GDPR Complicity: A Study of Cookies in the Legislated Internet

A.K.J.G. de Wit - *Department of Data Science and Knowledge Engineering*
Maastricht University
Maastricht, The Netherlands

Abstract—This thesis aims to analyse what kind of cookies are saved to a users browser before and after consenting to them trough the cookie banner. This goal is achieved by writing a webcrawler that visited a high amount of websites and interacted with the banner to get different cookies which could later be analysed. Using a cookie database and the zxcvbn password strenght estimator the type of each cookie could be determined and whether it contained personal information. From this analysis it was found that 90% of the websites analysed did provide a banner, that 19.5% of cookies saved in a users browser when visitng a website are related to targeting and advertising, and 35.67% of these cookies contain personally identifiable information. It was therefore concluded that most websites do provide a banner to communicate one's choice, and that this choice does in fact influence the cookies that the websites save. This influence is however smaller than expected as many websites do not ask for active consent to start storing personal information and targeting cookies in a users browser.

Index Terms—Security and Privacy Protection; Browser Cookies; GDPR

1 Introduction

Since the introduction of cookies in 1994 they evolved as one of the most important methods of storing and tracking users throughout the internet. ”It was a turning point in the history of computing: at a stroke, cookies changed the Web from a place of discontinuous visits into a rich environment in which to shop, to play – even, for some people, to live.” [1]

Earlier studies have shown that nowadays over 90% of the highest traffic websites utilize cookies to track users across the internet [2]. This long-term tracking of users, and the monetization of it, prompted the European Union to issue a big change in the privacy policies. In essence, when cookies can identify individuals trough their devices, it is considered personal data. Following the implementation of the GDPR, a website is obliged to give individuals the freedom of choice over the amount of personal information it can store in cookies [3]. This choice is usually presented in the form of a banner or pop-up, forcing the user to state their preference before entering the website. Previous studies have shown that even after passing the new legislation, more than 90% of researched websites create tracking cookies immediately when visiting the website [4], confirming the results of the 2016 study mentioned earlier.

In a 2019 study study, different cookie notices were presented to over 80.000 unique visitors of a German website, allowing developers to study and learn what type of banner makes a user more inclined to accept all cookies on a website. [5]

With the eye on increasing numbers of web developers and advertisers that manipulate the user to accept all cookies in order to visit a website [5], [6], this paper will focus on the options that cookie banners provide and whether the choices given are reflected in the cookie saved to a users device.

The existing solutions to websites using cookies to track a user on the web are several browser plugins with different functionalities. The first is 'I don't care about cookies' which finds the pop-ups or banners and uses CSS and javascript to simply hide them from the user. There are however a lot of websites that, against the GDPR law, assume consent when a user continues using the website and hence still save a lot of data of the user. The second solution is one that most browsers have built in, which is to delete all private data, including cookies, when exiting the application. This seems to give some privacy back to the user, making it unable for websites to track how often the user visits nor create a fingerprint based on the users personal information, but even to this solution there are downfalls. The most noticeable one is the fact that it will remove all saved logins from websites, forcing the user to log-in to every website they have an account on, every time after they close their browser. It is also important to keep in mind that deleting cookies does not prevent all data tracking. Many websites can reconstruct cookies unless the user deletes their cache and browser history as well, and even then companies can still fingerprint users based on their device, IP address, screen resolution, and plugins.

To answer the question of how many and what types of cookies websites utilize, an automated analysis on over 9671 websites was performed. Each site was visited, scanned for a cookie banner, and if there was one the program would interact with it while monitoring the cookies that are saved by the website. This created a large database of information about what kind of cookies a website stores after the user

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Knowledge Engineering, Maastricht University. Supervisor(s): A. Zarras, J. Spanakis has indicated some sort of preference, and the aim of this study is to show what influence these preferences have on the cookies utilized by websites in general.

The results show that cookies are still widely used on the internet and that more than 10% of the websites analysed did not even provide a banner for the user to consent to some form of the cookies the website used. For the other 90% that provide options, it is either difficult to deny or if the user does not even consent but just browse the website, 20% of the cookies saved in their browser are related to targeting and advertising.

