POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

The Risk of Web Tracking and the impact of user consent

Supervisors

Prof. Marco MELLIA

Dott. Ing. Martino TREVISAN

Dott. Ing. Nikhil JHA

Candidate

Antonino MUSMECI

Academic Year 2020-2021

Abstract

Nowadays, the internet and World Wide Web have become part of our lives by allowing its users to access vast amounts of information and resources. However, with the use of the internet, threats to our privacy have increased.

Along with the rapid increase in the number of Internet users, internet advertising has become a fast growing business. Data has become for companies a strategical asset to drive businesses, efficiently target products and services and obtain a more relevant position on the markets. To regulate the usage of tracking technologies and the usage of users personal data, governments have intervened and different legal instrument have been developed to safeguard privacy. In addition, in recent years there have been several solutions such as tracker-blockers to help users preserve their privacy during navigation. Although some of these tools offer good performance they also have many limitations and are often insufficient to guarantee the protection of users' privacy. This is also due to the fact that users are often not aware of the tracking. This work aims to understand how the scenario has evolved in recent years, providing an indication of the privacy risk a user faces when visiting a website. Using data from HTTPArchive dataset, we analyzed how the risk associated with a website may change over time. We also want to assess the impact of "Cookie Bars" on the tracking ecosystem and how user consent changes the presence of trackers on the web. To do that we designed a simple tool to automatically provide consent to the installation of cookies.

Acknowledgements

I would like to thank the supervisors, Prof. Marco Mellia, Dott. Nikhil Jha and Dott. Martino Trevisan, for the professionalism, and absolute availability that they have had towards me.
A special thank goes to my parents, my brother, my family and all my friends who have supported and encouraged me every day in these years.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Web tracking and privacy concerns

Nowadays, the internet and World Wide Web have become part of our lives by allowing its users to access vast amounts of information and resources. Especially in the last few months, they have also become essential tools for everyday activities such as shopping, studying, staying in contact with friends, working, spending their leisure time. However, with the use of the internet, threats to our privacy have increased.

Along with the rapid increase in the number of Internet users, internet advertising has become a fast growing business. Today's web advertising ecosystem heavily relies on data collection and tracking that allows advertising companies to profit from collecting a vast amount of data associated to the users.

Learning how an user spends his time allows these companies to more efficiently target products and services. These data are collected in order to determine which kind of customers would be most likely to be influenced by a particular ad. Ad targeting occurs when an advertiser selects a subset of potential viewers to show the ad to, and displays the ad online to that subset rather than to everyone using the media platform. For example we could choose to advertise only those social-network users who have a certain age, gender, education, income. Usually the traditional media can not offer this level of targeting. In the case of social networks, this data collection is visible because users freely share this kind of information, but in most of the cases the data collection system is not visible. While browsing the web users tracked by multiple companies, the tracking code is installed on web pages that have adverts as well as those that do not but it is important to highlight that not all the "tracking systems" are a concern for the user privacy; in fact, for example, cookies are legitimately used by a wide range

of websites to remember useful information like login details, preferences and items in a shopping basket.
While there are legitimate reasons for web site providers to track their users on their web site, these first-party cookies are useful to improve the user experience, the majority of company embed in their sites also external resources that vary from content that the user explicitly want to obtain, to implicitly loaded third-party services, ads, resources with the purpose of collecting user data without the user's need, understanding, explicit consent, or knowledge about what information they are gathering. Third-party cookies are a greater concern for privacy because cookies from the same tracking company can monitor various sites. Websites can fetch resources such as images and scripts from domains other than their own. This is referred to as cross-origin or cross-site loading, and is a powerful feature of the web. However, such loading also enables cross-site tracking of users. Imagine a user who browses two different websites. If both these sites load resources from the same tracker has a cookie stored in the user's browser, the tracker has the ability to know that the user visited both the website and the recipe website, the users behavior while on a site, how long the user spent on each site, what kind of web browser was used, and so on Social media sharing buttons are also responsible for tracking user's behavior across sites. When the user clicks one of the social media share buttons, the site can use that information to see what the visitor is sharing. The company can share information with the social platform as well.

1.2 Goal

In spite of the regulatory efforts made in recent years by various states, user privacy still remains an open issue. Still many sites collect data from users without consent and still many users are not aware of this. The literature states that cookies are the most common tracking mechanism on the web. Therefore, this thesis focuses on analyzing this type of tracking method. The goal of this work is to provide an indication of the privacy risk that a user faces when visiting a page on the web. This indication can help make users more aware of the risks and improve their experience when using the web. The thesis is divided into two parts. The first part evaluates how the scenario has evolved in recent years using a portion of the HTTPArchive [1] data set to study how the risk varies from 2015 to 2020. In the second part, on the other hand, we want to evaluate the impact of "Cookie Bars" on the tracking ecosystem and how consent changes the presence of trackers on the web. For this second part we designed a tool using Python3 and the Selenium library that automatically visits a website and gives consent for the installation of cookies by the website.

Chapter 2

Background

2.1 Http cookies

Users use software called web browsers, such as Google Chrome or Mozilla Firefox , to access web sites. When the user wishes to visit a website, he or she enters the name of the website into the browser, and the browser contacts the website, sending a request. The request includes an identifying string for a particular website, known as a URL (Uniform Resource Locator), and may also include other information, such as an identifier for the user. The website responds to the request with the content of the website, usually in a format known as HTML (HyperText Markup Language), which includes the content of the website. The HTML includes additional URLs that identify the location of other contents as images and videos. For this reason, opening a single web page may involve any number of separate connections to separate websites. The domain name that the user requests access to, and which provides the initial HTML webpage, is referred to as the first-party domain. Any other domains that provide portions of the webpage content are referred to as third-party domains.

Cookies are critical for tracking of users within the sites and across the web. They were introduced by Lou Montulli in 1994 to be a mechanism for websites to remember stateful information or to record the user's browsing activity.

When a user connects to a website, the website tells the users' browser to store a small piece of data that allows the website to distinguish one user from another.This piece of data is defined as a set of `name=value` pairs and may have different attributes such as `id=value`, `expires=date`, `domain=domain_name`. When the user visit the website a second time, the browser sends the stored cookie back to the site, so that the site can know that this user has already visited the website. Cookies are associated with the domain name of a website such that only that website has access to the information from that cookie. When the users goes to another site, the browser does not send the cookie. The process of storing and sending cookies is done automatically by the browsers when the user accesses web without explicit users interaction

Session, or temporary, cookies are those that are deleted or expire when the browser is closed, while persistent cookies have longer expiry dates from a few minutes to years depending on the functions they perform. These functions distinguish cookies into two categories: technical and non-technical or profiling. Technical cookies manage the data necessary for the pages to function and make navigation easier, for example by not having to re-enter your user name and password to access specific services, by remembering the display time of a video, by recognising the type of device you are using and adapting the size of page elements accordingly. Technical cookies also allow aggregate statistical analysis of the most visited pages and user preferences, but not used to analyse the behaviour or preferences of individual users. Profiling cookies may be installed by provider to make detailed analyses on: visitors to a website, search engines used, keywords used, language used, most visited pages. They can collect information and data such as IP address, nationality, city, browser, operating system, pages visited, , duration of the visit, number of visits made, device information, screen resolution and analyse the navigation habits of individual users, in order to provide, for example, content, including advertising, aimed at particular interests. This data are collected to create a profile of the user in order to determine which kind of customers would be most likely to be influenced by a particular ad. For this reason, data has become for companies a strategical asset to get a better position on the market.

Normally, a cookie's domain attribute will match the website domain. This is called a first-party cookie. A third-party cookie, on the other hand, is a cookie that belongs to a different domain. These cookies typically appears when web pages feature content from external websites, such as banner advertisements. This opens up the potential for tracking the the same user across multiple websites to build a more detailed profile of the user used by advertisers for more precise targeted advertising

2.2 Other Tracking Mechanisms

As the web evolved, so did the tracking system and websites start to use even more advanced tracking mechanisms than cookies. One of that is the web beacon also called tracking pixel,web bug. These are mechanisms in which the web page requires the user to send a request to download an object. Usually this is an invisible image from the tracking system's server. The image is a one pixel by one pixel image of the same color as the background of the page, so it does not change the page aspect. Downloading this object the user provides information as IP, type of web browser, type of device, that can be used for tracking the users' behaviour.

Another type of sophisticated method to track users across the web is fingerprinting. In order to correctly render the contents of a web pages the browser has to give some information about hardware and software to the web services. Eckersley in 2010 [2] showed that information like browser plugins, screen resolution, fonts, timezone and so on, can be combined to create a fingerprint for a specific device and that 94% of browsers with Flash or Java had an unique fingerprint. This fingerprint can be associated to an ID and used to identify a user across multiple websites. The most common and researched fingerprinting method is canvas fingerprinting Users' web browsers have unique characteristics such as the fonts, plugins, version, and many other parameters. In a website, canvas are areas designated to render bitmap images. The users' web browser unique characteristics make possible that canvas renders images in a unique way. Detecting this small differences in the rendering of a text or an image is possible to obtain a fingerprint without users' knowledge.

To regulate the usage of tracking technologies and the usage of users personal data, governments have intervened with legislation to protect users' privacy. Different legal instrument have been developed to safeguard privacy.

2.3 California Consumer Privacy Act

The California Consumer Privacy Act (CCPA) [3] is a state statute signed into law on June 28, 2018 and became effective on January 1, 2020. The law is a response to the increasing role data plays in business and to the privacy implications for the collection, protection and use of personal information. It is one of the most important privacy law in the United States establishing data privacy as a fundamental right for California residents. The users rights can be divided into four key parts that are protected under CCPA:
Right to know: Consumers have the right to know what information is being gathered about them.
Right to Deletion: Consumers have the right to request that a company delete any personal data they have about them. The act requires a business that collects any personal data about users to disclose the consumer's right to delete the personal data.

Right to Opt-Out: Consumers have the right to opt-out of the collection and sale of their personal data to third parties. Businesses must provide notice to consumers which information they sell to third parties and give them the possibility to opt-out of the sale of their personal data.
Right to Nondiscrimination: Businesses cannot discriminate against users who exercise their rights under CCPA, they cannot be refused services due to exercising their rights.

2.4 European Legislation for web privacy

One of this legal instrument relevant to tracking mechanisms is the E-Privacy Directive. The 2002 E-Privacy Directive, updated in 2009, is the regional instrument in European Union to safeguard data protection. The directive pays attention to cookies and tracking. The aim of the E-Privacy Directive is to increase and harmonize privacy protection across member states. "Member States shall ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her consent" [4]. The E-Privacy Directive only offered a set of rules, each member state must transpose EU Directives into their national legislation either providing less or more privacy protection to their citizens. For example, in some countries consent needs to be explicitly provided by the users, while in other countries consent might be implied. In any case, it follows that non-technical cookies cannot be installed on user's device without prior consent and this has become evident due to the presence of “Cookie Bar” on most websites. This bar informs users about the presence of tracking mechanisms and asks consent for their use.

On May 25 2018, the European Union adopted the General Data Protection Regulation (GDPR). It is a legal framework that focus the attention on the possible privacy issues that can emerge while browsing the Internet. The GDPR regulate how organisations must handle the information of those that interact with them, requires organization to protect the personal data and privacy of EU citizens, specify data subjects right, obligations and under what conditions personal data may be processed. These new guidelines increased the concern about how user data is managed and define which users rights each organization must ensure. Organizations that offer services in the European Union are required to be compliant with the GDPR, even if they are located outside.

