Investigating the Third party Web Tracking Eco System

Department of Informatics

Computer Science with Artificial Intelligence

Author

Yasas Karunatissa

132268

Supervisor

Dr. Martin Berger

Statement of Originality

This report is submitted as part requirement for the degree of Computer Science with Artificial Intelligence at the University of Sussex. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged.

Signed:

Yasas Karunatissa

Acknowledgements

I would like to acknowledge my supervisor, Dr. Martin Berger for his continuous support and guidance. The feedback received from my supervisor was both constructive and encouraging. This allowed me to overcome several problems I was having when completing my project.

Abstract

The goal of this project is to carry a thorough investigation towards online web tracking. Online web tracking is the art of measuring user behaviour on websites. In order to carry out this investigation, the project utilizes many tools such as selenium and mitmproxy to collect data. The data collected is evaluated using a machine learning algorithm to classify whether the data collected are tracking or non-tracking. The following report will analyse the methodology and the resources used to achieve this goal.

Contents

Figure table

1.0 Introduction

1.1 World Wide Web

The World Wide Web, also referred to as WWW is a network of online content created with the use of HTML. Hyper Text Markup Language, HTML is used to create the online content such as web pages; for example websites such as BBC, Youtube and Facebook. These websites can then be accessed by a Uniform Resource Locator, a URL. A URL is a unique web address given to a website which specifies its location on a computer network Figure 1. By entering the URL, the website can be viewed via a web browser, I.E. Google Chrome, Internet Explorer or Firefox. The webpages created using HTML can then be accessed via HTTP requests.

1.2 HTTP/HTTPS

HTTP stands for Hyper Text Transfer Protocol. This is the protocol used by the World Wide Web, it defines how messages are being formatted and transmitted. HTTP also defines which actions web servers and browsers should take in response for the various commands. HTTP requests are made to the server using the browser for HTML pages, images, scripts, style sheets amongst other data. Once the request is made to the server, the server will return a response containing the requested resource. This is known as a HTTP request-response cycle as displayed in Figure 2.

HTTPS is the secure version of HTTP. When a website uses HTTPS, this means that the communications between the browser and the website are encrypted by either Transport Layer Security (TLS) or Secure Sockets Layer (SSL).

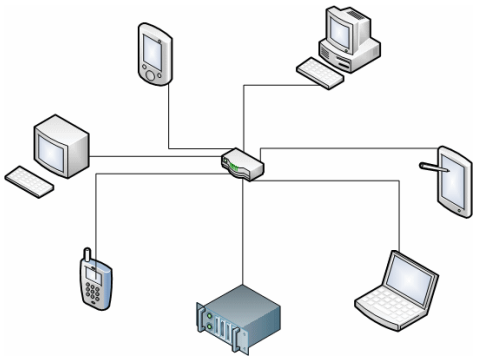


Figure 1 Network of computers [10]

Many websites in the modern day use cookies. The purpose of using cookies on websites is to store small amounts of data. Firstly, when the user accesses a website for the first time, it will generate a cookie for the accessed website. This cookie file will then be saved onto the user's computer. The cookie file will save

the web address of the website and any commands that the browser sends back to the website each time the user visits the website. The cookie files however do not save user information, they are mostly used to identify user behaviour on websites. This is so that the website can better cater for the user's behaviour. For example, when accessing a website, the web server will need to if the user requires logging in, adding items to a virtual shopping cart or completing any other process which requires the website to remember information as the user browsers different pages on the website. There are two different types of cookies known as first party cookies and third party cookies. The first party cookies are set by the website that the user visits, whereas the third party cookies are set onto the user's computer by a domain other than the website that the user is visiting.

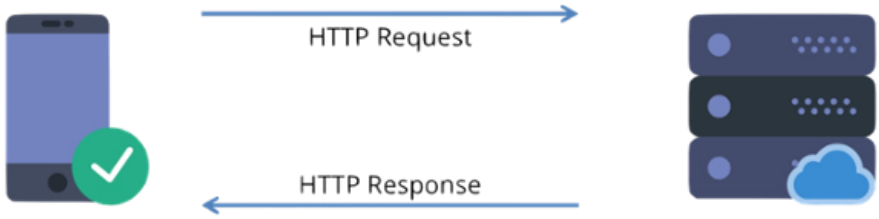


Figure 2 Diagram displaying the communication between the client and the server [11]

In the World Wide Web, there are two different types of websites, known as static and dynamic. A static website is a website which contains web pages with a fixed content. A static website developed using HTML. The developed webpage will display the same information for every user that visits the website. Static websites can differ from small to large website. This means that it can have one page or it could have several hundred pages.

Dynamic websites on the other hand, are web pages which are generated in real time. Such webpages include web scripting such as PHP and ASP. Once a dynamic web page is being accessed by the user, the page is then parsed on the web server, which results in HTML being sent to the client's web browser.

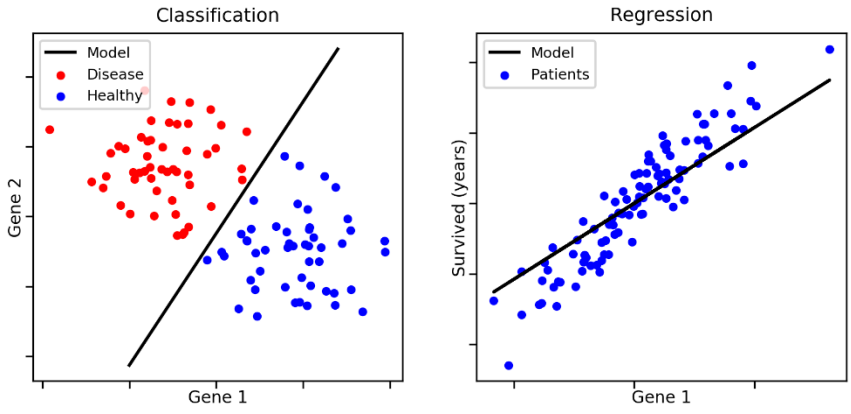
Both type of websites, static and dynamic use cookie for different purposes.

For this particular project, it extracts cookies through the HTTP requests send to the server from the browser and determines whether they are classified as tracking and non-tracking. Firstly, in order to gather cookies, the browser will need to access a list of websites. This process of the project is done through web-automation using the tool 'selenium'. Web automation is the process of replacing human interaction, this means that rather than having to access websites physically, the browser will iterate through a list of website and access them one by one. In order to pursue with this step of the project, the ideal tool selected is 'selenium'.

The next step of the process is to record the HTTP requests made from each website. To capture the HTTP requests made to each website, the tool used is MITMproxy, also known as Man In The Middle proxy. This tool will record all the HTTP requests made to the web server by each website, it will record the number of times this cycle occurs in one website. Using the collected HTTP requests, the cookies for the following HTTP requests will be extracted. This process is conducted via the use of a pythons request API. The request API, is used to call the cookiejar from each of the domains. The cookiejar will return a list of all the cookies that specific domain will use. When using tools such as selenium and MITMproxy, it requires a good, steady internet connection. Firstly, to access the given list of websites; secondly to request the cookiejar values from the domain. If the network connection is not steady and cause interruptions this means the request to the domain will not be made; this means some requests will be missed as the connection cannot be made. Hence missing cookie data from some domains.

Finally, the collected data is then assessed using a machine learning algorithm. Machine learning is an AI (Artificial Intelligence) based application which has the ability to learn and improve from experience rather than being explicitly programmed. Machine learning algorithms are often categorised as supervised and unsupervised. Supervised machine learning apply what has been learnt from the past data to new data with the labels which are being used in the algorithm in order to predict future events. On the other hand, unsupervised machine learning algorithms tends to learn from training data which are unlabelled, unclassified or uncategorised. Unsupervised learning identifies common features or common attributes in the given data and react to it based on the presence and the absence of those features attributes in the data.

Figure 3 Classification vs Regression [1]



With a machine learning algorithm, it will project a classification score as well as a regression score. Classification determines which class the data will belong to. On the other hand, regression is used to predict values of a given target as demonstrated on Figure 3.

This project uses a machine learning algorithm to justify if certain cookies are tracking or non-tracking. With the use of machine learning, it will produce an accuracy score as well as the loss function. The accuracy is will demonstrate which machine learning model is the best at identifying the relationships and patterns between variables based on the input data or training data given. The loss function defines how well the algorithm models the given data set. This means that if the predictions are wrong, the loss function value will be higher.