In summary, the main contributions of this paper are the following:

- A global analysis is performed on more than 9000 websites.
- The results show that a surprising high number of cookies that are saved on visit are related to tracking and advertising, against European laws.
- The results show that different options do in fact modify the amount of personal information saved in cookies, but its influence leaves a lot to be desired.

2 Background

2.1 HTML Cookies

Cookies [7] were first created in 1994 to enable stateful browsing over the stateless HTML protocol. This meant storing small pieces of data created by a website in a users browsers. The data could include information about user login, preferences, shopping carts and other complex web interactions that span multiple HTML pages. Each cookie has predefined set of information. It is characterized by a URL it is associated with, a name, a value, and and expiration date. Whenever a user requests a HTML page, the cookies with the corresponding URL are sent to the web-server issuing the web-page.

After the introduction of this technique to enable stateful navigation companies discovered that cookies are also a very efficient method to track users across different websites. This is most often for advertising and analytic purposes. To achieve this, websites that want to display advertisements and make money from that allow third-party domains to put uniquely identifiable cookies in the users browser, allowing these advertisement providers to track users across multiple websites that issue their cookies. This is how they are able to reconstruct a users online activity.

2.2 GDPR

The General Data Protection Regulation (GDPR) is a European law, implemented on May 25, 2018, that governs all collection and processing of personal data from individuals inside the EU [8]. The goal is *"to strengthen individuals' fundamental rights in the digital age and facilitate business by clarifying rules for companies and public bodies in the digital single market"* [9]. This legislation is regarded to have a major global impact as websites that collect or process data from users inside the EU will also have to comply with the GPDR. Cookie consent is indirectly one of the most important parts of the GPDR because, as mentioned earlier, the most common way to collect user data is trough cookies.

"Natural persons may be associated with online identifiers [...] such as internet protocol addresses, cookie identifiers or other identifiers." [3]

With this new law in place individuals will have control over their data, allowing them a choice and an insight into what is tracked, and whether they want that.

3 Methodology

The goal of this study is to analyse what cookies websites use when visiting the website and after selecting different options. This will be achieved by developing a flexible web crawler in Python which will visit the most popular websites of the internet. Using the Selenium package to visit the websites as a background process using the headless option of the chromedriver [10], the crawler can retrieve data from many websites in a relatively short time. When visiting a website, the crawler will choose whether to accept the cookies or not, and it will compare the information stored in the cookies between the different options. One of the obstacles will be to program the crawler in such a general way that it can visit most websites and interact with the banner for the cookie options. When all of this is achieved, an overview and conclusion is made to answer the questions of this paper, with a substantial amount of data to give relevant insights in the adherence of the GDPR by websites.

3.1 Domain Selection

To decide which websites should be analysed, it is preferable that websites that get a lot of traffic are chosen to show what kind of cookies most users will receive when browsing the internet. Therefore the list from The Majestic Million [11] was chosen which contains the million most visited websites. The crawler and web interaction is computationally heavy which disallowed for analysing every website on the index, however two independent searches, one from top to bottom and one randomly gave results of over 9671 websites, which still makes this research statistically relevant.

3.2 The Web-Crawler

Using the well-documented Selenium [12] python package and a list of the million websites with the most traffic (Majestic Million) a Web-Crawler was coded. After trial and error with several methods, one seemed to work on most of the provided websites. Using the XPath searching capabilities of Selenium, a search is done for any type of element in the web-page containing the word 'cookie' in text. Finding this element would indicate that somewhere on the web-page the user can read the word 'cookie' which indicates some sort of notice regarding the use of cookies on this website. Then iteratively moving up the DOM tree from this element, the crawler looks for sibling elements that have the ability to be clicked. If this is the case, it is safe to assume it has something to do with the cookie notice. A check is performed to see whether the element contains one of the keywords most found on the consent buttons. Another contingency is that the crawler checks whether the element that can be clicked will lead away from the current web-page with a different URL, which is not wanted. If the requirements are met, the crawler tries to click the item and when it detects the cookie has changed it saves the new cookie in a Pandas DataFrame. The crawler can be stopped at any moment and the DataFrame will be updated with the websites it was able to visit.