2.5 Other tracking countermeasures

The increasing number, pervasiveness and the evolution of tracking mechanisms, and the menace for user privacy have led to the birth of different tracker-blockers to help users preserve their privacy during navigation. The Popular browsers offer user configurable settings which can benefit the privacy of users and there are some actions that an user can take to protect his privacy but all of these mechanisms have some limitations. One option is to block cookies. Some browsers like Safari by default blocks third party cookies others allow users to determine which domains they want to accept cookies from and which do not, or allow cookies to be stored until the browser is closed. For example, in private browsing mode cookies are deleted when you exit private browsing mode The main limitation of disabling HTTP cookies, especially first-party cookies, is that many websites rely on cookies for their functionality and stopping cookies compromises this functionality Do Not Track (DNT) is standardized by the W3C and used in the most famous web browsers. DNT works by adding a field to HTTP request headers. The header field allows users to express preferences on being tracked or not. For now, DNT does not have much impact, since the technique requires compliance with the tracking parts and only a very small part of the trackers, considers user preferences [5, 6]. Another solution to disable tracking are the tracker-blockers. In most of the cases, tracker-blockers are available as browser extensions. Two kinds of browser extensions are interesting to look at from a privacy perspective: tracker blocking extensions and advertisment blocking extensions. Some of the most popular extensions are AdBlockPlus, uBlock, Disconnect, Privacy Badger, Ghostery. Most of these extensions we looked into, make use of blacklists periodically updated to block requests to known trackers or advertisers. Using blacklists with trackers to block requests has some drawbacks. Lists can lack trackers, or have domains on them, which are not trackers. Another limitation of blacklists is that domains that not only are trackers, but also provide content to the website, might be blocked because they are on a blacklist.

Chapter 3

Tracking risk measurement

3.1 Risk definition

Users browses dozens of websites everyday, encountering a large number of trackers. Still many sites collect data from users without consent and still many users are not aware of this, so user privacy still remains an open issue. We want to have a method to estimate how dangerous a website is in terms of tracking so as to improve the protection of users' data, their web experience by making them more aware of the presence and pervasiveness of web tracking. In order to reach this goal we assign to websites a score estimating the risk for users' privacy. The score depends on the number of trackers that each first party incorporates, their popularity and their pervasiveness. The higher this score, the more dangerous a website is considered to be. For a website, the risk score is based on the risk associated with each tracker j embedded on it.

$$S = c_1 f_1 + c_2 f_2 + c_3 f_3 = c_1 \sum_j f_{1,j} + c_2 \sum_j (f_{URL,j} + f_{Cookie,j}) + c_3 \sum_j C_j \quad (3.1)$$

The privacy score can be divided into three different components: the popularity of tracker j, the amount of information exchanged between site and tracker j, and the connection of tracker j with other trackers.

$$f_1 = \sum_j P_j \frac{log(P_j N)}{log(N)} \quad (3.2)$$

The probability P_j of encountering tracker j is computed as the number of websites trackers j embedded on divided by the number the number of first party websites. The logarithm is applied in order to prevent that score is based only on few popular third party.

The second part is based on the idea that the higher is the quantity of information exchanges with the third parties the higher is the possibility for the third party to collect personal data. To try to estimate the quantity of information exchanged between website and the third party, two quantity has been considered: The length of cookies set by the tracker j and the number of parameters in the URL of the requests from website i to third party j.

$$f_{URL} = \frac{\sum_j log(k_j)\frac{log(k_j)}{max(log(k_j))}}{max(\sum_j log(k_j)\frac{log(k_j)}{max(log(k_j))}} \tag{3.3}$$

$$f_{Cookie} = \frac{\sum_j log(x_j)\frac{log(x_j)}{max(log(x_j))}}{max(\sum_j log(x_j)\frac{log(x_j)}{max(log(x_j))}} \tag{3.4}$$

This component is computed as the sum of two function. In this functions x_j represent the total number of URL parameters and k_j represent the total lenght of cookies setted by tracker j on the first party. both this quantity are divided by the maximum over all the contacted trackers by the first party website in order to normalize the componets. The logarithms are used to smooth the function.

For the third component we build a graph with some nodes that represent the first parties and the others that represent the third parties. The first parties are linked with the embedded third parties. Starting from this graph we project it on trackers and we build a new graph where the trackers node are connected one another if they have at least one first party in common. Using this last graph we compute the closeness centrality for each third party.

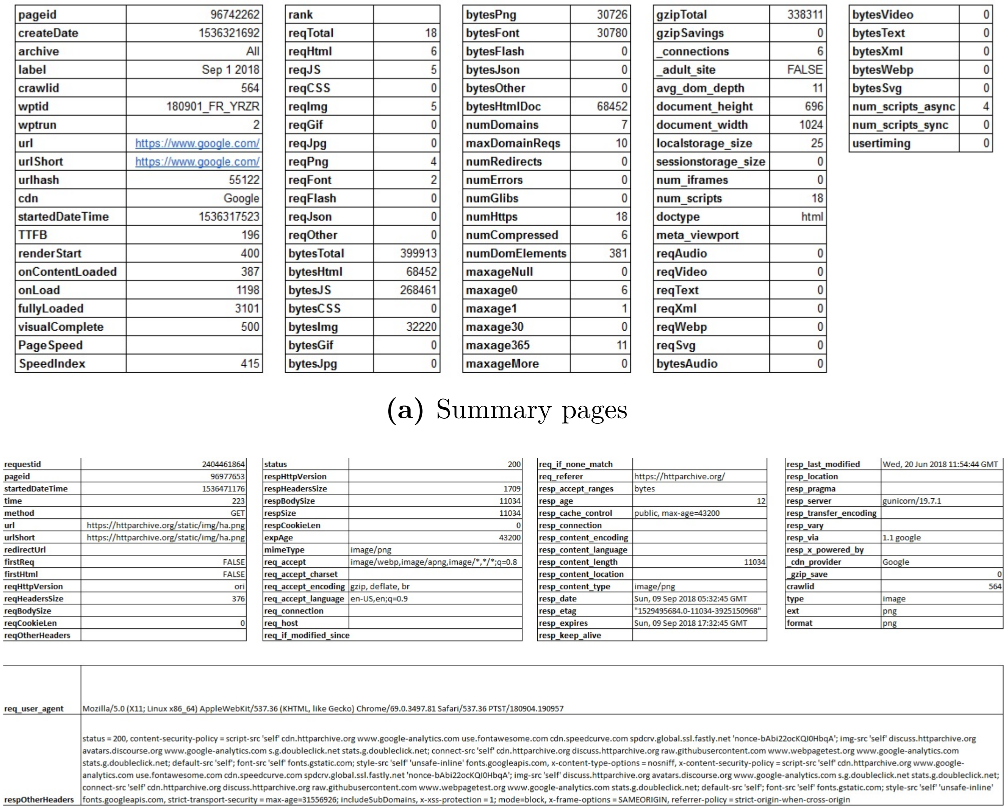
$$f_3 = \sum_j CC_j \tag{3.5}$$

This measure represents an estimate of how close each node is to all the others. The idea is that a tracker with a high CC being "at the center" of this network of trackers and being closer to all others, can exchange more information about the user. This third component is computed as the sum of closeness centrality of each tracker contacted by the first party website considered.

3.2 Dataset characterization

The HTTP Archive [1] is an open source project that tracks the evolution of the web. HTTPArchive makes a lot of information available via curated report. The data are also stores in BigQuery [7] where they are publicly available. To build the dataset, HTTPArchive periodically crawls millions of sites on the web on both desktop and mobile and record detailed information about fetched resources, used web platform APIs and features, request and response headers for each and every request on each page and execution traces of each page.
The URLs to crawl are taken from the Chrome User Experience Report, a dataset containing real users experience data on millions of websites. As of March 1 2016, the tests are performed on Chrome for desktop and emulated Android (on Chrome) for mobile. Prior to that, IE was used for desktop, and iPhone was used for mobile. The test agents are located in the Internet Systems Consortium data center in Redwood City, CA. Each URL is loaded three times with an empty cache and the data from the median run is collected.
HTTPArchive provides the data with file HAR [8]. The HTTP Archive format, or HAR, is a JSON-formatted file format for tracking of a web browser's interaction with a site. These files contain requests for each resource, and the response bodies for each request.
For this thesis we used the summary tables provided by HTTPArchive that contain summary information about the visited pages. We used the two tables called summary_pages, contain information about the visited pages, and summary_requests, contain information about all HTTP requests and responses made by the visited pages. The two tables contain a lot of information extracted from the har files, so the data has been processed by selecting only the information useful for the risk calculation. The summary_pages table, shown in Figure 3.1, is used to extract the domains that have been visited in that month.
The table contains among the various columns the URL of the page and a pageid used to join the summary_pages table to the summary_request table. From the table summary_requests (An example of the summary request table is present in Figure 3.1) have been extracted instead for every request the columns pageid, URL, respCookieLen used for the calculation of the risk. For this analysis we consider data collected by the HTTPArchive project each 6 months from January 2015 to July 2020, for a total of 37000 websites that are present in the data collected each year. We extracted from the URLs present in the summary_requests tables the second level domains. These domains have been classified as first party, tracking third party or non-tracking third party domains. The tracker domains are identified using Disconnect, EasyList and EasyPrivacy

tracker lists. Using these lists together we can identify a total of 24.664 distinct tracking domains. We labeled the second level domains as trackers according to these lists The domains extracted from the summary_requests tables are compared to the domains contained in EasyList, EasyPrivacy and Disconnect lists, and if there is a match, the domain is labelled as a third party tracker. If a domain is present in the list of second level domain extracted from the summary_pages tables, that domain is labelled as a first party. All the domains that are not identified as third party tracker or as first party are then labelled as non-tracking third party domains.



pageid	96742262	rank		bytesPng	30726	gzipTotal	338311	bytesVideo	0
createDate	1536321692	reqTotal	18	bytesFont	30780	gzipSavings	0	bytesText	0
archive	All	reqHtml	6	bytesFlash	0	_connections	6	bytesXml	0
label	Sep 1 2018	reqJS	5	bytesJson	0	_adult_site	FALSE	bytesWebp	0
crawlid	564	reqCSS	0	bytesOther	0	avg_dom_depth	11	bytesSvg	0
wptid	180901_FR_YRZR	reqImg	5	bytesHtmlDoc	68452	document_height	696	num_scripts_async	4
wptrun	2	reqGif	0	numDomains	7	document_width	1024	num_scripts_sync	0
url	https://www.google.com/	reqJpg	0	maxDomainReqs	10	localstorage_size	25	usertiming	0
urlShort	https://www.google.com/	reqPng	4	numRedirects	0	sessionstorage_size	0		
urlhash	55122	reqFont	2	numErrors	0	num_iframes	0		
cdn	Google	reqFlash	0	numGlibs	0	num_scripts	18		
startedDateTime	1536317523	reqJson	0	numHttps	18	doctype	html		
TTFB	196	reqOther	0	numCompressed	6	meta_viewport			
renderStart	400	bytesTotal	399913	numDomElements	381	reqAudio	0		
onContentLoaded	387	bytesHtml	68452	maxageNull	0	reqVideo	0		
onLoad	1198	bytesJS	268461	maxage0	6	reqText	0		
fullyLoaded	3101	bytesCSS	0	maxage1	1	reqXml	0		
visualComplete	500	bytesImg	32220	maxage30	0	reqWebp	0		
PageSpeed		bytesGif	0	maxage365	11	reqSvg	0		
SpeedIndex	415	bytesJpg	0	maxageMore	0	bytesAudio	0		

