Journal of Cyber Security
DOI: 10.32604/jcs.2022.029020
Article

Tech Science Press

Web Tracking Domain and Possible Privacy Defending Tools: A Literature Review

Maryam Abdulaziz Saad Bubukayr[1,*] and Mounir Frikha[2]

[1]College of Computer Science and Information Technology, King Faisal University, Alahsa, 31982, Saudi Arabia
[2]Department of Computer Networks and Communications, King Faisal University, Alahsa, 31982, Saudi Arabia
*Corresponding Author: Maryam Bubukayr. Email: 221445328@student.kfu.edu.sa

Received: 30 April 2022; Accepted: 01 June 2022

Abstract: Personal data are strongly linked to web browsing history. By visiting a certain website, a user can share her favorite items, location, employment status, financial information, preferences, gender, medical status, news, etc. Therefore, web tracking is considered as one of the most significant internet privacy threats that can have a serious impact on end-users. Usually, it is used by most websites to track visitors through the internet in order to enhance their services and improve search customization. Moreover, selling users' data to the advertising companies without their permission. Although there are more research efforts focused on third-party tracking to protect user privacy, there are still no comprehensive approaches to develop an efficient and accessible privacy protection method, even if more attention is paid to the topic. The main goal of this paper is to conduct a literature review on the web-tracking domain and possible privacy defending methods by presenting an overview of privacy issues, determining the possible tracking mechanisms that might be exploited, discussing the available privacy defense tools that could be utilized for improvement, and presenting the strength and weaknesses of each method.

Keywords: Web tracking; website privacy; cookies; security; anti-tracking; privacy defense tools; machine learning; and blacklist

1 Introduction

With every online activity and website visit, a huge amount of data is collected, including the pages we visit, the items we buy, the data we search for, the conversations we have with others, the people we contact, and more. This is legitimate according to the technical viewpoint of the website's administrator. When a website is requested, it downloads all files, including third-party content, from a server and uploads them to the end user's browser. This involves installing cookies that can perform multiple functions. There are several types of website cookies that are intentionally embedded by the website administrator and can have many purposes, e.g., Improving website performance, tracking users, and selling user data for targeted advertising. These ads are displayed on your favorite websites that provide data relevant to your preferences.

Throughout the history of the Internet, businesses have seen an increase in sales when they use Internet content on websites to market and promote their products [1]. In addition, when websites see the success of targeted advertising campaigns across the Internet, they seize the opportunity by collecting more data about users, analyzing it, and selling that data to targeted advertising to reduce operational costs [2,3].

On the other hand, users do not know what happens to their data and how their privacy is at risk. Accordingly, they are vulnerable to privacy risks. As a result, deleting or blocking third-party cookies is one of the most common challenges to protect users' privacy. The purpose of this paper is to provide an overview of web tracking domains and cookies, identify recent privacy defense tools used to detect third-party tracking behavior and cookies, and the advantages and limitations of each method. To achieve this, a literature search is carried out and a total of thirty primary studies are analyzed.

2 Related Works

Recently web tracking is one of the most significant privacy issues, every website utilizes it to present a high quality of services and track users by uploading third-party cookies in the users' browser.

2.1 An Overview of Web Tracking Domain and Third-Party Cookies

One of the most recent works in the literature, Ermakova et al. [3] provided a foundation for future studies by highlighting the methodologies and importance of the web-tracking field. Also, present a comprehensive literature review based on a structural framework. Moreover, the paper presents the utilized research methodologies and evaluates the web tracking papers with references to privacy, technology, and commercial aspects. The survey proposed that, there should be more future directions on mobile web tracking, and how mobile applications protect users against third-party tracking. Furthermore, there should be more agreements between privacy and commercial interest.

Also the authors in [4], Ishtiaq et al. presented the possible tracking mechanisms that could be used for uniquely identifying users while browsing the internet or making a purchase. Also, discuss how to defend users against these types of tracking mechanisms.

In [5], Re and Carpineto proposed a method that makes users aware of their potential web tracking profile across third-party cookies. The aim is to increase privacy and enhance the behavioral targeting process that keeps track of how users browse the internet.

Similarly, Bujlow et al. presented in [6] a survey on the web-tracking domain to educate users with various tracking mechanisms they may experience while browsing the internet regularly. These tracking mechanisms are diverse in coverages, scopes, and purposes. Moreover, discussed the available tools and techniques to protect users' privacy.

Moreover, Wills and Uzunoglu provided in [7] a comprehensive study on evaluating the effectiveness of existing anti-tracking methods in terms of detecting and blocking various types of third-party resources. Moreover, they described how third-party resources are identified and classified according to several defined categories. They classified them into six categories, which are Ad Trackers, Analytics, Beacons, Social, Widgets, and others. As future work, more research should be done to evaluate the effectiveness of anti-tracking tools using different methods. Also, classifying specific domains or exploring a particular set of categories about third-party domains.