3.3 Banner Providers

During the research into the cookie banners it became evident that there exist several companies that provide the service of a cookie banner for websites that don't want to bother with creating this temselves. The three major providers found were *'TrustArc', 'OneTrusts', and 'Cybot'*. In order for the crawler to work more efficient and also prevent unneccesary data collection, the crawler would record whenever it found a banner provided by one of these companies and refrain from interacting with it. As found, the cookies created by these banners while selecting different options were always the same (which is to be excpected if it is all handled by the same company), so a manual search for the possible options was performed, giving immedeate insight in the practices of these companies, showing what a user might expect when encountering a banner of one of these companies.

3.4 The Jupyter Notebook Analysis

The description of the DataFrame created and updated by the Web-Crawler can be found in the Table 1 below. Using Jupyter Notebooks this DataFrame is imported to conduct an analysis of the data found. Jupyter Notebooks were chosen for their ease of use and ability to display plots in the code, creating a better overview of the data. All the graphs in this paper are produced with the plotly python package.

Column Name	**Description**
Website	The name of the website
cookie_first	The cookies on first visit
cookie_second	The cookies after accepting
cookie_third	The cookies after denying
banner_provider	One of the three banner providers, otherwise NaN
options_available	Whether the user got options or just an accept button
click_found_accept	Accept button found
click_found_decline	Decline button found

TABLE 1
DataFrame description

A simple analysis of the amount of cookies found in total and per website has been made showing how many cookies a website saves in a users browser with one of the three options.

The next step is to find all the names of the cookies which identify what kind of cookie it is. The identification is done using the database of Cookiepedia [13]. Providing information such as the amount of websites the cookie is found on, whether it is a first- or third-party cookie, the lifespan of the cookie, and whether is is a persistent cookie or a session cookie. All these statistics are collected over the unique names found in the cookies of the crawler database.

The last experiment is a reproduction of a part of another research. In this research the authors used *zxcvbn*, a password strength estimator, to estimate whether the value of a cookie is unique enough to be identifiable information. As many cookies are encrypted by websites, this is a good metric to decide if the value might conntain personal information about the user or whether it is too generic to identify a single person. *"for example, a name that is common enough to be shared by several people and hence not useful to identify a single individual (e.g., "Bob Smith") falls below the* 10^9 *strength threshold, while one that is less likely to have omonyms (e.g., "Robert J. Smith-Johnson") has a score that makes us consider it a unique identifier."* [4]

4 Evaluation

4.1 The Cookies

The first experiments are about the amount of cookies, and what kind of cookies they are. In Figure 1 is shown how many websites the crawler was able to find cookies when visiting the website (*'initial'*), and how many websites it was able to find a difference in cookies after accepting them (*'accepted'*). The third column (*'declined'*) shows how many websites the crawler was able to decline the cookies with just one button. The last column (*'initial & accept'*) shows on how many websites the crawler was able to find both an initial cookie and one after accepting the cookies. The difference between the second and third column shows that some websites do not save any cookie on initial visit and only when accepting the cookies they do. There were 242 websites that showed this behaviour of not using any cookies on visit.

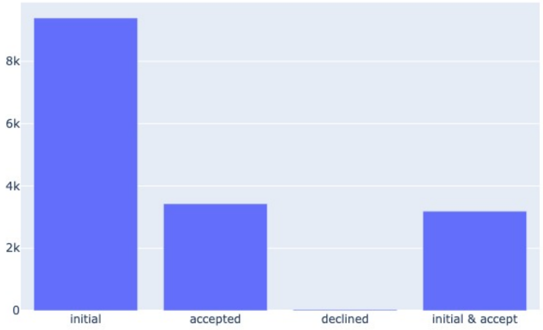

Fig. 1. Amount of cookies found

4.1.1 *Initial cookies*

The first experiment was to analyse what kind of cookies websites implement when a user visits the website without any interaction on the website whatsoever. This experiment shows what websites save from users when they have not yet consented (actively) to anything on the web-page. The first question is how many cookies websites save on initial visit. A small section of the top 25 largest cookie count websites is show in Figure 2.

