**COMP5355 Cyber and Internet Security**

**Group Project Report**

**Topic: Survey of web tracking techniques**

|  |  |  |
| --- | --- | --- |
| **Student** | **Contribution** | **Percentage** |
| **20003864G**  **Sze-to Cheuk Wang** | * Read and summarized 12 papers listed in the references, including 1-12 * Finished section 2 of the report * Finished slides 2-4 | **25%** |
| **21000018G**  **Li Yat Long** | * Read and summarized 11 papers listed in the references, including 1, 28-29, 31-34, 36-39 * Finished section 1 & 3.2 and contributed section 2, 4, 5 & 6 of the report * Finished slides 9-11 | **25%** |
| **21010116G Suen Man Kit** | * Read and summarized 16 papers listed in the references, including 1, 13-27 * Finished section 3.1 and contributed section 4 & 5 of the report * Finished slides 5-8 | **25%** |
| **21010703G**  **Mak Chun Wing** | * Read and summarized 11 papers listed in the references, including 1, 13-14, 26, 41- 47 * Finished section 3.3 and contributed section 4 & 5 of the report * Finished slides 12-14 | **25%** |

**CERTIFICATE OF ORIGINALITY**

**We hereby declare that this report is our own work and that, to the best of our knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.**

|  |  |  |  |
| --- | --- | --- | --- |
| **20003864G** | **21000018G** | **21010116G** | **21010703G** |
|  |  |  |  |

1. **Introduction**

As internet becomes an essential part of our daily life, collection of user habit and data become a compulsory for every business not only to offer advertisements or products that interest user, but also for analyzing the customer behavior or the whole market. This becomes a serious privacy problem as companies like Facebook can obtained the users data so easily by allowing other website like CNN news to embed the Facebook like button to the website [1]. Normal users are difficult and confused to understand what their data are being collected and tracked, how the companies will use those data and how to protect themselves from those collection. In this paper, multiple survey papers regarding web tracking defense methods and ways to improve privacy when browsing have been read, summarized, and analyzed with their technologies behind as well as their advantages and disadvantages.

1. **Problem Statement**

**2.1 Problem Definition**

In real world, footprints will be made if you walk alone between different locations. In the digital world, this is not an exception, different parties will base on their internet to track down every person who visit their site. This action is called “Tracking”.

Tracking is commonly used by various entities for a wide range of purposes. The obtained information is not always used directly by the tracker - a very common practice is that the collected data are sold to other parties (e.g., insurance companies or online stores) or accessed by government agencies and identity thieves.

Research has shown that nowadays third-party techniques can identify your browsing history even though by a pseudonymity way. Private mode cannot prevent the history private being unveiled by the website. [2]

Having said that, in the following section we will focus on the challenging issue that normal netizens will face when they are surfing the internet.

**2.2 Challenging Issues**

The first challenging issue is the user-oriented search problem, also known as filter bubble.

This issue is done by the search engine such as Google. As netizens surfing the internet, the browser will record down the website you have visited in order to deliver the user based on the search history. As a result, the user is gradually isolated from different points of view and form their own information bubble. However, research has shown that this advertising strategy aids the companies to reach their target audience more easily and able to boost their income. [8]

The second challenging issue is online advertising. The internet tycoons made use of users' browser history to classify target customers for the advertising companies. A research has shown that even user’s IP, plug-ins are also an element involved in the process of targeted advertisement. [3]

The third challenging user is web analytics and usability tests. Research have found that no matter government or business sector has keep track on the user behaviors. For example, a research shown that 7 out of 10 top identified websites append additional event-handlers to track the clicks and send the information to a third-party server. When it comes to government level tracking, a research has shown that Iran, Bahrain made use of web tracking tools like third-party cookies to surveillance purposes. [9] In Asia pacific region, only South Korea has specific regulation related to data privacy while America and Europe have a much tighten policy. [10]

The fourth challenging issue is accessing financial credibility. In 2009, Kevin Johnson reported to have his credit limit in American Express of 10800$ lowered to 3800$ after he shopped online in Walmart. American Express claimed that it was due to the fact that many other Walmart customers have problem with paying the credit back [1].

Some start-ups like Kreditech, Kabbage are starting to make use of users’ personal data which are registered in large companies such as eBay, PayPal to determine the credit risk in several minutes. Yet, more restrictions will be imposed in the long run due to the rising of public opposition and awareness. [4]

The fifth challenging issue is first party unintentionally provides identity.

If a website puts identifying information in a URL or page title, it may unintentionally leak the information to third parties. An aggregate of 48% leaked a user identifier in a Request-URI or referrer from 120 popular website. A search suggested that some third parties malicious JavaScript execution can stole the user first parties' information from the cookies. [5]

The sixth challenging issue is identity theft. A research from Carnegie Mellon University studies showed that data revealed by people in the Internet in many cases are sufficient to predict the social security number of these people. That opens the doors to steal some-one’s identity, as the social security number is widely used to authenticate on sensitive websites (e.g., banking or loan services)

In a recent paper, the authors and his teammate has built a model by LinkedIn scrapper, Pub 360 Scrapper and AnyMail finder to identify the top 100 most searched webpage of a user and display a risk score. Hence, prevention move can be taken to reduce the chance of identity theft. [6]

The seventh challenging issue is third-party tracking. Third-party tracking is a form of tracking performed by resources from other services that the one explicitly visited by the user. Third-party trackers (e.g., Doubleclick) are considered as a serious privacy threat, as they can collect and accumulate browsing statistics through many different websites.

In [1], the authors found that around 46% of home pages of the websites from the top 10000 Alexa ranking are monitored by at least one third-party tracker. In particular, one third of the requests sent to third-party websites was sent to a tracker. Google was responsible for tracking on 25% of examined websites, Facebook on 13%, and Twitter on 5%. Currently the third-party tracking problem is merely solved by human static blacklist update, not effective enough. [7]. Moreover, a recent research has shown that the existing tools are not able to let users to stay alert to third-party trackers. [11]

A study [12] specify 4 different types of trackers of third-party tracking. The first type uses only third-party cookies. The second type uses first-party cookies together with JavaScript. The third type uses first-party cookies, third-party cookies together with JavaScript and the fourth type is just used to serve advertisements while the actual tracking process is done by another services.

1. **Existing Solutions**
   1. **Tracking Defense Tools**
      1. **Microsoft tracking Protection List**

**Technique Introduction**

Microsoft tracking Protection List is a technique developed by Microsoft to defence web tracking. It can be used to defence against third-party web tracking by blocking the third-party advertisement services. This can be achieved by maintaining a list of blacklisted third-party and blocking all the content coming from those third parties from the blacklist [15].

Microsoft tracking Protection List relies on three main components to work: classification, enforcement, and mitigation.

Classification: One of the main components of Microsoft tracking Protection List. It is used to check and classify the URLs. Using this component, the Microsoft tracking Protection List will determine if the URL is classified as a tracker. In order to determine whether a URL is a tracker, Microsoft tracking Protection List uses the open-source lists from “Disconnect” [15].

Enforcement: After the URL is classified as tracker or non-tracker using the classification component, the enforcement component will enforce the protection mechanism to protect the users from those URLs that are identified to be trackers. The protection mechanism is enforced by using either restrict storage access or block resource loads [15].

For restrict storage access, it disables the tracking source from getting user data from the web storage, like to get or set cookies. Related APIs for accessing storage like “localStorage” or “IndexedDB” are also disabled [15].

For the mechanism block resource loads, the mechanism disables loading the sources that are identified as trackers.

Mitigation: Enforcing the protection mechanism for all URLs that are identified as trackers may not be appropriate for all users. Therefore, the mitigation component was introduced to allow users including their specifications based on their need and preference so that the websites they would like to use can still work as usual [15].

In order not to affect the users’ web browsing experience severely, there are three main modes provided in Microsoft tracking protection: basic, balanced, and strict[13]. The basic mode is the less rigorous and the strict mode is the most rigorous. Different levels of strictness can be chosen depending on the requirements of the strictness of the web tracking defence requirements [13].

Basic: This is the mildest version of the Microsoft tracking protection. It blocks only the trackers that are classified as harmful. Meanwhile, other web trackers that are not determined as harmful will still be able to track the user when using this mode. This mode of defence can be used by users who do not mind some of their activities are being tracked by web trackers if the web trackers will not pose threat to them [15].

Balanced: As the name suggests, this is a more balanced mode of Microsoft tracking protection. By enabling the balanced mode, other than those web trackers that will be blocked from the basic mode, the browser will also block the trackers originating from the websites that the user has not visited [15].

Strict: Strict mode is the most rigorous mode of the Microsoft tracking protection. It will block most of the web trackers from all sites, whether the user has visited them or not. Since it will block the web trackers from all sites so rigorously, some parts of the site may not work as usual [15].

**Experiment**

According to “Can Users Control Online Behavioral Advertising Effectively?” [14], the effectiveness of Microsoft Tracking Protection Lists and similar technology was experimented and examined. In the experiment, 62 users were recruited to examine the effective of Microsoft Tracking Protection Lists and similar technology [14]. In order to ensure the experiment to be able to examine the effectiveness of the technique on normal users, all the users participated in the experiment did not have any computer or IT related degree or jobs [14]. In the experiment, the technique of Microsoft Tracking Protection Lists and similar tool were introduced to the users, and the users were asked to configure the tools and use the tools to browse some websites, just like what a normal user would do [14]. The experiment results showed that Microsoft Tracking Protection Lists and similar technology can be used to protect the user from web tracking only if the users are able to install the tool, configure it to match their personal needs, and use it properly [14]. Failing to do one of these will result in ineffectiveness of the purpose. Therefore, the tool itself is effective for protecting against web tracking [14]. However, in order for the tools to be effective for the normal users, an easy-to-understand instruction should come along with the tool [14].