Mikhailovich et al. in [8] provided a deep analysis of the most effective machine learning models used to enhance information security problems in a web application. Moreover, they enhanced a methodology for introducing machine learning to construct a web-based security model using the proposed methodology. The paper outlined criteria for selecting the best method to train and identify the tasks of machine learning. A practical experiment was conducted using the developed safety model. An experimental assessment was performed including training time, accuracy, and linearity.

Some of the papers analyzed the existing privacy protection tools (PPTs). For instance, the authors in [9] analyzed and studied the behavior of ad-blocking browser extensions on a variety of websites, evaluated the advantages of using these extensions, and examined how the ad-blocker increases traffic. The result has shown that ad-blocking tools can prevent displaying advertisements on websites by blocking the third-party cookies and disabling them from storing on the end user's system. However, blocking all third parties can cause performance or functionality loss on websites if third-party images, JavaScripts, or flash files cannot be loaded.

Also, Dan and Golan in [10] analyzed one of the privacy-preserving tools that block all third-party tracking on web pages, called the Ghostery extension interface. The analysis method was performed in two phases, firstly a comprehensive review of the usage and execution of the extensions and secondly a heuristic analysis of the extensions interface. According to findings, researchers do not face any difficulties in using the Ghostery extension interface since they have a deep understanding of it. On the other hand, users who are unfamiliar with this extension do not benefit from its full features and capabilities. The researchers hope that developers and designers at Ghostery must focus more on developing an interface that is friendlier to a wider range of users this may help mitigate users' privacy breaches easily.

In [11], Likewise, Pujol et al. analyzed the benefit of AdBlock Plus that is utilized to detect ad traffic and web tracking from unbiased network measurements. Also, they assessed the spread of ad-blockers in this relevant network, and discussed the potential impacts of AdBlock Plus for Internet Service Providers (ISPs) and content providers.

Another privacy protection tool analyzed by Wu et al. in [12] was Private browsing mode, which is available on both desktop and mobile. Many contradictions were found between various browsers and between different versions of the same browser on different platforms. This is because of the tradeoff between privacy and security. Even if the user's private browsing mode does not reveal any sensitive information, it would still be possible to track the user based on the browser's fingerprint.

Younis et al. conducted a similar study in private and default browsing modes of four popular web browsers, including Google Chrome, Dolphin, Opera, and Mozilla Firefox, in [13]. The results show that users' personal information was better protected in Mozilla Firefox, while Google Chrome was the least secure web browser in both private and standard modes. Moreover, the result verifies that private browsing mode does not effectively protect users' privacy on the Internet. Also, the work in Tsalis et al. in [14] evaluated the private browsing mode in some windows' browsers like Chrome, Internet Explorer, Firefox, and Opera. The result emphasized that privacy threats still exist even if this protection method is activated.

Moreover, Krupp et al. in [15] analyzed tracking in IOS (iPhone operating system) applications to present more insight into how tracking is utilized and clarified the need for privacy in smartphone applications. They used the search engine DuckDuckGo as a case study to gather the data set to analyze smartphone applications on IOS. Moreover, they examined the most popular applications that provide data to users and expose personal data on the mobile such as messages, photos, contacts, and locations. As it's known, Facebook, Microsoft, Google, and Amazon-owned the most popular online tracker companies that receive personal information. The results show that 84% of IOS applications are connected with at least one tracking domain. Moreover, 95% of the IOS applications were categorized as trackers while most of them communicated with Google's services. Finally, the paper believes that there should be more transparency about how the IOS applications connected with third-party trackers and whether the personal information was sent to these trackers.

In addition, Englehardt et al. in [16] analyzed and measured 1 M websites using one of the tracking auditing tools, OpenWPM1. Also, 15 types of measurements were made on each website, including tasteful and stateless tracking, to study the impact of privacy protection tools (PPTs), and the syncing of tracking information between websites. The result confirmed that the suggested framework is effective in identifying, quantifying, and characterizing online tracking behaviors.

Gómez-Boix et al. carried out similar work in [17] where 2,067,942 stateless browser fingerprinting-based tracking techniques from a crawl of the top 15 French websites were analyzed. This technique could be exploited to track and identify users while browsing the internet.

Recently, several studies suggested anti-tracking methods to detect tracking behaviors and third-party cookies. In this context, Castell-Uroz et al. in [18] suggested a new anti-tracking method that analyzes the characteristics of URL strings to discover tracking resources and without using any external features. This method is called Deep Tracking Detector (DTD). The result of the study showed that over 5 million HTTPS coming from 100,000 websites, Deep Tracking Detector achieved 97% detection accuracy. Moreover, DTD can be easily executed in a browser plugin. However, still there is a need for future research to improve browser plugins that could help internet users to enhance their privacy.

Sun et al. in [19] suggested a new system called MFTracker Detector based on the theory of structural holes to discover third-party trackers by generating Jlist (JavaScript based list) and Flist (Flash based list).

This system achieved a high detection accuracy and the Jlist and Flist can be created automatically, while updating and maintenance are done manually which is passive and complicated.