These factors are highly considered throughout this project in order to monitor how well the algorithm justifies whether cookies are tracking or non-tracking.

2.0 Professional Considerations

2.1 BCS Code of Conduct

2.1.1 Public interest

This project will under no circumstances violate the public health, security or the wellbeing of others and the environment. If third party data is being used for this project, it will kindly ask for the users consent and inform the user of what types of data is to be gathered. In addition, any papers or research articles used in the report and any sources used towards completing the project will be carefully cited and given credit to. Moreover, under no circumstances will the project commit any type of discrimination i.e. sex, sexual orientation, marital status, nationality, religion, age or other requirement. The project will also honour the equal rights act of 1D of the BCS code of conduct.

2.1.2 Professional Competence & Integrity

This project will under no circumstances violate the professional competence & integrity. It will not use any other work that is not within my professional competence. The project will only use work which are relevant to the project and my level of study. Although, any work that will be used in the project will be referenced and given credit to. Building upon my project, I will develop my professional knowledge, skills and maintain awareness of other technological developments. Furthermore, I will respect and value others viewpoints. I will also accept honest criticism of my work in order to improve my project further. The project itself under no circumstances injure others, their property or reputation. When conducting this project, I will not make any offer of bribery or unethical inducement.

2.1.3 Duty to Relevant Authority

I will accept my personal duty to uphold the reputation of the profession. It will also honour and under no circumstances violate the code of conduct of the University of Sussex. The project will be completed to a level that no conflicts arise with the University of Sussex. Any other participant who has been involved in the project will be acknowledged and given credit to.

2.1.4 Duty to the Profession

This project will uphold the profession of Computer science and will not take any action to bring profession into disrepute. Moreover, this project will seek to improve the professional standards by enforcing and honouring them.

2.2 Ethical Implications of project

As the project mainly investigates how cookie behave on websites, it does not directly violate the ethical code conduct. The project will evaluate the cookies being extracted from a list of websites. In no way will the data gathered for this project will be sold onto other third party companies. The cookie data being collected will be assessed to determine the websites intentions in regards to the user. This will in no way discriminate the data or any individuals.

Moreover, as stated above, any use of resources such as research papers, online articles etc used towards completion of the project will be given the rightful acknowledgement and the credit it deserves.

3.0 Background research

In the modern day, there are many websites with cookies in place. A news article published by CBC addresses 'How they could be undermining your privacy' [2].
The articles begins by addressing the fact that every user online has a unique ID. Users are not saved as their name, users have simply been assigned a unique ID by an 'online ad company'. The personal unique ID is saved on to the user's computer along with the user's web history. The article further describes how an 'evesdropper' can use cookies from advertising and tracking companies to link 90% of a user's web page visits to the same 'pseudonymous' ID. This 'pseudonymous' IDs can often be linked to real-world identities. The article clearly states that the users name, username and email will all be linked to the pseudonymous ID. This means that using this ID, the eavesdropper can easily identifier the real life user.

As described above, many websites in the modern day make use of cookies. There are two types of websites, known as static and dynamic. A static website is a website which contains web pages with a fixed content. A static website developed using HTML. The developed webpage will display the same information for every user that visits the website. Static websites can differ from small to large website. This means that it can have one page or it could have several hundred pages.

Dynamic websites on the other hand, are web pages which are generated in real time. Such webpages include web scripting such as PHP and ASP. Once a dynamic web page is being accessed by the user, the page is then parsed on the web server, which results in HTML being sent to the client's web browser.

In comparison, dynamic websites benefit more with the use of cookies in comparison to static websites. This is mainly as most website allow users to login, thus with the use of cookies, the website can save session or user details for easy access; rather than having to query the server each time, the website can easily access the cookie to retrieve the information needed. This is beneficial for the website as it will take less time to retrieve information from the cookie file rather than querying the server for user information. On the other hand, static web pages most display the same content for every user; thus the use of cookies are not necessary. However, static websites can still use cookie in order to measure the number of user visits it gets. Also, static website may use cookies when displaying ads on websites. This means that user-targets ads will be displayed on the website.

More to the point, the article describes how the technique 'cookie-linking', which works 'ubiquity of third-party trackers'. The article says these trackers are 'everywhere' and largely invisible to the web users. In order to test this theory, the author, conducts his own test. The article states by visiting cbcnews.ca, it reports that the domain is connected to 28 third party sites. These include advertising networks, tracking services and social media site. When visiting the buzz feed website, it also connects to 17 third-party sites. After visiting both these websites, the author found that there are common

third party trackers which are in place on both websites. These being doubleclick.net, scorecardresearch.com and adnxs.com. This means that both websites have an overlap of the trackers being used; this overlap is key to cookie linking.

A paper named, 'cookies that give you away: Evaluating the surveillance implications of web tracking' [3] conducts an examination of the alexa's top 50 websites which require account creation. This examination is conducted in order to measure the number of attributes leaked when creating an account.

Plaintext Leak Type	Percentage of Sites
First Name	28%
Full Name	12%
Username	30%
Email Address	18%
At least one of the above	60%

Figure 4 List of leaked data from alexa top 50 websites supporting user accounts

The paper states, from the tested list of websites, 34 of the 50 websites tested used HTTPS to secure login pages. From the 34 websites, only 15 kept using HTTPS for future interactions. It also states that, majority of the sites tested secure the user credentials on login pages and personally identified information such as name, username and email address is however transmitted via HTTP and not HTTPS. The results showed that over half of the sites leak at least one type of identifier and 44% leak either username or email address. These identifiers can then be used to cross reference with the real-world identity. The table displayed in Figure 4 shows the percentage of user information leaked by the tested websites.

In addition, the paper states a web identity leak problem by www.youtube.com. As www.youtube.com uses a google property for the users to sign in, this means that other google owned domains such as gmail and google drive can still interact with the browser by default with the use of HTTPS. However, depending on the users preset setting on the browser, youtube.com will default to a non-HTTPS page after login. Also, the attack conducted by the paper, identities that the information leakage from popular website occurs due to clusters. Clusters are a group of website linked together which use the same login-in property, similarly to googles. The results gathered from the attack show that 6 of the 30 websites attacked leaked at least one type of identity linked to the user's real life identity. The paper states that there was a data leak found in every one of the 45 AOL user browsing profiles. There also many other websites apart from the alexas top 50 ranked websites to leak identities via clusters.

The paper, 'Online-Tracking: A 1-million-site Measurement and Analysis' conducts online tracking measurements on the top 1 million websites [4].

The article begins by describing the two types of tracking parties, ‘first party’ and ‘third party’. First party tracker is the websites the user visits directly. Whereas, third party trackers are trackers which are hidden trackers. Third party trackers are capable of obtaining the users browsing history with the use of cookie and other mechanisms. Third party trackers often communicate with the ‘referer’ which tells the tracker the first party sites the user is visiting. With the use of the ‘referer’, third party trackers are capable of obtaining user email addresses and other sensitive information regarding to the user. The referrer request header contains a link from the previous web page to the current web page the user has visited.

Previous findings as listed in the article describes the ‘largest third party organisations’ growing from 10% to 20-60% of the top websites between the years 2005 to 2008. Later studies discuss the ‘prevalence’ and the ‘complexity’ of trackers. This being said, it discusses how the trackers have evolved through the years from when it first began. A study carried out by Libert, identifies third-party HTTP requests on the top 1 million sites. This study showed that, the tracker Google is capable of tracking nearly 80% of sites through its ‘various’ third-party domains. The study also discusses on how trackers have since evolved from simple HTTP cookies to using cache E-tags and HTML5 local storage. These findings then lead to media backlash and legal settlements against the perpetrating companies. The primal usage of trackers is being able to adversities online ads related to the users online behaviour. This means that the site will display ads based on the user past activity. This can be of help to the user as it may help them find the products they need online. However, a further study identifies the trackers raising ‘privacy or ethical concerns’. Such as ‘price discrimination’, this means that a product will display different prices to different customers than what the product is being sold for in the market. Another raised concern, the ‘filter bubble’, which occurs when online information systems ‘personalise’ what is being displayed on the website due to the users online behaviour and what they have viewed in the past.