“A Closer Look at Third-Party OSN Applications: Are They Leaking Your Personal Information?” [13] conducted experiment on the effectiveness of Microsoft Tracking Protection Lists on two major Online Social Network (OSN) applications – Facebook and RenRen. Since Facebook and RenRen are two major OSN applications that used by many normal users, the evaluation of effectiveness of Microsoft Tracking Protection Lists on them can greatly reflect the effectivness of the Microsoft Tracking Protection Lists on the OSN as a whole [13].

In the experiment, 997 of the facebook applications and 377 of the RenRen applications were tested [13]. From the 997 facebook applications that were tested, researchers found that 98% of them tracked users’ basic information [13]. The basic information including the user’s full name, their hometown, and the list of their friends. Moreover, about 75% of the Facebook app obtained the user’s email address. For RenRen applications [13], more than 69% them provide the users’ personal information to fourth parties. The table below summaries the most common information tracked by Facebook with their frequency.

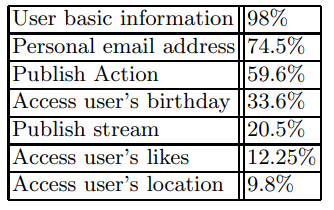


Table 1: Most common information tracked by Facebook

As the experiment results showed, after using Microsoft Tracking Protection List, 410 trackers from Facebook and 120 trackers from RenRen were identified and handled by Microsoft Tracking Protection List [13]. The experiment results show that it is useful to use Microsoft Tracking Protection List to identify trackers that track users’ information.

**Advantages**

One advantage of Microsoft Tracking Protection List is that its principle is easy to be understood by normal users – block the source of trackers. When compared to other technologies like Tor, the idea of Microsoft Tracking Protection List is simple and straightforward. Therefore, it is easier for normal users who do not have certain background to understand this technique and start using it.

Also, there are different level of restriction that the user can choose: basic, balanced and strict [15]. For users who want less restriction on the trackers and want most websites work as expected, they could choose the basic mode. For general users, they could simply use the default mode (balanced mode) and they will already have certain level of protection. For users who want highest level of restriction, they could choose the restrict mode. Therefore, Microsoft Tracking Protection List provide a wide range of flexibility for different users.

**Disadvantages**

From experiment, the researchers found that the interface of Microsoft Tracking Protection List was confusing leading to improper usage of the technique [14]. For example, most of the users in the experiment wanted to opt-out all of the trackers but only opted out one [14]. It was also the problem of inappropriate defaults used by the technique. As a result, the effectiveness of the technique was seriously affected despite the technique itself can provide the desired effect.

Another disadvantage is that Microsoft Tracking Protection List can only be used when using Microsoft Edge [15]. It cannot be installed in other browsers like Google Chrome and Firefox. Therefore, the usability of Microsoft Tracking Protection List is limited to only one browser, and it is not flexibility.

* + 1. **Tor**

**Technique Introduction**

When visiting websites, TCP/IP is the most used protocol. The source & destination port included in the TCP header and the source & destination IP address included in the IP header allows the web servers to know who is communicating with them [27]. Hence, it is possible for the web servers to track the users’ activities.

The purpose of Onion Routing Protocol (Tor) is to achieve anonymous web browsing [19]. When using Tor, only the sender can know both the information of the sender and the receiver of the message [17]. By using Tor, even the Internet Service Provider (IPS) of the sender and the major routers could not know both the information of the sender and the receiver [16].

When the user uses Tor to visit a website, the request message on the onion network are encrypted in many layers like an “onion” [21]. The encrypted message will then be transmitted through a series of “onion routers”, which is the routers that implement the Tor protocol [21]. Each “onion router” will have the key to decrypt one layer of the encrypted message. After decrypting a layer of the message, the router will obtain another encrypted message to be sent to the next node, and an IP address that the message needs to be sent to [21]. The last node will be able to decrypt the last layer of the encrypted message and get the original request from the user and the receiver’s IP address. Between any two “onion routers”, there will be a new TCP/IP connection created [18]. Therefore, there will be no way the web server can know who the sender is. The web server will only know that the request is coming from the last “onion router”. Therefore, anonymous web browsing can be achieved by using Tor [18].

There are three main types of “onion routers” when using Tor – entry node, bridge node and exit node [16]. The entry node is where the sender initializes the connection with; the exit node is where to send the message to the receiver; the bridge nodes are the nodes between the entry node and exit node [16]. Since the sender initializes the TCP/IP connection with the entry node, it will be able to know the sender’s IP address and port used [18]. However, it will not know the receiver’s IP address since it can only obtain the IP address of the next router after decrypting a layer of the message [18]. Similarly, the exit node will only know the destination’s IP address but not the sender’s IP address. As a result, there is no way the destination can know the IP address of the sender [18].

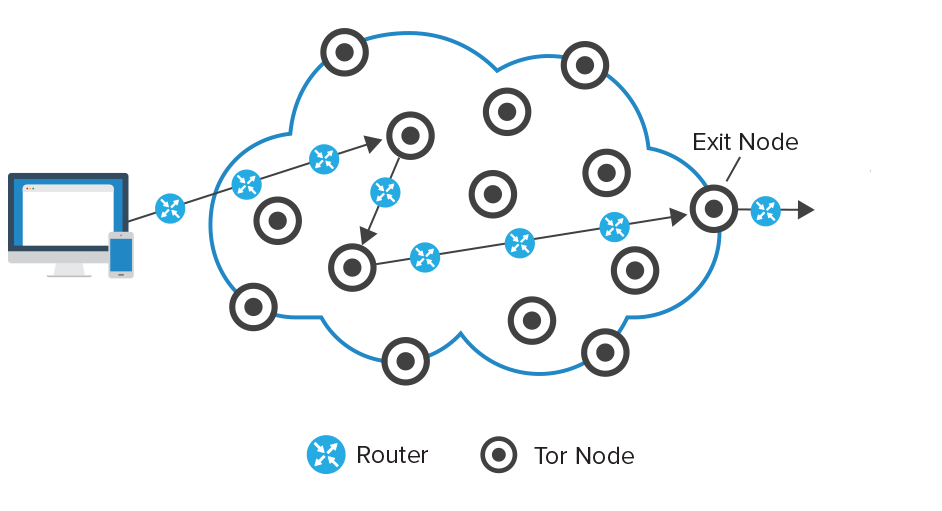


Fig. 1

The choices of the “onion routers” are mostly randomized since we do not want anyone to be able to guess the routing path of the message [20]. However, there are some necessary requirements for choosing the routers, like whether the router supports the port of the application and if the routers have enough resources to handle the routing [20].

**Experiment**

According to [18], it conducted experiment to evaluate the effectiveness of onion network on anonymous web surfing. In the experiment, the participants used the onion routing and tried to visit website, send e-mail, and log in to websites anonymously [18]. In the experiment, five onion routers were set up and the internet traffics went through the five routers randomly based on the principle of onion network. For the encryption used in the onion routing, 1024 key was used [18]. The time the traffic went through five onion network including encryption operation was under 0.5s [18]. The experiment result showed that no onion routers could be able to know both the source and destination information of the traffic [18]. The first onion router would be able to know the sender’s IP and port used, and the last onion router would be able to know the receiver’s IP and port used [18]. The intermediate nodes would not be able to know either of those information. The protection is valid since it is not possible for modern technology is not able to crack the 1024 bit RSA encryption [18]. As a result, the websites will not know who the real requestor & who browses the website and therefore they will not be able to conduct web tracking to the users who use the onion network [18].

**Advantages**

When compared to similar services for hiding users’ identity like using a VPN server, the main advantage of using Tor can achieve a higher level of anonymity. When using a VPN server, the sender and destination’s information are known by the VPN server. Therefore, the users have to assume that the VPN server is trustable, and it will not extract our information. The IP information that we are trying to hide is transparent to the VPN server. But when the users are using the Onion network, the experiment demonstrated that no single onion node can know both sender and destination’s information [18]. Therefore, using the Onion network is safer and more effective in terms of anonymity since the users do not need to assume any node is trustworthy so that we can let it know our information.

**Disadvantages**

When using onion networks, we need to choose several onion nodes as the routing path for sending the message to the destination. The selection of onion nodes is not based on efficiency but is used to employ anonymity when sending messages [20]. Therefore, the time used for data transfer using onion networks will be slower than directly finding the routing path using routing algorithms like Bellman-Ford algorithm.

Also, the entry node of the onion routing path will know the source information and the exit node of the onion routing path will know the destination information [22]. So, if someone manages to be both entry and exit node of the routing path, they can know the information of both source and destination. If this happens, the anonymity of data transfer using onion networks will be broken.

* + 1. **Adblock plus**

**Technique Introduction**

Adblock plus is a free source that was developed in 2006 by Wladimir Palant. It supports a wide range of web browsers, including but not limited to Google Chrome, Firefox, Microsoft edge, Safari, and so on [23].

One of the main purposes of web tracking is for the companies to provide personalize advertisement for the users to enhance the effectiveness of the advertisement. The main idea of Adblock plus is to block the advertisements from the websites so that the users will not be annoyed by the advertisements. Therefore, it can eliminate the major effect of web tracking [23].