Furthermore, the authors in [20] implemented a Canvas-based tracking method that detects the use of canvas fingerprinting tracking and canvas-font tracking based on 10 K popular websites ranked by Amazon Alexa.com. The proposed method can also detect tracking even if code obfuscation is used. It is based on dynamic code analysis to discover obfuscated tracking by observing the JavaScript calls that the browser creates to the website and comparing them with the original source code of the website. The presence of obfuscated canvas-based fingerprinting is confirmed by the experimental results. Furthermore, tracking methodologies can exist more in secondary pages than on home pages.

In addition, Yu et al. produced in [21] a well-designed and more flexible rule-set that allows users to customize their privacy protection to suit their needs. They used the Word2Vec method to provide a new framework that may help mitigate third-party tracking. Several actions were taken based on the privacy level of the websites. According to research findings, an error rate decreases from 71% to 24% after using the proposed framework. In addition, the paper showed a new way of thinking about blocking third-party tracking. As future work, a need to improve the protection of the common web pages and the extension of the research data set to get a more satisfactory outcome were mentioned.

Finally, Beigi et al. in [22] designed an effective system for anonymizing web-browsing histories called Pbooster. The main purpose of this scheme is to ensure the privacy of users while preserving the utility of their Web browsing history. However, this work does not collect real data and evaluates the efficiency of the proposed Pbooster system in terms of both privacy and utility in practice.

The literature presented different methods used to detect web tracking and protect user privacy. However, these tools are inefficient and most of them applied rules based on elements and domains that need to be blocked. Therefore, this may result in blocking all access tracking as when anti-tracking methods are implemented, it blocks all the third-party tracking that users may like and dislike.

2.2 Blacklist and Machine Learning-Based Technique

The subject of blacklist extension is well studied in various papers that define all their characteristics and review all relative methods. However, in the situation of using an automated blacklist to classify third-party tracking and improve users' privacy, a limited amount of research has been conducted which we can cite.

Ikram et al. in [23] proposed a machine learning technique to filter out malicious JavaScript programs in web browsers automatically. They separated functioning scripts from tracking scripts based on only generating a small-labeled data-set consisting of tracking scripts. This data-set consists of 2,612 JavaScript programs that can be extracted from existing blacklists of anti-tracking tools. The results indicate that the accuracy achieved by existing anti-tracking tools is less than or equal to 78% while the proposed classifiers achieved 99% accuracy and offer the potential for improved detection rates. The limitation of this work only detects tracking through JavaScript programs and other methods such as HTTP requests are not included.

Mughees et al. in [24] proposed a machine learning method to analyze anti-ad blockers used by most websites to discover which users employ content blockers on their browsers and display notifications accordingly. Those notifications request users to switch off ad-blockers, pay a service fee or contribute a donation. As reported in the article, 686 out of 100 K websites utilize anti-ad blockers on their web pages. Therefore, ad-blockers continue to use filter lists to disable anti-ad blockers using web request blocking and page element removal. Finally, more future research should counter the rate between ad blockers and anti-ad blockers.

Cozza et al. in [25] proposed a hybrid method called GuardOne that utilized blacklisting (commonly used by anti-tracking methods) and machine learning to automatically detect the privacy-intrusive required while surfing the internet based on whether an Ad Tracker is active or not. As compared with classical systems, the GuardOne mechanism can filter out malicious resources effectively and without a drop-in performance, this can decrease personal data leakage. The limitation of the result is that it used Disconnect and Ghostery only to construct the data-set. Thus, it depends on their behaviors. As future work, the paper recommended further research in studying the accuracy when various classifiers are utilized, one for each type of web resource to classify.

Safae et al. in [26] adopted a comprehensive review of the most popular machine learning models utilized for web page classification and compared them according to relevant characteristics. For web page classification, the authors assign each web page to one or more categories. This classification is useful in data extraction systems, contextual advertising on the web, search engines, and others. Furthermore, it has a high influence on classifiers accuracy, as well as the decision on which classifiers to employ.

Cuzzocrea et al. in [27] proposed a machine learning-based technique (ML) that can be used to improve advanced detection and analysis of web phishing. In particular, they used decision tree algorithms to detect whether a website is susceptible to conducting phishing activities or not. When a positive result is reported, the website is classified as phishing. The developed ML method relies on analyzing the URL-based features such as URL length and domain age. The experimental result identifies the benefits of adopting the ML model to address the web-phishing detection problem. This ML model, as stated by the authors, was not tested using real data or implemented into a web browser extension. Moreover, JavaScript code is not analyzed for web-phishing detection.

Similarly, Odeh et al. presented in [28] a survey on recent protection techniques that were used to detect phishing attacks on websites. They are deep learning, automated techniques, heuristic, and machine learning-based techniques. The results demonstrated that machine learning-based techniques are the most effective way in eliminating phishing attacks on the web. Several useful machine-learning techniques were examined in the paper, including Support Vector Machines (SVM), Random Forests (RF), Ada Boosting, and Naive Bayes (NB). Almost all of the approaches examined focused on traditional methods. It was recommended that more research should be done in the future to improve ML performance on a large set of data and images, over-fitting, websites with captcha information, poor accuracy, and hyper tuning of ML techniques.