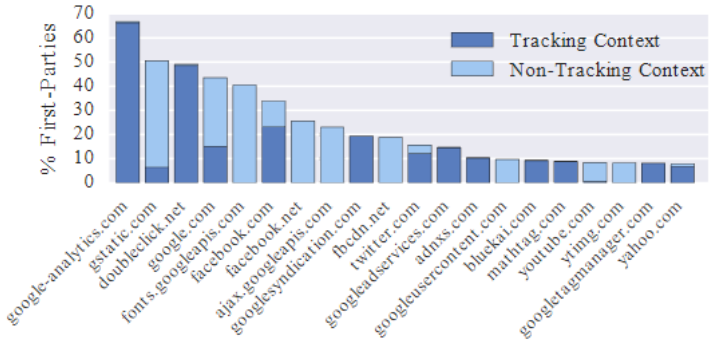
The article further discusses on the web security measures it has taken in order to detect such trackers. It states that, Yue and Wand modified a Firefox browser source code to measure any insecure JavaScript implementations of any websites.

Moreover, headless browsers have been used to measure third party JavaScript implementations in many popular websites. Headless browsers are also capable of detecting the vulnerabilities which are likely to arise from such scripts. Many studies have justified that using selenium-based frameworks, it has been able to measure and categorise malicious advertisements being displayed, to measure malware and other vulnerabilities on live streaming websites.

The paper reports on the results taken from January 2016 of the top million sites. It begins by describing that the online trackers having a ‘long tail’ followed by a quick ‘drop off in scale’ for individual trackers as Figure 5 displays. This means that the number of online trackers being used by websites have decreased suddenly. The paper aims to answer questions such as ‘who are the largest trackers?’, which sites embed the

largest number of trackers?, which tracking technologies are used and who is using them?'.

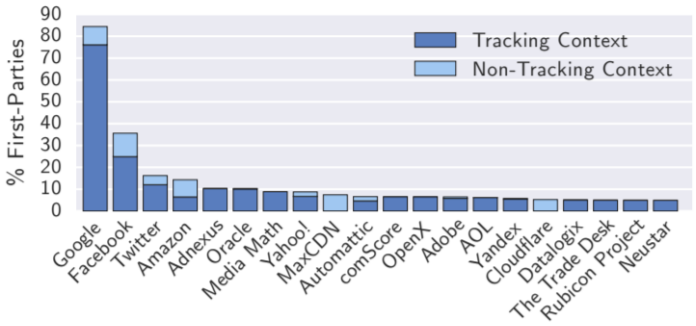
Figure 5 the long tail of Online Tracking



As before mentioned, the study is conducted of the top 1 million sites, creating over 90 million requests. The results gathered show that not all third parties are classified as trackers.

Moreover, the paper analyses the top third party organisations many of which contains multiple entities. The study conducted shows that Facebook and Liverail appear as separate entities for third party tracking. However, Liverail is owned by Facebook. The study uses the domain-to-organisation mappings provided by Libert and Disconnet to carry out the experiment.

Figure 6 Top third party organisation trackers



The results gathered from the study shows that Google, Facebook, Twitter, Oracle and Amazon are the leading third party organisation trackers displayed on more than 10% of the sites, as displayed on et to carry out the experiment.

Figure 6. Google appear to be significantly higher compared to any other third party organisation tracker.

A paper named, characterizing the use of browser-based blocking extensions to prevent online tracking illustrates on how many users use trackers in order to block any unwanted content. Such ad blockers, track blockers and content blockers.

This papers goes into detail about a study that was conducted between tracker uses and non-users. The study consisted of three major questions as listed below:-

- How much do users understand online tracking and does heightened knowledge about online tracking relate with users adopting such blocking extensions.
- Do users consciously adopt various blocking extensions to protect themselves from online tracking?
- When and how do users disable their extensions and accept being tracked?

The results gathered from the study are very mixed. Most people have said that they do not use any online blocking extensions. On the other hand, some say that they use blocking extensions to improve their online experience. Furthermore, the study also shows that some of the users disable the blocking extensions when the content of the website is restricted. This is an interesting way for the website to inform the user that they need to disable the extension, if the user wants to access the full contents of the website. [5]

3.1 Online tracking

The paper suggests that within the website, there is a first party tracker as well several third party trackers.

The paper also suggests that first parties often use third party trackers to gain information about their customers and assess their behaviour on the website.

Taking the New York Times paper website for example, the first party tracker that the users knowingly interact with is 'New York Times'. However this is not the only tracker that this website contains. There is another trackers, a third party tracker called the 'Google tag manager'. This tracker will provide the New York Times about their user analytics, such as the number of visitors on their website and marketing support. Another third party online tracker which is embedded on the website is called, 'Google Publisher Tags'. This tracker is particularly for 'targeted advertisements'. This is also called 'Online Behavioural Advertising (OBA), this is basically based on peoples interests and their demographics and browsing histories.

The paper also says that users to do not interact with the trackers knowingly, this means they are not aware of the trackers which are active on the website. This is considered 'violating privacy'.

Third party web trackers are embedded across the websites, the trackers are then able to link these websites together and see the browsing history. Also, by visiting certain websites people can reveal their sensitive information such as their interests, demographics and the devices they use in order to access the websites. Third party trackers are also able to track people by 'largely employing stateful tracking'. This involves the use of HTTP cookies, to track the number of website visits. Regardless,

some trackers engage in a more persistent and stateless way by ‘re-spawning flash cookies’ and ‘fingerprinting’. Both of which is capable of tracking people even when the HTTP cookies are cleared.

3.2 Perceptions of online tracking

This paper examines some studies which have taken place, asking people how the internet works as well as the online privacy and security attitudes and behaviours. The study showed that people with a stronger technical background was able to understand the security and privacy threats better than people with no technical background. However, the people with a stronger technical background took no extra steps in order to avoid it. In another study, people reported in regards to data aggregation from third parties than first parties.

When conducting a study in regards to people’s opinion on online tracking driven OBA (Online Behavioural Advertising), people’s attitude was ‘naunced’, i.e. people had mixed feeling. This is mainly because at times OBA provided good, useful adverts that the users liked. Other times, it provided with bad, embarrassing adverts which the user didn’t want to see.

3.3 Data Collection

As stated above, the project collects cookie data in order to analyse whether their tracking or non-tracking. In order to collect cookies from websites, the project will visit the most visited websites by users. For this purpose, the project uses alexas top 250 visited websites by users. The list of website can be found here [6].

3.4 Machine Learning

Previously mentioned, the project uses machine learning to classify where cookies are tracking or non-tracking. Thus, this process requires a machine learning algorithm. The two ideal machine learning models are SVM and MLPC.

3.3.1 SVM

The first chosen model is SVM, Support Vector Machine. This is a set of supervised learning methods used for classification and regression. However, SVM is most used for classification purposes. With the SVM algorithm, it plots each data item as a point in n-dimensional space; n being the number of features. The data will be plotted with the value of each feature being the value of particular coordinate. Then, classification is performed by finding the hyper plane which separates the data into two, in this case tracking and non-tracking.

3.3.2 MLPC

The other model taking into consideration is the MLPC, Multi-Layer Perceptron Classifier. MLPC is known to be a feedforward artificial neural network. A multi-layer perceptron consist of three lters of nodes such as an input layer, a hidden layer and an output layer. MLP also uses a supervised learning technique for training called backpropagation. Backpropagation is used to calculate the gradient that is needed to calculate the weights which are to be used by the network.

3.3.3 Grid Search

The machine learning aspect of the project performs a grid search. A grid search is performed in order to find the best hyper parameters for the model. The parameters return by the grid search will help the model increase its accuracy towards its predictions.

3.3.4 Training data

To train a machine learning model, training data is used. Training data is used in an algorithm in order for it to understand the patterns of different data. Furthermore, identify any relationships in the data.

For this particular project, an extensive research was carried out to find training data for the machine learning models. From the research carried out found a website named ‘Block List Project’ had many lists of URLs which belong to several different categories. The website contained a list of URLs which has captured using browser plugins such as ad blocker and ghostery. These plugins are used on browsers in order to remove any ads/popups on websites.

3.3.3.1 Browser blocking extensions

The purpose of browser blocking extension also referred to as a plugin is to remove any unwanted or addition content from the webpage that is currently being displayed on that particular webpage. For example, this could include any additionally displayed ads. Once any additional content is detected by the extension, it will add the domain of the content being blocked to a list. This list will grow over time, meaning that the list of blocked domain will keep adding to the list. The list is then used for future reference by the extension in order to block unwanted content. In other words, if the domain in the list matches a content domain being displayed on the website, the extension will restrict this content from being displayed.