The top website that saved 60 different cookies on visit was *'clover.com'*. The rounded average amount of cookies

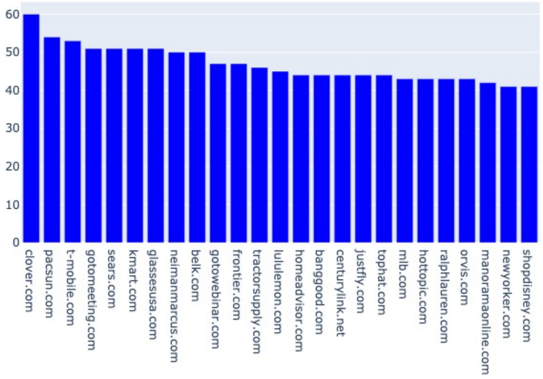

Fig. 2. Top number of cookies on visit

saved on initial visit is 6. The next question is what the type of cookie is that is saved on first visiting a website. The crawler found *59147* names of cookies of which *18367* were unique (*31.05%*). Using the Cookiepedia [13] database to query each unique name found and get the statistics about the type of cookie, an overview of the type of cookies in initial visit can be given as shown in Figure 3.

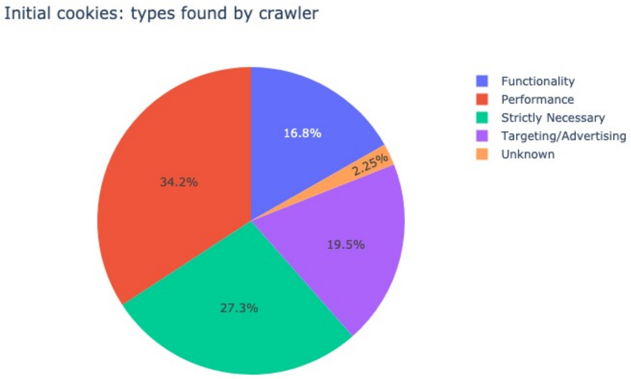

Fig. 3. Percentages of cookie types

Using the identification of Cookiepedia we get a clear overview of the types of cookies that websites save when a user soley visits the website without interaction. It is interesting to notice that already 19.5% of the cookies saved to the users computer when visiting a website are targeting or advertising related, while this is something that under the GDPR websites should ask permission for.

The website that created the most cookies on visit was *'clover.com'* and in Figure 4 we can see the percentages of the types of cookies the website generated. Something that stood out during the analysis of these cookies was that of the 60 cookies, 35 contained the name *'GA Connector'*. This turns out to be a website that integrates both, *Customer Relation Management* part of the Salesforce suite, and *'Google Analytics'*. Their privacy policy reads:

'When you access our websites or use our mobile applications, we, our service providers, and our partners may automatically collect information about you, your computer or mobile device, and activity on our websites or mobile applications.' [14].

However this is not in line with the GDPR where a user has to actively consent to the use of these types of cookies, and continuing to browse a website is not considered 'active' consent.

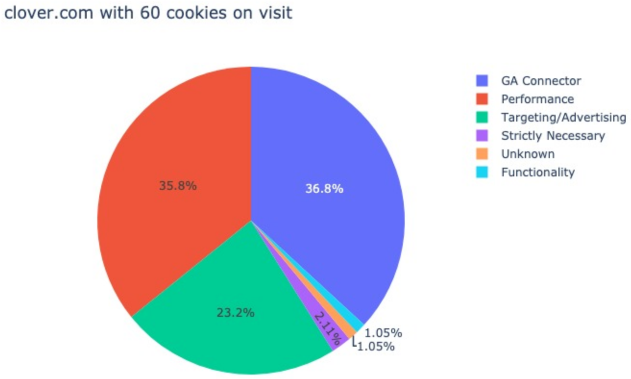

Fig. 4. Cookie types for Clover.com

There were also 484 websites that saved only one cookie on visit. The percentages of the types of these single cookies can be found in Figure 5. Notice here that the majority is necessary or at least for performance or functionality. Still 20.5% , 99 in total, of these cookies are targeting or advertising. The majority of these single cookies related to advertising are Google Analytics '_gat' or the Facebook Pixel '_fbp', indicating their presence even on websites that use just one cookie.