In [29], Wu et al. proposed a tracker detector model to classify third-party resources automatically. They utilized the blacklist technique by exploiting the Ghostery labeling as a dataset in combination with the BFTree algorithm which is one of the machine learning classifiers. The crawl was performed from 6441 Web pages and the JavaScript programs of 33,366 elements were analyzed to distinguish third-party trackers from non-trackers. As a result of the implementation, the model can specify JavaScript programs with high accuracy of 97.34%. However, there is a need to improve the proposed detection system in the future to avoid attackers from modifying the rules of the classifier by adding or removing classification rules to access through some JavaScript APIs. Moreover, they only employed JavaScript APIs to classify, while other web resources are not considered.

The work in Dudykevych and Nechypor in [30] was based on extracting HTTP features, traffic collection crawler, and machine learning method to automatically detect web-tracking HTTP requests. Using the proposed technique, invisible third-party trackers were detected with known platforms.

Moreover, Thu and Chetan proposed in [31] a new model called AdRemover based on Random Forest classification, blacklists, and whitelists. The decision trees were trained by determining which URLs are likely to contain ads or non-ads to create the filter lists automatically. Five main features were considered in the dataset generation, which are Lexical Feature, External Request Resources, Site Popularity Feature, Ad keywords Feature, and Host-Based Feature. With Random Forest classification, the accuracy percentage improved to over 98%. It is necessary to add more features to the proposed model to make it more robust and efficient in the future.

The authors in [32] trained Naive Bayes machine learning techniques using the five HTTP features (%3rdPartyReq, %cookies, #referers/req, #rec/sentBytes, #referers) and AdBlock Plus blacklists from August 2013. They examined which features and classifiers are most effective in identifying privacy-invasive services. The accuracy and recall of the result were up to 83% and 85%, respectively. Another finding is that shopping sites providing promoting content were mainly found among other services. Furthermore, the authors believe that organizations and users can directly benefit from the proposed approach by implementing it in the same way.

3 Research Methodology

In order to perform this literature review, multiple steps have been conducted to address the current literature of web tracking domain and recent privacy defense tools. A brief description of the review steps are as follows:

3.1 Planning

For a successful literature review, we ensured that our steps were formulated in an organized manner. This step identified the major steps needed to achieve the literature review's objectives.

3.2 Determining the Search Terms and Methods

An appropriate search method should be strictly followed. Therefore, this method defined how each article has been selected for implementing the literature review study. A comprehensive search about the web tracking domain was conducted. Various English databases such as Springer, Elsevier, and the IEEE Digital Library were searched. These databases were searched between 2016 and 2021. To search these electronic databases, the following terms were considered when searching:

Web tracking (tracking OR website tracking OR webtracking

OR third-party cookies OR website cookies). AND Possible (available OR recent) AND (Privacy defending methods OR anti-tracking methods OR privacy protection tools OR privacy-preserving tools OR PPTs). Fig. 1, Shows the PRISMA flow diagram for the research selection process.

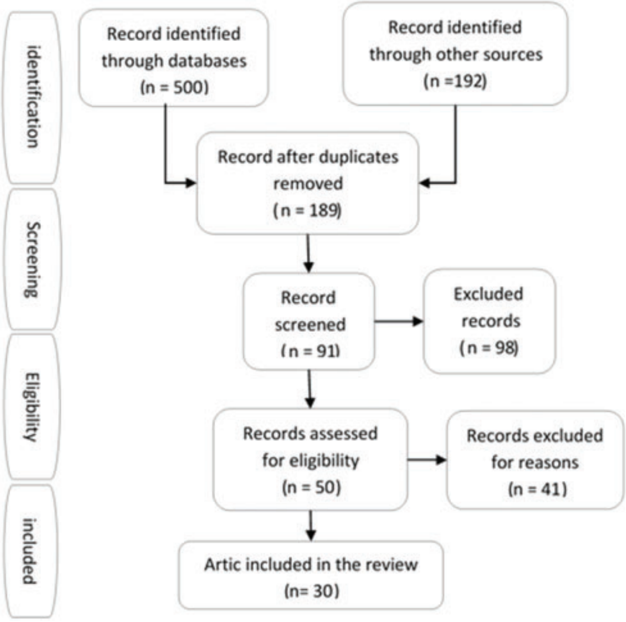
identification
Screening
Eligibility
included
Record identified through databases (n = 500)
Record identified through other sources (n =192)
Record after duplicates removed (n = 189)
Record screened (n = 91)
Excluded records (n = 98)
Records assessed for eligibility (n = 50)
Records excluded for reasons (n = 41)
Artic included in the review (n= 30)

Figure 1: PRISMA flow diagram

3.3 Specifying the Eligibility Criteria

The search was conducted in IEEE Explore, Saudi Digital Library, and Google Scholar databases using the following inclusion and exclusion criteria:

3.3.1 The Exclusion Criteria Included

- Papers not written in English language.
- Papers for workshop or PowerPoint presentations.
- Papers that are not accessible.
- Papers with no focus on web tracking domain.

3.3.2 The Inclusion Criteria Include

- Recent papers published within the period 2016–2021.
- Papers that address the web tracking field with keywords matching the search title.