In order to make sure that the model produces accurate results, pre-processing is performed on the data. This includes feature extraction.

3.3.5 Pre-processing

The process of pre-processing data is an integral step in machine learning that has to be taken. This is mainly because, the data has to be purified in order for the machine learning model to learn better. Thus, for the machine learning to predict accurate results the data has to be pre-processed. The data can be pre-processed in many ways such as feature extraction, standardizing data and normalising data.

3.3.5.1 Feature Extraction

Feature extraction is the process of automatically selecting the features in the data set that will contribute the most to the prediction value. I.e. selecting the most relevant attributes. Using feature selection in the algorithm will have many benefits. Such as reducing overfitting, improving accuracy and reduces training time.

3.3.5.2 Benefits of feature extraction

- Reduce overfitting
  Overfitting is when the training data used in the model negatively impacts the performance towards its new data. This means that the model will learn noise patterns or any random fluctuations from the training data as demonstrated on Figure 7. The problem is that news noise pattern and random fluctuations are irrelevant to the new data, impacting the new data negatively.

  By having less redundant data means having less opportunity to make decisions based on noise. In addition, less redundant data mean that the algorithm will run much more efficiently.

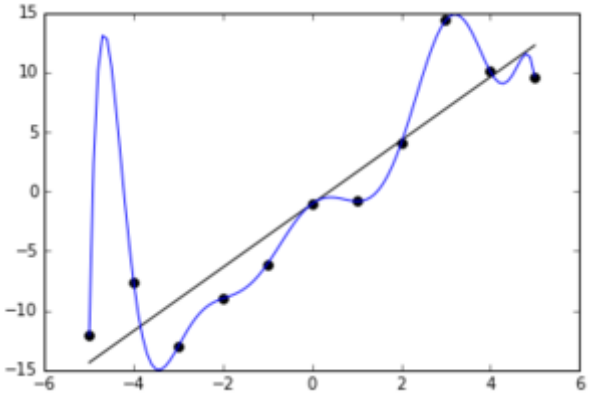


Figure 7 A graph plotted using over fitted data [7]

- Improves accuracy
  As mentioned previously on the report, having a higher accuracy means that the model will produce better predictions. Hence with having less miss-

leading data, it means that the algorithm will be able to have a better accuracy. I.e. modelling accuracy improving.

- Reduces training time
  Training time is the time take for the model to learn from the data it's provided. The training time can vary for each model and depending on the data being provided. As stated above, feature extraction will remove any redundant data and remove any mis-leading data. This means that the algorithm will train on less data than expected, which will improve the time it takes to train the machine learning algorithm.

3.3.5.3 Standard Scaler

Standardizing the dataset means rescaling the distribution of values, this is so that the mean of observed values is 0 and the standard deviation is 1. This can be looked at in a way where it is subtracting the mean value or centering the data.

3.3.5.4 Normalising Data

Normalisation of data is rescaling the data from its original range. This is so that all values are within the range of 0 and 1. This type of data processing can be useful when the time series data has input values with differing scales.

3.3.5.5 Min Max

Min Max, in other words scaling data to fit between a given minimum and maximum value. Usually, this means that data is fitted between 0 and 1.

4.0 Implementation

4.1 Browser Automation & Collecting HTTP requests

Firstly, in order to collect the cookies, the project collects HTTP requests from a list of websites. To pursue with this step of the process, Selenium and Mitmproxy is used. As previously stated on the report, selenium is used to automate the browser. This means that, using selenium, the browser will automatically visit a list of given URLS without any human interaction. Selenium will access Alexas top 250 visited websites.

With the use of Mitmproxy, HTTP requests are being collected. This means, as selenium access the list of websites, Mitmproxy will record all the HTTP requests made to the server by the client when accessing these websites and records the requests made to a list. For example, when accessing a website, the website is likely to make multiple HTTP requests. As said, these HTTP requests will be listed in a CSV to later be accessed.

4.2 Cookie Collection

Secondly, after getting the HTTP requests via MITMproxy, these HTTP requests are then queried. This is to retract the cookie jar that this specific HTTP request holds. For this process, python and its request library is used. Using python, the request to the HTTP is made. This will return the cookiejar belonging to that HTTP.

Cookie data is being collected in order to feed the data into the machine learning algorithm, the algorithm will then justify whether these are tracking or non-tracking cookies. Figure 8 shows the cookie jar requested from the website, www.youtube.com. As the figure shows, once requested the cookie jar, it displays all the cookies which are being used on that particular URL.

Below list the properties that the cookie jar returns.

- URL
  To whom the cookie belongs to
- Value
  Cookie value
- Domain
  The specific URL the cookie taken from
- Path
  Location in the website directory
- Secure

This justifies whether the cookie is safe or not

- Expires

  HTTP date time stamp of the cookie

```
<RequestsCookieJar[Cookie(version=0, name='GPS', value='1', port=None, port_specified=False, domain=
'.youtube.com', domain_specified=True, domain_initial_dot=True, path='/', path_specified=True, secur
e=False, expires=1555682940, discard=False, comment=None, comment_url=None, rest={}, rfc2109=False),
 Cookie(version=0, name='VISITOR_INFO1_LIVE', value='DDeDCWKBY7s', port=None, port_specified=False,
domain='.youtube.com', domain_specified=True, domain_initial_dot=True, path='/', path_specified=True
, secure=False, expires=1571233140, discard=False, comment=None, comment_url=None, rest={'httponly':
 None}, rfc2109=False), Cookie(version=0, name='YSC', value='KEKQ0x4hLVU', port=None, port_specified
=False, domain='.youtube.com', domain_specified=True, domain_initial_dot=True, path='/', path_specif
ied=True, secure=False, expires=None, discard=True, comment=None, comment_url=None, rest={'httponly'
: None}, rfc2109=False)]>
```

Figure 8 shows the cookie jar retracted from 'www.youtube.com'

From the received cookie jar, cookie properties will then be extracted. The properties extracted are URL, secure, path, path_specified, expires, discard, port, port_specified as listed on Figure 9. Once the properties are extracted, they are written into a CSV.

4.3 Machine learning

Finally, as mentioned in the background research, a machine learning algorithm is used to classify which cookies are classed as tracking. To begin with, there are two machine learning algorithms used. As stated previously, two models are being used to measure the accuracy score as well as the loss function from each mode. The model that gives the best accuracy and the lowest loss function is the best model as it will produce accurate results.

Figure 9 List of training data for machine learning, extracted from the cookie jar

Url	secure	path	path_spec	expires	discard	port	port_specified	Tracking
mktoresp.com	FALSE	/	TRUE		TRUE		FALSE	0
.mailerlite.com	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.evergage.com	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	0
mystats.in	TRUE	/	TRUE		TRUE		FALSE	1
.octonet.com	TRUE	/	TRUE	1.59E+09	FALSE		FALSE	1
officemart.ru	FALSE	/	TRUE		TRUE		FALSE	1
.onstate.co.uk	TRUE	/	TRUE	1.59E+09	FALSE		FALSE	0
onstate.co.uk	FALSE	/	TRUE	1.56E+09	FALSE		FALSE	1
.google.com	FALSE	/	TRUE	1.56E+09	FALSE		FALSE	0
.google.com	FALSE	/	TRUE	1.57E+09	FALSE		FALSE	1
.ooyala.com	FALSE	/	FALSE		TRUE		FALSE	0
.pardot.com	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.episerver.com	TRUE	/	TRUE	1.59E+09	FALSE		FALSE	1

As stated above, training data is used in the model for it to learn better about the patterns and relationships in the data. For this process, there are two lists being used provided by the ‘block list project’. *[8]* From the ‘block list project’, there are two lists used. Both lists are a list of URL being captured using browser based plugins. The first list is of all domians which are blocked by adblocker

and ghostery. This is known as a combined list which contains URLs related to Ads, Crypto, Drug related, Malware, pornographic content as well as tracking.

The second list includes a list of tracking URLs captured by browser based plugins. Using python, it evaluates whether the tracking domains are included in the combined list of domains. If so a binary value (0,1) will be added next to the corresponding URL. The binary value 0 will indicate that the URL is not tracking, i.e. not being included on the tracking list of URLs. Whereas 1, will represent the link as tracking. This means that this particular URL is included on the tracking list of URLs.

As discussed earlier, pre-processing is conducted on the data in order to improve the models accuracy score and the loss function. One of the pre-processing steps done is feature extraction.