(a) Summary pages

requestid	2404461864	status	200	req_if_none_match		resp_last_modified	Wed, 20 Jun 2018 11:54:44 GMT
pageid	96977653	respHttpVersion		req_referer	https://httparchive.org/	resp_location	
startedDateTime	1536471176	respHeadersSize	1709	resp_accept_ranges	bytes	resp_pragma	
time	223	respBodySize	11034	resp_age	12	resp_server	gunicorn/19.7.1
method	GET	respSize	11034	resp_cache_control	public, max-age=43200	resp_transfer_encoding	
url	https://httparchive.org/static/img/ha.png	respCookieLen	0	resp_connection		resp_vary	
urlShort	https://httparchive.org/static/img/ha.png	expAge	43200	resp_content_encoding		resp_via	1.1 google
redirectUrl		mimeType	image/png	resp_content_language		resp_x_powered_by	
firstReq	FALSE	req_accept	image/webp,image/apng,image/*,*/*;q=0.8	resp_content_length	11034	_cdn_provider	Google
firstHtml	FALSE	req_accept_charset		resp_content_location		_gzip_save	0
reqHttpVersion	ori	req_accept_encoding	gzip, deflate, br	resp_content_type	image/png	crawlid	564
reqHeadersSize	376	req_accept_language	en-US,en;q=0.9	resp_date	Sun, 09 Sep 2018 05:32:45 GMT	type	image
reqBodySize		req_connection		resp_etag	"1529495684.0-11034-3925150968"	ext	png
reqCookieLen	0	req_host		resp_expires	Sun, 09 Sep 2018 17:32:45 GMT	format	png
reqOtherHeaders		req_if_modified_since		resp_keep_alive			

req_user_agent	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/69.0.3497.81 Safari/537.36 PTST/180904.190957
respOtherHeaders	status = 200, content-security-policy = script-src 'self' cdn.httparchive.org www.google-analytics.com use.fontawesome.com cdn.speedcurve.com spdcrv.global.ssl.fastly.net 'nonce-bAbi22ocKQI0HbqA'; img-src 'self' discuss.httparchive.org avatars.discourse.org www.google-analytics.com s.g.doubleclick.net stats.g.doubleclick.net; connect-src 'self' cdn.httparchive.org discuss.httparchive.org raw.githubusercontent.com www.webpagetest.org www.google-analytics.com stats.g.doubleclick.net; default-src 'self'; font-src 'self' fonts.gstatic.com; style-src 'self' 'unsafe-inline' fonts.googleapis.com, x-content-type-options = nosniff, x-content-security-policy = script-src 'self' cdn.httparchive.org www.google-analytics.com use.fontawesome.com cdn.speedcurve.com spdcrv.global.ssl.fastly.net 'nonce-bAbi22ocKQI0HbqA'; img-src 'self' discuss.httparchive.org avatars.discourse.org www.google-analytics.com s.g.doubleclick.net stats.g.doubleclick.net; connect-src 'self' cdn.httparchive.org discuss.httparchive.org raw.githubusercontent.com www.webpagetest.org www.google-analytics.com stats.g.doubleclick.net; default-src 'self'; font-src 'self' fonts.gstatic.com; style-src 'self' 'unsafe-inline' fonts.googleapis.com, strict-transport-security = max-age=31556926; includeSubDomains, x-xss-protection = 1; mode=block, x-frame-options = SAMEORIGIN, referrer-policy = strict-origin-when-cross-origin

(b) Summary requests

Figure 3.1: HTTPArchive summary tables

3.3 Risk change over time

The scatter plots in Figure 3.2, 3.3, 3.4 show for each considered period the correlation between the different components. The values of the components appear on axes, and each individual website appears as a point on the graph, so from this plots we can see how each component varies with respect to each other. The slope provides information on the strength of the relationship. from the scatter plots in Figure 3.2 we can observe that f_1 and f_2 have a weaker correlation so for most of the website they return different information. The correlation between f_1 and f_3 is more moderate (Figure 3.3) while the scatter plots in Figure 3.4 show a positive slope near to 1, so there is a strong positive correlation between the components f_2 and f_3. This means that this two components return the same kind of information for most of the considered websites.

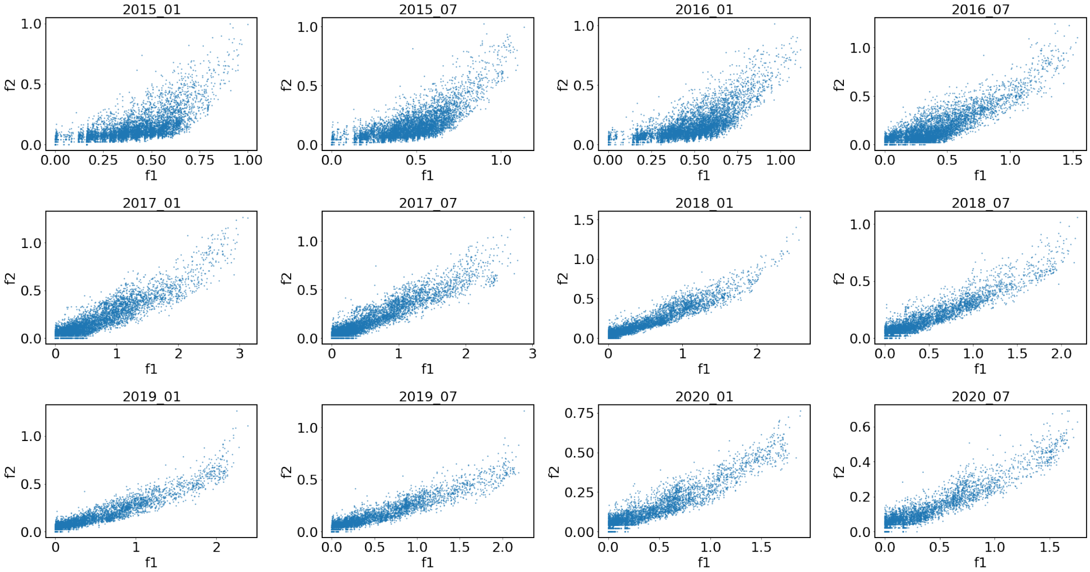


Figure 3.2: f_1 vs f_2

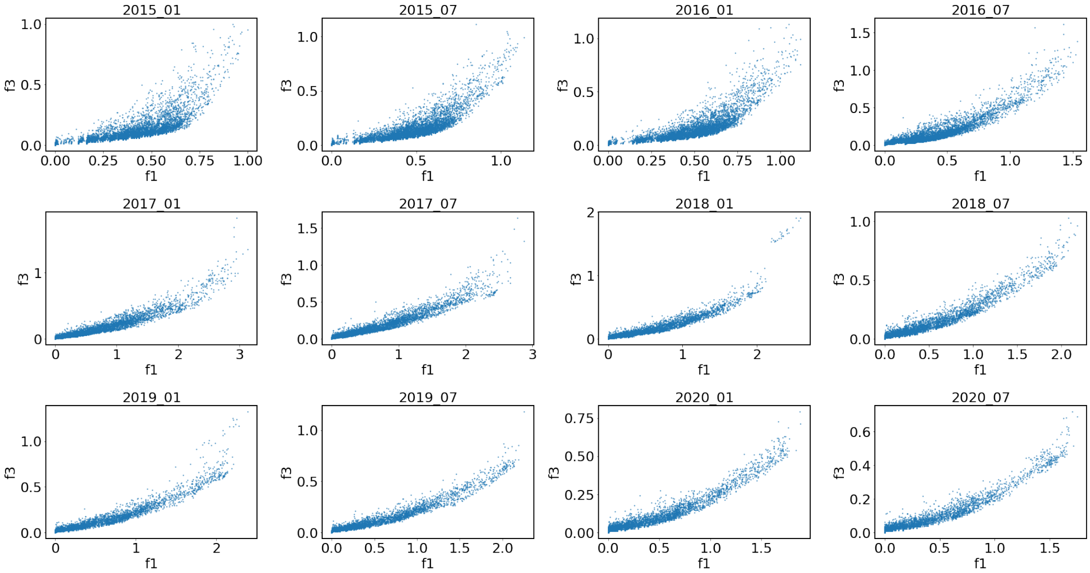


Figure 3.3: f_1 vs f_3

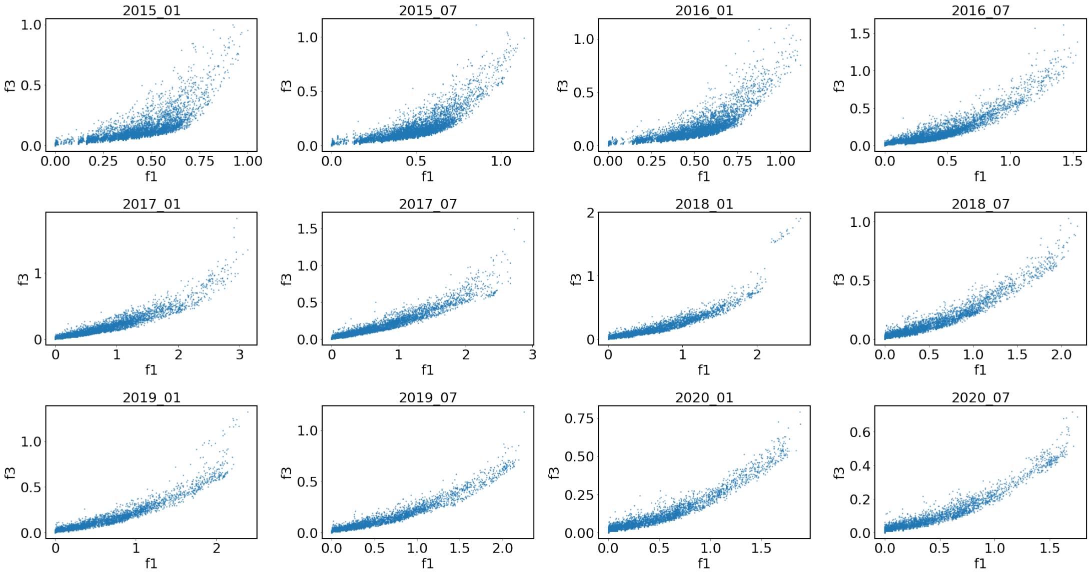


Figure 3.4: f_2 vs f_3

The boxplots in Figure 3.8 show the risk trend over the period from January 2015 to June 2020. In order to compare the data of the different periods, we choose to normalize the risk value with respect to the first reference year.

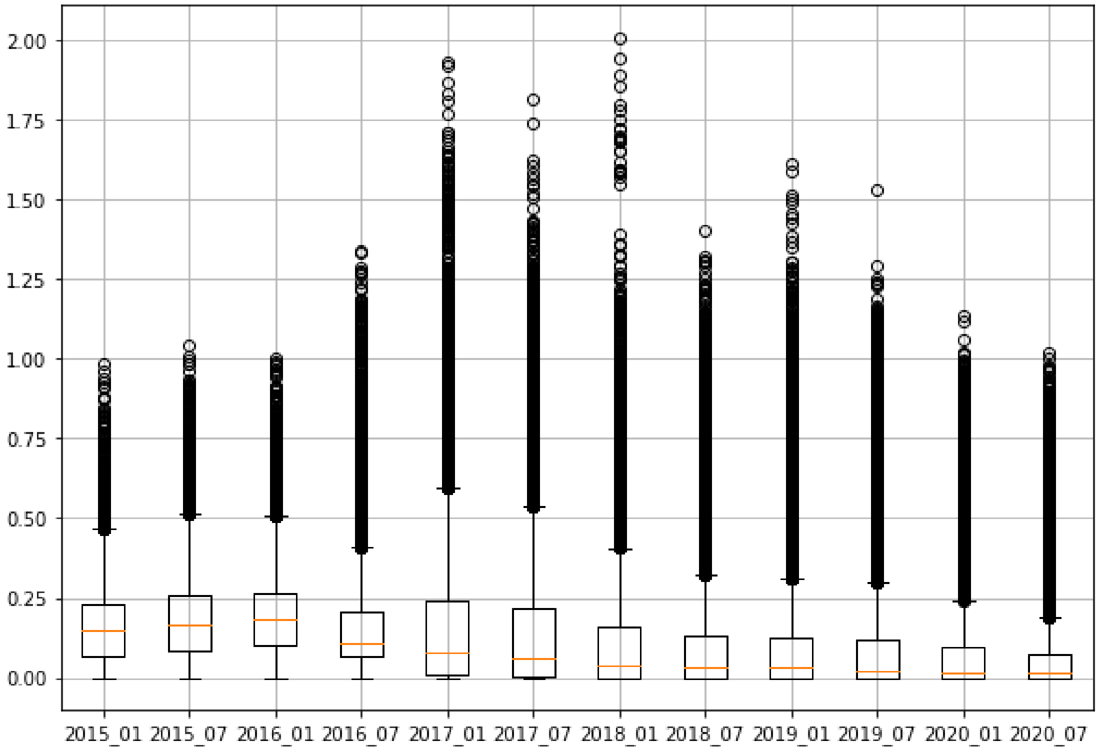


Figure 3.5: Total risk from January 2015 to July 2020

We can therefore see that the risk thus calculated in this subset of sites initially grows until it reaches its maximum value in January 2016 and then begins to decrease over time due to the diminishing presence of third-party cookies within these sites.
As show in January 2016 the 80% of first party websites considered embed some kind of tracker, and therefore have an estimated risk greater than 0, the remaining 31% do not embed trackers according to the HTTP requests considered. Instead in July 2020 only the 50% of first party websites embed some kind of tracker.