The mechanism of Adblock Plus is that it has a list of blacklisted advertisements and a list of acceptable advertisements. Based on the unobtrusiveness that is defined by AdBlock Plus, the advertisements may go to the acceptable list and Adblock Plus will not block it if they are determined as unobtrusive [23]. After having the lists, Adblock Plus would determining whether the traffic coming from the advertisement server is acceptable or not [23]. If it is not acceptable, there are two possible ways Adblock Plus may handle the traffic, either block the browser request or simply do not render the advertisement [23].

Due to the convenience and avilability of Adblock plus, it is the most popular Internet browser extensive for many browsers including Mozilla, Firefox and Google Chrome [24]. Overall, it has 85% market share among all the advetisment blocking tools [24].

**Experiment**

According to “Adblock Plus Efficacy Study” [23], it conducted experiment to evaluate the effectiveness of Adblock Plus. In order for the experiment to be more representative of the situation of normal users, the default configuration of Adblock Plus was used [23]. In the experiment, the participants needed to browse 32 different websites, each for 5-15 minutes [23]. The websites were websites commonly used by normal users that usually contain advertisements like youtube, yahoo, etc [23]. All the traffics were analyzed by the application called Manage Engine’s Netflow Analyzer [23]. The data was then evaluated and analyzed. As a result, if the users were not browsing websites with video traffic, the bytes downloaded by a user when browsing a website was reduced by 40.0%, which is very significant [23]. Even when video traffic was taken into consideration, there was still 25.0% of reduction [23]. Therefore, the experiment showed that using Adblock Plus is effective in achieving the objective of blocking advertisements [23].

**Advantages**

Using web tracking to obtain users information and provide more accurate advertisements is a very common way of target advertisement. It makes many users feel annoyed and insecure. In fact, there are some efforts trying to deal with this issue. For example, the US Federal Trade Commision forced companies to let users to choose whether they accept the online behavioral advertising or not [14]. However, this is not enough to handle the problems caused by all of the websites. Therefore, Adblock Plus provided another option for dealing with this problem [24]. By maintaining an acceptable and unacceptable lists of advertisement, the users can choose to block those advertisements that are obnoxious and allow advertisements that are acceptable [24].

**Disadvantages**

One disadvantage of using AdBlock Plus is that it will significantly reduce the legitimate advertisement income of the websites. Especially for the free websites, they main rely on the advertisement as source of income to maintain operation. In fact, the income from online advertisement is an important part of nowadays economy. In 2013, the total income of advertisement was $57.42 billion dollars, and dropped to $49.5 billion dollars in 2014 [23]. We can see that advertisement is a vital part of today’s economy, and it is affected by the ad blocking tools like Adblock Plus. This is bad for the website content creators and the economy. It could be also bad for the normal users since the free website content creator may be less willing to share content for free due to the ad blocking tools like Adblock Plus.

Another disadvantage of Adblock list is that it may not block the advertisements that are willing to pay [25]. As mentioned, there is a list of acceptable advertisement list that is maintained by Adblock Plus so that the acceptable advertisements can be exempted from blocking. However, commercial factors are also in the consideration of the blocking exemption [25]. Any publisher can request to be in the exception list in Adblock Plus and if successful, they have to pay for licensing fee [25]. Therefore, it is doubtful that whether Adblock Plus would compromise the integrity of the exception list for money.

* 1. **Privacy Enhancement Techniques** 
     1. **Opt-out cookies**

**Technique Introduction**

A screenshot of a computer

Description automatically generated with medium confidence

Fig. 2. Implementation of AdChoices program on the advertisement [28]

Graphical user interface, application

Description automatically generated with medium confidence

Fig. 3. Webpage for advertisement opt-out setting [29]

Opt-out cookies is introduced to offer users an option to stop the advertisement agencies (ad agencies) using the collected data for online behavioral advertising (OBA) [29]. This program has been supported by the Digital Advertising Alliance (DAA) and Network Advertising Initiative (NAI), which the DAA are joined by multiple advertising associations in the US, covering more than 80% of national advertising [29,30]. In practice, the user can get into the opt-out page which allows the user to opt out advertisements from each ad agencies like Fig. 2 by clicking the icon shown on the top right of the advertisement that joined the program.

**Experiment**

According to experiment done by Sakamoto, T., and Matsunaga, M. [29], the effectiveness of the program has been evaluated and it involves the following steps. First, the Amazon Alexa’s top hundred news category sites were visited, and their cookies are stored as Dataset 1, which aimed to collect both identifiers set by ad domains and cookies of ad agencies joined DAA or NAI [29]. Second, both DAA and NAI opt-out page were visited, activate the opt-out cookies automatically and stored the cookies as Dataset 2 [29]. Finally, the top hundred sites were visited again with cookie data in step 2 and the cookies were saved as Dataset 3 when finished, which aims to analysis if the ad agencies joined DAA or NAI changed their behavior after using opt-out cookies [29].

Their findings as well as other advantages or disadvantages are listed in the following section.

**Advantages**

Unlike private browsing which the functionality depends on each browser and its version, opt-out cookies performed the same when the user visit the same website from different browsers with the same opt-out cookies setting.

**Disadvantages**

First, few disadvantages can be found on setting up the opt-out cookies. The opt-out cookies expired periodically with most of them have at least 5 years limit while a few only last for less than five years, which means that the user need to update their preference periodically and manually [28, 29]. In addition, user may unnoticedly remove the opt-out cookies when clearing the cookies on their browser [28]. On the other hand, study by [31, 32] shows that majority of user do not understand the opt-out feature, which believes that it can stop both tracking and OBA, clicking the “AdChoice” icon will trigger more advertisements with same category or brings up pop-up advertisement, or small portion of people even though the opt-out page is a scam.

On the other hand, the program only stops the ad agencies to offer OBA based on the user data collected, they can still collect data for user tracking which one of the NAI members Yahoo did, and same behaviour can be shown from [29] experiment that half of agencies keep tracking when opt-out [31, 33].

* + 1. **Private browsing mode**

**Technique Introduction**

The goal of private browsing mode is to hide the users identify from both offline and online attackers, such that the websites cannot link the user activity in public browsing mode and private browsing mode, and the person taken control the machine should not be able to obtain any information regarding the user activity in private browsing section [34]. This goal is achieved by a separate temporary browsing mode that deletes browsing records and cookies when terminating the browsing section. [35]

**Experiment**

|  |  |
| --- | --- |
| Model | Dell Precision PWS 490 |
| RAM | 3.25 GB |
| Storage | 300 GB hard disk (NTFS format) |
| OS | Windows XP Professional Service Pack 2 |
| Tested software  version | Internet Explorer version 8.0.6001.18702/  Firefox version 3.6.11/  Chrome version 7.0.517.41 |

Table 2. Specification of machines used on [36] experiment

A test done by Said, H., et al. [36] explored the effectiveness of private browsing mode in three common browser including Chrome, Firefox and Internet Explorer and they were executed on one of the three workstations with same specification shown on Table X. In addition, three unique list with URLs or keywords for searching were prepared to be inputted to one of the workstations, ensuring multiple machines won’t have same input and ensure the accuracy of the test. The steps of the test were as follows:

First, private mode was activated, and the URL in unique list were inputted and searched to each browser. Second, all links within the visited pages were visited through new tab or windows. Third, all keywords from the unique list were searched through a variety of search engines or forum search options to simulate real-world internet usage. After these stage, the private browsing sections were terminated, and the physical memory was captured for analysis. The whole process was repeated for three times on each machine.

On the other hand, the experiment done by Aggarwal, G., et al. [34] conduct a survey on Firefox which allows extensions or plug-ins functional on private mode, evaluating if any JavaScript-only extensions violated the mode. They tested the most popular 40 extensions listed on the official Mozilla website one by one with a period of time by simulating daily usage of browser. During the test, they track the files and data written to see if the extension violated the mode.

Their findings as well as other advantages or disadvantages are listed in the following section.

**Advantages**

First, private browsing was proved to be effective on clearing the cache or history file, which could not be found on the common place where browsers store those data, unless the attacker can access the RAM of the user device [36]. This also preventing the website linked user in public section to user in private section, as well as user in one private section to user in another private section [34].

In addition, private browsing mode is available on most of the web browsers such as Google Chrome, Firefox or Safari by simply activate the mode in the menu bar and it provides the same protection across all webpages visited within the mode unlike opt-out or DNT which the protection or functionality depends on each website.

**Disadvantages**

First, unlike opt-out cookies and DNT, the protection of private browsing mode highly depends on the design of each browser, and protections from different versions of same browser may vary. For example, the safari version tested by [34] did not disable the previous cookies, history and HTML5 storages during private mode, which offers the attackers or ad agencies collect the link the user in private mode to the same user in public mode, while other browser like chrome or Firefox did not have the same vulnerable design.

Meanwhile, browser extensions may also affect the protection of private browsing. Although some browser like Chrome disables extensions while private browsing, Firefox do not and most of the extension did not consider the private browsing mode guidelines when designed like maintaining a list of special URLs, storing timestamp or allow specific processing in some website, which let attackers understand the private browsing activity. [34, 37] To make the situation worse, official SDK for developing extensions cannot ensure full compliance with private mode [37].

In addition, both disadvantages mentioned above may results in user difficulties on understanding the current protection and configuring the settings of their currently used browser’s private browsing mode.

Finally, although private browsing mode offers higher protection compared to opt-out and DNT by removing cookies which is commonly used for tracking the user habits or preferences, when the section terminates, it still cannot protect against fingerprinting – another common tracking technique [34, 38].

* + 1. **New type of HTTP header field – Do Not Track (DNT)**

**Technique Introduction**

Supported by multiple common browsers including Firefox and Safari, DNT is just a HTTP header to let the website know the user preference regarding the web tracking [28].