4.3.1 Feature Extraction

By viewing the data, one distinct way of feature selection is by removing any redundant data columns. Redundant data means data that is not unnecessary. This might be because the data is the same through the column as does not change in anyway. As Figure 9 displays, there are several columns where redundant data is presented. The columns that provide redundant data are Url, path, port and port_specified. As the table in Figure 9 clearly show these fields are unnecessary when training. The column port is empty throughout the list of URLs, this means that by using this column towards training the model will nor increase predictions or decrease predictions.

The columns, path and port_specified display the same data. This means that these columns will be redundant as well. By removing these columns it will improve the model. This is because the model will have use less data to train with. Hence, increasing the time taken for the model to train. The final column of data that is removed from the list of training data is the URL column. This is mainly because the values held in this column is a string. Moreover, the string changes constantly; meaning that this column will also be classed as redundant data.

For the training data being used for this project, feature selection will be demonstrated on these columns justified above.

4.3.2 Normalisation, Standardization & MinMax

After feature extraction has been performed on the dataset, normalisation and standardization will be performed on the same data set. The normalised data will be used by the SVM and the MLPC along with the standardized dataset.

The final pre-processing step used is MinMax. As stated above, this is to scale features to lie between the minimum and the maximum value.

4.3.2 Training the model

To train the model, the data extracted from the cookie jar is used. However, when training the model, it cannot accept strings. Thus, these will have to be converted into numerical values. These will be secure, path_specified, discard and port_specified. These fields hold Boolean values i.e. True or False. These values are then converted to binary values i.e. (1,0).
The data as displayed on Figure 9, shows nine fields. However, some of the fields on the data list are redundant. This means that these fields can be pre-processed.

5.0 Data analysis

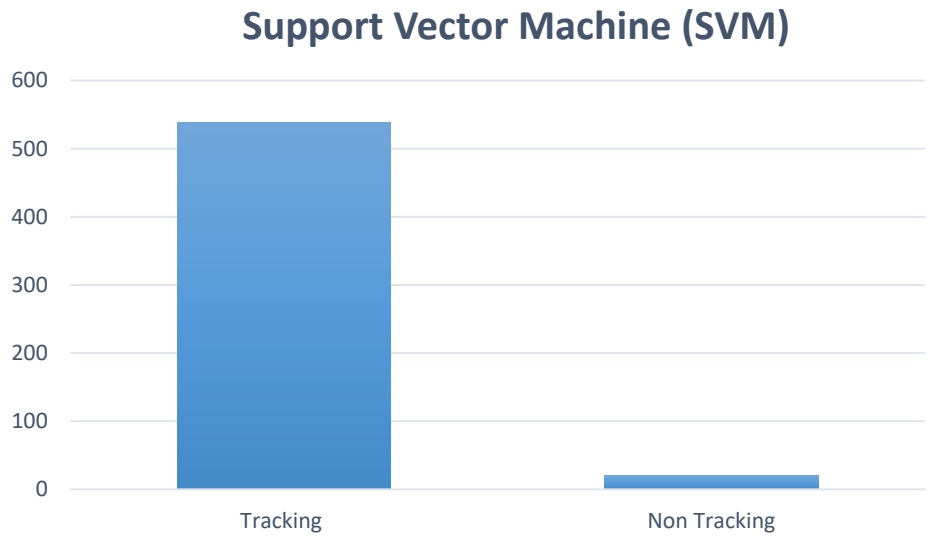
As mentioned previously on the report, the project uses two different models to test the data (SVM, MLPC). Each model was tested three times. The first test is with minimal pre-processing, second test is with normalised data and the third test is with standardised data.

Firstly, the study is carried out on the top 250 website list produced by alexa. When using the request API to gather cookies from 250 websites, a total of 558 cookies were extracted from list of websites.

Secondly, as previously mentioned in the implementation, the project cross referenced a tracking list of domains against a combined list of domains. However, the number of values in that list accumulated to over 30,000. This meant that in order to get cookie data, over 30,000 cookie requests need to be made. This means that the time taken to request over 30,000 cookies will be very significant. In order to improve the time take to make the requests, the dataset was reduced to 2000 requests. This meant that only 2000 cookie requests will be made. Reducing the data is a key aspect as it will improve the time taken to get all the valid cookie data.

5.1 SVM

Figure 10 Results from the SVM model expressed in a bar chart



The results gathered from these tests were surprising. This is mainly because SVM (Support Vector Machine) only provided results for the data set with minimal pre-processing applied to it. The results gathered back from the algorithm indicated that 4% of the cookies analysed out of 558 were non-tracking. The results indicate that

most of the URLs contained tracking cookies and hardly any URLs contained non-tracking cookies as indicated on Figure 10.

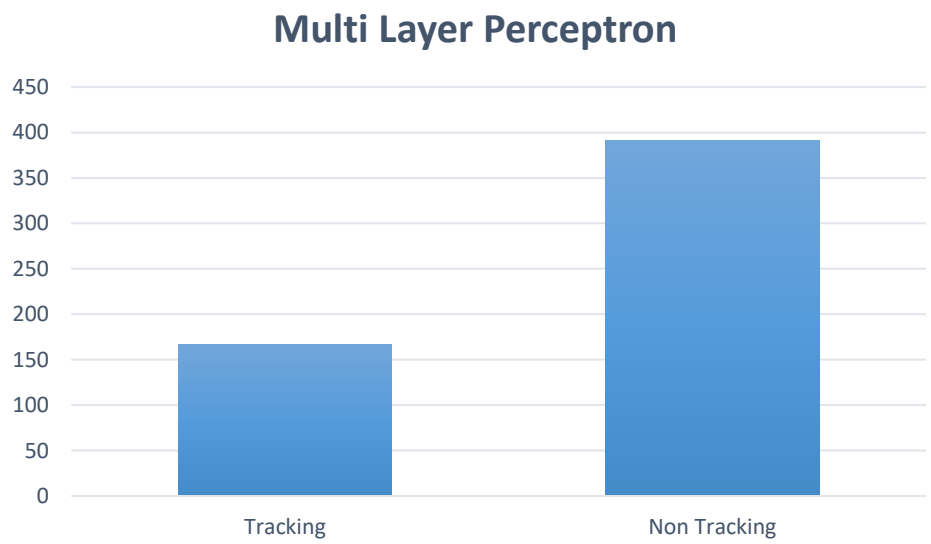
In comparison to other SVM model results, these results are slightly realistic. This is mainly due to the fact that other SVM models did not produce great results. A total of three SVM models were tested, two out of three indicated that all 558 cookies were tracking. One SVM model was using the normalised dataset whereas the other was using the standardized data set. However, both predicted the same result.

5.2 MLPC

The results gathered from the MLPC (Multi-Layer Perceptron Classifier) vary between the three models.

The first model used a dataset with minimal pre-processing. The results gathered from this model indicated that 30% of the 558 cookies are tracking. A significant amount of cookies were classed as non-tracking, approximately 391 cookies as expressed on Figure 11.

Figure 11 Results from the first MLPC model



The second MLPC model recorded similar results. It suggested that out of 558 cookies, 74% were non-tracking. This MLPC model uses the normalised dataset. Few cookies were classed as tracking when using this model in comparison to the first MLPC model. The data gathered from this model is expressed in Figure 12 .

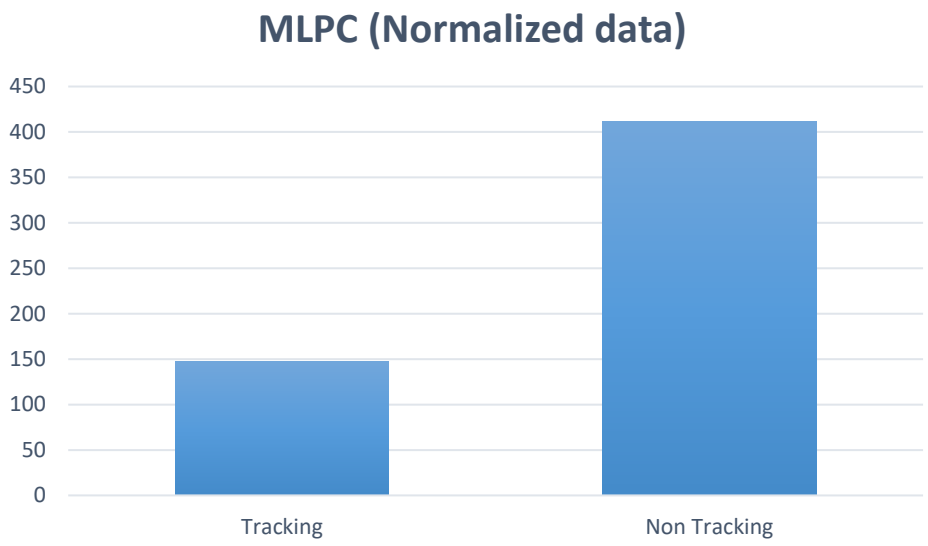


Figure 12 Results show only 26% of cookies classed as tracking

Having said that, the third MLPC model, which uses standardized data argues that over 85% of the cookies assessed are tracking cookies. Whilst, the remaining 15% are classed as non-tracking. The results gathered from the third model are significantly different compared to the other two models.