This effect of decrease of trackers could have among its reasons the increasing diffusion of new tracking techniques such as those mentioned in chapter 2 of this thesis and the adoption by the providers of the various regulations for the protection of privacy. For example, the increasing presence of cookie-banners that should prevent the installation of cookies from third parties before the

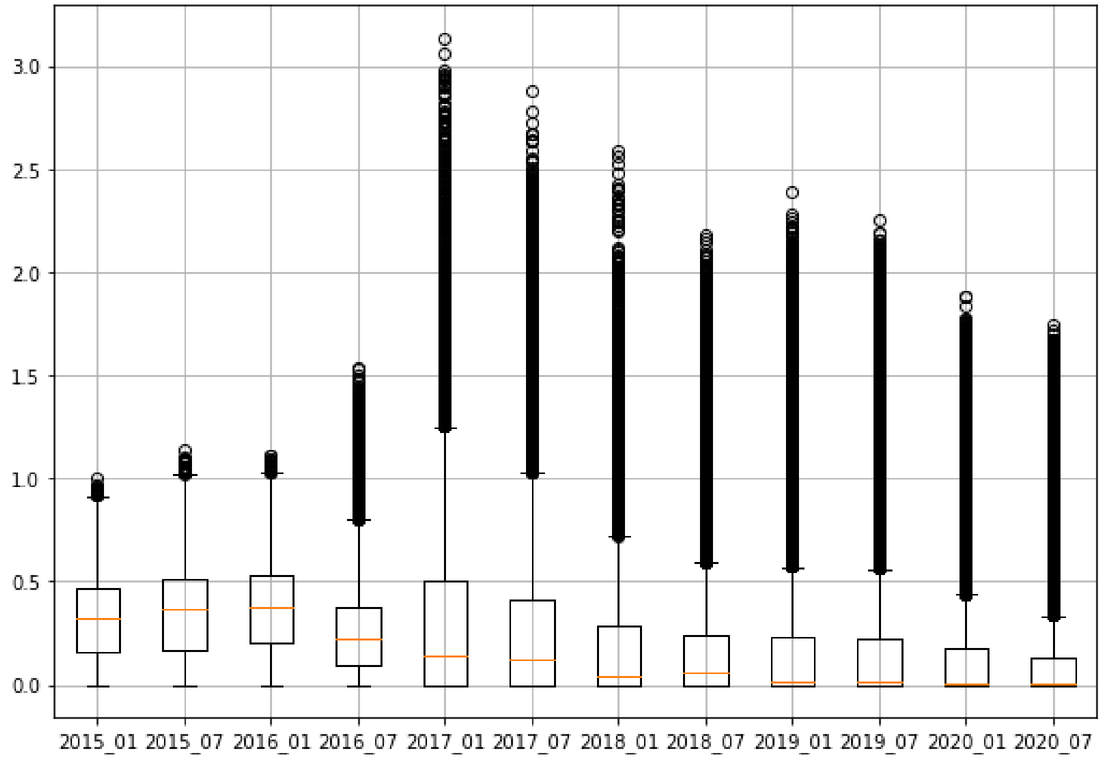


Figure 3.6: f_1 component from January 2015 to July 2020

consent of the user.

We want to evaluate the impact of providing consent to the installation of cookies. To perform a first analysis we had selected the 100 sites showing the highest risk value in 2020 and we manually visited the website in this list to provide consent to the usage of cookies if a Cookie Bar was shown. We stored HAR files before and after provide consent to the installation of cookies. The result of this first analysis is shown in Figure 4.16, we can see a substantial difference in the value of risk before and after giving consent.

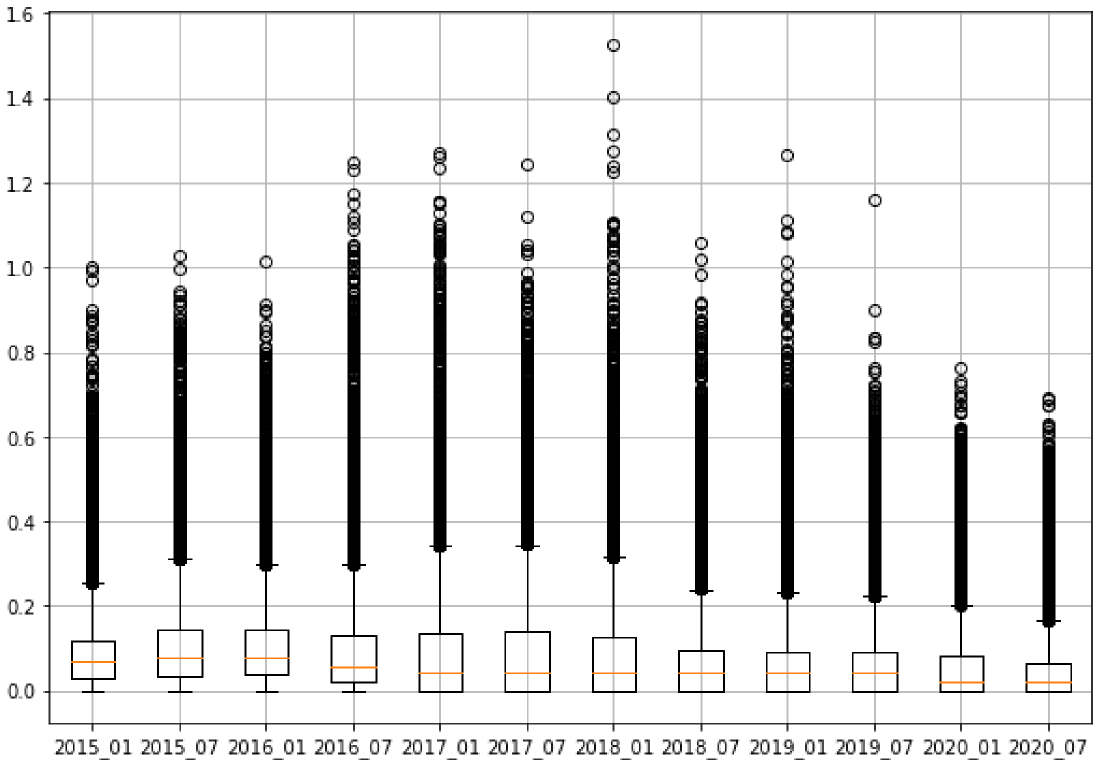


Figure 3.7: f_2 component from January 2015 to July 2020

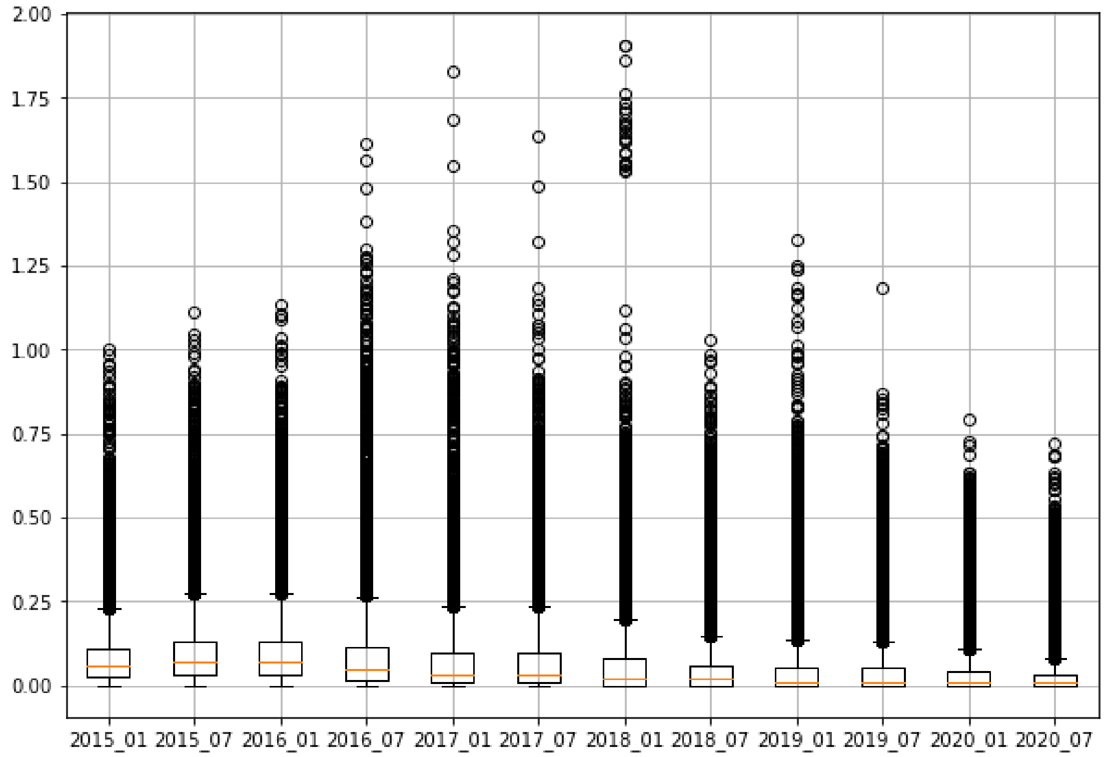
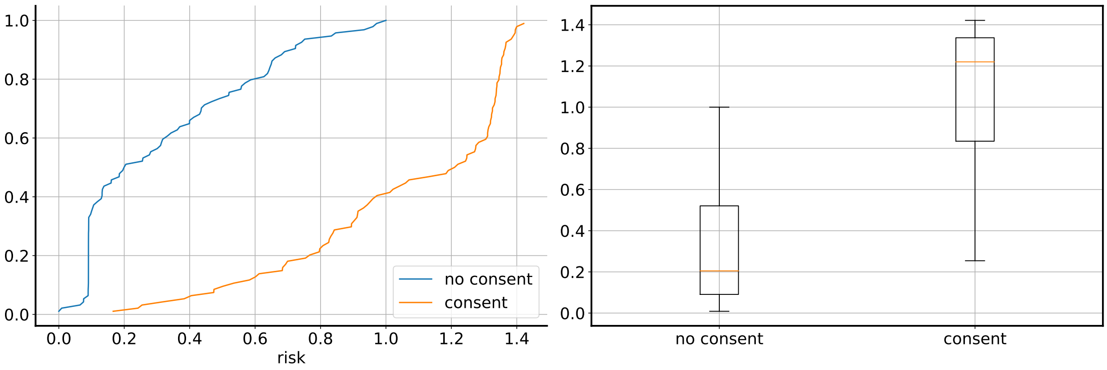


Figure 3.8: f_3 component from January 2015 to July 2020



(a)

(b)

Figure 3.9: Risk value before and after giving consent

Chapter 4

Provide consent to a Cookie Bar

4.1 Automatically accept Cookie Bar

As mentioned on the second chapter, in order to safeguard the privacy of users surfing the web, some countermeasures have been taken over time. Considering the particular pervasiveness that profiling cookies, especially third-party ones, can have on users' privacy, some legislations provide that users must be adequately informed about cookie usage and that users have to express their consent to the installation of cookies on their device. The most evident aspect of this law is the presence on websites of the so-called Cookie Bar or Cookie Banner. For example, with regard to Italian legislation when a user visits the website for the first time, the user must be shown immediately a suitably sized banner, the size of the banner must be such as to cause a perceptible discontinuity in the user's experience of the visited webpage and this banner must be an integral part of the action through which the user signifies consent [9].
Consent is usually given by means of a button embedded in the Cookie Bar. When the button is clicked, the website refreshes and the installation of cookies is activated (Figure 4.1).