**Experiment**

An experiment done by Acar, G., et al. [39] tested the effectiveness of DNT by visiting top 3000 domains from Amazon Alex for multiple times via Amazon EC2 instances with three Firefox configurations – all cookies, all cookies with DNT turned on, blocking third-party cookies.

Similarly, Balebako, R. [33] ran a test to compare DNT to other privacy-enhancing tools like opt-out. They ran the test on virtual machines with Firefox 7.0.1 and Windows 7 installed, applying 5 different topics for training and test on five test pages with behavioral training [33].

Their findings as well as other advantages or disadvantages are listed in the following section.

**Advantages**

Compare to Opt-out and private browsing mode, DNT is the easiest to be activated by simply clicking a toggle to turn on in browser setting for a single time and it also implemented by most browsers [28].

**Disadvantages**

Among these three solutions, DNT have the worst or even questionable effectiveness on protection from web tracking problem. There were only two out of hundreds website support DNT based on the [33] test and DNT did not limit behavioural targeting of advertisements. Results of [39] also supports this with its results shown only 2.9% reduction on number of domains involved in synchronization and 2.6% reduction on number of IDs involved in synchronization when activating DNT. In addition, among the large enterprise, only Pinterest and Medium prominently support DNT and claims the user data won’t be tracked or sent to third party [40]. As there are no penalty for ignoring DNT, companies like Yahoo and Twitter can regret respecting DNT, and companies like Google and Facebook ignored it from the beginning [40]. Finally, the DNT header can even identify the user more uniquely due to the limited of user based has this HTTP header set [38].

* + 1. **Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Opt-out cookies | Private Browsing | DNT |
| How to activate | Select ad agencies that do not wish to offer OBA  through opt-out website | Open new private browsing window through browser’s menu | Toggle DNT in browser setting |
| Protection coverage | Only on the ad agencies selected by the user | History, cookies are deleted when private mode terminated | Always effective if DNT toggled on |
| Effectiveness | Webpage can still track user if they wish  Limited ad agencies support this feature | Website cannot link to profile of one user in public mode to user in private mode, as well as one user in private mode to user in another private mode | Questionable effectiveness which most website ignore this feature |

Table 3. Comparison on all Privacy Enhancement Techniques

**3.3 Tracking Audit Tools**

There are different types of Web tracking techniques which can be essentially classified into stateless or stateful tracking. Stateless tracking sets identifiers for distinguishing users by using means such as cookies and ETages while stateful tracking makes use of the fingerprint of Web browsers or operating systems for identification of users. In view of that, we have surveyed three of existing tracking audit tools for users to recognize the web-trackers as follows, which include TainDroid, FPDetective and MindYourPrivacy.

**3.3.1 TaintDroid**

**Technique Introduction**

TaintDroid is an efficient system-wide dynamic taint tracking and analysis system capable of simultaneously tracking multiple resources of sensitive data on smartphones and an information-flow tracking system for real-time privacy monitoring. It can be used to monitor the behavior of 30 popular third-party Android applications. Monitoring sensitive data with TaintDroid provides informed use of third-party applications for phone users and valuable input for smartphone security service firms seeking to identify malicious applications.

TaintDroid is an extension to Android mobile-phone platform that tracks the flow of privacy sensitive data through third-party applications, and it assumes that the downloaded, third-party applications are not trusted and monitors-in real-time how these applications access and manipulate users’ personal data by detecting when sensitive data leaves the system via untrusted applications and facilitating analysis of applications by phone users or external security services.

TaintDroid automatically labels data from privacy-sensitive sources and transitively applies labels as sensitive data propagates through program variables, files, and inter-process messages. When tainted data are transmitted over internet network, or otherwise leave the system, TaintDroid logs the data’s labels, the application responsible for transmitting the data and the data destination. The real-time feedback gives users and security services greater insight into what mobile applications doing and can potentially identify malicious applications.

**Advantages**

There are several advantages on using TaintDroid for privacy data.

First, TaintDroid can log the outflow or inflow of any sensitive data to protect the privacy-sensitive data for smartphone users by automatically labelling data from privacy-sensitive sources and applies labels when sensitive data propagates through program variables, files and inter-process messages or transmits over network or leaves the system. Such real-time feedback can provide users with greater insight into what mobile applications are doing and identification of malicious or misbehaving applications.

Second, the performance overhead of TaintDroid’s runtime is of minimum. TaintDroid leveraged Android’s virtualized architecture to integrate four areas of taint propagation in variable-level, method-level, message-level and file-level. It can make the appropriate trade-off between performance and accuracy of data propagation detection.

**Disadvantages**

Apart from benefits that TaintDroid can provide, there are also several limitations for TaintDroid.

First, smartphones are resource constrained. The limitations of smartphones preclude the use of heavyweight information tracking systems such as Panorama.

Second, third-party applications are entrusted with several types of privacy sensitive information. The monitoring system must distinguish multiple information types requiring additional computation and storage.

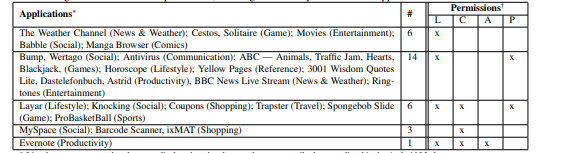
Third, context-based privacy sensitive information is dynamic and can be difficult to identify even when sent in the clear. Geographic locations are pairs of floating-point numbers that frequently change and are hard to predict.

Finally, applications can share information. Limiting the monitoring system to a single application does not account for flows via files and IPC between applications, including core system applications designed to disseminate privacy sensitive information.

**Experiments**

The experiment is conducted 30 randomly selected popular applications (an 8.4% sample size) in Android Market out of 358 applications which require Internet permissions along with permissions to access either location, camera, or audio data. The experiment includes starting the application, performing any initialization or registration that was required amd manually exercising the functionality offered by the application. The system logs recorded using TaintDroid include tainted binder messages, tainted file output, and tainted network messages with the remote address.The experiment is conducted in May 2010 for around 100 minutes generating 22,594 packets(8.6MB) and around 1,130 TCP connections. The experiment also checks if either implicit or explicit user consent is required for exporting sensitive information to identify any possible violations of privacy is involved.

**Results**



^All listed applications also require access to Internet

Table 4.1 Applications grouped by the requested permissions (L:location,C:camera,A:audio,P:phone state).

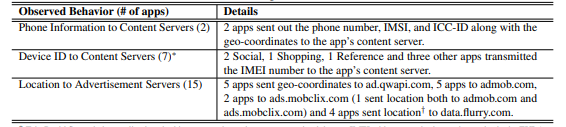


Table 4.2 Potential privacy violations by 20 of the studied applications.

Note that three applications had multiple violations, one of which had a violation in all three categories

As shown in Table 4.2, the findings are summarized. TaintDroid flagged 105 TCP connections containing tainted privacy sensitive information. The study of 30 popular applications shows the effectiveness of the TaintDroid system in tracking applications’ use of privacy sensitive data accurately. There were no false positives generated except the IMSI taint source which is disabled for experiments. The flags from TainDroid can help identify the potential violation of privacy for the tested applications. Half of the studied applications share location data with advertisement servers and one third of the applications export the device ID, sometimes with the phone number and the SIM card serial number. There was no perceived latency, running experiments with TaintDroid.

**3.3.2 FPDetective**

**Technique Introduction**

FPDetective is designed as a flexible, general purpose framework that can be used to conduct the web privacy studies which are large in scale. FPDetective focuses on the detection of fingerprinting of web.