3.4 Extracting the Data and Result

The keywords of the selected papers using word cloud are presented in Fig. 2. Furthermore, Fig. 3 shows the method used for the data extraction process.

Figure 2: Keywords of the selected papers using a word cloud

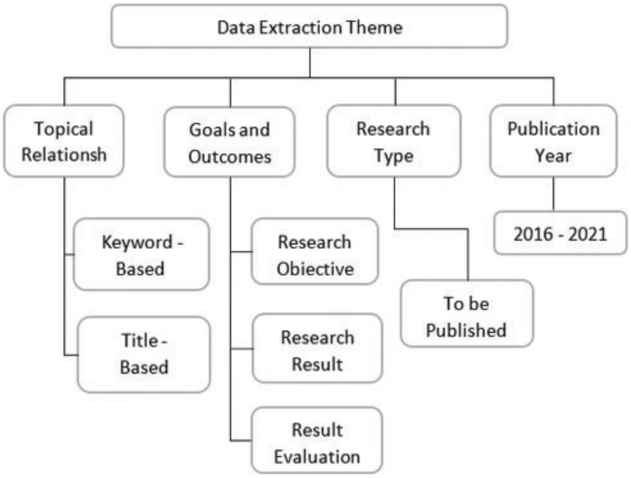

Figure 3: Data extraction theme

3.5 Review Analysis

3.5.1 Finding from the Literature

This section provides a structured and more detailed overview of different potential tracking mechanisms that might be exploited by the tracker, as well as possible defense strategies and privacy defense tools (Summarized in Tab. 1) over the period 2015–2021.

Table 1: Web tracking taxonomy from differ-perspectives

Tracking mechanism	Tracking defense strategies	Privacy defense tools
■ Session-based. ■ Storage-based. ■ Cache-based. ■ Fingerprinting.	■ Block JavaScript, Flash, Java, or Silverlight. ■ Block Flash execution. ■ Block Silverlight execution. ■ Use Tor Browser. ■ URL-based. ■ Clearing of browser web cache ■ Disable cookies (third-party cookies, all cookies, or selectively). ■ Disable the userData storage in IE. ■ Remove the additional HTTP headers.	■ Opt-out mechanism. ■ Private browsing mode. ■ Do Not Track Header. ■ Anonymous Search Engines. ■ Content-blockers.

Tracking Mechanisms

The tracking mechanisms can be differentiated based on how to bypass privacy settings, being difficult to detect, and their resistance to being blocked. Among the most common tracking methods, we can include:

Session Based

Session-based is a mechanism that is used for recording and memorizing a series of user requests on a specific website with the aim of recognizing these preferences for future requests.

Storage Based

Storage-based is the most common and more advanced approach. Generally, the tracking of users' behavior is not restricted to one website, but it can be tracked across several websites that contain multiple third-party services. whenever a user visits a website, the data is being stored in small files called cookies, these cookies are shared among third-party services so that it's more consistent and precise. This approach posed the greatest threat to the privacy of users. The most common mechanisms of this approach are HTTP cookies, Silverlight Isolated Storage, Internet Explorer user Data storage, Flash LocalConnection object, and HTML5 Global, Local, and Session Storage.

Cache Based

Cache-based or client-based is a method that stores temporary web files (or caches) in order to identify the visited websites and recognize browser instances. Using this method, DNS response time for websites will be reduced, as well as it may serve as another method of tracking.

Fingerprinting

Another recent way of tracking methods for uniquely identifying users is to use fingerprinting. Typically, it builds up a user history by identifying the system, network, geographic area, operating system, browser name and version, or instance. Therefore, whenever a user visits a website, the user's preferences are matched within the history in order to determine that it is the same user. That way, tracking can be performed across multiple websites and without any cookies to be set. The fingerprinting method includes several mechanisms such as browser version fingerprinting, Operating System instance fingerprinting, canvas fingerprinting, Network and location fingerprinting, and Device fingerprinting.

Tracking Defense Strategies

The ability to detect tracking and non-tracking websites can be achieved by analyzing several strategies against multiple tracking methods. As a result of the review (see Tab. 2), we can summarize some of the most popular tracking strategies as follows:

- Block JavaScript, Flash, Java, or Silverlight execution.
- Block Flash execution.
- Block Silverlight execution.
- Use Tor Browser.
- URL-based.
- Clearing of browser web cache.
- Disable cookies (third-party cookies, all cookies, or selectively).
- Disable the userData storage in IE.
- Remove the additional HTTP headers.

Table 2: Summary of the major research findings with possible privacy defending methods

Ref	Tracking defense strategies	Privacy defense tools	Limitations
(Pujol et al. 2015)	TCP/HTTP header	Content-blockers (AdBlock Plus)	The method does not protect the privacy violation of the users efficiently.
(Wu et al. 2015)	JavaScript programs	Content-blocker with machine learning technique called tracker detector model	The attacker can modify the rules of the classifier to get access through some JavaScript APIs.
(Gugelmann et al. 2015)	HTTP features	Content-blocker with machine learning technique	The classifier is trained based on some HTTP features and AdBlock Plus blacklists from August 2013.