Figure 4.1a shows a Cookie Bar on a web site. After clicking the Cookie Bar on the following visits the site looks different and various ads are shown (Figure 4.1b). In this chapter we want to analyze the impact that the Cookie Bars have on the presence of cookies in a website. In particular through this analysis we highlight that some websites,even though they show a Cookie Bar, install third party cookies before obtaining consent while other websites may set profiling cookies but do not display a Cookie Bar. Moreover we can also see the effect of the consent on the amount of cookies present in the web-sites and the effect of the consent on the privacy risk for the users. To perform this analysis we have created a tool called cookie-consent. This tool accept automatically the Cookie Bar to provide consent for cookies installation and to allow automated measurements on web tracking. It visits a URL and uses a heuristic to find the Cookie Banner and allow cookies. It is based on the I DON'T CARE ABOUT COOKIES 3.2.4 [10] CSS selectors and on a set of keywords to find the right button/link to click. The tool is implemented using Python3 and uses the selenium [11] library and chromedriver [12] to allow Selenium from Google Chrome to automate the process. Selenium load a clean browser profile, runs Google Chrome, visits the website and clicks on the Cookie Accept button, if one is found. Then Re-visit the URL after the consent is given and dump statistics in a JSON file, including all the HTTP requests fired at each stage, the cookies that are installed and some information about the found banners. The tool also stores screenshots of the website before and after clicking the Cookie Bar and of the Cookie Bar if one is found.

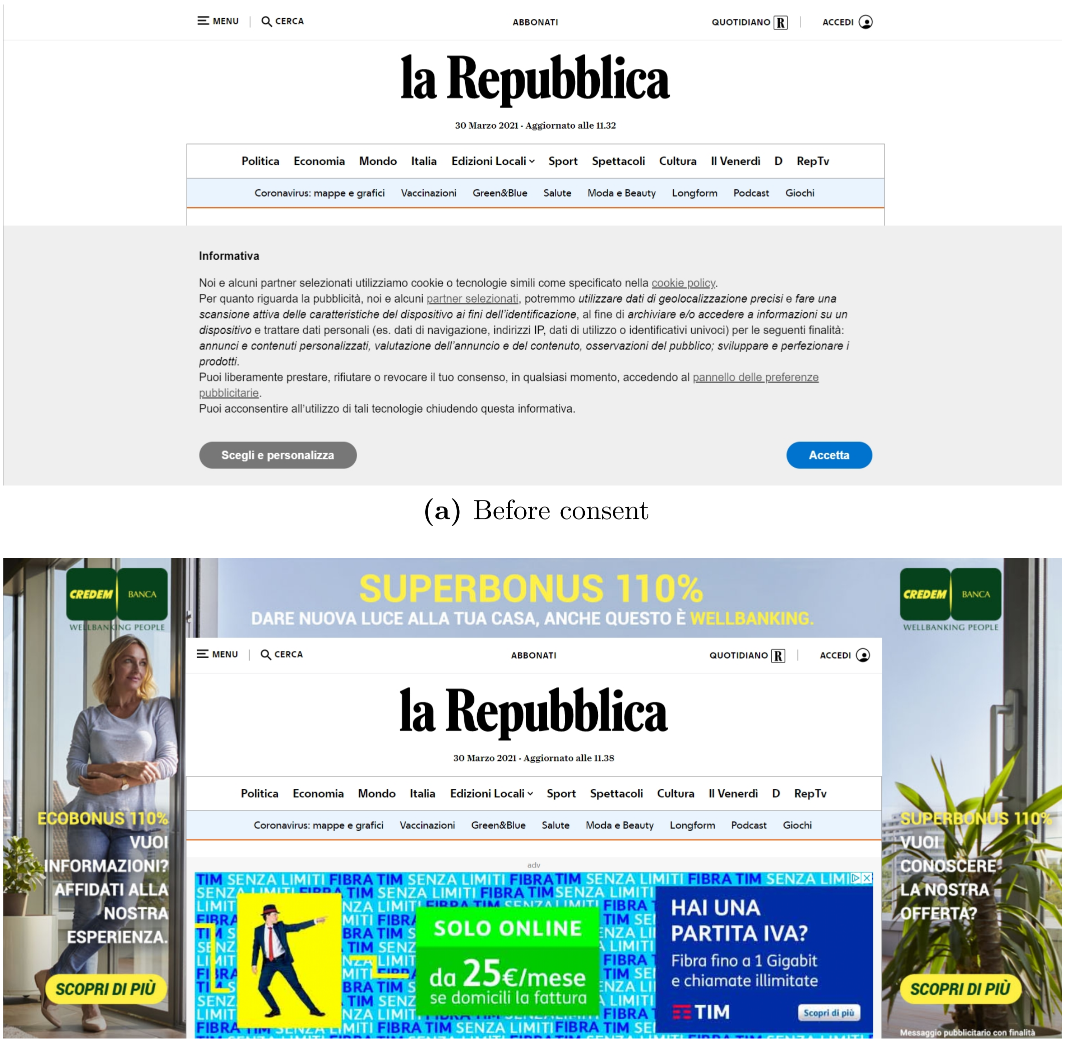
The list of URLs to crawl is taken from SimilarWeb [13], a website that provides web analytics services. SimilarWeb provide a website ranking service that shows per country and per category website ranks. We pick the most popular websites in 6 countries: Germany, France, Spain, Italy, United Kingdom and United States.

Each URL is loaded one times and the data from the run are collected via JSON file. The first visit is done on the landing page of the web site. After collecting the request/responses and cookies installed the tool tries to click the Cookie Bar in order to give consent for the cookie. Alternatively, if the click fail, a scroll action is performed on the page. After giving consent to the installation of cookies, the tool revisits the page, re-storing requests/responses and installed cookies.

For the keyword collection, we took from Similarweb lists the top 200 sites for each country that show a Cookie Bar. This list of sites was divided into two lists even-list and odd-list. From the even-list sites we retrieved the keyword. We manually visited the website on the even-list and collected 146 keywords belonging to 5 different languages (a subset of these keywords is shown in table 4.1)

4.2 Tool performance evaluations

The functioning of the tool with these collected keywords was then evaluated on the group of even websites. We can observe (Figure 4.2) that on this list of



Informativa

Noi e alcuni partner selezionati utilizziamo cookie o tecnologie simili come specificato nella cookie policy.
Per quanto riguarda la pubblicità, noi e alcuni partner selezionati, potremmo *utilizzare dati di geolocalizzazione precisi* e *fare una scansione attiva delle caratteristiche del dispositivo ai fini dell'identificazione*, al fine di *archiviare e/o accedere a informazioni su un dispositivo* e trattare dati personali (es. dati di navigazione, indirizzi IP, dati di utilizzo o identificativi univoci) per le seguenti finalità: *annunci e contenuti personalizzati, valutazione dell'annuncio e del contenuto, osservazioni del pubblico; sviluppare e perfezionare i prodotti*.
Puoi liberamente prestare, rifiutare o revocare il tuo consenso, in qualsiasi momento, accedendo al pannello delle preferenze pubblicitarie.
Puoi acconsentire all'utilizzo di tali tecnologie chiudendo questa informativa.

(a) Before consent

(b) After consent

Figure 4.1: Example of Cookie Bar on www.repubblica.it

French	Italian	Spanish	German	English
j'accepte	Accetto	aceptar y cerrar	akzeptieren	allows
j'ai compri	abilita tutto	de acuerdo	erlauben	agree
accepter	accettare	prosseguir	zustimmen	fine by me
autoriser	accetto	enté	fortfahnre	confirm

Table 4.1: Subset of keywords

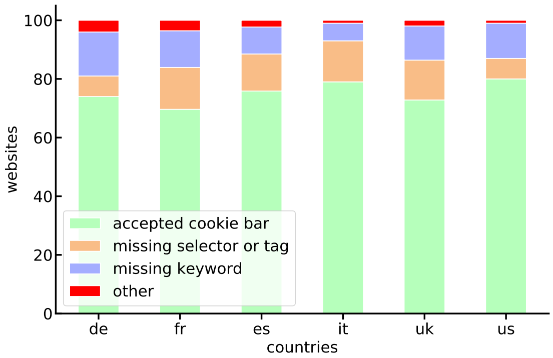
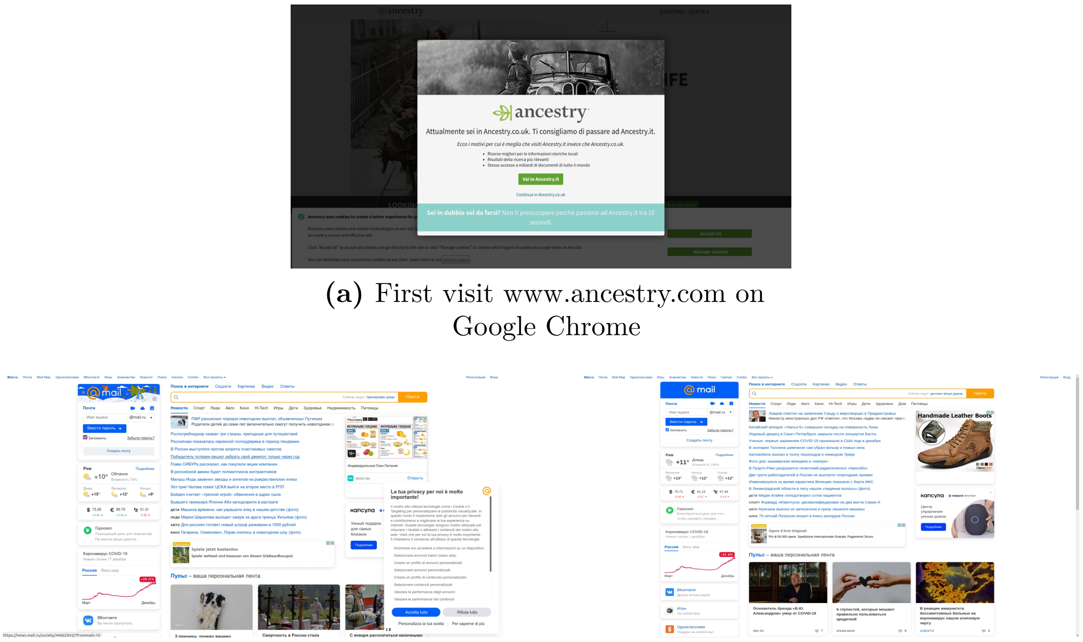


Figure 4.2: Per country result on even-list

websites the tool succeeds in giving consent to cookies between 70% and 80% of the times. There are two cases presented where the script fails to click the Cookie Bar. The tool cannot find one of the chosen tags or the tool cannot find any of the keywords specified in the list of keywords. The case named as other contains all the other cases in which the script cannot accept the Cookie Bar because of problems related to the site itself. Some of this cases are for example when the Cookie Bar is not the first element shown on the webpage as we can see in Figure 4.3a, the Cookie Bar is not shown when the site is visited in chrome headless mode (Figure 4.3b) or the site is not accessible through this mode.

After collecting the keywords the script was used on the top 1000 of each country.
The plots in the figure 4.4 represent the percentage of accepted Cookie Bars on the top 1000 for each considered country. We see that initially the percentage of accepted Cookie Bars is around 60% and 70% and then goes down until it stabilizes around 50%. This is due to the fact that in these lists there are also websites that do not show a Cookie Bar and it is not possible without a manual visit to know in advance if a website has a Cookie Bar or not. The graphs present two curves realized through two different configurations of the script. For the blue line we used the CSS selectors and the keywords extracted from the even-list and we limited the research of the button on the webpage to <a> and <button> tags. For the second line, the orange one, we used the keywords taken from both lists of sites, even and odd list, and moreover we removed the limit of the selectors making a research on the whole webpage and on more tags: <a>, <button>, <div>, <form> and <span>. In this way we limit the failure of the script to the lack of the keyword in the keywords list but we have to accept the risk of false positives.



