Computers & Security 122 (2022) 102873

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Web-tracking compliance: websites' level of confidence in the use of information-gathering technologies

David Martínez [a,*], Eusebi Calle [a], Albert Jové [b], Cristina Pérez-Solà [c,d]

[a] Institute of Informatics and Applications, Universitat de Girona, Girona, Spain
[b] Universitat Oberta de Catalunya Spain
[c] K-riptography and Information Security for Open Networks, IN3, Universitat Oberta de Catalunya Spain
[d] CYBERCAT-Center for Cybersecurity Research of Catalonia Spain

ARTICLE INFO

Article history:
Received 2 February 2022
Revised 22 July 2022
Accepted 8 August 2022
Available online 9 August 2022

Keywords:
web tracking
cookies
web beacons
level of confidence
privacy
compliance

ABSTRACT

With the emergence of new technologies and the generalized use of social media, corporations have an invested economic interest in employing web-tracking techniques, but there is also the issue of protecting users' privacy. In this field of research, only few articles introduce methods to evaluate website compliance in the use of current web-tracking techniques. Moreover, evaluating the level of compliance requires, in the majority of cases, manually implementing an extensive data analysis. In this paper, we present four new algorithms (CIA, CDA, BDA, and SCA) and a novel measure (WLoC) to evaluate user tracking compliance in websites and the level of confidence in the use of information-gathering technologies, by employing the recently published Website Evidence Collector (WEC) software from the European Data Protection Supervisor (EDPS). The paper also showcases a case study of the top 500 websites most visited by Alexa in Spain to evaluate the performance of the presented algorithms and metrics. Results reveal a novel procedure for obtaining categorized websites' compliance and confidence levels of a set of websites under the current European legislation, thus updating and enhancing some of the previous research work.

1. Introduction

The Hypertext Transfer Protocol (HTTP) defines how the modern Internet works, giving rise to the websites as we know them today (Fielding et al., 1999). Cookies provide additional pieces of information exchanged in each communication between web browsers and web servers, incorporating the concept of "state" to the HTTP protocol (Kristol, 2001). This enables web servers to remember, or distinguish users, and maintain sessions, which is essential for the proper functioning of modern websites. However, it also introduces other controversial usages such as web tracking through so-called "tracking cookies" or "third-party cookies", which have become the leading information-gathering technology, transparent to average users (Gomer et al., 2013). The usage of web beacons, which is a more recent web-tracking and information-gathering technology, is rising as they are more challenging to detect and neutralize (Sipior et al., 2011).

In the research field, the quintessential Warren and Brandeis (1890) formally introduced the right to privacy. The concept of privacy had been evolving and was finally accepted legally and morally; albeit generating discussion on an ethical level (Moor, 1991). Websites have been causing concern in the field of user privacy since the formation of the Internet. In particular, the Internet Explorer browser has allowed the systematic use of tracking cookies since its incorporation in 1995. Tracking cookies and web beacons significantly affect users' privacy, as they are used to gain personal data and other relevant information (Lin and Loui, 1998).

Because corporations have a vested economic interest in using web-tracking techniques, regulation of the technology has become essential (Acquisti et al., 2016). User data has become the main asset for big corporations, even forming brand new economic models such as the "personal data economy" (PDE) and "pay-for-privacy" (PFP) (Elvy, 2017), substantially increasing their economic impact. Although regulation is always a step behind the emergence of new technologies (Marchant et al., 2009), government in-

* Corresponding author.
E-mail addresses: david.martineza@udg.edu (D. Martínez), eusebi.calle@udg.edu (E. Calle), ajove@uoc.edu (A. Jové), cperezsola@uoc.edu (C. Pérez-Solà).

https://doi.org/10.1016/j.cose.2022.102873

stitutions around the world are making a special effort to limit the impact on users' privacy of web-tracking technologies. One of the most advanced data privacy regulations is the European Union's (EU) General Data Protection Regulation (GDPR), effective from May 25th, 2018. The Spanish Data Protection Agency (SDPA) has also developed a comprehensive updated guide to cookie usage (SDPA, 2020), offering guidance on how to fulfill informed consent requirements for the collection of user device data according to the GDPR. Although this guide is made specifically for Spanish legislation, it has to be consistent with European regulations, and therefore we have used this guide as a compliance reference (see Appendix Appendix A).

Nevertheless, Hu and Sastry (2019) found that the leap forward in data privacy regulation provided by the GDPR has not resulted in a material reduction in long-term numbers of web tracking. The authors speculate that users may be exhausted from having to choose their privacy preferences on every website visited. In the end, users accept the default choices offered by the websites by simply turning on the tracking. Furthermore, the authors concluded that the websites that request user consent appear to set even more cookies after gaining that consent than those who do not request it. Although the majority of average users are not even aware of the web-tracking behaviors, there is a growing awareness among users of the need to protect their privacy. It is expected that privacy protection will become increasingly harder as the technology becomes more widely deployed (Khalifa et al., 2011).

The logic of web-tracking technologies works silently on users' web browsers. Nevertheless, in general terms, users trust that the information provided by the website is correct and trust that their consent decisions will be respected. The basis of that trust is the confidence that data privacy regulations provide the users with, although there is no easy way to prove it. Consequently, web tracking techniques present two sides of the same coin regarding user awareness: i) visible elements, defined by clearly noticeable elements such as how the websites inform users about the usage of web-tracking technologies and obtain user consent (e.g., cookie consent banners); and ii) non-visible elements, defined by hidden information-gathering technologies being silently executed on the user's web browsers. Unquestionably, the socioeconomic impact of information-gathering technologies and their effect on users' privacy is critical enough to study the relation between these visible and non-visible behaviors.

In this paper, we combine and analyze both behaviors of visible and non-visible elements into a single research framework, as the relation between them has not been previously studied. Our analysis of non-visible elements includes not only the detection of tracking cookies but also web beacons and browser fingerprinting, as a difference from existing works, which are the most commonly used web-tracking methods at the present time (Sanchez-Rola et al., 2017). Moreover, we present a new measure with which to evaluate the Websites Level of Confidence (WLoC) of a set of websites in the use of information-gathering technologies obtained from the result of the analyses of both element behaviors. Websites typically provide compliant visible elements (such as a valid informed consent collection), although the actual behavior or internal website logic (non-visible elements) executes web-tracking or information-gathering techniques regardless of the provided user consent. Hence, there is a need for the users to know if websites truly meet their data privacy choices and, therefore, provide confidence. Given a sample of websites, we present several automatic and semi-automatic algorithms to perform the analyses and formally introduce the WLoC measure. Furthermore, we provide the first published academic results based on the Website Evidence Collector (WEC) software. The WEC is a novel and powerful tool published by the European Data Protection Supervisor (EDPS) that supports the automation of privacy and personal data protection inspections of websites (EDPS, 2021).

The rest of the paper is organized as follows. Section 2 summarizes the review of related literature. Section 3 presents the methodology followed, including a description and the algorithms used on the visible (Section 3.1) and non-visible (Section 3.2) elements analyses, the definition of the novel WLoC measure (Section 3.3), and the website sample categorization procedure (Section 3.4). Then, in Section 4, we introduce our case study (Section 4.1) and illustrate the results, the effectiveness and discuss the approach described in the paper, including the results of the visible (Section 4.2) and non-visible (Section 4.3) elements analyses, together with the observed confidence (Section 4.4). Finally, Section 5 summarizes the paper and provides guidelines for future work.

2. Review of related literature

To the best of our knowledge, Degeling et al. (2018) is the first article that measures the GDPR impact on web privacy. They categorize and analyze the behavior of the 500 most popular websites for each EU member state, including how websites collect user consent. They conclude that the web has become more transparent since the GDPR entered into force, in terms of informing users about web-tracking behaviors, whereas there are still relevant limitations on user mechanisms to consent or deny the processing of their data. Although their study provides an in-depth analysis, it only focuses on the visible elements and does not analyze the actual non-visible behavior.