(Continued)

Table 2: Continued

Ref	Tracking defense strategies	Privacy defense tools	Limitations
(Dudykevych and Nechypor 2016)	HTTP features	Content-blocker with machine learning technique	Other tracking mechanisms were not included, and it is considered as complementary technique rather than an independent tool.
(Garimella et al. 2017)	HTTP features	Content-blockers	All communication with the tracking server is blocked. Therefore, blocking all third party can cause performance or functionality loss on websites. Furthermore, ad-blocking is not sufficient and can be easily detected by trackers.
(Wu et al. 2017)	Fingerprinting	Private browsing mode	Even if the user's private browsing mode does not reveal any sensitive information, it would still be possible to track the user based on the browser's fingerprint.
(Tsalis et al. 2017)	Exploitation of the artefacts	Private browsing mode	Artefacts that are recorded during web browsing cannot be kept confidential by this method.
(Le et al. 2017)	Canvas font and fingerprinting	Canvas-based tracking method	The obfuscated canvas-based fingerprinting still exists.

(Continued)

Table 2: Continued

Ref	Tracking defense strategies	Privacy defense tools	Limitations
(Ikram et al. 2017)	JavaScript programs	Content-blocker with machine learning technique	Only detects tracking through JavaScript while other tracking methods were not included.
(Gómez-Boix et al. 2018)	Browser fingerprinting	—-	As a result of the analysis, browser fingerprinting mechanism fails in differentiate users belonging to a particular demographic group.
(Vo and Jaiswal 2019)	URL-based	Content-blocker with machine learning technique called AdRemover	It focused on advertisement content. Also, it is necessary to add more features to the proposed model to make it more robust in the future.
(Cuzzocrea et al. 2019)	URL-based	Machine-learning framework	It is not tested using real data or implemented into a web browser extension.
(Castell-Uroz et al. 2020)	URL-based	Content-blocker based on deep learning (called deep tracking detector)	As a limitation of this method, only the characters of the URL string are used to identify web tracking resources.
(Cozza et al. 2020)	JavaScript programs and HTTP requests	Content-blocker with machine learning technique called GuardOne	The classifier is trained based on a labeled dataset. Therefore, it depends on its behaviors which is inefficient.

(Continued)

Table 2: Continued

Ref	Tracking defense strategies	Privacy defense tools	Limitations
(Younis et al. 2021)	Web history or email communications	Private browsing mode	The result verifies that private browsing mode does not effectively protect users' privacy on the Internet.
(Sun et al. 2021)	Flash and JavaScript	Content-blocker called MFTracker Detector	The maintenance of the blacklists are done manually which is passive and complicated.

Privacy Defense Tools

As the tracking methods continue to evolve rapidly and are almost used by every website, the need for more sophisticated privacy defense tools has risen. Nowadays, several privacy defense tools are available in order to safeguard against various tracking methods and ensure users' privacy online. Tab. 2 shows the major privacy defense tools that were identified in this literature review. They are summarized as follows:

Opt-out Mechanism

Using this tool, websites are prevented from collecting or storing cookies, but it can be ignored, and third parties are still tracked [15,16].

Private Browsing Mode

It is like activating a temporary session where the search history will not be saved, and the searched pages cookies will be all cleared after closing the session.

Do Not Track Header

This tool gives the site visitors the preference to choose if they want to be tracked or not by the site and whether they want to share any collected data from their activities or not. However, it is useless and can be ignored [19].

Anonymous Search Engines

In fact, the majority of search engines often track users' activities. Therefore, there is a growing demand for search engines that offer reliable results with private versions and without storing queries or tracking online activity. Various alternative browsers exist that hide the HTTP header or IP address, and disable websites from receiving the used search string such as DuckDuckGo, MetaGer, Swisscows, etc. However, some of them do not offer the privacy that they claim [1], while others are not user-friendly.

Content Blockers

Currently, the most popular anti-tracking mechanism is content blockers which are web browser extensions that are used to prevent malicious content, third-party tracking links, and other threats based on blacklists (predefined lists) [4,5]. However, they do not effectively block web tracking, cause performance issues, and are difficult to manage by end-users. Moreover, there are multiple problems associated with their maintenance and performance.

3.5.2 Discussion

As a result of the Literature review, many papers proposed some anti-tracking mechanisms to detect and block third-party cookies in order to protect users' privacy. Some papers analyzed the URL string, used the Do Not Track Header, private browsing mode, Anonymous communication, or opt-out mechanism. All these mechanisms are inefficient, and the trackers can easily bypass and still track users. Therefore, the privacy level continues to be unacceptable. Other anti-tracking methods provided in the previous studies lacked more accuracy when various classifiers were utilized. Moreover, many papers have agreed that there is still no integrated solution with high efficiency to address privacy protection in web browsers. Several studies [4, 5, 7, 8, 10, 25 and 26]confirmed that the most common anti-tracking method applied in web browsers to detect tracking is content blockers that are based on blacklists (pre-defined lists).