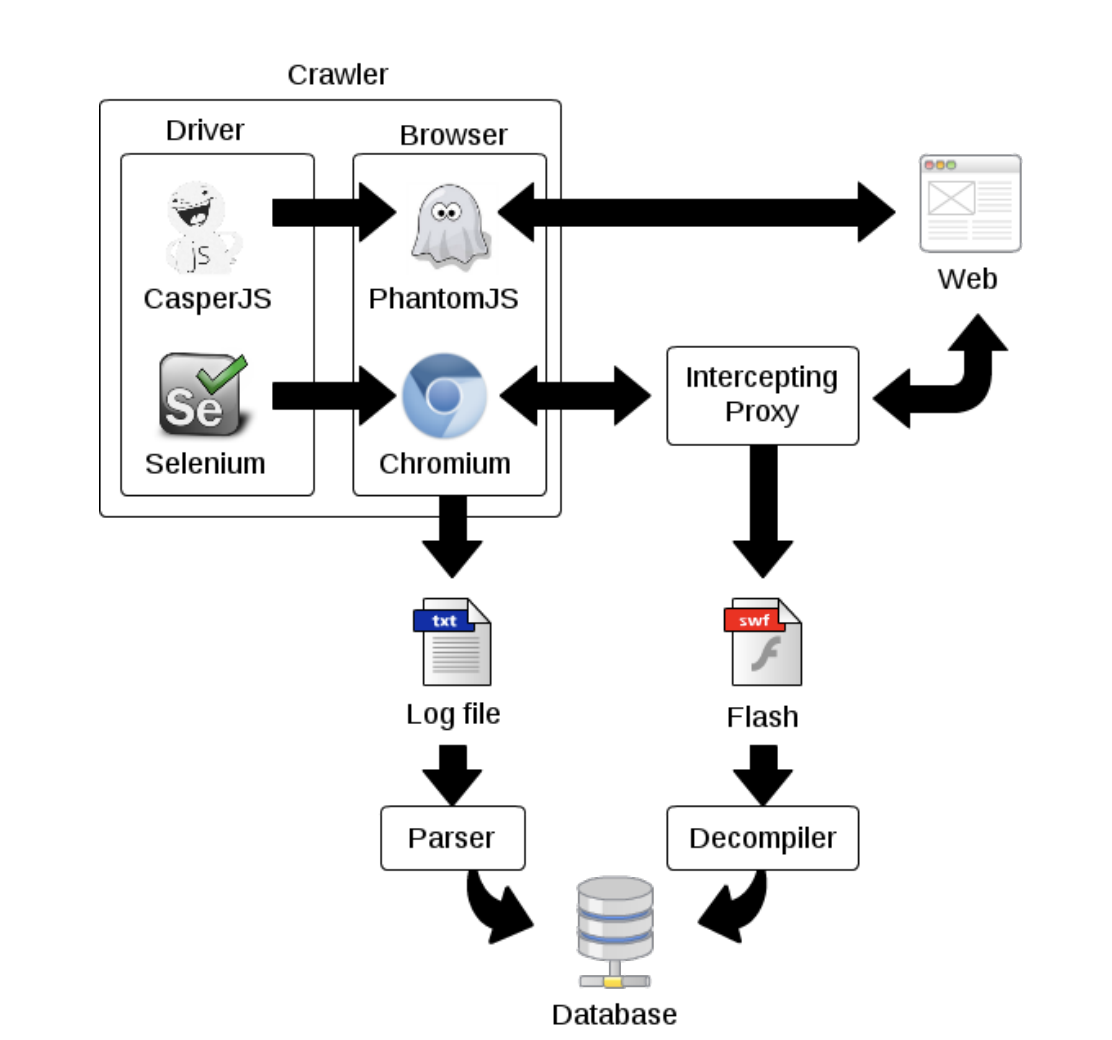


Table 5.1 FPDetective Framework

As shown in Table 5.1, the components of FPDetective include crawler, parser, intercepting proxy, decompiler and central database. The crawler’s purpose is to visit websites and collect data about events that might be related to Flash-based fingerprinting, such as the loading of fonts or accessing specific browser properties. The parser is used to extract relevant data from the logs generated from the crawler, and store the data into the database. The parser also tags sites with a label if a known fingerprinting script is found in the HTTP requests made for this visit. For the intercepting proxy, to obtain Flash files from static analysis, it is redirected traffic through mitmproxy which is an SSL-capable intercepting proxy. The mitmdump module to log all the HTTP traffic passing through the proxy, and the libmproxy library to parse and extract Flash files based on content sniffing. The decompiler is used to decompile Flash files and obtain the ActionScript source code. The source code can then be searched for fingerprinting related function calls, for example enumerateFonts and getFontList to obtain a binary occurrence vector. For central database, crawls are run with several machines, but using a central database to store, combine, and analyse the results of different crawls with minimal effort. The data stored in the central database include the set of JavaScript function calls, the list of HTTP requests and responses, and the list of loaded or requested fonts. For the Flash experiments, a binary vector is stored and it represents the occurrence of ActionScript API calls that might be related to fingerprinting.

**Advantages**

For using FPDetective, it can identify 16 new fingerprinting scripts and Flash objects (including commercial fingerprinting including in-house solutions).

First this app can be used as a means to protect users and web applications against malevolent actors. For example, by detection of the use of stolen credentials and identification of Sybil attacks, some security issues like identity theft or unauthorized actions can be prevented in advance.

Secondly, Stateless user tracking allows advertising companies to sidestep the limitations imposed by regulation on cookies in Europe and the United States. With many smartphones and tablets, fingerprinting allows advertisers to augment previously gathered user-data and track the user across devices. By identifying the fingerprinting of user devices through the browser, stateless user tracking can be prevented.

**Disadvantages**

Using FPDetective, it is able to identify companies engaging in web-based device fingerprinting. However, it is not possible to infer the purposes for which fingerprinting is being put to use by these companies. Companies can express that they deploy fingerprinting in the context of variety of web services with good intention like fraud detection, or protection against account hijacking, anti-bot and anti-scraping services and DDOS attacks.

**Experiments**

The FPDetective framework is used to crawl the top Alexa websites while searching for instances of web-based device fingerprinting. The experiments were seperated into the ones geared towards the discovery of JavaScript-based fingerprinting attempts and the remaining ones towards the discovery of Flash-based fingerprinting. For each type of experiment, Font-detection attempts is first automatically searched and used for selection of candidates for manual analysis. For the JavaScript experiments, it is attempted to find JavaScript-based font probing attempts, for which the number of requested fonts Is used as a measure. To filter out websites that load high numbers of fonts, but not used for probing, it is checked if websites measure the width and height of displayed text by using the number of calls to the offsetWidth and offsetHeight properties of the corresponding HTML elements. For Flash experiments, the sites are crawled with Chromium and Flash objects are intercepted with mitmproxy. The discovered Flash files were decompiled and a list of ActionScript API calls are searched for that might be relevant to fingerprinting technique. The binary vector representing the occurance of each function call in the Flash object is stored into the central database and used for selection of objects for manual analysis.

**Results**

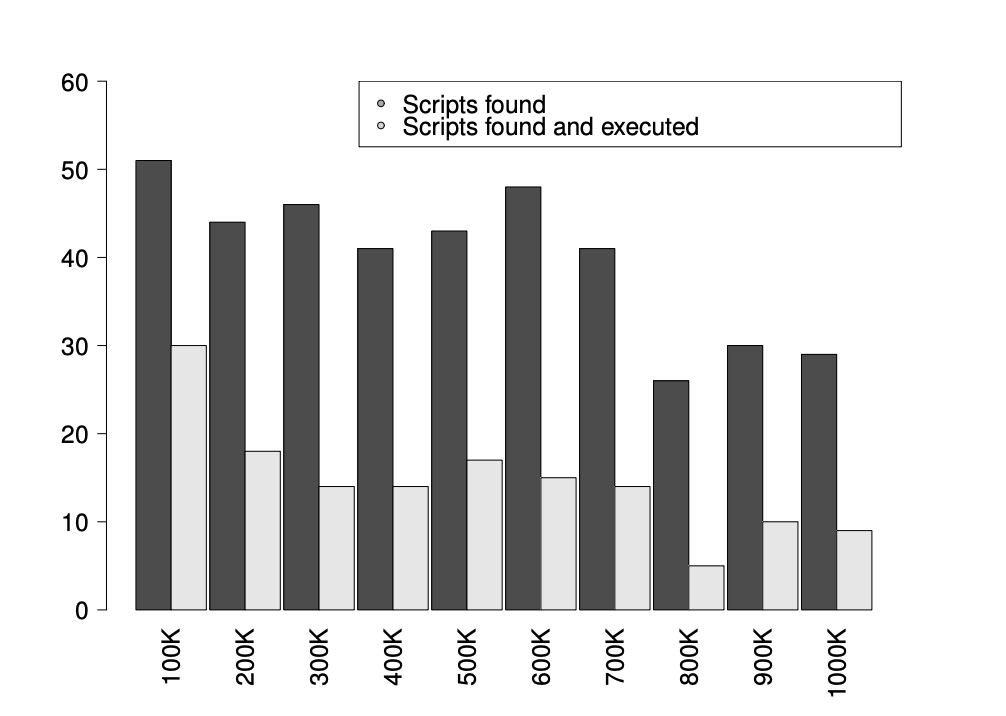


Table 5.2 JavaScript-based font probing scripts on homepages of Top 1 million Alexa sites

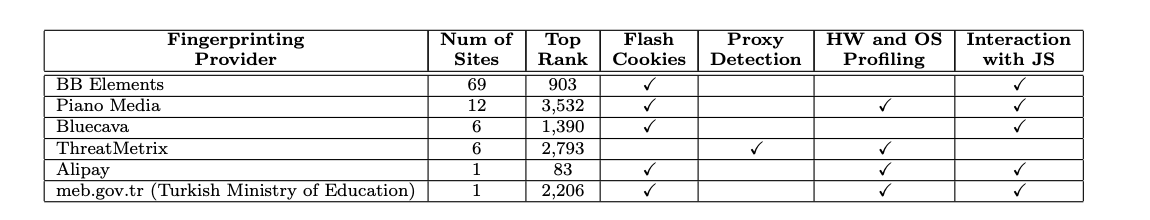


Table 5.3 Flash Fingerprinting objects found on Top 10K Alexa websites

For JavaScript-Based Font probing, the top Alexa websites with FPDetective were crawled to find out the extent of the JavaScript-based font probing. The sites that fingerprint users on the inner pages and the sites that fingerprint users on the homepage are compared. The histogram in Table 5.2 represents the distribution of sites using JavaScript-based font probing among the Top 1 million Alexa sites. The histogram is divided in intervals of 100,000 sites according to their rank in Alexa, which are divided further into two bins. The darker bin refers to the total number of sites in that popularity range that served a fingerprinting script. These scripts are counted by matching previously discovered script URLs or regular expressions to URLs of HTTP requests made while visiting the site.

For Flash-based font enumeration, the top 10 thousand Alexa websites using FPDetective and for each site, the homepage is visited and waited for 10 seconds to allow for resources to load completely. The function calls used to send the collected information to a remote server or functions for opening JavaScript interfaces to either call JavaScript functions from the Flash file or allow the scripts in page to make access into Flash functions. In Table 5.3, it is shown for the results of Flash-based fingerprinting. For example, BB Elements which offers tools for managing ad campaigns were found on 69 sites. Piano Media is mainly found on newspaper sites and claims to emply fingerprinting for paywall enforcement. Apart from these, there are also some fingerprinting providers found in different sites, and they are Bluecava, ThreatMetrix, Alipay and meb.gov.tr, contained into number of sites 6,6,1 and 1 respectively.