Attualmente sei in Ancestry.co.uk. Ti consigliamo di passare ad Ancestry.it.

Ecco i motivi per cui è meglio che visiti Ancestry.it invece che Ancestry.co.uk.

- Risorse migliori per le informazioni storiche locali
- Risultati della ricerca più rilevanti
- Stesso accesso a miliardi di documenti di tutto il mondo

Sei in dubbio sul da farsi? Non ti preoccupare perché passerai ad Ancestry.it tra 10 secondi.

(a) First visit www.ancestry.com on Google Chrome

(b) First visit www.mail.ru on Google Chrome

(c) First visit www.mail.ru on Google Chrome headless mode

Figure 4.3: Example of "other" errors

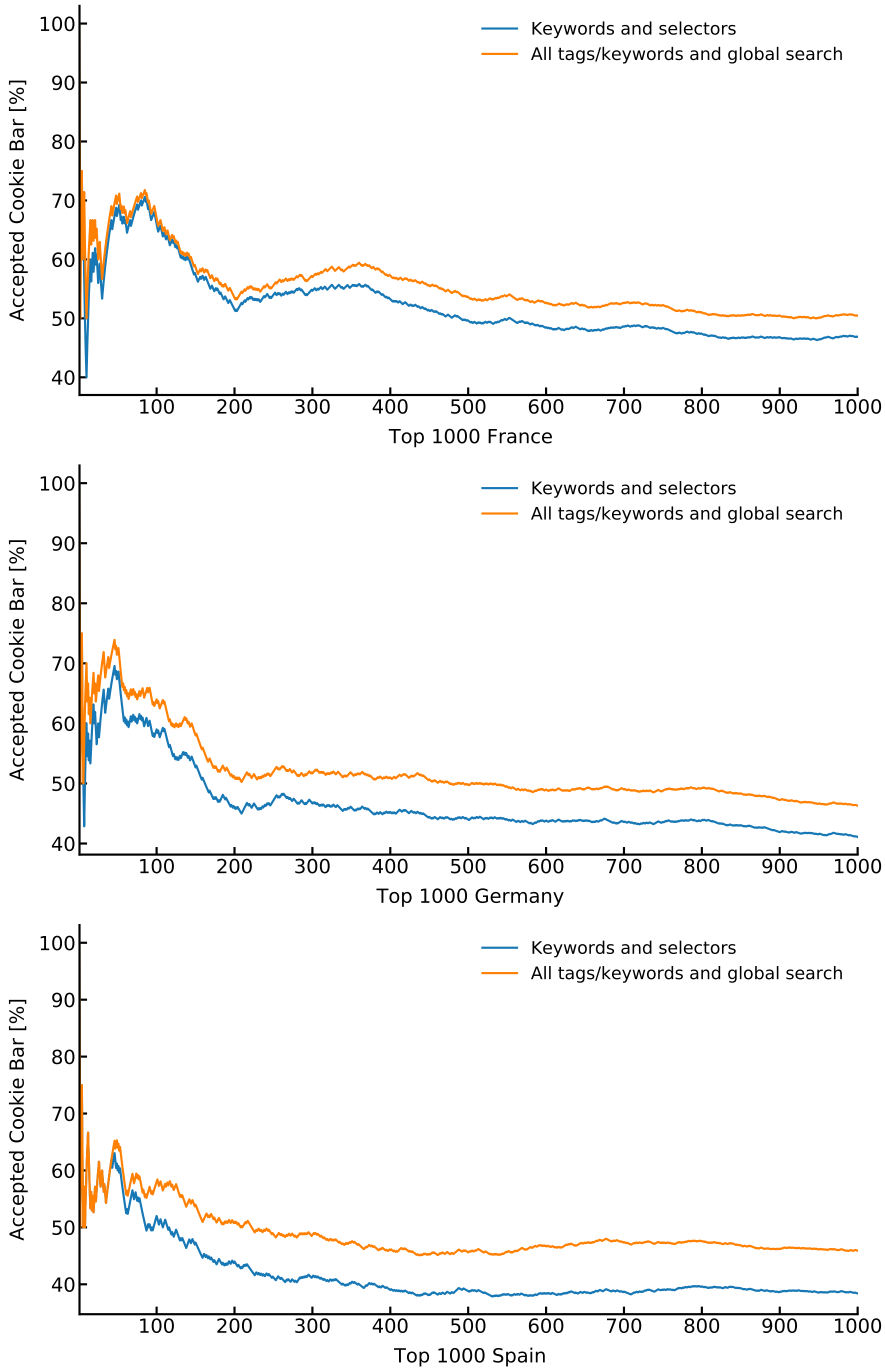


Figure 4.4: Percentage of accepted Cookie Bar on website ranking

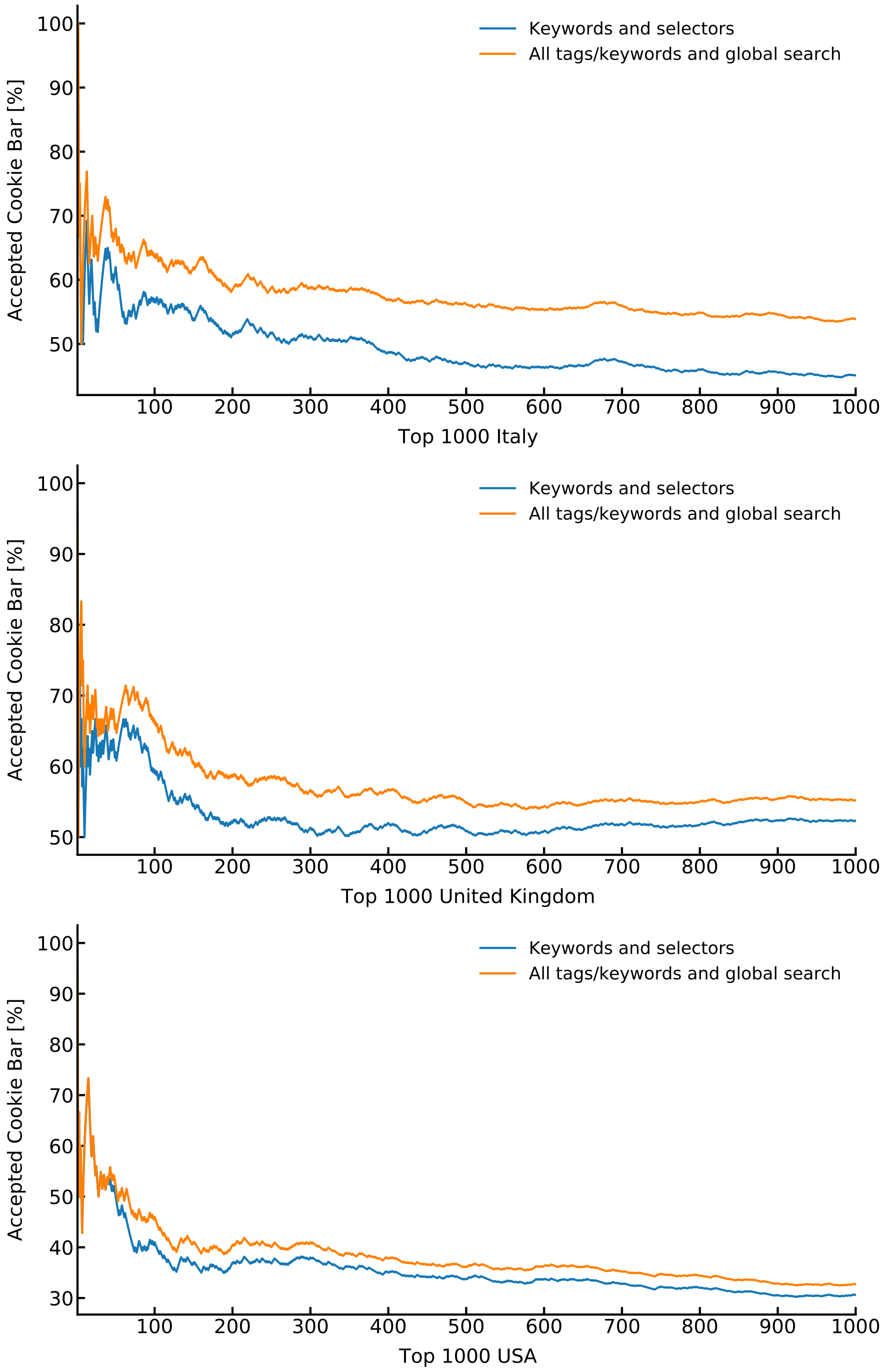


Figure 4.4: Percentage of accepted Cookie Bar on website ranking

Why does the script fail? As we can see in the plots of each country, initially the number of times the tool works correctly is higher and then the percentage

of success decreases. This effect has probably two reasons: The first one is that the choice of the keywords has been made by visiting the first sites in the ranking and the second one is that probably the less visited sites are also those with a lesser attention to the compliance with the cookie law.
To study in more detail what happens when the tool fails, we took a list of 100 websites for three of the countries considered. These website are selected randomly among those not taken into account for the keywords. From the screenshot taken by the tool we can see what happen when the script fail. The bar plot in Figure 4.5 shows that a significant part of the failures is due to the lack of a clickable Cookie Bar inside the site. The 10% of the failure is caused by the lack of the keywords in the keywords list and in another 10% the tool fails because of the websites are offline. The case other collect the same cases shown in Figure 4.3. The result is that the tool is able to provide consent in most cases. When the tool fails the main reason is the lack of a clickable Cookie Bar on the website while only a minor part of the failures is due to the lack of keywords on the list.

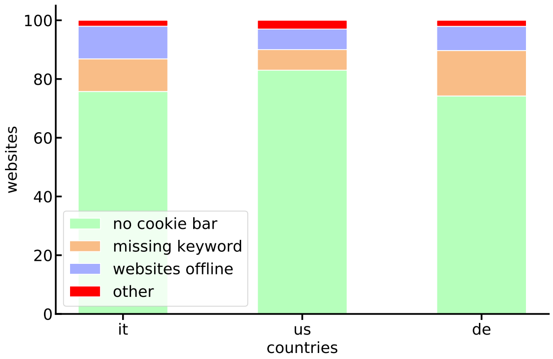


Figure 4.5: Per country result on 100 random websites that do not show a cookie bar

4.3 Data Collection

We crawled our tool on the first 1000 websites of the most visited websites in similarWeb ranking for each of the 6 considered countries. The web sites are divided in 24 different categories:

- Adult
- Arts and Entertainment

- Business and Consumer Services
- Community and Society
- Computers Electronics and Technology
- E-commerce and Shopping
- Finance
- Food and Drink
- Gambling
- Games
- Health
- Heavy Industry and Engineering
- Hobbies and Leisure
- Home and Garden
- Jobs and Career
- Law and Government
- Lifestyle
- News and Media
- Pets and Animals
- Referance Materials
- Science and Education
- Sports
- Travel and Tourism
- Vehicles

For each country and category, we have the 100 most popular websites that corresponding to 8362 unique websites to visit because the lists overlap. About 6% of visits failed due to websites being offline.