4 Recommendations

This section provided a discussion about the current and most popular anti-tracking method, which is content blockers. They are web browser extensions that are used to prevent malicious content, third-party tracking links, and other threats based on blacklists (predefined lists) [4,5]. However, they cannot completely block web tracking, cause performance issues, and are difficult to manage by end-users. Moreover, there are multiple problems associated with their maintenance and performance. According to maintenance issues, users will not be able to maintain and update the blacklists manually every time visiting the websites to make it effective against the new third-party cookies that download the advertising content and keep track of users. According to performance issues, the blacklists need to utilize a large space of memory in the web browsers in order to store cookies and determine whether they are malicious or not.

Due to this, researchers have combined blacklisting with machine learning approaches to detect privacy-intrusive activities automatically. However, the papers are quite limited related to this research area and their result still needs to be addressed to deal with multiple web resources and a high accuracy rate.

5 Conclusion and Future Work

The paper outlines the literature review of main studies related to the web tracking domain and cookies. Moreover, it classifies the most common privacy defense tools used to ensure privacy. Finally, it evaluates the advantages and limitations of each tools.

Since many tracking mechanisms are available, it is not easy to avoid being tracked at all. Therefore, there is a pressing need to improve the protection of users' privacy, mitigate the risks of third-party tracking, and extend the research data set in order to get a more satisfactory outcome. Thus, a proper combination of privacy defense techniques could help mitigate the risks that users are most concerned about.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Parra-Arnau, D. Rebollo-Monedero and J. Forne, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, vol. 33, pp. 53–63, 2014.

[2] S. Schelter and J. Kunegis, "Tracking the trackers: A large-scale analysis of embedded Web trackers," in *Proc. of the Tenth Int. AAAI Conf. on Web and Social Media (ICWSM 2016)*, Cologne, Germany, vol. 10, No. 1, ISSN 2334-0770, pp. 679–682, 2016.

[3] T. Ermakova, B. Fabian, B. Bender and K. Kerstin, "Web tracking–A literature review on the state of research," 2018. https://doi.org/10.24251/HICSS.2018.596.

[4] A. Ishtiaq, S. H. Abbasi, M. Aleem and M. A. Islam, "User tracking mechanisms and counter-measures," *International Journal of Applied Mathematics, Electronics and Computers*, vol. 5, pp. 33–40, 2017.

[5] D. L. Re and C. Carpineto, "Enhancing user awareness and control of Web tracking with ManTra," in *IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI)*, NE, USA, pp. 391–398, 2016. https://doi.org/10.1109/WI.2016.0061.

[6] T. Bujlow, V. Carela-Español, J. Solé-Pareta and P. Barlet-Ros, "A survey on Web tracking: Mechanisms, implications, and defenses," *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476–1510, 2017. https://doi.org/10.1109/JPROC.2016.2637878.

[7] C. E. Wills and D. C. Uzunoglu, "What Ad blockers Are (and Are Not) doing," in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, Brno, Czech Republic, pp. 72–77, 2016. https://doi.org/10.1109/HotWeb.2016.21.

[8] K. M. Mikhailovich, M. A. Valerievna, P. P. Andreevich, U. I. Alexandrovich and K. A. Vladimirovich, "Guidelines for using machine learning technology to ensure information security," in *2020 12th Int. Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 285–290, 2020. https://doi.org/10.1109/ICUMT51630.2020.9222417.

[9] K. Garimella, O. Kostakis and M. Mathioudakis, "Ad-blocking: A study on performance, privacy and counter-measures," in *Proc. of the 2017 ACM on Web Science Conf., WebSci '17*, New York, NY, USA, ACM, pp. 259–262, 2017.

[10] D. Bouhnik and G. Carmi, "Interface application comprehensive analysis of ghostery," vol. 5, pp. 4–10, 2018.

[11] E. Pujol, O. Hohlfeld and A. Feldmann., "Annoyed users: Ads and Ad-block usage in the wild. in *Proc. of the Internet Measurement Conf.*, Tokyo, Japan, pp. 93–106, 2015.

[12] Y. Wu, D. Meng and H. Chen, "Evaluating private modes in desktop and mobile browsers and their resistance to fingerprinting," in *2017 IEEE Conf. on Communications and Network Security (CNS)*, Las Vegas, NV, USA, pp. 1–9, 2017. https://doi.org/10.1109/CNS.2017.8228636.

[13] L. B. Younis, S. Sweda and A. Alzu'bi, "Forensics analysis of private Web browsing using android memory acquisition," in *2021 12th Int. Conf. on Information and Communication Systems (ICICS)*, pp. 273–278, 2021. https://doi.org/10.1109/ICICS52457.2021.9464591.

[14] N. Tsalis, A. Mylonas, A. Nisioti, D. Gritzalis and V. Katos, "Exploring the protection of private browsing in desktop browsers," *Comput. Secur.*, vol. 67, pp. 181–197, 2017.

[15] B. Krupp, J. Hadden and M. Matthews, "An analysis of Web tracking domains in mobile applications," in *13th ACM Web Science Conf. 2021 (WebSci '21). Association for Computing Machinery*, New York, NY, USA, pp. 291–298, 2021. https://doi.org/10.1145/3447535.3462507.