**3.3.3 MindYourPrivacy**

**Technique Introduction**

It is a visualization tool used for clarifying the entities threatening users’ privacy and visualization of third-party Web tracking. It adopts deep packet inspection, DNS-SOA-record-based categorization, and HTTP-referred graphical analysis to visualize collectors of Web browsing histories without device dependency. It can be used to show and visualize collectors of private information. MindYourPrivacy shows users’ web tracking information by using a tag cloud technique and provides graphical files for Web tracking analysis.

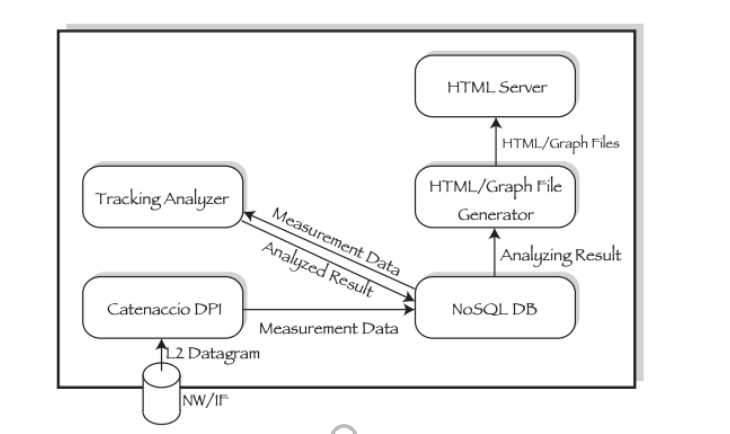
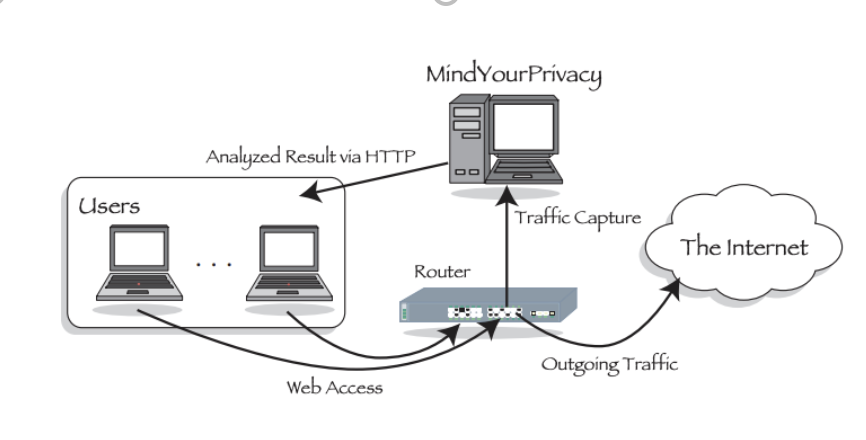


Table 6.1 MindYourPrivacy’s Implementation Architecture

As shown in Table 6.1 above, the system model of MindYourPrivacy captures users’ Web browsing traffic at gateways, analyzes the browsing traffic and displays the results of analysis to users by HTML/Graph files via HTTP server.

**Advantages**

First It is device-independent using a deep packet inspection technique. It can be not only used for gain information from personal computers but also mobile devices for web tracking and web-tracking counter-technologies must be device-independent to gain holistic information. The deep-packet-inspection-based approach has an advantage for deployment and offers reduced implementation costs because of browser and device independence.

Second, MindYourPrivacy provides results of analysis in the form of an HTML file via an HTTP server to facilitate users’ access. Users are therefore not required to install any specific applications or browser add-ons to access and use the results. MindYourPrivacy expresses results visually in tag cloud format. In Table 6.2, it is showing the analyzed results of a Web browsing history by the Web user interface provided by MindYourPrivacy. It shows Facebook, Twitter, Google and Hatena potentially collected the browsing history. Also, a Graphviz .dot file and a Cytoscape .sif file are available to download for the more sophisticated analysis.

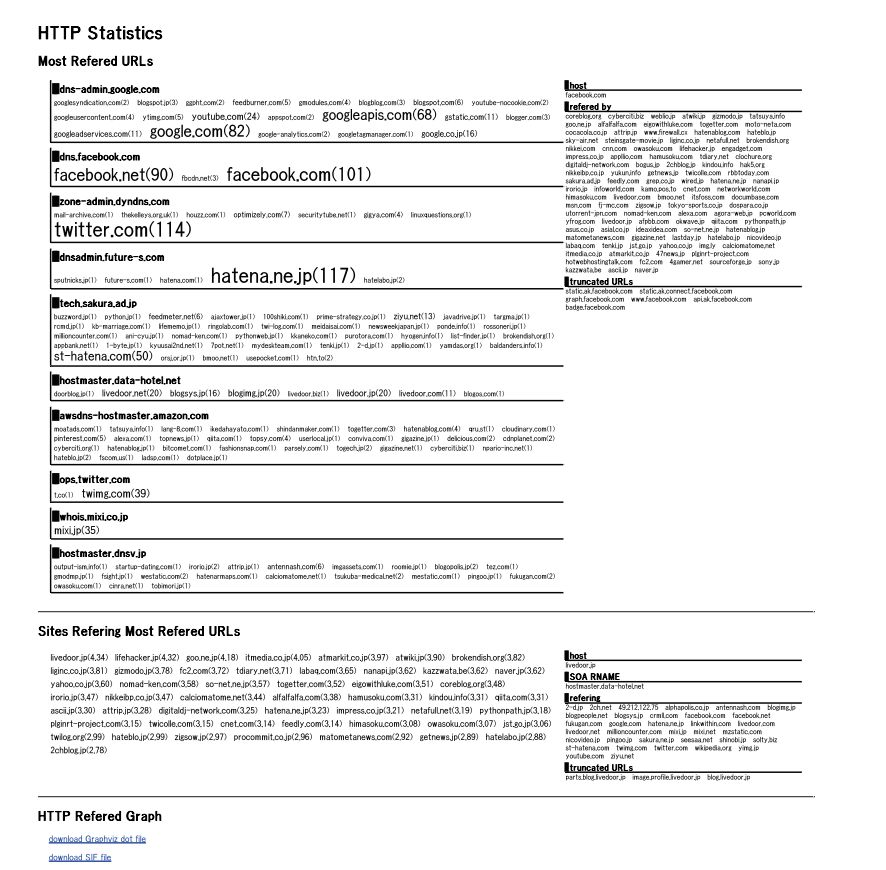


Table 6.2 MindYourPrivacy’s Web User Interface

**Disadvantages**

There are cases the actual operator differs from the registered entity for the DNS. For example, one can have a private domain name and run his or her own website under that name. The person can ask a hosting service provider to register the domain name under its own name rather than that of the registrant. The scheme cannot analyze HTTP traffic correctly if the referrer field of the HTTP header is omitted.

**Experiments**

An experiment in an IT technology camp called WIDE camp is conducted and has 129 attendees discussed IT technologies for four days. They were either IT specialists or students majoring in IT. There were two types of experiments conducted including user traffic analysis and questionnaire-based use analysis.