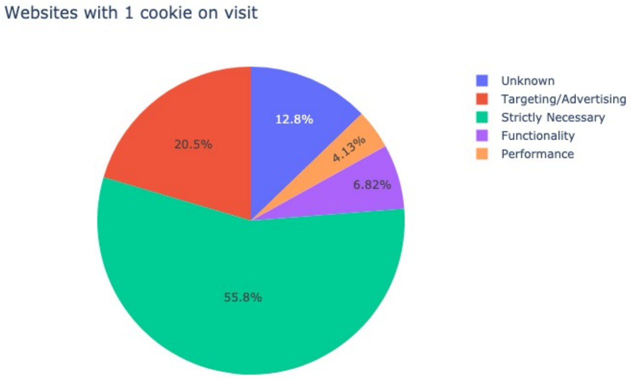

Fig. 5. Cookie types with one cookie on visit

4.1.2 Accepted cookies

The second experiment was to see what cookies websites use when users actively click the accept banner on their website. Previous research has shown that websites try to design their banner in such way that users will probably accept all the cookies the website has to provide. This experiment will show what cookies most users who are not concerned about their privacy will receive on their computer.

The maximum amount of cookies after accepting came once again from the website *'clover.com'* with 4 additional cookies on top of the 60 saved from the initial visit, so 64 in total. The rounded average amount of cookies after accepting is 9, indicating that on average 4 more cookies are saved to the users device when accepting the cookies. In Figure 6 the percentages of the cookie types after accepting the cookie banner are given. There is a small increase to be noticed in the percentages of functionality and targeting/advertising and a small decrease in necessary cookies, the percentages for performance are about the same, all with regards to the cookies from the initial visit.

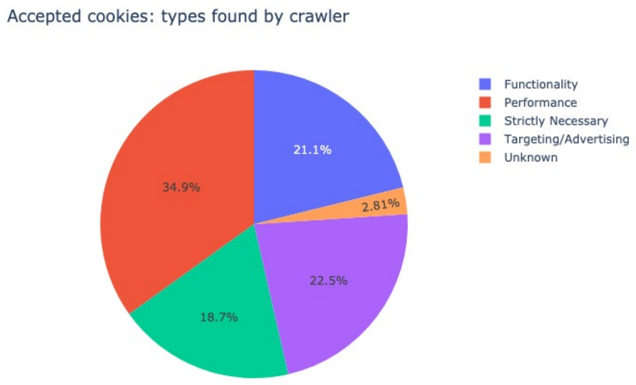

Fig. 6. Percentages of cookie types after accept

To see the change in cookie types after accepting the cookie banner, the websites for which both the initial cookies and the accepted cookies were found, were selected and the amount of cookies per type were displayed in Figure 7. Notice that for the 9671 websites that were analysed, the biggest increase after allowing a website to store any cookie was in performance cookies and strictly necessary cookies. This result is somewhat unexpected as the cookies that fall under these two categories usually do not contain any personal information [15].

4.1.3 Denied cookies

The third experiment was to see the effect of denying the cookie banner. This experiment however does not give a significant insight as websites usually provide an *'accept all'* or *'choose settings'* as option, so the crawler was not able to find an interesting number of websites that allow for plain denial of all cookies.

For the websites where the crawler was able to deny all cookies the percentages of the cookie types found after denying can be seen in Figure 8. The two most found types are targeting/advertising and performance cookies, where it is noticeable that the cookies with the former category should not have such a high presence after denying.