4.4 Tracker analysis after consent

We start by analyzing the third-party cookies on web pages before the user consent is given. For this analysis we consider also the cookie lifetime. Cookies can have an expiration date. Usually session cookies, that are removed when the browser is closed, do not have an expiration date specified. Other type of cookies like profiling cookies have an expiration date specified and cookies with an expiration date in the past will be deleted from the browser. Cookies with a long lifetime are more likely to be used for tracking purposes. Figure 4.6 shows the cumulative distribution of cookies' lifetimes. This distribution is computed over the third party cookies embedded on the 8362 website we visited. Also in this case we used Disconnect, EasyList and EasyPrivacy tracker lists. As shown only 10% of third party cookies last less than 1 day and 20% last less than one month. For the following analyses, we consider only cookies with a lifetime greater than 1 day

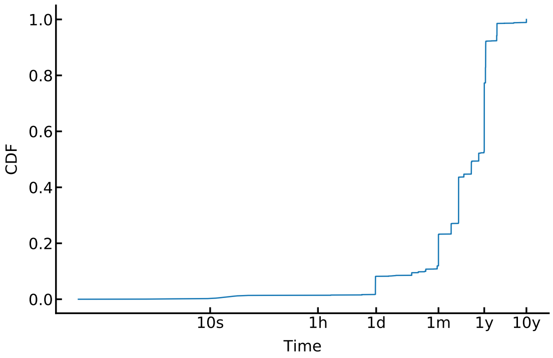
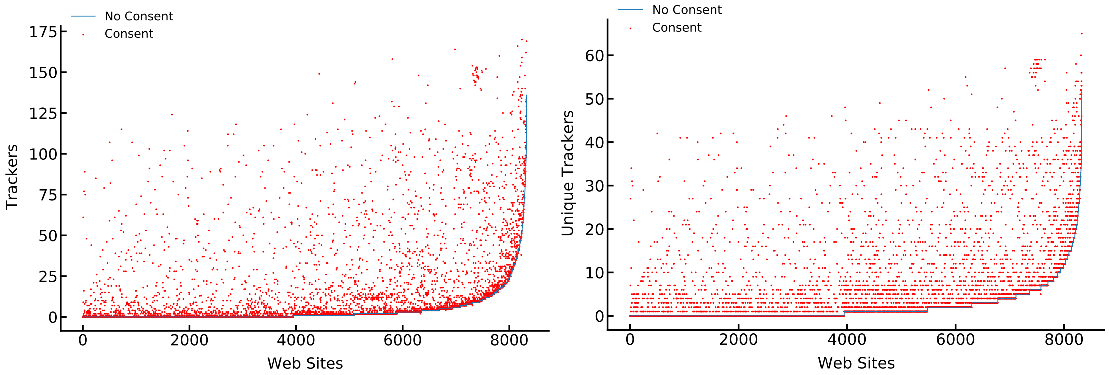


Figure 4.6: Cumulative distribution of lifetimes of third party cookies.

We consider the 8362 first party websites to analyze the number of third party cookies which are installed before and after we provide consent. Figure 4.7b represent the number of contacted third parties and Figure 4.7a the number of third parties cookie installed. The websites are sorted by the number of third party cookies installed before consent and are represented with a blue line. The red dots represent the situation after consent has been given. As shown, 53% of sites install third party cookies before obtaining consent and this percentage becomes 64% after obtaining consent. Even some of the first party websites that initially do not install or install a few cookies after obtaining consent install a very high number of cookies.

Now we want to see how these third party trackers are distributed among our websites. In Figure 4.8 we can see the number of third parties installing

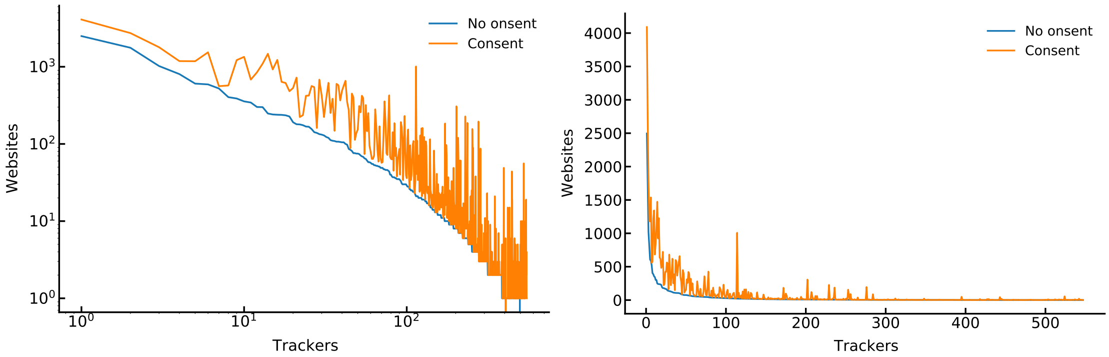


(a) Number of third party cookies embedded before and after giving consent.

(b) Number of third parties contacted before and after giving consent.

Figure 4.7: Third party cookies and third parties contacted before and after giving consent

cookies in these 8362 websites and how this number change after we provide consent. Third parties are ranked based on the number of websites they are contained in before consent.



(a) Logarithmic scale

(b) Linear scale

Figure 4.8: Third-party embedding cookies. Website are ranked based on the number of websites they are contained in before consent

In Figure 4.9 we focus on the top 10 third party installing cookies before and after consent. For example we can see that before consent is given doubleclick.net embeds its cookies on 30% of the website considered. After providing consent this tracker is present on 50% of the pages which is more than 4000 website. This means that 78% of websites that install third-party cookies install cookies from doubleclick.net.

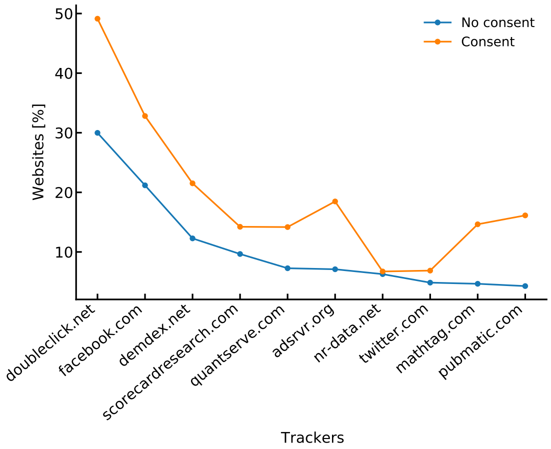


Figure 4.9: Rank of the 10 third-party embedding the largest number of cookies in websites. Tracker are ranked based on the number of websites they are contained in before consent.

The heat map in Figure 4.10 shows the situation before cookie consent. The figure shows the percentage of sites that install at least one third party cookie before consent. We can see that in many websites third party cookies are installed before obtaining users' consent. Law_and _government and science_and_education websites are the most compliant with the cookie law. While the News_and_media category is the one that, on average, complies less with the cookie law. in fact in many cases the sites of this category install cookies without waiting for the user consent. On average the 72% of News_and_media websites embed third party cookie before obtaining consent and this percentage reach 85% for the Italian news and media websites. There are also differences between countries. The websites in the top 1000 of Germany are those that install less cookies before obtaining the consent. In Figure 4.11 we can see how the number of sites installing third party cookies change after obtaining consent. The biggest variation, in line with what we observed above, occurs in the top 1000 of Germany while it is less significant for the other countries since most of them already install third party cookies.

The third of these heat map (Figure 4.12) instead shows the impressive increase in the average number of cookies installed on these sites. For some categories this number becomes even 10 times higher.

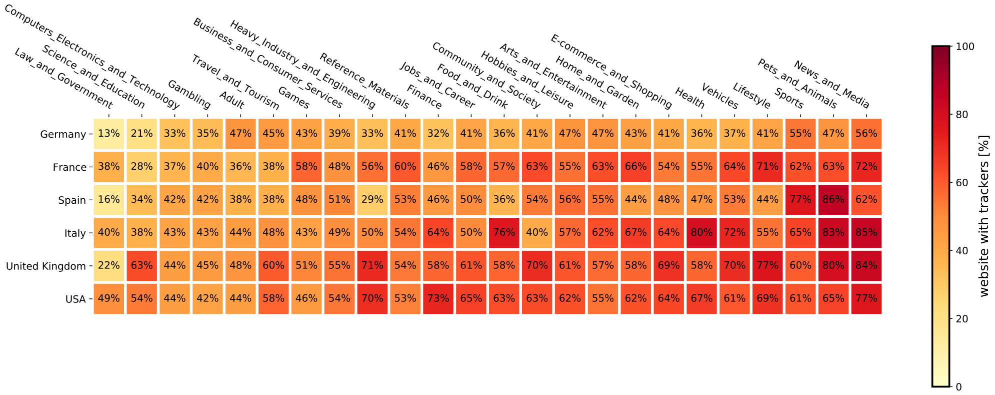


Figure 4.10: Per country and per category websites installing at least one third-party cookie

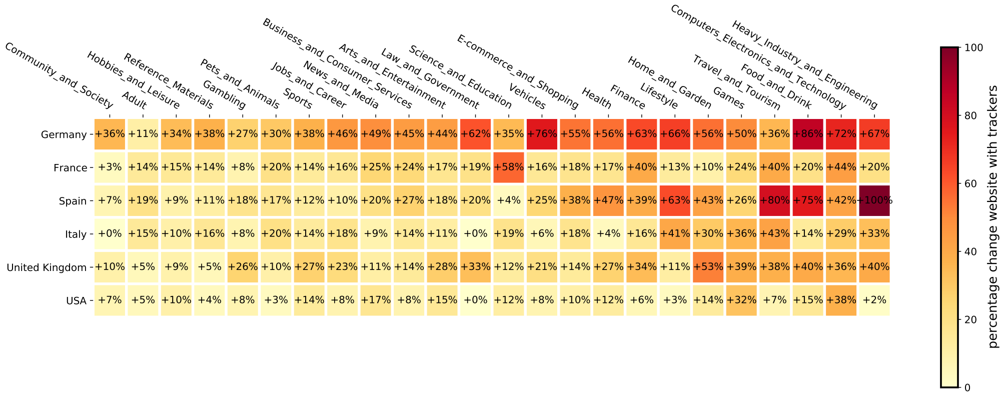
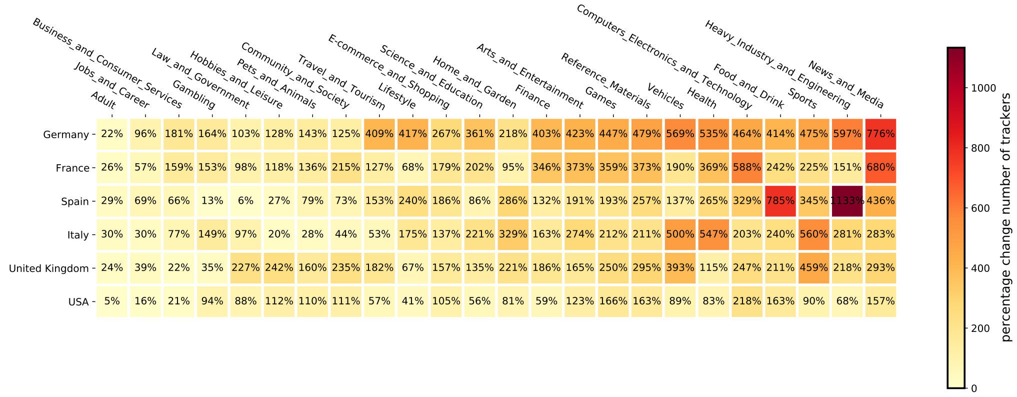


Figure 4.11: Increase in the number of sites with at least one third-party cookies