The study of Aladeokin et al. (2017) provides a more practical approach for non-visible elements and evaluates compliance by considering some of the Commonwealth countries and their privacy protection laws (PERC for the United Kingdom, PIPEDA for Canada, PoPI for South Africa, and the Information Technology Act for India). The authors provide a methodology to detect the usage of tracking cookies, collecting user browser data, and analyzing each website's traffic. In their study, the websites are considered as compliant with the Commonwealth countries privacy protection laws if they do not set cookies, or they either set cookies informing the user and obtain consent or provide an option to opt-out before setting the cookies. However, they do not consider the GDPR and the use of alternative web-tracking methods such as web beacons or browser fingerprinting.

Third-party cookies are arguably the most extended technique for web tracking (Sanchez-Rola et al., 2017), especially after the end of Flash Cookies or Local Shared Objects (LSO). LSO are pieces of data that websites that use Adobe Flash may store on a user's computer), the usage of which has become limited after the Adobe Flash Player discontinued support in 2020. However, users have become more aware of the use of tracking cookies (Jayakumar, 2021), and modern web browsers provide increasingly effective countermeasures against them. Therefore, alternative web-tracking methods such as web beacons and browser fingerprinting have become popular. Web beacons are used on web pages to unobtrusively allow checking that a user has accessed some content, and browser fingerprinting is a powerful method that websites use to collect information about users' browsers. Users or browser protections may disable cookies, although web beacons and browser fingerprinting techniques are highly challenging to detect, allowing websites to bypass browser protections.

As stated, we combine and analyze both behaviors of visible and non-visible elements into a single research framework to evaluate compliance and, in addition, the confidence of a website sample through the novel WLoC measure. We consider a website com-

pliant if it is in agreement with our legislation background (see Appendix Appendix A).

3. Methodology

This section first presents the analytical algorithms to monitor a website's usage of information-gathering technologies (visible and non-visible elements analyses). The novel Websites Level of Confidence (WLoC) measure is also defined. Moreover, the presented website categorization is justified, and the techniques used to categorize the website sample are introduced. This paper introduces the analytical algorithm descriptions, while their details and implementation can be consulted on the dedicated external public repository (Martínez, 2021).

3.1. Visible elements analysis: Consent collection

Websites require explicit user consent in order to use web-tracking technologies, i.e., either cookies or alternative information-gathering technologies. Visible elements analysis is focused on the user's initial selection, usually obtained through cookie banners. Cookie banners are a website's most common consent collection mechanism, which are currently evolved to include not only the use of cookies, but also other web-tracking technologies. Second-level cookie banners are a common way for websites to gain explicit user consent. They are composed of: (i) first-level banner, where the legally required cookie usage information is available to the user alongside an accept button, a "Cookie Settings" or similar button that links to the second-level banner, and occasionally a reject button; and (ii) second-level banner, where the user can customize cookie options to accept the desired ones, save their selection, or reject all non-essential ones.

Visible elements consider six forms of gaining user consent (Figure 1): a) *without option*, where websites only inform about cookie usage without asking for any consent; b) *confirmation*, where the users are informed about cookie usage and asked for acceptance without any option to reject them; c) *binary*, where the users are informed and asked for consent with an option to reject non-essential cookies; d) *selection*, where the user is able to select which cookies, grouped by types or purposes, they want to accept or reject; e) *slider*, where the user is able to select which cookies, grouped by types or purposes, they want to accept or reject hierarchically; and f) *full control*, where websites provide the users with the ability to accept or reject highly atomic groups of cookies (ten or more provided groups) and even individual cookies. It is worth noting that, usually, forms d), e), and f) may be provided on a second-level cookie banner.

The Consent Inspector Algorithm (CIA) provides a semi-automatic procedure to capture clear images (i.e., website screenshots) of the website's cookie banners. Each website may be classified into one of the previously defined forms of gaining user consent through a manual inspection of the website screenshots. The CIA gains access through an automatized browser (with its default configuration) to each website. Once the website is loaded, the CIA captures a screenshot of the site's main page. Then, it detects second-level cookie banners by identifying clickable buttons matching predefined strings from a string dictionary, including, e.g., "Cookie Settings" or "Personalize" (example available at Martínez (2021)). If it detects a second-level banner, the CIA clicks the button and captures another screenshot (Figure 2). The second screenshot provides the information on the second-level cookie selection, which is usually needed to identify the form of gaining user consent.

3.2. Non-visible elements analysis: Identifying consentless web-tracking techniques

The proposed non-visible elements analysis detects the defined behavior and usage of web-tracking techniques on websites. Two algorithms process the data gained from running the Website Evidence Collector (WEC) software (EDPS, 2021). After the WEC execution, we store the "inspection.json" output file that saves all the retrieved data in JSON format, including the use of web-tracking techniques before the user consent selection (i.e., consentless web tracking). This file will be subsequently processed by the Cookies Detector Algorithm (CDA) and the web Beacons Detector Algorithm (BDA). Therefore, the accuracy of the results relies on the WEC inspection data, trusted by the European Data Protection Supervisor (EDPS). The WEC and the algorithms use filter lists (the well-known "easyprivacy" and "fanboy-annoyance"), also called rule lists, which contain rules to determine what should be blocked and hidden on websites (often used by ad blockers). Both CDA and BDA algorithms provide simple and effective processing of the WEC data to extract tracking cookies, web beacons, and browser fingerprints.

3.2.1. Cookies Detection Algorithm (CDA)

First, we identify and extract the tracking cookies used on websites without the users' explicit consent from the WEC execution output (i.e., before any user consent selection). The CDA implements a simple procedure to precisely filter the tracking cookies from the obtained WEC cookies. Figure 3 shows a randomly selected website that displays the aspect of a tracking cookie detected by the WEC, including detailed technical information about the cookie and its attributes. The cookie technical data show the name, value, domain, expiration timestamps, size, security parameters, and a large log stack for each detected cookie. In particular, the "domain" and "expires" attributes are essential to detect tracking cookies, as the domains of tracking services are well known and the expiration time of tracking cookies is extensive, frequently two years or more.

The Cookies Detector Algorithm (CDA) provides an automatic procedure to categorize the cookies that websites use in the browsers without user consent. From the WEC execution output, the algorithm identifies the tracking cookies applying the same "easyprivacy" and "fanboy-annoyance" filter lists used by the WEC software to detect other tracking-related elements. These filter rules are applied to: i) the cookie origin URLs, which usually target tracking scripts; and ii) the cookie origin domains (Figure 4). In essence, the WEC software provides the data, and the CDA identifies and extracts the cookies that can be classified as tracking elements.

3.2.2. Web Beacons Detection Algorithm (BDA)

The WEC provides an extensive analysis of the web beacons usage, including all the elements suspected to be part of tracking behaviors. Figure 5 shows an example of the web beacons detected by the WEC for a randomly selected website from our sample (adoptauntio.es). Two beacons have been shown to be detected and matched the well-known filter lists of *fanboy-annoyance* and *easyprivacy*. In particular, the first web beacon detected appears to be a simple API call of the browser to get a "back to top" image. However, it shows that an unnecessary query parameter is added to the request, composed of a hash value and suspected for tracking purposes. The second web beacon detected is a script called "gtm.js", downloaded from Google Tag Manager (GTM) website, generally used for tracking purposes.