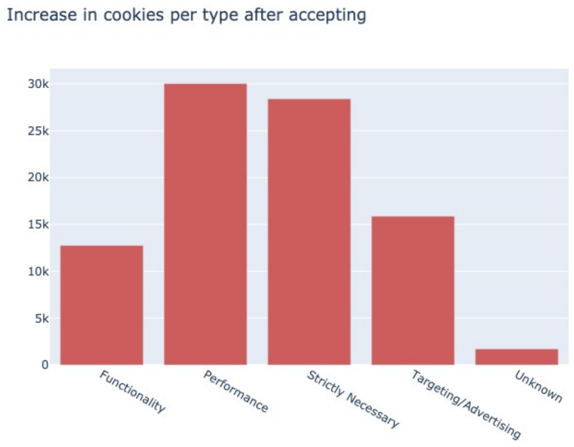

Fig. 7. Change in cookie type after accepting

It is logical that the crawler was not able to find a lot of websites that provide a button to deny everything at once as most websites make it a multiple step process, however it would have been interesting to compare the effect of accepting and denying between a significant number of websites. For a more extensive research in the effect of denying cookies see the 2019 study in which 2000 websites were analysed manually [4].

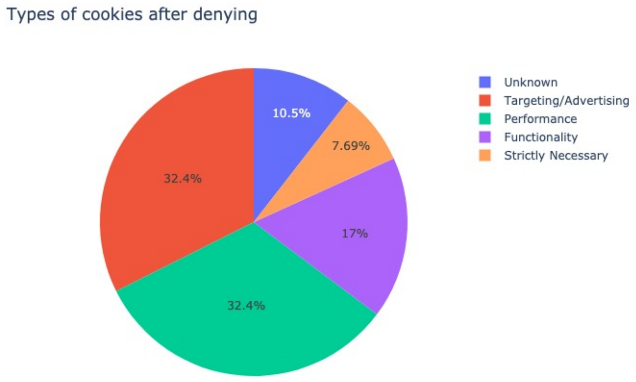

Fig. 8. Percentages of cookie types after denying

4.2 Banner providers

During the research it became evident that many websites utilize a plug-and-play type banner to handle their GDPR complicity. The three biggest found were 'Cybot', 'OneTrust', and 'TrustArc', the number of websites they were found on can be found in Table 2.

Of these providers, OneTrust is the most prominently present with almost 3% of the websites the crawler went trough, hence this provider was chosen to analyse further. The OneTrust banners provide the user with the option to choose to enable or disable cookies for each category discussed in earlier sections, Strictly Necessary, Performance,

Provider	Times found
Cybot	32
TrustArc	10
OneTrust	289

TABLE 2
Amount of websites using banner provider

Targeting(/Advertising), Analytics (in this paper Functionality). With a manual version of the crawler, a number of websites that implemented the OneTrust banner were visited and each option was either turned on or off manually. After 10 websites it became clear that the company provides exactly the service they promise as the cookies that were saved and then run trough cookiepedia all matched the categories that were turned on.

4.3 Personal information

The last experiment was to use the *zxcvbn* password strength estimator to estimate whether the value of a cookie could contain any personal information or is too general. Using the insight from another study [4] values with a log-10 value higher than 10^9 would be considered a possibility to contain personal information. This provides understanding of the amount of personal information in the cookies retrieved by the crawler for the different options discussed in Section 4.1.

In Figure 9 the total amount of cookies and the amount of cookies that are assumed personal information for the cookies on visit is shown. 35.67% of these cookies are assumed to be personal information which is a really high number considering these are cookies that are saved whenever the users visits the websites in question.

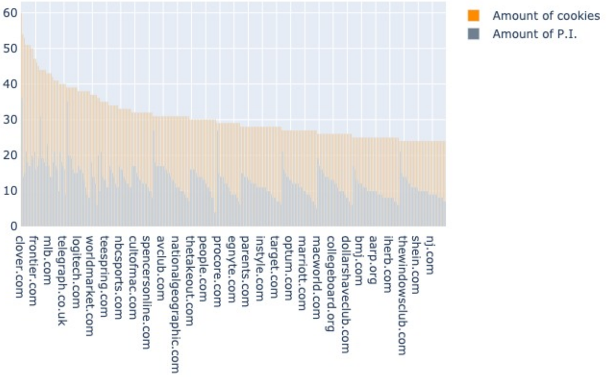

Fig. 9. Cookie count and amount of cookies with personal information; Initial visit

After accepting the percentage went up to 41.84%. Which is not the big increase as expected, however the amount of personal information saved upon visit is already thus far high that this lack of increase is not surprising. The amount of cookies and the amount of cookies with personal information after accepting the banner can be seen in Figure 10.