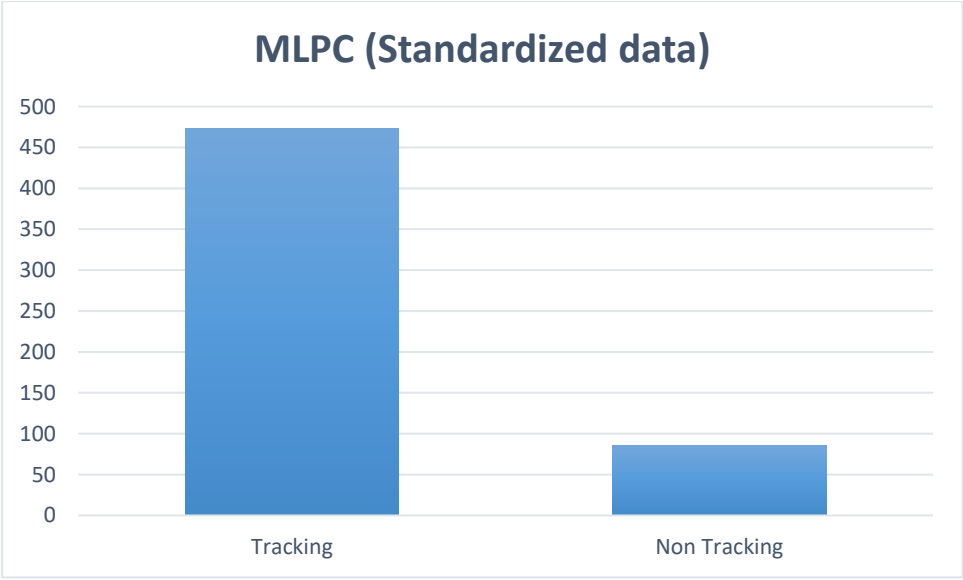


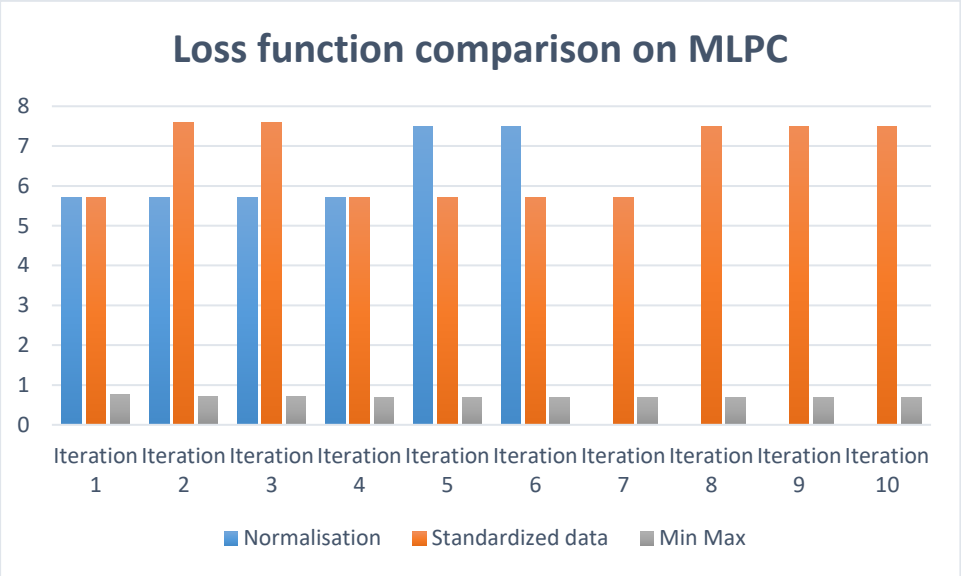
Figure 13 Third MLPC model results

As Figure 13 demonstrates, this is a significant difference compared to results produced by other MLPCs.

The final step of pre-processing is Min Max. Min Max was applied to the dataset and was tested on the two models. The results gathered from both models did not differ much in comparison to previous results. The predictions with the SVM model said that 4% of the 558 cookies were not tracking. This prediction is identical to the prediction made on the first SVM test.

Moreover, when using Min Max on the MLPC dataset the results were yet again disappointing. The MLPC model recorded the results with all the cookies known to be tracking. However, the loss function on that particular test was recorded to be 0.69 (the loweset). Figure 14 demonstrates the loss functions for each MLPC test. As the chart demonstrates, using Min Max had the best loss function. As stated in the background research section, having a lower loss function is better as it means that the algorithm is modelling the data well. Figure 14 demonstrates three out of the fours tests performed. The loss function was not generated for the first test.

Figure 14 this chart lists the loss function comparisons of each MLPC test



5.3 Grid Search

As SVM produced very disappointing results, a grid search was conducted in order to find its best hyper parameters. As previously stated, conducting a grid search will return the best parameters that could improve the models predictions.

After conducting the grid search, the best hyper parameter suggestion it produced was setting the 'gamma' value of the SVC to 1. The 'gamma' value is known to be the decision region. For example, when the gamma is low, the decision region is very broad. Whereas, if the gamma value is high, the decision boundary is also high. This creates massive decision-boundaries around data points.

Figure 15 SVM with the normalised dataset, with gamma value set to 1

When setting the 'gamma' value to 1 in the SVC models, some difference in the predictions were found. For example, the SVC model using normalised pre-processing data predicted that all the cookies were tracking before setting the gamma to 1. However, after setting the 'gamma', the predictions from the model improved. The results show 4% of the cookies assessed are known to be tracking as demonstrated in Figure 15. Though this model predicts similarly to the first SVM model tested, it is still considered an improvement. This is because it is making a better prediction compared to before when the 'gamma' was set to default.

Setting the 'gamma' to 1 on the SVM which uses the Min Max pre-processed data also had better predictions compared to last time. Before setting the gamma, it predicted that 4% on the cookies are non-tracking. After setting the gamma, the model predicted that 8.6% of the cookies are non-tracking. This meant that 91.4% are classed as tracking as displayed Figure 16. Even though that the percentage of tracking cookies are still very high, it is a better result compared to 96%.

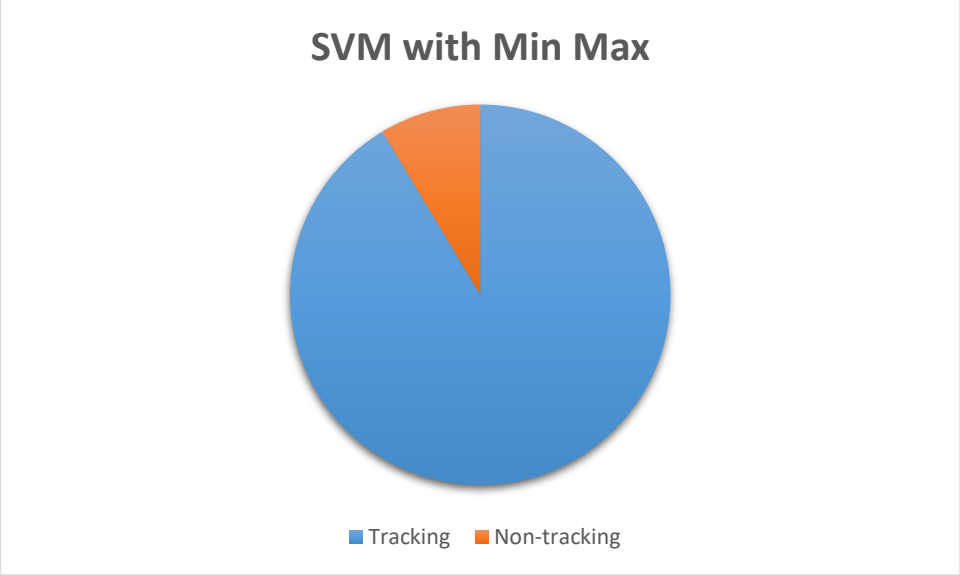


Figure 16 Gamma set to 1; SVM with Min Max processed data

6.0 Evaluation

Taking into account all the data being collected, the more realistic predictions are made by the first two MLPC models. As said before, one of the models only use feature extraction as a pre-processing method, i.e. minimal pre-processing whereas the second model uses feature extraction as well as data normalisation. Both models predicted similar results. On the other hand, the first SVM model and the third MLPC predicted similar results. Both models indicated that the majority of the cookies are tracking. Having said that, in comparison to the third MLPC model and the first SVM model, it shows that the MLPC classifies a larger number of cookies are non-tracking as displayed on Figure 17. This means that out of all the model which were successful to produce results, SVM is the only model to classify a larger number of cookies as tracking. It predicted that 538 of the cookies evaluated were tracking.