The Web Beacons Detector Algorithm (BDA) not only extracts the WEC detected beacons, but also identifies scripts and detects

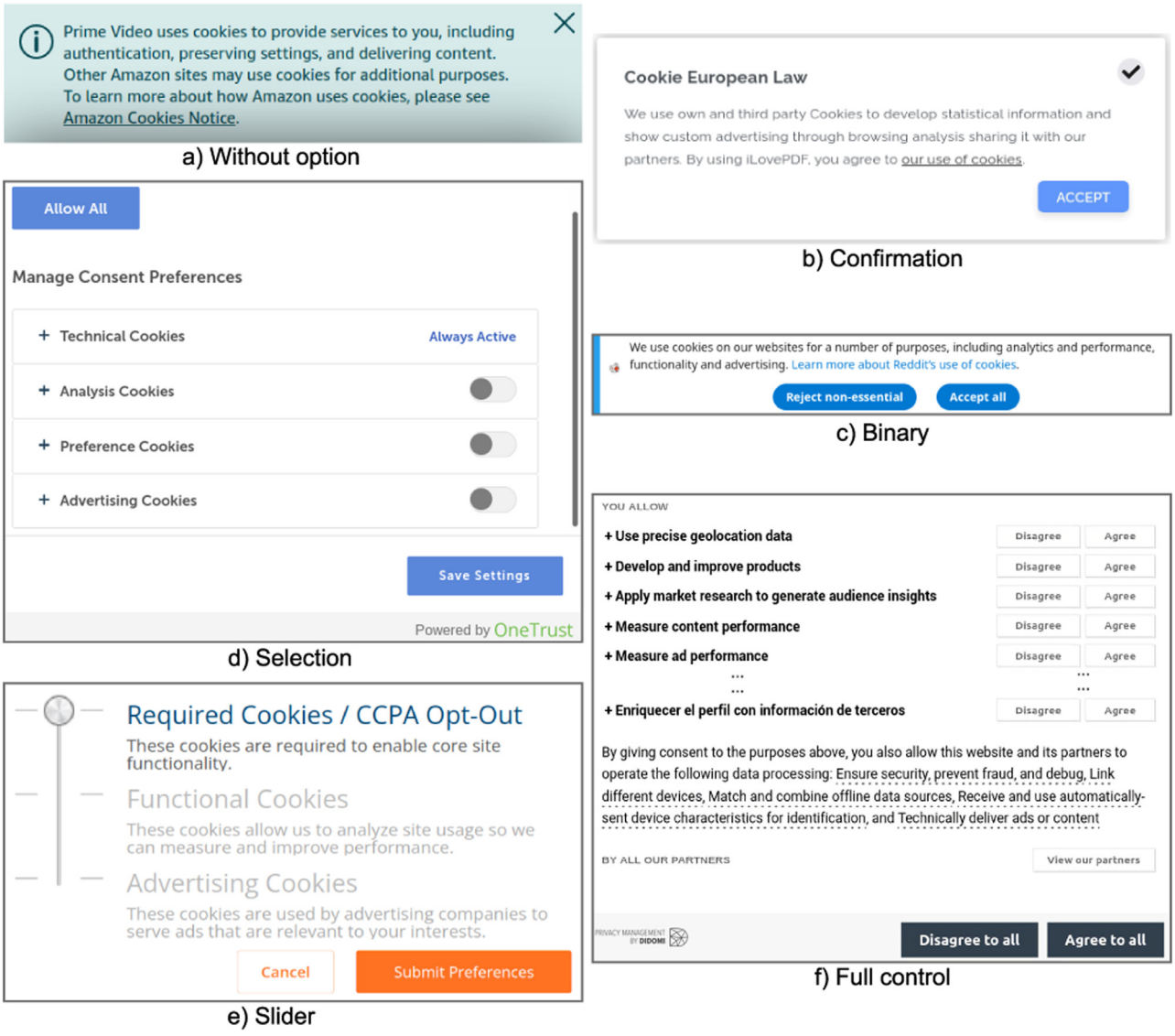


Fig. 1. Forms of gaining user consent.

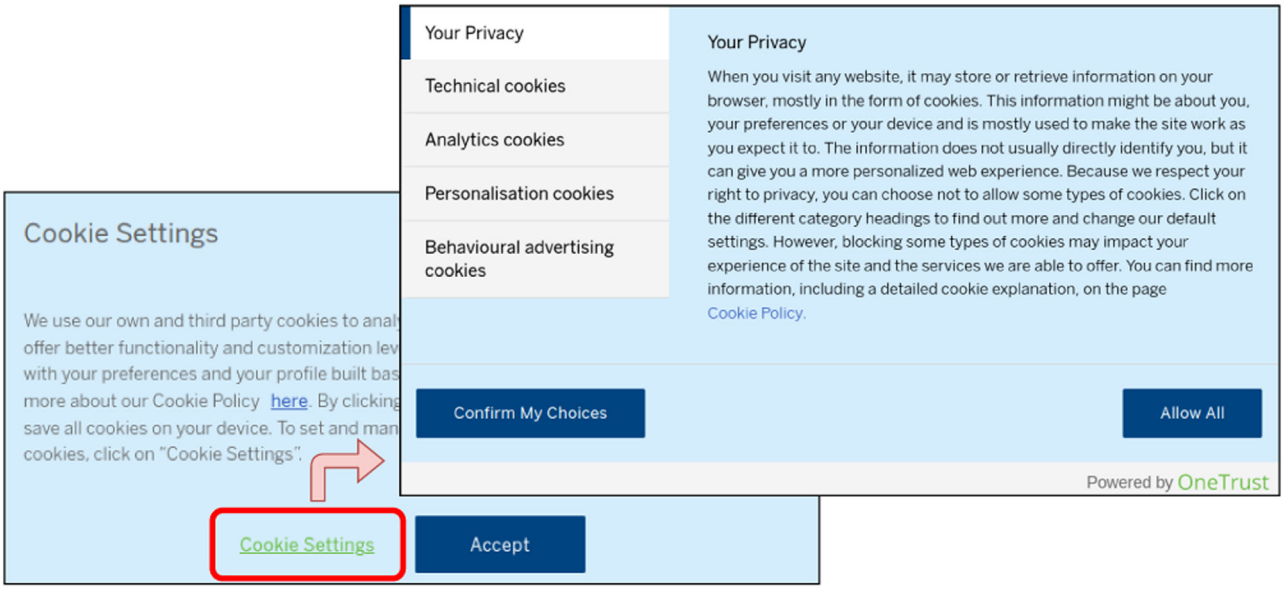


Fig. 2. Detection of second-level cookie banners (bbva.com/es).

```
{
  "name": "g__gdic","value": "kox0e0s8uanfmylj51","domain": "
      advanced-web-analytics.com","path": "/","expires": 1684594096
      ,"size": 25,"httpOnly": false,"secure": false,"session":
      false,"expiresUTC": "2023-05-20T14:48:16.000Z","expiresDays":
       730,
  "log": {
    "stack": ["..."],
    "type": "Cookie.JS",
    "timestamp": "2021-05-20T14:48:16.332Z",
    "location": "https://bancsabadell.com"
  },
  "firstPartyStorage": false
}
```


Fig. 3. Cookie detected by the WEC on a website arbitrarily selected as an example (bancsabadell.com).

Fig. 4. Steps of the cookies detector algorithm.


```
"beacons": [
  {"url": "https://s.adoptauntio.es/es/www/img/v3/commons/pictos/
      back_to_top.svg?238d24b8d724ac703bb5ca06458cd78c","query": {"
      238d24b8d724ac703bb5ca06458cd78c": null},"filter": "/
      back_to_top.","listName": "fanboy-annoyance.txt","log": {"
      stack": [{"fileName": "https://www.adoptauntio.es/","source":
       "requested from https://www.adoptauntio.es/ and matched with
       fanboy-annoyance.txt filter /back_to_top."}],"timestamp": "2
      021-05-20T16:52:30.607Z"},"occurrances": 2},

  {"url": "https://www.googletagmanager.com/gtm.js?id=GTM-NSQFRH2",
      "query": {"id": "GTM-NSQFRH2"},"filter": "/gtm.js","listName"
      : "easyprivacy.txt","log": {"stack": [{"fileName": "https://
      www.adoptauntio.es/","source": "requested from https://www.
      adoptauntio.es/ and matched with easyprivacy.txt filter /gtm.
      js"}],"timestamp": "2021-05-20T16:52:29.617Z"},"occurrances":
       1}
]
```


Fig. 5. Web beacon detected by the WEC on a website arbitrarily selected as an example (adoptauntio.es).

browser fingerprinting techniques. In Iqbal et al. (2020), the authors propose a robust algorithm that detects browser fingerprinting behaviors. On the one hand, they provide a list of well-known browser fingerprinting script MD5 hashes. Our algorithm downloads the web beacon scripts identified by the WEC and compares them with the hash list provided by Iqbal et al. (2020). On the other hand, the authors provide a list of JavaScript API keywords frequently used by fingerprinting scripts, measuring the relative prevalence of API keywords by computing the ratio, where a higher value of the ratio for a keyword means that it is more prevalent in fingerprinting scripts than non-fingerprinting scripts. Our algorithm also detects those JavaScript API keywords with a ratio higher than 95 on Iqbal et al. (2020) from the downloaded beacon scripts identified by the WEC. Figure 6 shows the steps of the web beacons detector algorithm.