[16] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Vienna, Austria, pp. 1388–1401, 2016.

[17] A. Gómez-Boix, P. Laperdrix and B. Baudry, "Hiding in the crowd: An analysis of the effectiveness of browser fingerprinting at large scale," in *Proc. World Wide Web Conf. (WWW)*, Lyon, France, pp. 309–318, 2018.

[18] I. Castell-Uroz, T. Poissonnier, P. Manneback and P. Barlet-Ros, "URL-Based Web tracking detection using deep learning," in *2020 16th Int. Conf. on Network and Service Management (CNSM)*, Izmir, Turkey, pp. 1–5, 2020. https://doi.org/10.23919/CNSM50824.2020.9269065.

[19] J. Sun, Z. Huang, T. Yang, W. Wang and Y. Zhang, "A system for detecting third-party tracking through the combination of dynamic analysis and static analysis," in *IEEE INFOCOM 2021-IEEE Conf. on Computer Communications Workshops*, Vancouver, BC, Canada, pp. 1–6, 2021. https://doi.org/10.1109/INFOCOMWKSHPS51825.2021.9484564.

[20] H. Le, F. Fallace and P. Barlet-Ros, "Towards accurate detection of obfuscated web tracking," in *2017 IEEE Int. Workshop on Measurement and Networking (M&N)*, Naples, Italy, pp. 1–6, 2017. https://doi.org/10.1109/IWMN.2017.8078365.

[21] S. Yu, D. V. Vargas and K. Sakurai, "Effectively protect your privacy: Enabling flexible privacy control on Web tracking," in *Fifth Int. Symp. on Computing and Networking (CANDAR)*, Aomori, Japan, pp. 533–536, 2017. https://doi.org/10.1109/CANDAR.2017.26.

[22] G. Beigi, R. Guo, A. Nou, Y. Zhang and H. Liu, "Protecting user privacy: An approach for untraceable web browsing history and unambiguous user profiles," in *Proc. of the Twelfth ACM Int. Conf. on Web Search and Data Mining, ACM*, Melbourne, VIC, Australia, pp. 213–221, 2019.

[23] M. Ikram, H. Asghar, M. Kaafar, A. Mahanti, and B. Krishnamurthy, "Towards seamless tracking-free Web: Improved detection of trackers via one-class learning," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 1, pp. 79–99, 2016, https://doi.org/10.1515/popets-2017-0006.

[24] M. H. Mughees, Z. Qian, and Z. Shafiq, "Detecting anti ad-blockers in the wild," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 3, pp. 130–146, 2017.

[25] F. Cozza, A. Guarino, F. Isernia, D. Malandrino, A. Rapuano *et al.,* "Hybrid and lightweight detection of third party tracking: Design, implementation, and evaluation," *Computer Networks*, vol. 167, no. 106993, ISSN 1389–1286, pp. 1–18, 2020. https://doi.org/10.1016/j.comnet.2019.106993.

[26] L. Safae, B. E. Habib and T. Abderrahim, "A review of machine learning algorithms for Web page classification," in *2018 IEEE 5th Int. Congress on Information Science and Technology (CiSt)*, Marrakech, Morocco, pp. 220–226, 2018. https://doi.org/10.1109/CIST.2018.8596420.

[27] A. Cuzzocrea, F. Martinelli and F. Mercaldo, "A machine-learning framework for supporting intelligent web-phishing detection and analysis," in *Proc. of the 23rd Int. Database Applications & Engineering Symp., ACM*, New York, NY, USA, Article 43, pp. 1–3, 2019, https://doi.org/10.1145/3331076.3331087.

[28] A. Odeh, I. Keshta and E. Abdelfattah, "Machine learning techniquesfor detection of website phishing: A review for promises and challenges," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conf. (CCWC)*, NV, USA, pp. 813–818, 2021. https://doi.org/10.1109/CCWC51732.2021.9375997.

[29] Q. Wu, Q. Liu, Y. Zhang and G. Wen, "Trackerdetector: A system to detect third–party trackers through machine learning," *Computer Networks*, vol. 91, pp. 164–173, 2015. https://doi.org/10.1016/j.comnet.2015.08.012.

[30] V. Dudykevych and V. Nechypor, "Detecting third-party user trackers with cookie files," in *2016 Third Int. Scientific-Practical Conf. Problems of Infocommunications Science and Technology (PIC S&T)*, Kharkiv, Ukraine, pp. 78–80, 2016. https://doi.org/10.1109/INFOCOMMST.2016.7905341.

[31] T. Vo and C. Jaiswal, "Adremover: The improved machine learning approach for blocking Ads," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON)*, New York, NY, USA, pp. 1–4, 2019. https://doi.org/10.1109/UEMCON47517.2019.8993052.

[32] D. Gugelmann, M. Happe, B. Ager and V. Lenders, "An automated approach for complementing Ad blockers' blacklists," in *Proc. on Privacy Enhancing Technologies*, vol. 2015. no. 2, ISSN 2299-0984, pp. 282–298, 2015. https://doi.org/10.1515/popets-2015-0018.