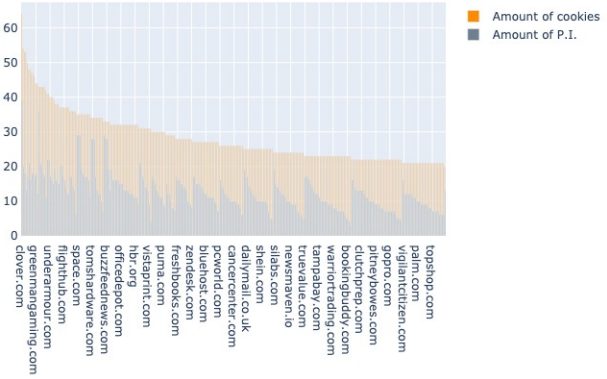

Fig. 10. Cookie count and amount of cookies with personal information; Accepted cookies

5 Discussion

From the experiments are some points to take away. The first thing to notice is that the majority of websites already save cookies to the users browser, without the user giving consent yet. From these cookies we were able to identify the names of the cookies and cross-reference these with the database of Cookiepedia to see whether the type of cookie could be found. This showed what types of cookies are used upon visit and something to note is that already 19.5% of these cookies are targeting and advertising related. The websites that only saved one cookie for the first time consisted mainly of strictly necessary cookies, which makes sense when websites go for the bare minimum. Even so, a lot of small websites try to monetize the traffic on their website by implemening Google Advertisements or the Facebook Pixel as their only cookie, explaining the 20.5% of single cookies being targeting and advertising related. The results above came only based on the names of the cookies that websites used, but the most important part of a cookie is the value that is saved in it. Using the password strength estimator it was possible to make an educated guess which values were complex enough to contain user information. On average with the cookies that were saved initially by visiting the website, about 36% of these cookies were estimated to contain user information, which is a very high number considering the user has not agreed to anything.

The next experiment was the same analysis but this time on the cookies that were saved after accepting the cookie banner. This time the percentage of cookies that was targeting and advertising related was 22.5%, a small increase compared to the cookies on visit but not by much.

It is important to note that this comparison is difficult as the results from the first experiment are based on a lot more websites due to the fact that many websites did not provide a banner, 10% of the websites that were visited by the crawler did not mention anything about cookies on their page, or the crawler was not able to find an accept all option. To combat this difference, a selection was made in which the crawler found both the cookie in the first experiment as well as in the second, providing a fairer comparison between the cookies per website. This difference shows the increase of the different types of cookies between the initial visit and when the crawler accepted all cookies. The result of this comparison was however, quite surprising. Cookies that are strictly necessary do not need consent from the user as they are needed to make the website work, still it was the second largest increase as category. The other category that showed a big increase after accepting the banner was performance, these cookies collect information about how users interact with websites, for instance which pages visitors go to most often, and if they get error messages from web pages. It makes sense that although the user information is aggregated, websites need to ask permission for this usage, and will want to utilize this type of cookie as one of the most to see how users interact with their website and what can be improved.

To see whether the cookies contain more user information when the user consents to the use of cookies, the same password strength estimator as before was used on the accepted cookies. The percentage of estimated user information in these cookies is almost 42%, combined with an increase of 4 more cookies on average, this shows that users will have more of their personal information saved in cookies when accepting a cookie banner.