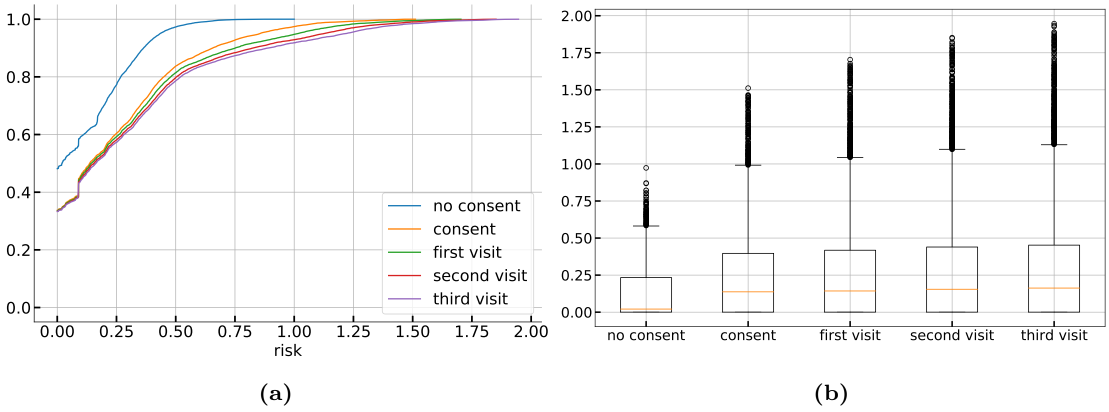


	Adult	Jobs_and_Career	Business_and_Consumer_Services	Gambling	Law_and_Government	Hobbies_and_Leisure	Pets_and_Animals	Community_and_Society	Travel_and_Tourism	Lifestyle	E-commerce_and_Shopping	Science_and_Education	Home_and_Garden	Finance	Arts_and_Entertainment	Games	Reference_Materials	Vehicles	Health	Computers_Electronics_and_Technology	Food_and_Drink	Sports	Heavy_Industry_and_Engineering	News_and_Media
Germany	22%	96%	181%	164%	103%	128%	143%	125%	409%	417%	267%	361%	218%	403%	423%	447%	479%	569%	535%	464%	414%	475%	597%	776%
France	26%	57%	159%	153%	98%	118%	136%	215%	127%	68%	179%	202%	95%	346%	373%	359%	373%	190%	369%	588%	242%	225%	151%	680%
Spain	29%	69%	66%	13%	6%	27%	79%	73%	153%	240%	186%	86%	286%	132%	191%	193%	257%	137%	265%	329%	785%	345%	1133%	436%
Italy	30%	30%	77%	149%	97%	20%	28%	44%	53%	175%	137%	221%	329%	163%	274%	212%	211%	500%	547%	203%	240%	560%	281%	283%
United Kingdom	24%	39%	22%	35%	227%	242%	160%	235%	182%	67%	157%	135%	221%	186%	165%	250%	295%	393%	115%	247%	211%	459%	218%	293%
USA	5%	16%	21%	94%	88%	112%	110%	111%	57%	41%	105%	56%	81%	59%	123%	166%	163%	89%	83%	218%	163%	90%	68%	157%

Figure 4.12: Increase in the number of embedded third-party cookies

4.5 Risk assessment after consent

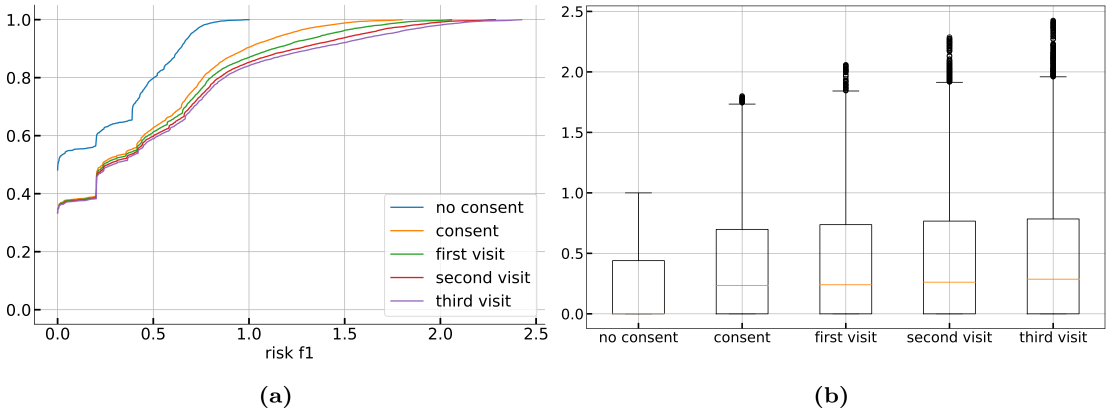
After we have analyzed how the web tracking environment varies when consent to the installation of cookies is given, in this last part we focus on the risk assessment as done in chapter 3. Figure 4.13a shows the cumulative distributions of the risk. The data has been normalized with respect to the risk values before consent. The 53% of first party websites considered before consent and the 64% of first party websites considered after consent embed at least one tracker, and therefore have a value risk greater than 0. The blue line that represent the risk distribution before we provide consent is shifted towards the left side of the graph, so towards small values of risk, in fact, we can see from the plot that before providing consent to the installation of cookies 80% of the first party websites have a risk value lower than 0.25 and only a small number, about 3%, of these website have a risk higher than 0.5. The situation changes when we consider the risk after giving consent to the installation of cookies. This situation is represented in the graph by the orange line. We can see that the orange line is shifted towards higher values of risk. In this case 60% of the first party websites have a risk value lower than 0.25 while about 17% of these website have a risk value higher than 0.5. The other lines represent the risk value when we visit other link in the page, this other visit are performed on random link in the page after we provided consent. Also in this case the line are shifted towards higher value so the risk increase for the considered websites but this increase is much smaller than the previous one because most of the trackers are contacted already after the first visit following the acceptance of cookies.



(a)

(b)

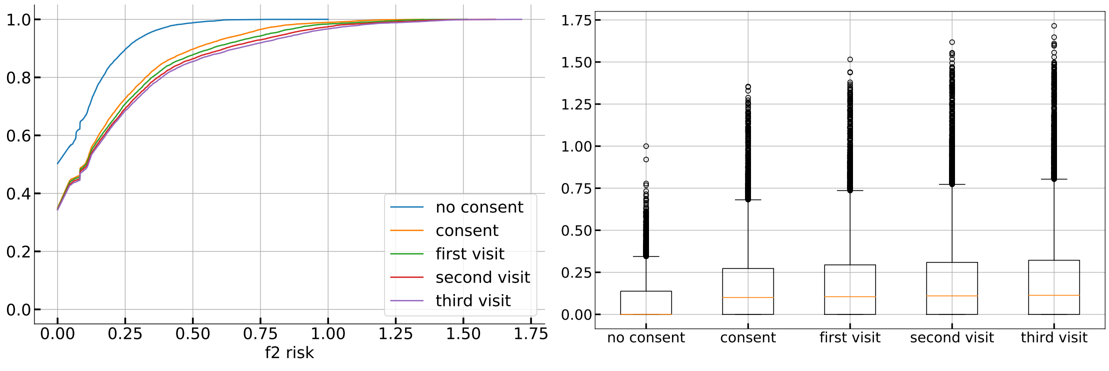
Figure 4.13: Risk value before and after giving consent



(a)

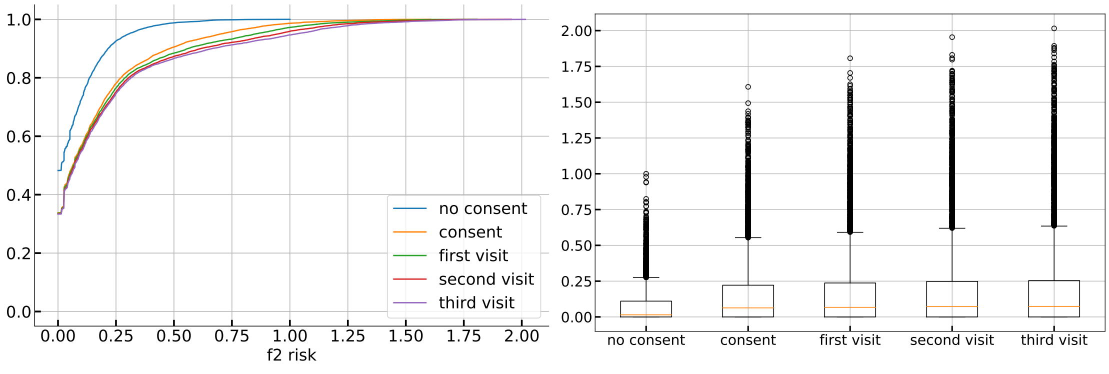
(b)

Figure 4.14: f_1 component before and after giving consent



(a) **(b)**

Figure 4.15: f_2 component before and after giving consent



(a) **(b)**

Figure 4.16: f_3 component before and after giving consent

Chapter 5

Conclusions and future works

The goal of this thesis was to study the tracking ecosystem and provide an evaluation of the privacy risk users face when visiting a web page. To do that we assigned an indicator of tracking risk to first party websites based on three different components: the popularity of the embedded trackers, the quantity of information exchanges with the third parties trackers and the quantity of information trackers can exchange about users.
For this task HTTP requests and trackers list provided by trackers-blocker was used. Using a subset of the HTTPArchive dataset, we analyzed how the risk associated with a website may change over time. The result obtained shows that using this definition of risk, the risk tends to decrease over the years. We can say that probably this effect is caused by the use of new and advanced tracking mechanisms and is also due to the implementation of Cookie Bars on sites that require user consent to the installation of cookies.
In order to evaluate the effect of Cookie Bar on the presence of trackers, we designed a simple tool to automatically provide consent to the installation of cookies. We showed that the tool is effective at automatically accepting the Cookie Bar and we analyzed, using this tool, 8362 website from Similarweb ranking. The results obtained show that about 47% of sites incorporate cookies from trackers before obtaining some kind of consent. After consent, the number of third-party trackers contacted increases with 64% of sites incorporating cookies from trackers and the average number of third-party cookies installed per page increases from 4 to 12. As a result of this, the risk of tracking increases.
Based on the results presented in this thesis, a possible future work could be to use a larger dataset to better represent the monitoring ecosystem. For example,

it might be possible to improve the results of the tool by expanding the list of keywords and also considering websites from other EU and non-EU countries. Additional studies may also be needed to better define the risk of tracking, for example by taking into account the history and preferences of the individual user and thereby providing a more accurate indication of individual risk, improving his awareness of web tracking and helping him to protect his personal data.

Bibliography

[1] *HTTPArchive*. URL: https://httparchive.org/ (cit. on pp. 9, 18).

[2] Peter Eckersley. «How Unique is Your Web Browser?» In: vol. 6205. Jan. 2010, pp. 1–18. ISBN: 978-3-642-14526-1. DOI: 10.1007/978-3-642-14527-8_1 (cit. on p. 13).

[3] *California Consumer Privacy Act (CCPA)*. Tech. rep. URL: https://www.oag.ca.gov/privacy/ccpa (cit. on p. 13).

[4] *Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws.* Tech. rep. URL: http://eur-lex.europa.eu/legalcontent/en/TXT/?uri=CELEX:32009L0136 (cit. on p. 14).

[5] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gurses, Frank Piessens, and Bart Preneel. «FPDetective: Dusting the web for fingerprinters». In: Nov. 2013, pp. 1129–1140. DOI: 10.1145/2508859.2516674 (cit. on p. 15).

[6] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. «Detecting and Defending Against Third-Party Tracking on the Web». In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX Association, Apr. 2012, pp. 155–168. ISBN: 978-931971-92-8. URL: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/roesner (cit. on p. 15).

[7] *BigQuery*. URL: https://cloud.google.com/bigquery (cit. on p. 18).

[8] *Har 1.2 spec*. Tech. rep. URL: http://www.softwareishard.com/blog/har-12-spec (cit. on p. 18).

[9] *Simplified Arrangements to Provide Information and Obtain Consent Regarding Cookies - 8 may 2014.* Tech. rep. URL: `https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/3167654` (cit. on p. 26).

[10] *I don't care about cookies.* URL: `https://www.i-dont-care-about-cookies.eu/` (cit. on p. 27).

[11] *Selenium Web Browser Automation.* URL: `http://www.seleniumhq.org/` (cit. on p. 27).

[12] *ChromeDriver - WebDriver for Chrome.* URL: `https://chromedriver.chromium.org/` (cit. on p. 27).

[13] *Similarweb.* URL: `https://www.similarweb.com/` (cit. on p. 27).