3.3. Websites Level of Confidence (WLoC)

It is expected that websites hide the actual tracking behavior from the users, hence the comparison between visible and non-visible element analyses on a set of websites could reveal the level of confidence that websites deliver to them. The study of this confidence may show which websites, apparently complying with the consent collection requirements (visible elements), are really

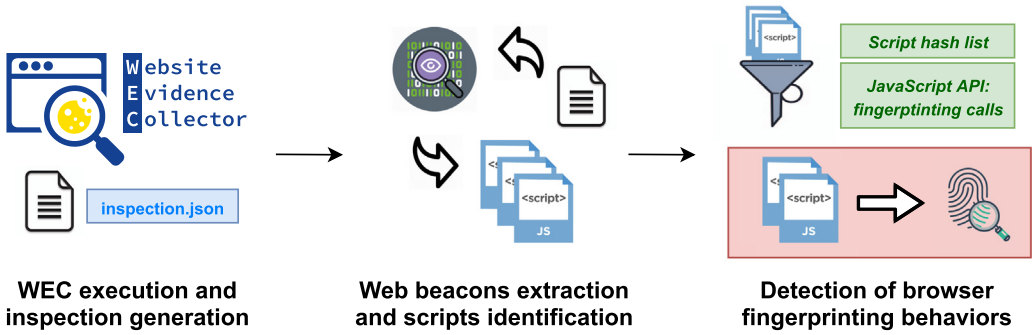


Fig. 6. Steps of the web beacons detector algorithm.

Table 1
Proposal of WLoC qualitative assessment.

WLoC value	Qualitative assessment
WLoC $\geq$ 0.8	The sample degree of compliance confidence is high. A large percentage ($\geq$ 80 %) of the websites request user consent properly, hence do not execute web-tracking techniques regarding users without previous consent.
0.8 >	The sample degree of compliance confidence is suboptimal. In practice, between 20 % and 40 % of the websites that apparently comply with user consent collection execute web-tracking techniques without user consent invisibly.
0.6 >	The sample degree of compliance confidence is poor. In practice, approximately half of the websites that apparently comply with user consent collection execute web-tracking techniques without user consent invisibly.
0.4 >	The sample degree of compliance confidence is very poor. The sample is unreliable as, in practice, most websites ($\geq$ 60 %) that apparently comply with user consent collection execute web-tracking techniques without user consent invisibly.

enforcing them in practice (non-visible elements). Therefore, we present the so-named Websites Level of Confidence (WLoC) measure to formally evaluate the level of non-compliance of a sample. This novel measure not only represents a normalized numeric value quantifying the level of compliance, but also enables us to compare the results obtained in this paper with other research work or future research. Although it is a quantitative measure, Table 1 shows a proposal of qualitative assessment, which assists in interpreting the WLoC measure results.

The WLoC measure is calculated based on the visible and non-visible analyses over the same sample of websites. If any website presents no results or results for only one of the analyses (i.e., some of the analyses fail), it has to be removed from the sample. Mathematically, we define the sample of websites as M. Then, four additional sets are obtained based on the results of the analyses:

- A: Set of websites that do not require user consent, according to the analysis of visible elements ($A \subset M$).
- A_o: Set of websites that do not require user consent, according to the analysis of visible elements, and that the use of web-tracking techniques is not detected in the analysis of the non-visible elements ($A_o \subset A$).
- B: Set of websites that properly require user consent, according to the analysis of visible elements ($B \subset M$, $B \cap A = \emptyset$).
- B_o: Set of websites that properly require user consent, according to the analysis of visible elements, and that the use of non-compliant web-tracking techniques is not detected in the analysis of the non-visible elements ($B_o \subset B$).

As the WLoC measure evaluates confidence, the visible element analysis for non-compliant websites ($M - (A \cup B)$) is not considered (i.e., websites that are not compliant with consent collection). On the basis of the defined sets, the WLoC measure is computed by dividing the union of the sets A_o with B_o by the union of the sets A and B. Mathematically, the measure is defined as:

$$\text{WLoC} := \frac{|A_o \cup B_o|}{|A \cup B|} \quad (1)$$

3.4. Sample Categorization Algorithm (SCA)

On the Internet, we can find boundless types of content. Hence, it is clear that a governmental website behaves differently to a news or streaming website. For this reason, we defined a semi-automatic procedure to categorize websites by topics, where each website is classified in one or more topic categories, to provide more accurate results. Given a sample of websites, we present the simple but effective Sample Categorization Algorithm (SCA). The SCA consists of downloading each website's content (HTML code) and automatically extracting its categorization based on strings placed on the website domain name and the HTML description or keywords metadata. The algorithm loads the string sequences from a dictionary for each defined category, and is fully customizable (an example of SCA string dictionary is available on the dedicated public repository). It is worth noting that the quality of those string dictionaries will affect the efficiency of the algorithm and also contribute to reducing the number of false positives. Figure 7 shows the complete categorization progress stages. A manual inspection will be necessary on those websites where the algorithm cannot extract the categories automatically.

4. Results and discussion

In this section, we introduce our case study and illustrate the results, including those from the visible (Section 3.1) and non-visible (Section 3.2) elements analyses, together with the observed confidence (Section 4.4), the effectiveness and discuss the approach described in the paper. Additionally, the applicable regulations background for our case study is presented in Appendix Appendix A.

4.1. Case study

The usefulness of the presented algorithms and measures is illustrated in the top 500 websites most visited by Alexa in Spain (Alexa Internet, 2021). From those, we categorized 464 websites and discarded the rest as they were inaccessible (36 offline websites, 7.2 % of the original sample). Concerning the categorization, we consider the following 19 categories (adapted from

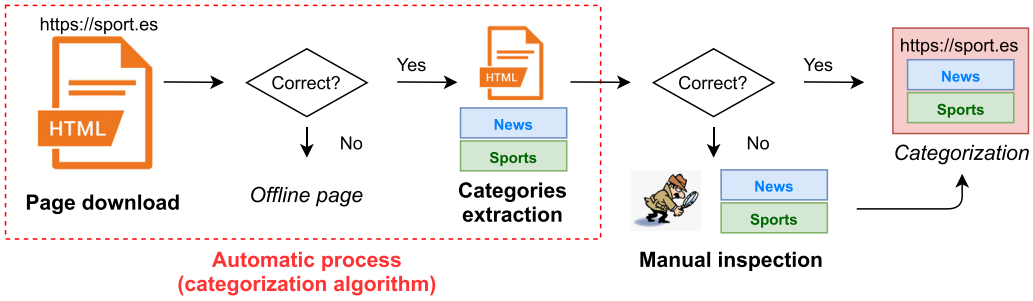


Fig. 7. Categorization process stages.

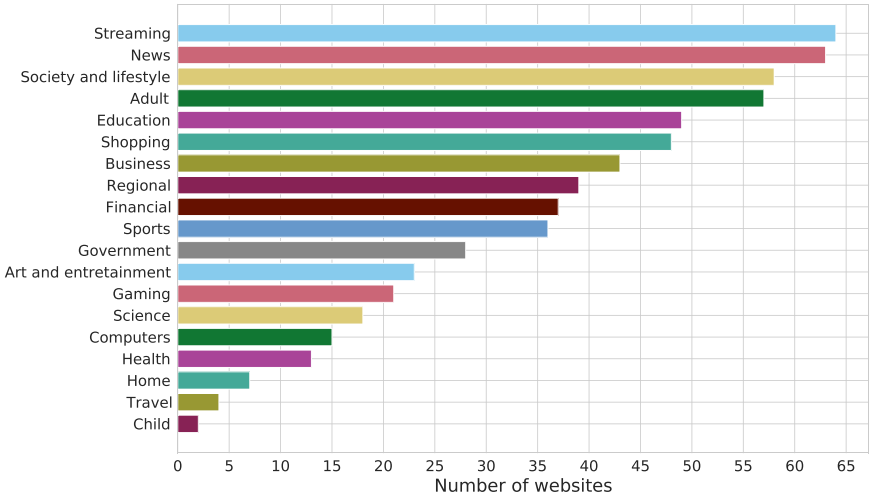


Fig. 8. Frequency of categories in the sample of websites.