After conducting tests with the Min Max pre-processing step, the predictions did not improve as much. In other words, the models were predicting the same values other models already had predicted.

On the other hand, when the grid search was conducted to find the better hyper parameters, two SVC models significantly improved. This being said the model was making better predictions when the 'gamma' was changed to 1.

Referring back to the background research conducted, it says that most websites are embedded with third party trackers to monitor user behaviour; whether it is to measure a websites metrics using google analytics or to measure a user online shop. Third party trackers will be embedded onto websites, this is that the website can better cater for its users as mentioned above. If this is the case, then SVM produced very accurate results, identifying most of the cookies as tracking. This makes it clear that most cookies used by websites are tracking user behaviour.

On the other hand, the predictions made by the model may be incorrect. This could possibly lead to the training data being scaled down. As said in the analysis, the training data took a significant amount of time to request the cookies from the URLs. Thus, it was scaled down to a 2000. However, out of the 2000 requests only 99 URLs had valid cookies. This means that some of the URLs listed were either empty or couldn't connected to the server. Which meant that python overlooked at these URLs and only collected the cookies from URLs which contained actual values in the cookie jar.

More to the point, the machine learning algorithms were trained on minimal data. This is likely to produce inaccurate predictions. As stated previously, machine learning algorithms are very data hungry, meaning that it needs a large amount of data to understand and learn the patterns and relationships between the data. Without having a

significant amount of data to learn from, the algorithm is likely to produce inaccurate predictions.

Looking back at the results gathered from the models, it is hard to say that it produced accurate results due to the lack of training data it used to learn from.

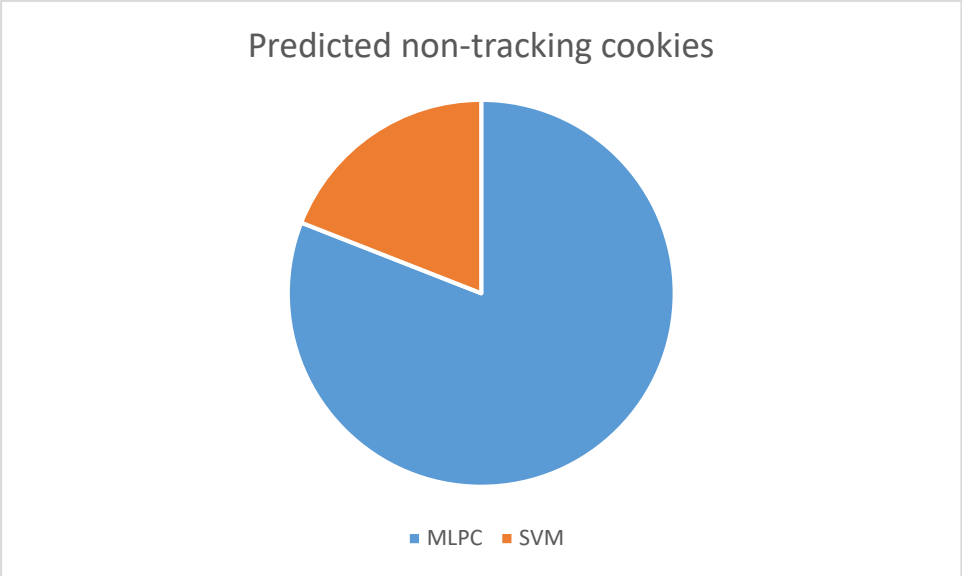


Figure 17 Number of non-tracking cookies predicted by SVM and MLPC

Taking into account all the results gathered, the predictions made by the results are not accurate. This is mainly as all the models have very different predictions and some models were not able to predict any values. Some predictions made from the models differed a lot from the others or through each iteration. This is highly related to the lack of training data being used to train the algorithm.

7.0 Conclusion

To conclude, there are several alterations that can be made towards this project. This being said however, the projects goal was to determine whether cookie were tracking or non-tracking and the results gathered from the machine learning clearly justified that some tracking cookies were found in the top 250 websites visited. Although the results could have been improved with better quantitative training data.

7.1 Current implementation improvements

As discussed before, one of the biggest flaws of the project was python. This is mainly due to the time the request function takes. When analyzing a large sum of data it is essential that the run time is less. However, with the python request API, the requests being made to URLs took a significant amount of time. That being said, the algorithm used to collect the data can also be optimized to increase performance.

Moreover, implementation of machine learning. Given more time, the project look towards more ways to classify data with the use of machine learning. This means testing other machine learning models.

7.2 Future work

This project mainly focuses on cookies being saved to a computer. Nevertheless, it would be interesting to find out how tracking/non-tracking cookies work when in contact with a smart devices, i.e. smart phones and tablets. As the number of smart devices grow over time, this would be a interesting study to pursue. As stressed above in the report, in order to predict better results the machine learning algorithm needs to have more quantitative data. By having quantitative data, the algorithm will be able to understand the patterns of data better and any relationships in the data. Thus, making better predictions.

7.3 Final thoughts

Overall, as mentioned before, the project itself has been successful. It was able to justify how many cookies were tracking and non-tracking. Thus, achieving the goal of the project. Nevertheless, working on this project allowed me to enhance my skills as well as learning how to use valuable tools which may benefit me in the future.

8.0 References

[1] [Online]. Available: https://aldro61.github.io/microbiome-summer-school-2017/figures/figure.classification.vs.regression.png. [Accessed 5 5 2019].(IMAGE)

[2] D. Misener, "Browser cookies: How they could be undermining your privacy," CBC, 8 April 2014. [Online]. Available: https://www.cbc.ca/news/technology/browser-cookies-how-they-could-be-undermining-your-privacy-1.2602095. [Accessed 3 May 2019].

[3] S. E. C. E. P. Z. a. A. N. D Reisman, Cookie that give you away: Evaluating the surveillance implications of web tracking, Princeton NJ USA: Princeton University, 2014.

[4] S. Englehardt and A. Narayanan, Online Tracking: A 1-million-site Measurement and Analysis, Princeton University, 2016.

[5] J. V. A. N. a. M. C. Arunesh Mathur, Characterizing the use of Browser-Based Blocking Extensions To Prevent Online Tracking, Princeton University, 2018.

[6] Amazon, "Alexa," Amazon, 2019. [Online]. Available: https://www.alexa.com/topsites. [Accessed 04 05 2019].

[7] [Online]. Available: https://media.springernature.com/original/springer-static/image/chp%3A10.1007%2F978-1-4842-3564-5_1/MediaObjects/463052_1_En_1_Fig1_HTML.jpg. [Accessed 5 5 2019].(IMAGE)

[8] "Block List Project," 2019. [Online]. Available: https://tspprs.com/. [Accessed 3 5 2019].(IMAGE)

[9] J. R. M. a. J. C. Mitchell, Third-Party Web Tracking: Policy and Technology, Stanford, CA: Stanford University, 2012.

[10] [Online]. Available: http://1.bp.blogspot.com/-1E9sXGo8DfI/UU5HjOOvCLI/AAAAAAAAAJI/ufdT3mAsqdQ/s400/h.png.(IMAGE)

[11] [Online]. Available: https://res.cloudinary.com/di2vaxvhl/image/upload/v1545849277/HTTP_txch7g.png. [Accessed 5 5 2019].(IMAGE)

9.0 Appendix

9.1 Appendix A

Project Proposal

Aims and objectives

The aim of this project is to investigate how cookies behave on different websites. This project will focus on how a users activity will be tracked within the website. The project will also focus on different types of cookies from different websites. It will assess the information the cookie takes in and evaluate what is done to the information.