The last experiment to discuss is the detection and analysis of cookie banner providers. During the research phase the three providers most found were *'TrustArc', 'OneTrusts', and 'Cybot'*, so the crawler was coded to detect whether a website used one of these to provide their users with a banner. Something that stood out was the fact that the majority of banners that were implemented came from OneTrust. The Cookiepedia database used in earlier experiments is also maintained by OneTrust and their banner on multiple websites proved very reliable in the way the types of cookies reflected the users choice.

6 Conclusion

The question whether the users choice to allow or disallow cookies relates to the actual personal data stored can be answered from the results in two ways. The first is to look at the types of cookies the user receives on their device, and from the results we see that the amount of personal information saved upon visiting a website is already higher than expected. After that the choice to accept or deny cookies is reflected in a change in the amount of cookies, the type of cookies, and the amount of cookies carrying personal information. The increase after accepting is not as big as expected but this is due to the already high amount of personal data stored upon visiting websites.

The percentage of websites providing a banner is about 90% of the websites that were analysed, there were however many websites that provided a banner stating their use of cookies while not actively asking for consent before saving cookies to the users browser, which is not what the intention of the GDPR laws were.

It is hard to analyse what information is saved in cookies as websites use their own encryption to protect the users data, login details, or other personal information. This is why the password strength estimator was used, allowing for a reliable prediction whether a cookie contained personal information. The highest count of cookies carrying personal information fell under the categories performance, and targeting/advertising. These two categories both need personal information to track users on the website and over multiple websites.

When the research was started one of the questions was whether storing cookies from a high amount of websites would be an issue with storage but as cookies are all small text files this proved to be the least of the issues in the process of data gathering. The biggest bottleneck was designing the crawler in such way that it was able to interact with any type of cookie banner. With many websites having a different structure and a different text for their buttons this process was the most time intensive part of the research.

The overall conclusion of this research is that most websites comply with the GDPR in the way that they provide a banner notifying the user that cookies are used and allowing the user to modify what types of cookies are saved to the users browser. However, with 20% of these initial cookies related to targeting and advertising and 35% containing personal information, the amount of cookies that are saved upon visit, even before the users consent is higher than expected and shows that many websites provide the banner but do not follow the GDPR guidelines to the fullest.

References

[1] J. Schwartz, "Giving web a memory cost its users privacy." https://www.nytimes.com/2001/09/04/business/giving-web-a-memory-cost-its-users-privacy.html, Sep 2001. "[Online; accessed 24-July-2020]".

[2] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 1388–1401, 2016.

[3] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)," in *Official Journal of the European Union*, 2016.

[4] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, "Can i opt out yet?: Gdpr and the global illusion of cookie control," pp. 340–351, 07 2019.

[5] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(un) informed consent: Studying gdpr consent notices in the field," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 973–990, 2019.

[6] O. Kulyk, A. Hilt, N. Gerber, and M. Volkamer, ""this website uses cookies": Users' perceptions and reactions to the cookie disclaimer," in *3rd European Workshop on Usable Security (EuroUSEC), London, England, April 23, 2018*, Internet Societa, Reston 8VY), 2018.

[7] L. Montulli and D. M. Kristol, "Http state management mechanism." https://www.rfc-editor.org/rfc/rfc2965.txt, 2000.

[8] "What is gdpr?." https://www.cookiebot.com/en/gdpr/. "[Online; accessed 2-August-2020]".

[9] "European commision, the general data protection regulation (gdpr)." https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en, 2018. "[Online; accessed 28-July-2020]".

[10] "Chromedriver." https://chromedriver.chromium.org. "[Online; accessed 2-August-2020]".

[11] "Majestic million." https://majestic.com/reports/majestic-million. "[Online; accessed 1-April-2020]".

[12] "Selenium documentation." https://www.selenium.dev/documentation/en/. "[Online; accessed 2-August-2020]".

[13] "Cookiepedia." https://cookiepedia.co.uk/. "[Online; accessed 28-July-2020]".

[14] “Clover privacy policy.” https://www.clover.com/privacy-policy. ”[Online; accessed 5-August-2020]”.

[15] “How we classify cookies.” https://cookiepedia.co.uk/classify-cookies. ”[Online; accessed 28-July-2020]”.