Hu and Sastry (2019)): adult; art and entertainment; home; science; shopping; education; sports; financial; government; child; gaming; business; news; computers; regional; health; society and lifestyle; streaming; and travel.

The categorization of the 464 website sample resulted in 355 websites categorized automatically by the algorithm (76.5 %) and 109 websites inspected manually (23.5 %), which is remarkable considering the task complexity. Figure 8 shows the frequency of the categories in the sample. Overall, 625 categorizations have been extracted where streaming (64), news (63), society and lifestyle (58), and adult (57) categories stand out with more than 50 websites each.

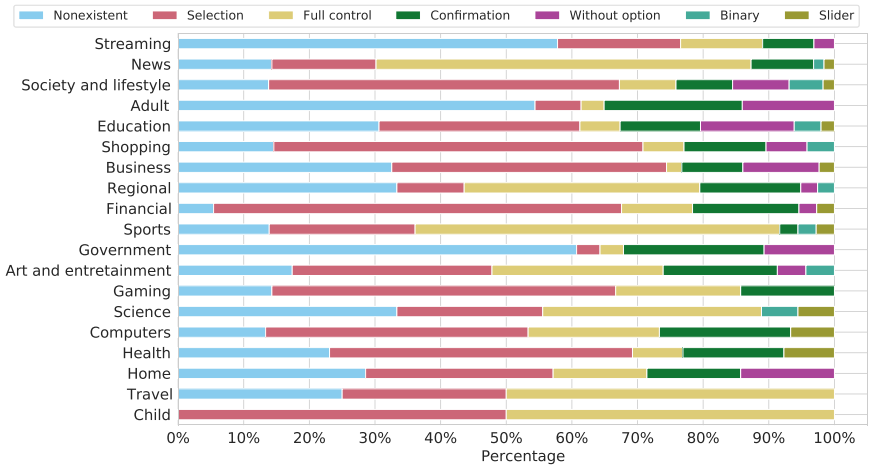
The procedure to obtain the confidence (WLoC) of the defined website sample is as follows: first, the Consent Inspector Algorithm (CIA) is computed on the websites acquiring the results of the analysis of the visible elements. Then, the Cookies Detector Algorithm (CDA) and the Web Beacons Detector Algorithm (BDA) are computed over the sample obtaining the results of the analysis of the non-visible elements (i.e., consentless web tracking techniques). Finally, the novel WLoC measure of the website sample is calculated through the visible and non-visible analysis results (CIA, CDA, and BDA).

4.2. Consent Inspector Algorithm (CIA) results

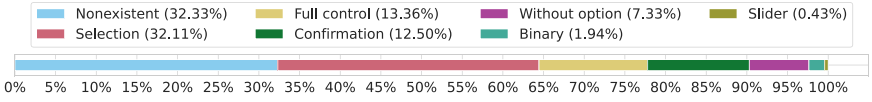
This section presents the results of the analysis of visible elements, in particular from the execution of the Consent Inspector Algorithm (CIA) over the sample. Figure 9 illustrates the observed forms of gaining explicit user consent in terms of the use of web-tracking techniques aggregated by categories and their average distribution. It is interesting to note that the websites categorized as "government", "streaming", and "adult" follow a distinct pattern compared to the rest. They pointed out a high probability of not presenting any form of user consent (greater than 50 %). In the case of the "streaming" and "adult" categories, the reason for this lack of user consent collection is the relation of their websites with content of questionable legality. However, this is not that clear in the case of "government" categorized websites because apparently, they should be the most compliant and trusted ones. We observed that this situation is caused by two factors: (i) the non-use of web-tracking techniques, which is detected by the CIA; and (ii) the exemption of user consent collection requirements for the public administration. It is worth noting that observations are linked to user consent as technical, considering the GDPR's "valid consent" concept (see Appendix Appendix A).

In general, the rest of categories follow similar patterns close to the average distribution, which can be seen in Figure 9b. Concerning the average distribution, consent management does not exist in almost one-third (32.33 %) of the analyzed websites. However, it is worth noting that this does not mean they are not compliant. Some websites may not use tracking cookies or alternative information-gathering technologies or use exempted cookies without tracking purposes (i.e., essential or session cookies). The two last categories shown in Figure 9a are unrepresentative due to their low number of websites (less than five).

The applicable regulations background of this paper presented in Appendix Appendix A restricts websites that present non-essential cookie selections as default. Hence, and according to the Spanish Data Protection Agency guide to cookie usage (SDPA, 2020), only "binary", "selection", "slider", and "full control" forms of user consent with non-essential cookie selections disabled by default are considered compliant. Considering the subset of websites with these forms of user consent (47.84 % of the whole sample), it is remarkable that 17.57 % of them are consid-



(a) Aggregated by categories.



(b) Average distribution.

Fig. 9. Consent Inspector Algorithm (CIA) results.

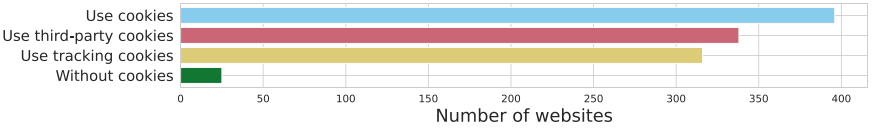


Fig. 10. Number of websites that use cookies without user consent.

ered non-compliant as they present non-essential cookies selected as default. Keeping aside the websites with nonexistent consent management, almost one-fifth (19.83 %) of the sample websites are non-compliant with consent collection.

4.3. Identifying consentless web-tracking techniques

This section depicts the results of the analysis in the case of non-visible elements (i.e., consentless web-tracking techniques). Considering the 464 websites available from the analysis of the visible elements, the Website Evidence Collector (WEC) tool was able to analyze a total of 421 websites ($\approx$ 91 %). The other 43 websites could not be analyzed by the WEC tool due to either being offline at the moment of the analysis or providing effective protection against robots.

4.3.1. Cookies Detector Algorithm (CDA) results

The execution of the Cookies Detector Algorithm (CDA) over the WEC analyzed websites proves that the use of tracking cookies is widespread (see Figure 10). More than three-quarters (75.06 %) of the websites use tracking cookies without previous user consent. This percentage increases even further if all the third-party cookies are considered as tracking cookies (80.29 %, including the third-party cookies that the algorithm could not classify as tracking cookies). Overall, almost all the WEC analyzed websites (94.06 %) use some kind of cookies without previous user consent.

It is also interesting to evaluate the previous data aggregated according to the website categories. Figure 11 shows the percentage of cookie types (tracking cookies and other types) aggregated by categories, wherein no category shows the use of tracking cookies below 60 %. All the observed websites of the categories “child”, “travel”, “home”, and “health” use tracking cookies without consent, and the rest follow similar patterns. Only two categories, “society and lifestyle” and “financial”, present a percentage of tracking cookies below 65 %.

The vast majority of the websites analyzed by the WEC use tracking cookies illegitimately. Figure 12 illustrates a histogram of the number of tracking cookies detected for each website. The distribution presents a mean of $\mu = 6.88$ tracking cookies per website, represented in the figure with a dotted line, with the maximum value of 60 being from the “ukr.net” website. Most websites use between 0 and 5 tracking cookies, and then the frequencies reduce exponentially until 30 tracking cookies. However, two slight increases can be seen from 35 to 50 tracking cookies and from 55 to 60; figures that are alarming indeed. These increases were caused by the top 10 websites that use the most tracking cookies (see Figure 13) i.e., all of them use more than 35 tracking cookies each. The first three positions of the top 10 are occupied by websites that present between 55 and 60 tracking cookies, followed by 53 in the case of fourth position. The other top ten positions show a gradual but consistent reduction from 46 to 39 tracking cookies.