Moreover, I will be creating an application in order to measure the information processed by the cookies on different websites.

Relevance

For my course, Computer Science, this is relevant as it goes into detail on how most website act with user data. Personally, this is a very interesting topic for me as I am a very keen web developer. I believe this project will allow me to better understand how web tracking works.

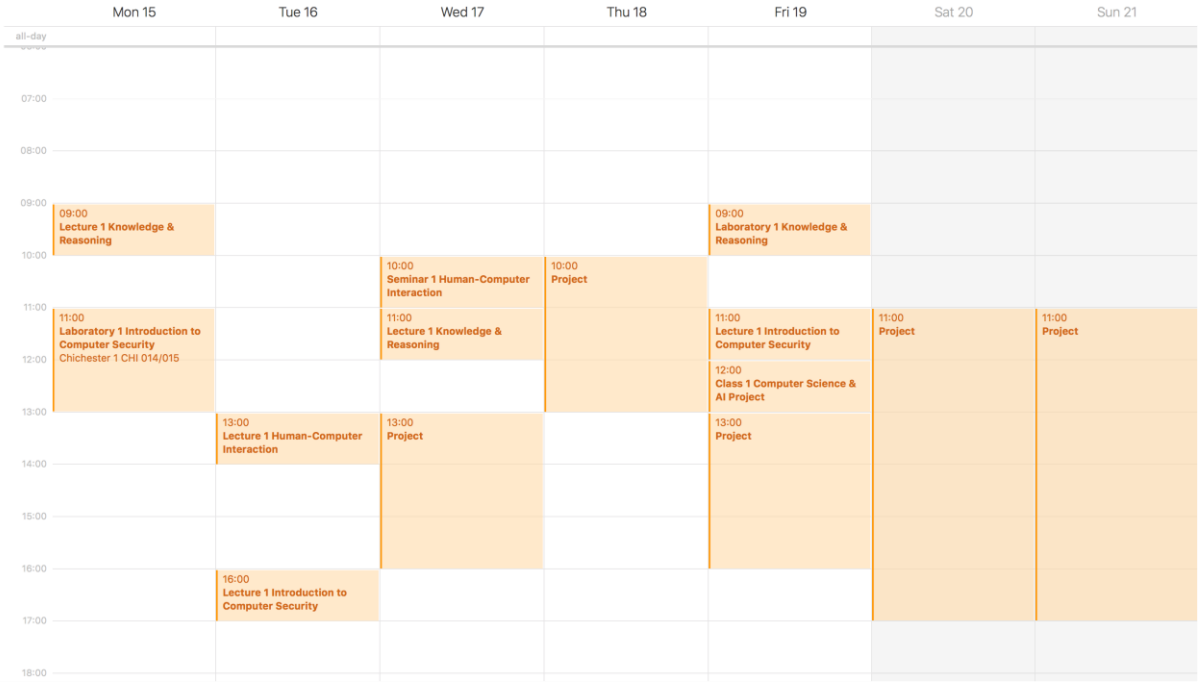
Furthermore, considering that data privacy is an issue of high importance at the present time, this project will be very interesting. There will be many day to day issues which I will be able to talk about when writing the report.

Resources Required

The project will be written in either Haskell or python, thus I will be needing the IDE for python and the compiler for Haskell. These have already been downloaded and installed to my personal computer.

In order to conduct research in regards to my project, I will be needing reading material. I believe that the reading material will provide me with more information on the aspects of web tracking. Also, providing me with information in regards to the implementation of the project.

Timetable



Looking at my timetable, I currently have 10 hours of contact time for my modules. For my project, my aim is to spend between 15 to 20 hours working on my project.

Background reading

- When cookies meets the blockchain: privacy risk of web payments via cryptocurrencies – Steven Goldfeder, Harry Kaldoner, Dillon Reisman, Arvind Narayan – August 2017
- Who left open the cookie jar? A comprehensive evaluation of third-party cookie policies – Gertjan Franken, Tom Van Goethem and Wouter Joosen – August 2018
- Characterizing the use of Browser-based blocking extensions to prevent online tracking – Aruneth Mathur, Jessica Vitak, Arvind Narayanan, Marshini Chetty

9.2 Appendix B

Meeting Log

Skype call meeting

Date: 28/8/2018

Duration: 10mins

Minutes

This meeting was mainly to discuss my intentions towards the project and get initial feedback from the project supervisor. We spoke about how to progress through the project and what applications can be used to investigate.

Meeting 1

Date: 24/9/2018

Duration: 20mins

Minutes

Basically discussed what the project is briefly.

I asked whether I should build a bot for my project in order to investigate the cookies.

He advised me on how it is better to have regular meeting and to keep on top on the workload. Also, told me to assess the behaviour of a browser – use headlist chrome. Told me to looking into supervised machine learning to assess the cookie behaviour and to see how and what information is gathered by the cookies.

Told me to have a look at a variety of cookies. For example, good cookies and bad cookies. This means looking into cookies gathered by the BBC, telegraph (i.e. professional website, good cookies) and bad website (i.e. bad cookies).

Told me to make a setup to get cookies, rather than having to get cookies manually.

Advised to learn a new language, either Haskell or scarlar.

Told me to look into git – githook. (githook pipeline for presentation – for testing).

Meeting 2

Date: 18/09/2018

Duration: 20mins

Minutes

In this meeting, we spoke about Haskell and my difficulties with it. Martin explained the Haskell functions to me, the inputs and the outputs.

I asked about my project proposal, in regards to adding to it and changing the project proposal. He said not to worry about the project proposal just yet and that it was fine.

Meeting 3

Date: 30/10/2018

Duration: 10mins

Minutes

This meeting was mainly to discuss my current progress with the project. In the meeting we also discussed about the interim report.

In addition, we spoke about my progress with Haskell. It was decided that the better option was to move further with my project using python rather than Haskell.

Meeting 4

Date: 25/2/2019

Duration: 10mins

Minutes

This meeting was about the update of my project. To talk about the current status of my project. Due to exams and other coursework, not much progress has been made towards the project.

Meeting 5

Date: 18/3/2019

Duration: 10mins

Minutes

This meeting was in regards to using MITMproxy in my project and how to collect cookies.

Meeting 6

Date: 25/2/2019

Duration: 15mins

Minutes

This meeting was a skype call with Martin, in regards to the final steps of my project. I.e machine learning aspect of the project. Martin advised that I should look for training data gathered from browser plugins in order to train my algorithm.

9.3 Appendix C

Raw collected data

Url	secure	path	path_spec	expires	discard	port	port_spec	tracking
.google.cc	FALSE	/	TRUE	1.56E+09	FALSE		FALSE	1
.google.cc	FALSE	/	TRUE	1.57E+09	FALSE		FALSE	1
.statcount	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.hugedom	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.9gag.com	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.adtracker	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
stats.imm	TRUE	/	TRUE		TRUE		FALSE	1
.adtracker	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	0
.iplogger.(	TRUE	/	TRUE	1.56E+09	FALSE		FALSE	1
iplogger.c	FALSE	/	TRUE		TRUE		FALSE	1
.bravenet	FALSE	/	TRUE	1.56E+09	FALSE		FALSE	1
.bravenet	FALSE	/	TRUE	1.87E+09	FALSE		FALSE	1
.soundanc	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.luckyorar	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
stat.syner	FALSE	/	TRUE		TRUE		FALSE	1
.atdmt.co	FALSE	/	TRUE	1.62E+09	FALSE		FALSE	1
.geocount	FALSE	/	TRUE	1.58E+09	FALSE		FALSE	1
217.160.0.	FALSE	/	TRUE		TRUE		FALSE	1
.soundanc	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.hugedom	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.casinotro	FALSE	/	TRUE		TRUE		FALSE	1
stats.berk	TRUE	/	TRUE		TRUE		FALSE	1
.webcoun	FALSE	/	TRUE	1.58E+09	FALSE		FALSE	1
.www.hub	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.www.hub	FALSE	/	TRUE		TRUE		FALSE	1
.ioffer.cor	FALSE	/	TRUE	1.59E+09	FALSE		FALSE	1
.ioffer.cor	FALSE	/	TRUE	1.56E+09	FALSE		FALSE	1