**Results**

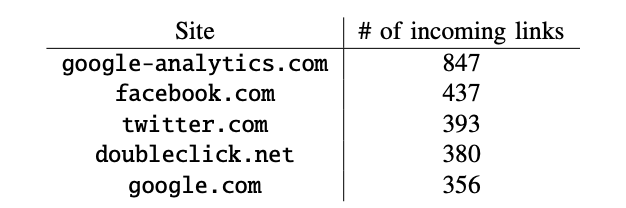


Table 6.3 Top-five Most-referred Sites

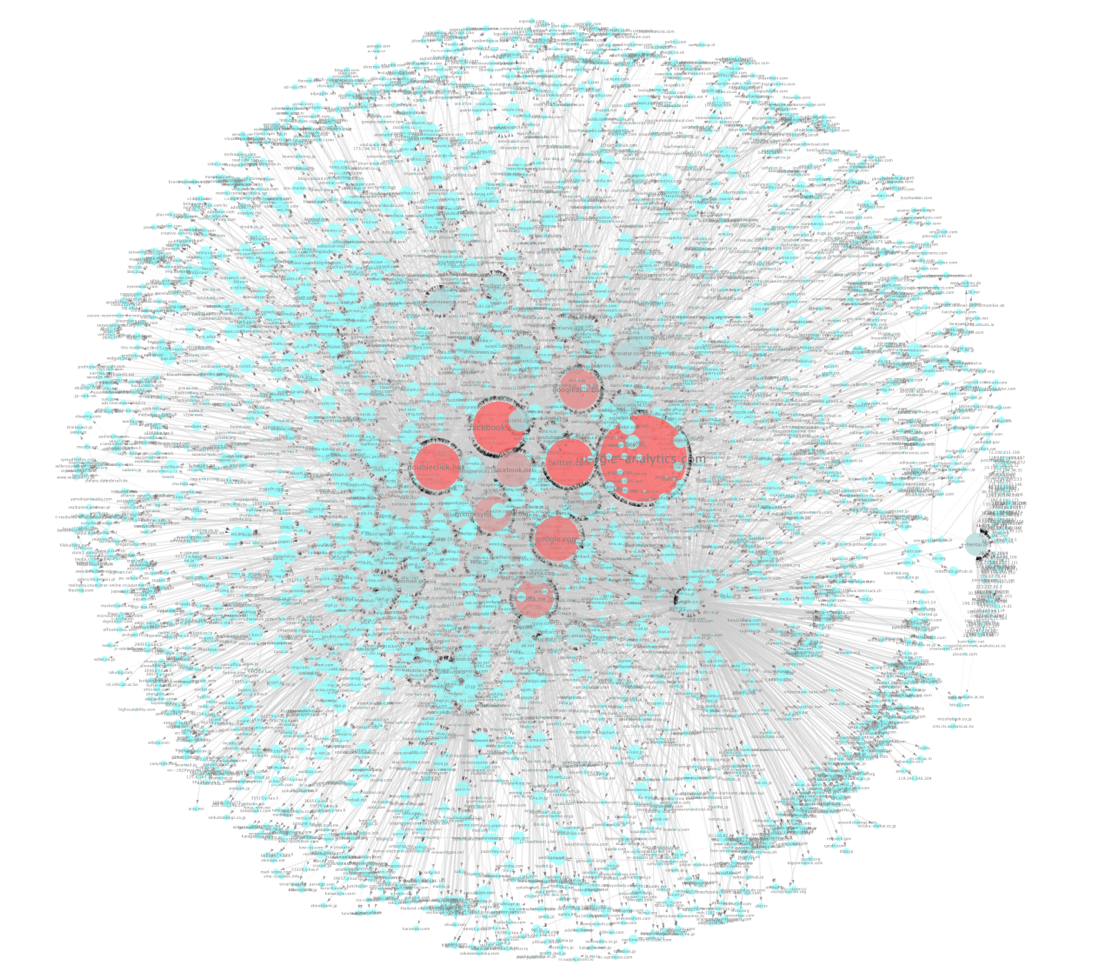


Table 6.4 HTTP Referrer Graph of WIDE Camp Attendees

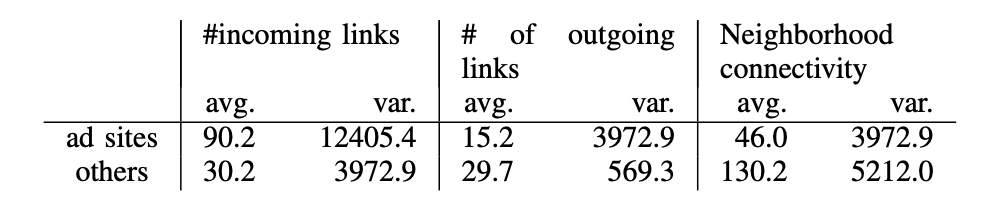


Table 6.5 Feature Vector of Rank 1 Cluster’s Edge

The results of experiment showed that visualizing Web tracking effectively affects users ‘perception of privacy. Analysis of the user data collected at the camp also showed that MCODE clustering and some features derived from graph theory are useful for detecting advertising sites potentially threatening users by collection of user information by Web tracking for their own usage.

The attendee’s network traffic is captured for analysis of cookie and ad sites. As shown in Table 6.3 the number of links for each site are listed and presented as the HTTP referrer graph as in Table 6.4. For the ad sites, they tend to have more incoming links more than outgoing links. By studying the ratio of outgoing and incoming degrees, it can be helpful to detect ad sites as shown in Table 6.5.

The questionnaire-based study is also conducted. A total of 34 (34/56 ~ 60%) people said that they mind Web tracking. Only one person who had minded Web tracking before answered that he or she did not mind it after the experiment.

1. **Future work**

For the tracking defense tools, there are some aspects that can be worked on to improve the effectiveness and efficiency of the tools. Some studies showed that despite the tracking defense tools themselves are effective in defending web tracking, they might not be effective to be used in reality due to the difficulty of using the tools. For example, a paper [14] showed that none of the participants in their study was able to use the tools to achieve the privacy objectives they needed even though the tools had the proper capabilities to achieve them. This shows that the usability of the tools is an important aspect to be worked on. It is because one of the main purposes of tracking defense tools is for normal users to be able to prevent themselves from being tracked by the web trackers, but not only for those who are already expert in the tools.

Also, as the web tracking techniques are evolving, current tracking techniques that can be handled by the tracking defense tools may not be sufficient in the future. Therefore, it is also important for the tracking defense tools to improve to meet the new challenges. For example, a paper [1] listed some possibilities of web tracking techniques in the future. For example, the new tracking algorithms can be device based, client based, network based, and so on [1]. These techniques are being studied by many experts like Google, Amazon, eBay, and so on [1]. They are trying to invent new methodologies to increase the effectiveness of web tracking and pose a great threat to the tracking defense tools that are effective today [1]. Therefore, it is important to monitor the new trend of web tracking and foresee the possible tracking techniques in the future and develop techniques to defend them if the current tools are not sufficient to defend them.

For the privacy enhance techniques category, multiple future works have been suggested by the papers. First, a paper [34] pointed out that the private browsing mode only defend limited local or online attackers, future research is necessary to offer stronger privacy guarantees with no degrading on user experience. For example, some browses allowed extension operated during private mode while extension often designed without the consideration of private mode, resulting the degrade of privacy protection which the normal user would not noticeable. Second, education to normal user is necessary as most non-regulatory methods including opt-out cookies required users own understanding and decision, which many users found difficult or even misunderstanding on the functionality [31]. Finally, the investigation of the effectiveness of opt-out cookies mainly focuses on the US through the ad agencies joined the DAA or NAI, the effectiveness on other regions remains questionable [29].

For the tracking auditing tools, the mentioned auditing tools can help in find out fingerprints leakage, or some malicious webs or Apps to invade users privacy intrinsically and without explicit requests of users’ consents. According to the experiment in the paper about TaintDroid, there were an 8.4% samples from 358 well-known Apps from Android Market which include some commonly-installed Apps like News &Weather or Social Media or Personal Health selected. The astonishing results show that 20 of these Apps potentially violate users’ privacy and secretly obtain users’ sensitive information like device ID or SIM card serial number to identify users. It shows the fact that the security risk is everywhere. The future improvements can focus on a more comprehensive security auditing software that audits the traffic with every inputs or outputs from users devices to arouse the awareness of users about any leakage of security and provide some intelligent advices for users better to protect their privacy.

1. **Related Work**

While web tracking is becoming a serious security concern for normal users, a variety of techniques that protect the users against web tracking were introduced [26]. Three main categories of defense tools that are commonly used against web tracking are tracking defense tools, privacy enhancement techniques, and tracking auditing tools [1]. Various research on this topic was done due to the importance of web tracking and defense mechanisms [1].

For tracking defense tools, much research on this topic focused on the generic overview of the tools and discuss their theoretical effects. However, the details of the tools, their advantages, disadvantages, and how they perform in real practice are usually not discussed. Some other research [13, 14] focused on the study of a particular tool such as Microsoft tracking defense tools without categorizing the tools and providing an overview of the tools. In this paper, the category of tracking defense tools was studied and various tools in the category were selected and explained in detail regarding their principles, advantages, and disadvantages. Research on their effectiveness was also conducted.

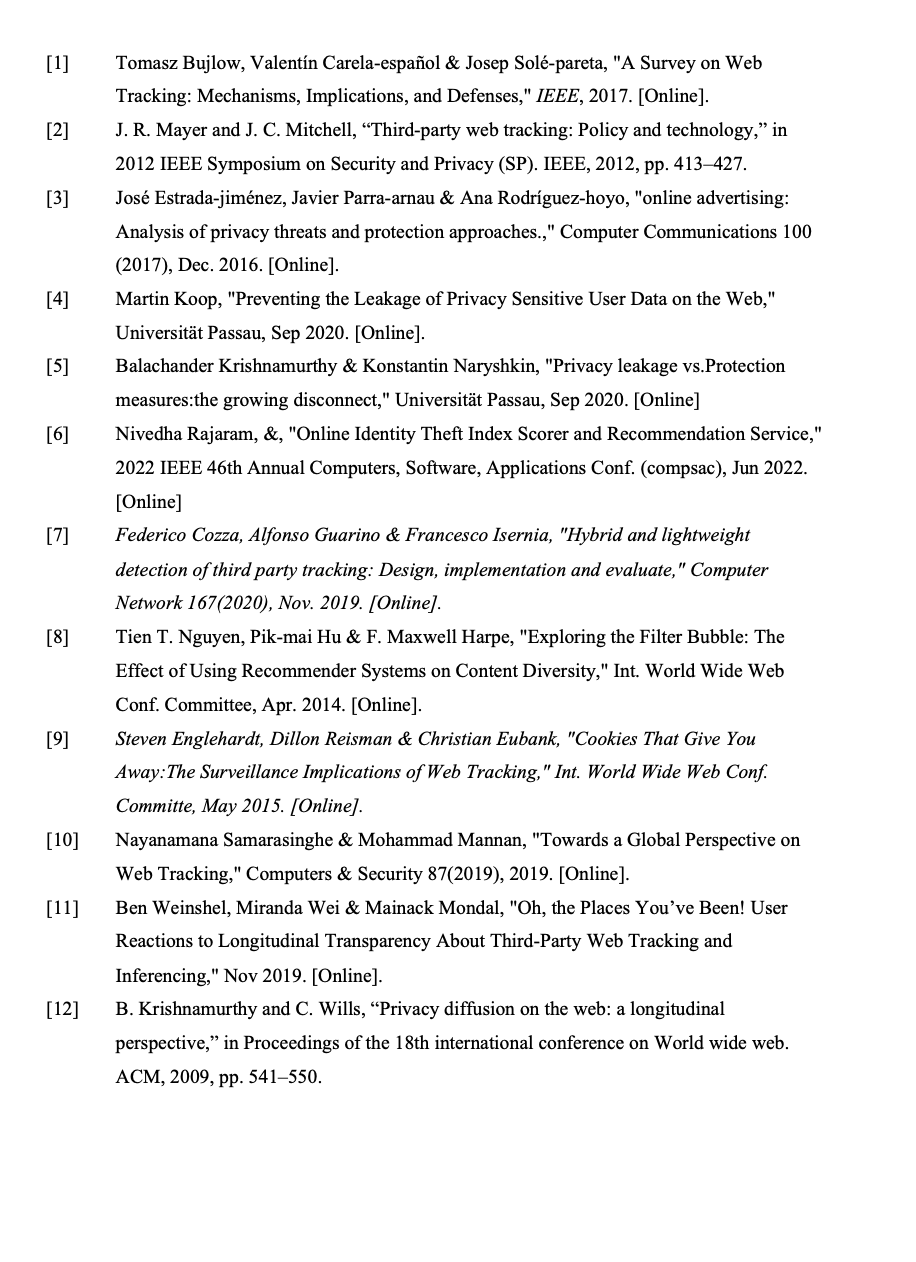
For the privacy enhance techniques category, [28] done a similar survey on third party tracking, but with a larger focus on government and companies’ regulations, economics on tracking and the tracking technologies behind. In addition, both [31, 32] did a comprehensive study on how people react and understand on the ad choice mechanics, which is another significant factor besides the effectiveness of the mechanics. Finally, [38] done research on how the data sent from browser including DNT will generate a fingerprint and offer insights on how to stop from fingerprinting.