In terms of the domains that possess the tracking cookies, Figure 14 illustrates how the control of tracking cookies is monopolized by large multinational companies, as only a few domains control most of the cookies detected. The strongest domain is “doubleclick.net”, which is Google’s advertising services. The second and fourth domains at the top, “google.com” and “youtube.com”, are also owned by Google, presenting a total of 122 websites that use their tracking cookies without previous consent (almost a third of the entire group of websites analyzed).

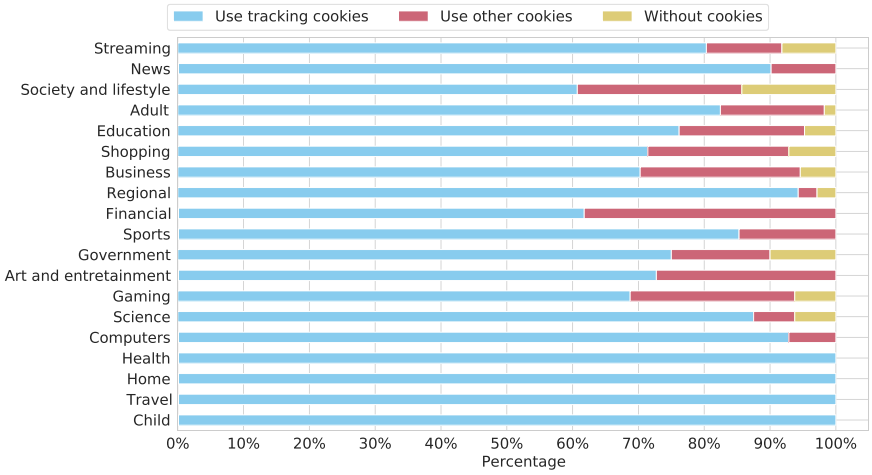


Fig. 11. Number of websites that use cookies without user consent aggregated by categories.

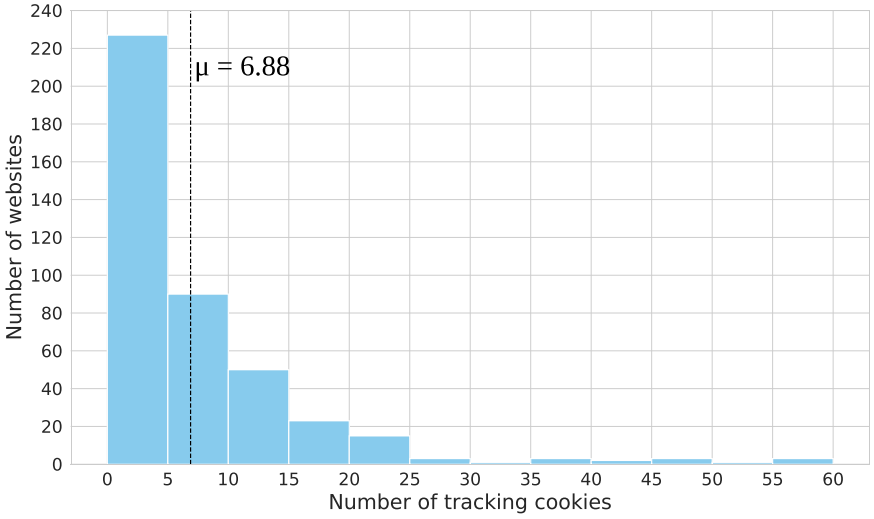


Fig. 12. Number of tracking cookies per website histogram.

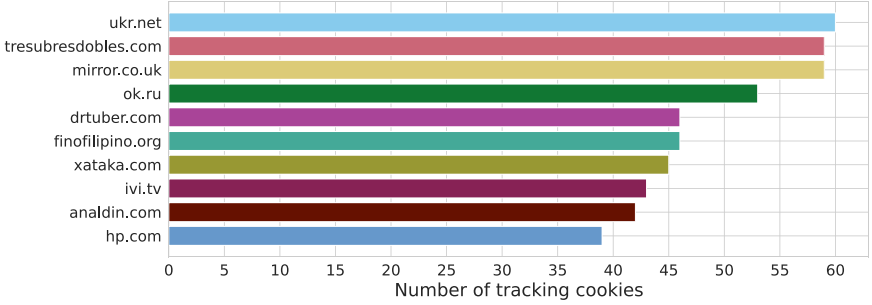


Fig. 13. Top 10 websites that use the greatest number of tracking cookies.

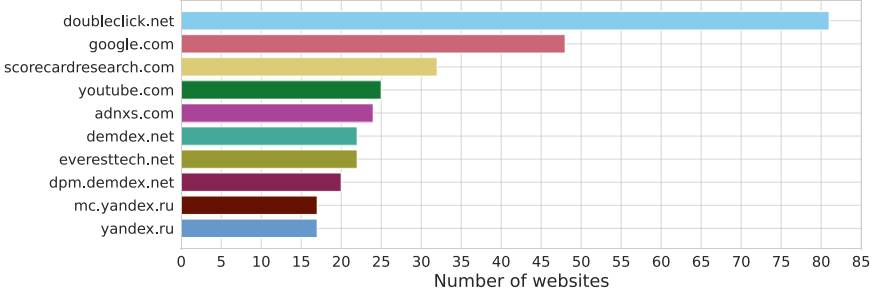


Fig. 14. Top 10 domains that own the greatest number of tracking cookies.

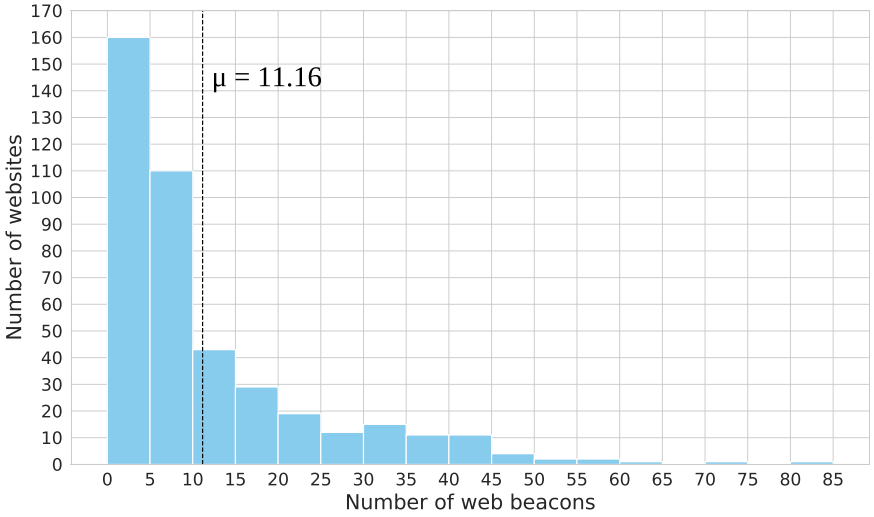


Fig. 15. Number of web beacons per website histogram.

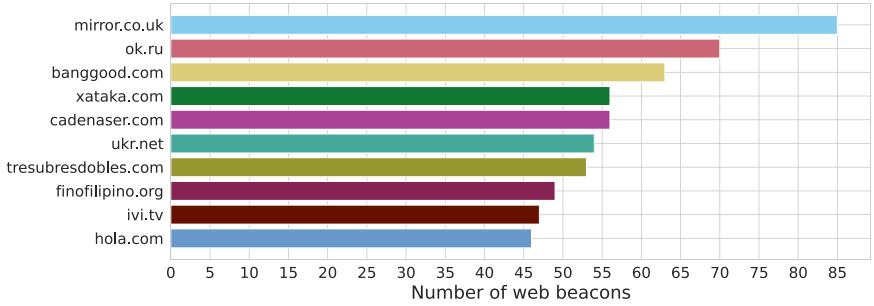


Fig. 16. Top 10 websites that use the greatest number of web beacons.

