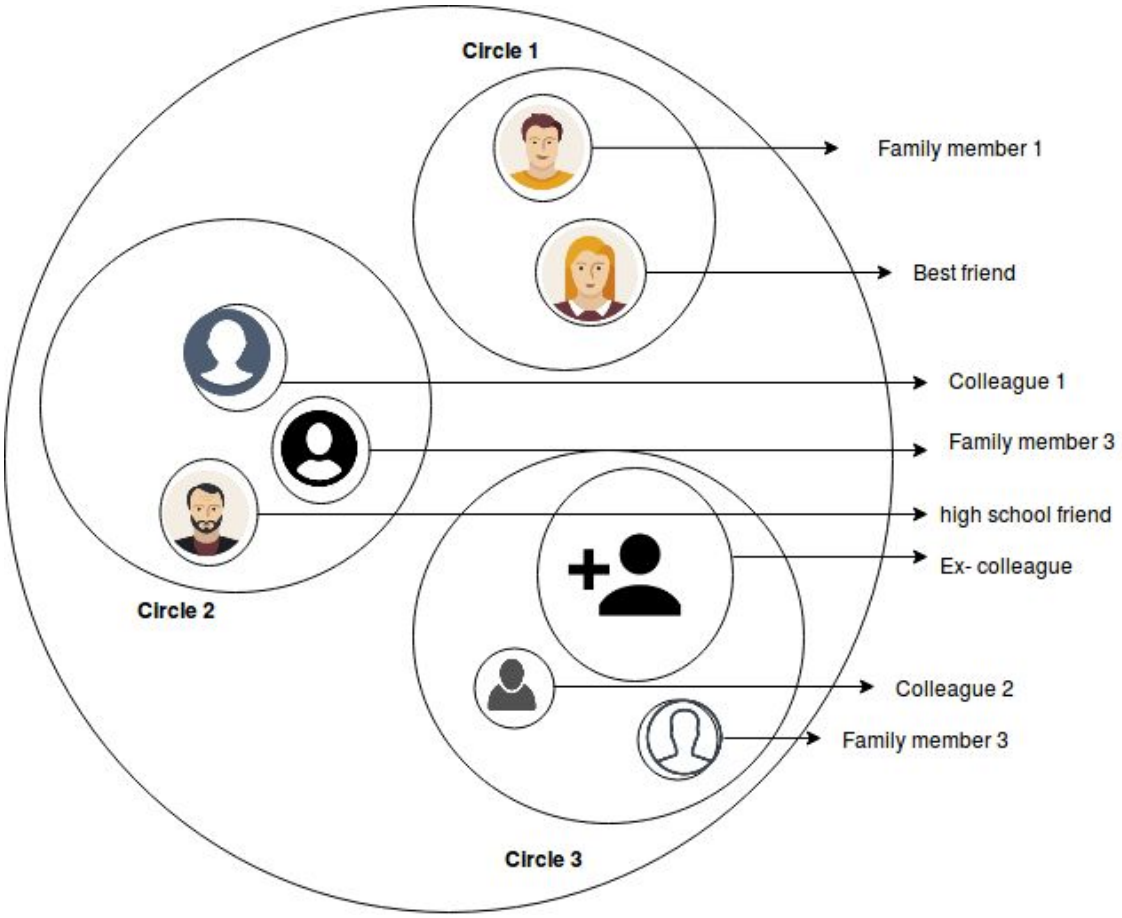

Figure 13: Representation of pseudo-social circle derived from social media related URLs

We all have different social circles - family members, childhood/high school friends, friends from work place, ex-colleagues , etc. Our affinity towards them is not necessarily unique. Even though they are not directly connected with each other, it is highly likely that our affinity towards them is similar. By capturing the social media URLs (Facebook and Twitter), Figure 13 represents one such social sphere. In circle 1, I saw a family member and my best friend; in circle 2, one of my colleagues, highschool friend and a family member were seen. This means I weigh them differently, but they can be grouped based on my affinity towards them.

Just by knowing the browsing history, now the ISPs can tell who are my close friends, how much do they matter to me and who all have equal importance in my life.

To summarize:

I built a small/ naive tool to replicate the similar graphs shown in this article for almost anyone who is a Linux+Firefox user, browses Internet including social media like anyone else and most importantly stores the browsing history for a decent period of time. While making this tool as generic and simple as possible, I had to omit digging more information that could have been gathered from my own browsing history and exclude use of APIs (as they require individual users to obtain the API tokens). However, to know more about what browsing history could reveal about your personalities, refer to the case study by Share Lab. This provides lot more insights on what one can dig from your browsing history.

Whether or not the culprit ISPs as depicted in this article evade your privacy by doing all these analytics, it is indeed important to realize the power of metadata and your contribution to big data processing in the wild. Since privacy of the metadata can not be protected by merely encrypting it, we need stronger policies to defend our digital rights.

The tool which I call as Haukana metadata can be downloaded from **here**. Once you download it, follow these instruction:

- Unzip the folder a → right click on a blank area → Click on “open in terminal”.
- In the terminal, type **sh requirements** and **press Enter.**
- This will download all the necessary modules needed to run the tool.
- Once it is completed, type python **tool.py** and **press Enter.**
- It will take some time to process your browsing history. So, be patient until it opens a new browser tab as a result. Everything will be processed within your computer and hence, the tool does not send the data anywhere.

- The newly opened tab will contain some instructions and links to the visualizations derived from your browsing history.
- It is important to note that some of the functionalities may not work as it is shown in this article, mainly because there are no reference data about browsing history. So, I had to build it based on my own browsing history.
- It goes without saying that the code is open source, and any contribution to the code to improvise and add more functionalities are more than welcome. Even otherwise, in case of issues, do not hesitate to contact me, either by sending an email with subject line “Hakuna Metadata” to sidtechnical@gmail.com or by raising an issue on Github.