4.3.2. Web Beacons Detector Algorithm (BDA) results

The execution of the Web Beacons Detector Algorithm (BDA) illustrates that the use of web beacons is even more widespread than the use of tracking cookies, with almost all the WEC analyzed websites (90.26 %) use web beacons. Figure 15 illustrates a histogram of the number of web beacons detected for each website. The distribution presents a mean of $\mu = 11.16$ web beacons per website, represented in the figure with a dotted line, where the maximum value of 85 is from the "mirror.co.uk" website, and the majority of websites use between 0 and 10 web beacons. The histogram generates an exponential distribution similar to the tracking cookies histogram, although with an increased mean number of web beacons per website. The top 10 websites that use the most tracking cookies, which can be seen in Figure 16), show that all of them use more than 45 web beacons. The top three positions are occupied by websites that present more than 60 web beacons each. The other positions in the top 10 show a constant reduction from 56 to 46 web beacons.

In terms of the domains that own the web beacons, Figure 17 illustrates how the control of web beacons is also monopolized by the same owners as the tracking cookies. The web beacons that Google owns, which are the ones from the domains "www.google-analytics.com", "www.googletagmanager.com", "stats.g.doubleclick.net", and "www.googleadservices.com", are present in 268 websites, i.e., 63.66 % of the WEC analyzed websites.

Additionally, the BDA verified that 44 websites (10.45 %) use browser fingerprinting techniques. From these, 36 have been detected from the JavaScript API calls that the website scripts execute. The remaining 16 websites have been obtained from comparing the files MD5 hashes with the list obtained from Iqbal et al. (2020). Both JavaScript API calls and MD5 hash matches have been detected on six of these websites.

4.4. Websites Level of Confidence (WLoC) results

The visible and non-visible element analyses ably reveal a website's confidence in terms of web-tracking compliance, through the so-named Websites Level of Confidence (WLoC) measure (see Section 3.3). A total of 464 websites from the sample were processed successfully in the analysis of visible elements (i.e., consent collection) in contrast with the analysis of non-visible elements (i.e., consentless web-tracking techniques) where only 421 websites were able to be processed.

The WLoC measure is performed on the website sample M and the previous analyses results (see Equation 1). The sets A and B are obtained from the consent collection analysis, where $|A| := 134$ and $|B| := 169$. The consentless web-tracking technique's results the websites of A and B result in the subsets A_o and B_o, where $|A_o| := 18$ and $|B_o| := 9$. From this data, we obtain a final value for the WLoC measure of 0.0891 (8.91 %).

The obtained WLoC value clearly states a nearly nonexistent level of confidence, categorized as "very poor" according to our proposal of qualitative assessment from Table 1. This value means that only 8.91 % of the websites that passed the consent collection analysis successfully also passed the consentless web tracking techniques analysis. The rest of the websites use tracking techniques without user consent, either because the consent collection is nonexistent or because it exists and, although it is apparently compliant, tracking techniques are used in any case.

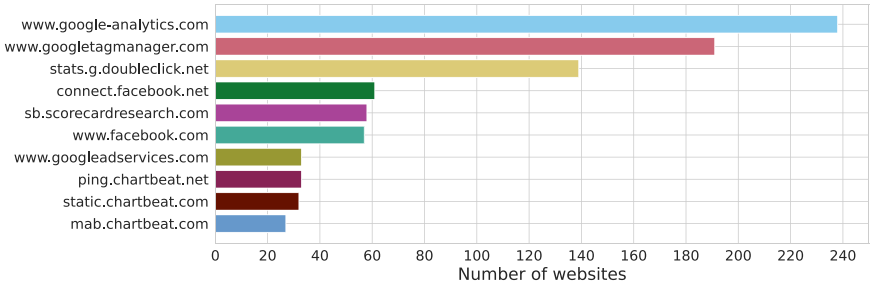


Fig. 17. Top 10 domains that own the greatest number of web beacons.

Considering the subsets A_o and B_o, only 6.41 % of the website sample are fully compliant. This value is calculated considering the compliant websites (i.e., subsets A_o and B_o) over the total number of analyzed websites (421).

5. Conclusions and future work

This paper presents new algorithms, methods, and metrics to evaluate the level of web-tracking compliance considering current European legislation, including tools such as the novel Web Evidence Collector recently published by the European Data Protection Supervisor (EDPS). Compliance is evaluated by taking into account not only what websites apparently intend to do (visible elements analysis), but also what they actually do (non-visible elements analysis).

The previously-mentioned tools are tested in a case study and highlight several remarks. Websites categorized as "government", "streaming", and "adult" deliver a high probability (greater than 50 %) of not presenting a form of user consent, mainly caused due to the non-use of web-tracking techniques and the exemption of user consent collection requirements for the public administration. Our Consent Inspector Algorithm (CIA) pointed out that 17.57 % of the total analyzed websites are considered non-compliant due to non-essential cookies selected as default, and roughly one-fifth of the sample websites present an invalid user consent form. Moreover, our Cookies Detector Algorithm (CDA) revealed that almost all WEC analyzed websites (94.06 %) use cookies without previous user consent, with more than three-quarters (75.06 %) using tracking cookies. CDA identifies a mean of 6.88 tracking cookies per website, while most websites use between 0 and 5 tracking cookies, and also illustrates that the control of tracking cookies is monopolized by large multinational companies such as "Google". Concerning the results of our Web Beacons Detector Algorithm (BDA), a similar distribution as the CDA is accomplished with a higher mean number of 11.16 web beacons per website. BDA endorses the monopoly of web tracking as web beacons are controlled by the same owners of tracking cookies, for instance, "Google" web beacons are present in 63.66 % of the websites. In addition, another interesting result is that browser fingerprinting techniques have been detected on a total of 44 websites (10.45 %).

In this paper, we have also presented a novel measure to evaluate and categorize confidence levels of website samples in terms of web tracking called WLoC. Furthermore, this metric allows for the monitoring and comparing of different case study confidence levels. The main conclusion is that only 8.91 % of the websites that properly collect user consent are successfully enforcing it in practice, a figure categorized as "very poor" according to our qualitative proposal.

The presented algorithms are prepared to be easily adapted to different regulations. Moreover, the WLoC measure is applicable independently of the pertinent privacy regulation. Consequently, it would be interesting, as future work, to consider the applicability of the proposed algorithms to other privacy regulation frameworks. Furthermore, it would be interesting to monitor the WLoC measure for the Spanish Alexa's top 500 websites in order to find out if it gradually improves with websites adapting to the new regulations or, otherwise, remains the same or even worsens.

Declaration of Competing Interest

There is no conflict of interest

CRediT authorship contribution statement

David Martínez: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Eusebi Calle:** Formal analysis, Methodology, Writing – review & editing, Supervision. **Albert Jové:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Cristina Pérez-Solà:** Formal analysis, Methodology, Resources, Supervision, Writing – review & editing.

Acknowledgements

UdG researchers thank Red temática Go2Edge (Ref.: RED2018-102585-T) and Ajut PontUdG2020/23 for their partial funding. This work was also supported in part by the Spanish Government under Grant RTI2018-095094-B-C22 "CONSENT", and is based in the first author's Master's thesis from the Cybersecurity and Privacy Master's degree from the Universitat Oberta de Catalunya.

Appendix A. Regulations

In our case study, we use the described methodology to evaluate the compliance of the top 500 websites visited by Alexa using European legislation with respect to the use of cookies. The European legislation establishes a common framework for cookies regulation, that can be further particularized by the member States. As stated in Section 1, we use the Spanish Data Protection Agency guide to cookie usage (SDPA, 2020) as a compliance reference, which offers guidance on how to fulfill informed consent requirements for the collection of user device data according to the GDPR.

A1. Applicable regulations

The regulation of cookies in the EU is given by the 2002 directive on privacy and electronic communications (ePrivacy directive, updated in 2009) European Union (2002). Moreover, whenever cookies are used to collect personal data, as is usually the case, they will also be affected by the General Data Protection Regulation of 2016 (RGPD) European Union (2016).

These regulations not only apply to cookies, but also to other technologies that store or retrieve data from a terminal equipment of a natural or legal person who uses a service of the information

society, whether it is for professional reasons or not. Such technologies include (and are not limited to) local shared objects, flash cookies, web beacons, bugs or fingerprinting techniques.

A2. Informed consent

Article 5.3 of the ePrivacy directive establishes the requirement to obtain the express consent of the user when storing or recovering information in the terminal equipment of users. This obligation applies thus to cookies. The requirements of the consent to be valid must be interpreted as established by the GDPR: it must be free, specific, informed and unequivocal (i.e., the user has accepted the use of cookies through a clear affirmative action).

A3. Exceptions to consent

Consent is not mandatory in all cases. ePrivacy in its article 5.3, allows cookies to be exempt from the requirement of informed consent if they meet any of the following criteria:

- Criterion A: the cookie is used *for the sole purpose of transmitting a communication through an electronic communications network*.
- Criterion B: the cookie is *strictly necessary for the provider of an information society service to provide a service expressly requested by the subscriber or user.*

A cookie that satisfies the exemption criteria must have a lifespan directly related to the purpose for which it is used and it must be programmed to expire once it is no longer necessary. This implies that cookies that meet criteria A and B will probably expire at the end of the browsing session.

However, if these cookies are also used for non-exempt purposes (for example, for behavioral advertising purposes), they will no longer be exempt and the consent of the interested party must be obtained.

A4. Requirements for informed consent

For consent to be considered informed, the Article 29 Working Party [1] recommends the use of layered privacy statements or notices, so that the user is allowed to go to those aspects of the statement or notice that are of greatest interest to them, thus avoiding information fatigue, without prejudice to the fact that all the information is available in a single place or in a complete document that can be easily accessed by the interested party. Therefore, this system may display the essential information in a first layer, when the page or application is accessed, and complete the information in a second layer, by means of a page that offers more detailed and specific information about cookies.

Consent may be made effective through various formulas. One approach is by clicking on a section that states “I consent”, “I accept”, or any other similar sentence. Consent may also be obtained by inferring it from an unequivocal action carried out by the interested party. In any case, clear and accessible information has to be provided about the purposes of the cookies and whether they are going to be used by the same editor and/or by third parties.

It is important to stress that consent must be given before the use of cookies: cookies can not be installed until the user takes a valid action expressing their agreement. Moreover, in no case the mere inactivity of the user (or the option to ‘continue browsing’) will be considered a valid form of consent. Neither will the examination of a second information layer (in the case that the information is presented by layers) or the inspection of the cookies preferences section.

[1] As of 2018 the Article 29 Working Party ceased to exist and has been replaced by the European Data Protection Board (EDPB).

A5. Consent revocation

Consent management platforms must allow the user to withdraw it easily, that is, withdrawing consent must be as easy as granting it. However, this paper will focus on how consent is granted and whether websites comply with users’ preferences. We will leave for further work, verifying the existence and suitability of the consent withdrawal system, as well as other legal obligations such as the accessibility of information about cookies within the website, the legal notice or the privacy policy.

References

Acquisti, A., Taylor, C., Wagman, L., 2016. The economics of privacy. Journal of economic Literature 54 (2), 442–492.

Aladeokin, A., Zavarsky, P., Memon, N., 2017. Analysis and compliance evaluation of cookies-setting websites with privacy protection laws. In: 2017 Twelfth International Conference on Digital Information Management (ICDIM). IEEE, pp. 121–126.

Alexa Internet, 2021. *Top Sites in Spain*. Top 500 Alexa most visited websites in Spain.https://www.alexa.com/topsites/countries/ES. Extracted 22th April 2021.

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., Holz, T., 2018. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. arXiv preprint arXiv:1808.05096.

EDPS, 2021. *EDPS Inspection Software: Website Evidence Collector*. Programari d'extracci d'evidȩncies de les pgines web.https://edps.europa.eu/edps-inspection-software_en.

Elvy, S.-A., 2017. Paying for privacy and the personal data economy. Colum. L. Rev. 117, 1369.

European Union, 2002. Directive 2002/58/ec of the european parliament and of the council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications). Official Journal of the European Union L201 37–47.

European Union, 2016. 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, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union L119 1–88.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T., 1999. Hypertext transfer protocol–http/1.1.

Gomer, R., Rodrigues, E.M., Milic-Frayling, N., Schraefel, M., 2013. Network analysis of third party tracking: User exposure to tracking cookies through search. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Vol. 1. IEEE, pp. 549–556.

Hu, X., Sastry, N., 2019. Characterising third party cookie usage in the eu after gdpr. In: Proceedings of the 10th ACM Conference on Web Science, pp. 137–141.

Iqbal, U., Englehardt, S., Shafiq, Z., 2020. Fingerprinting the fingerprinters: Learning to detect browser fingerprinting behaviors. arXiv preprint arXiv:2008.04480.

Jayakumar, L.N., 2021. Cookies nconsent: An empirical study on the factors influencing of website users attitude towards cookie consent in the eu. DBS Business Review 4.

Khalifa, O.O., Chebil, J., Abdalla, A.-H., Hameed, S., 2011. Ethical issues in monitoring and based tracking systems. IIUM Engineering Journal 12 (5).

Kristol, D.M., 2001. Http cookies: Standards, privacy, and politics. ACM Transactions on Internet Technology (TOIT) 1 (2), 151–198.

Lin, D., Loui, M.C., 1998. Taking the byte out of cookies: privacy, consent, and the web. ACM SIGCAS Computers and Society 28 (2), 39–51.

Marchant, G.E., Sylvester, D.J., Abbott, K.W., 2009. What does the history of technology regulation teach us about nano oversight? Journal of Law, Medicine & Ethics 37 (4), 724–731.

Martínez, D., 2021. Websites level of confidence: extended version of the algorithms and results of the research work “Web tracking compliance: websites level of confidence in the use of information-gathering technologies”. https://doi.org/10.5281/zenodo.5793920. 10.5281/zenodo.5793920

Moor, J. H., 1991. The ethics of privacy protection.

Sanchez-Rola, I., Ugarte-Pedrero, X., Santos, I., Bringas, P.G., 2017. The web is watching you: A comprehensive review of web-tracking techniques and countermeasures. Logic Journal of the IGPL 25 (1), 18–29.

SDPA, 2020. *Guḑa sobre el uso de las cookies*. https://www.aepd.es/sites/default/files/2020-07/guia-cookies.pdf.

Sipior, J.C., Ward, B.T., Mendoza, R.A., 2011. Online privacy concerns associated with cookies, flash cookies, and web beacons. Journal of internet commerce 10 (1), 1–16.

Warren, S.D., Brandeis, L.D., 1890. Right to privacy. Harv. L. Rev. 4, 193.

David Martínez is an Adjunct Professor at the University of Girona and a Research Technician at ICRA (Catalan Institute for Water Research). He holds a Master's degree in Cybersecurity and Privacy from the Universitat Oberta de Catalunya and is also starting a Ph.D. at the University of Girona, where he furthers his research on the detection and mitigation of attacks in telecommunication or environment modeled networks using artificial intelligence techniques.

Eusebi Calle receives his PhD in Computer Science by Universidad de Girona (UdG) in 2004. Associated Professor and leader of the Broadband Communications and Distributed Systems (BCDS) research group in UdG. His topic of interest covers Telecom- munication networks: Optical networks, network management algorithms, robustness and survivability analysis, etc. He has more than 80 papers in international congresses and international journals. Currently he is TPC in different well-known congresses such as DRCN, RNDM, IFIP Networking, etc. He was also the TPC chair of OPTICS 201 and RNDM 2017. He has also participated and coordinate more than 20 national and European projects.

Albert Jové obtained a BSc in Biology from the Universitat de Barcelona in 1984, postgraduate in Computer Auditing from the Universitat Politecnica de Catalunya in 1994, and Computer Engineer from the Universitat Oberta de Catalunya (UOC) in 2007. CISA, CISM, and CGEIT certifications. He works as a freelance professional in projects to adapt companies to data protection regulations and as an associate lecturer at the UOC.

Cristina Pérez-Solà obtained her PhD in 2016 with a thesis focused on privacy in online social networks (OSN). Her current research is focused on security and privacy, with special interest in security and privacy of blockchain based cryptocurrencies and networks. She has co-authored 39 scientific publications and 3 divulgative books. She has been cited 776 times and has an h-index and i10-index of 13 (google scholar data).