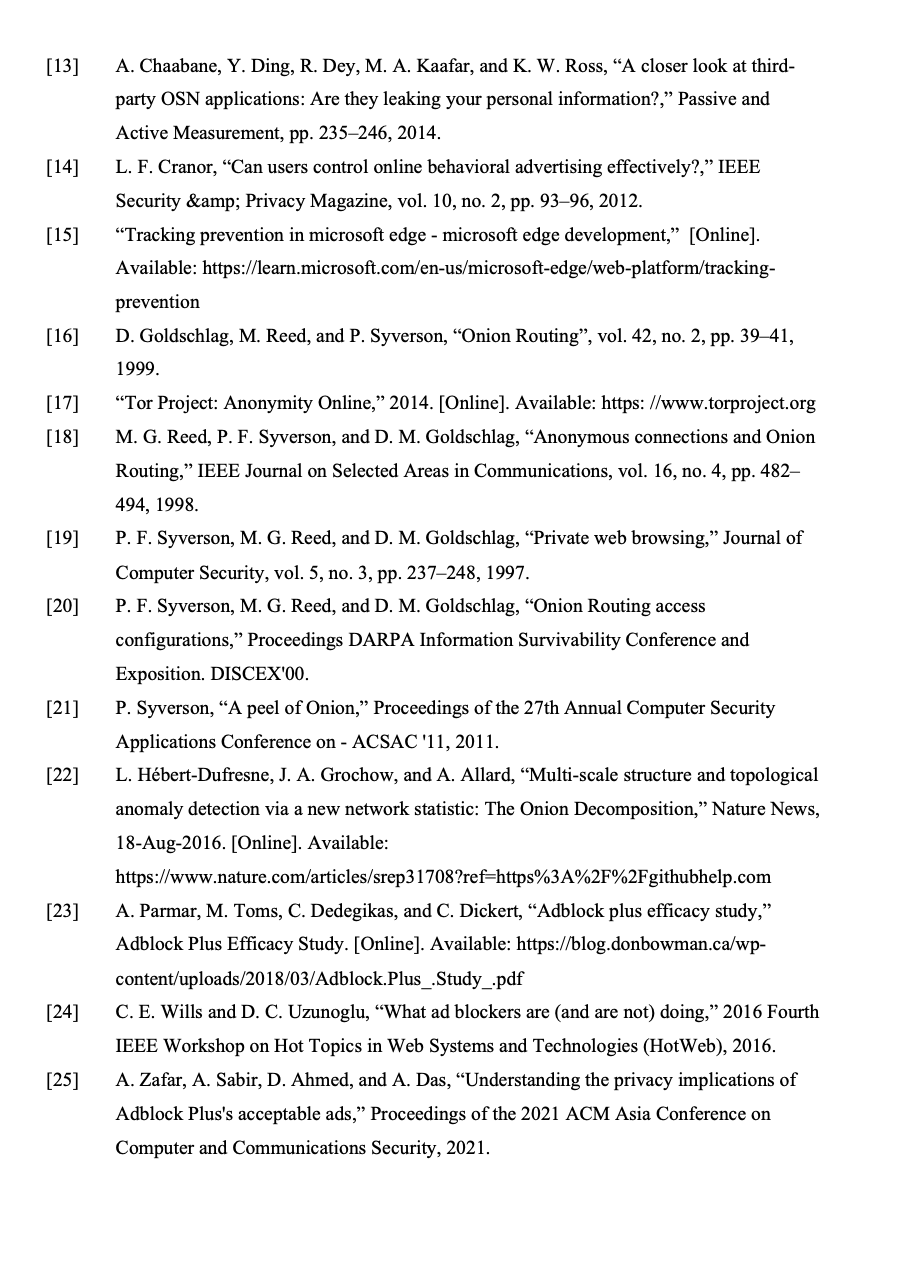
For the tracking auditing tools, we have studied the tracking auditing tools that can recognize mobile devices flow of sensitive information, detect the fingerprinting of web or capture the browsing traffic of web. Apart from these areas, we have still some regions of auditing tools not studied deeply in this paper. According to a paper [47], Roesner et al. makes a discussion on their-party SNSs like Twitter or Facebook that these Web trackers can link real names of individual person to the tracking information they collect. The widgets such as “Tweet” or “Like” can be used for tracking purpose. In this paper we do not mention too much on the mechanism of these social media for tracking of information via webs.

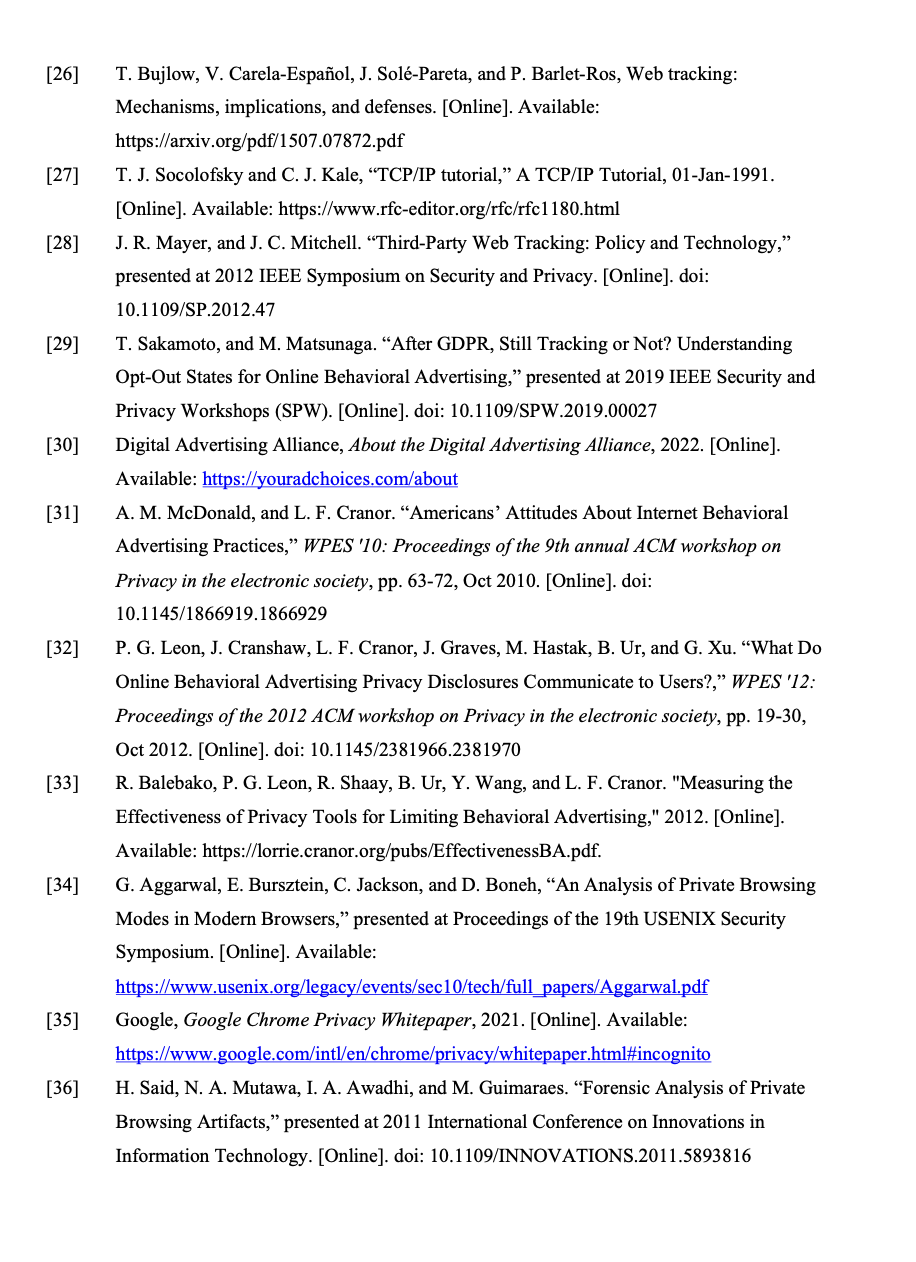
1. **Conclusion**

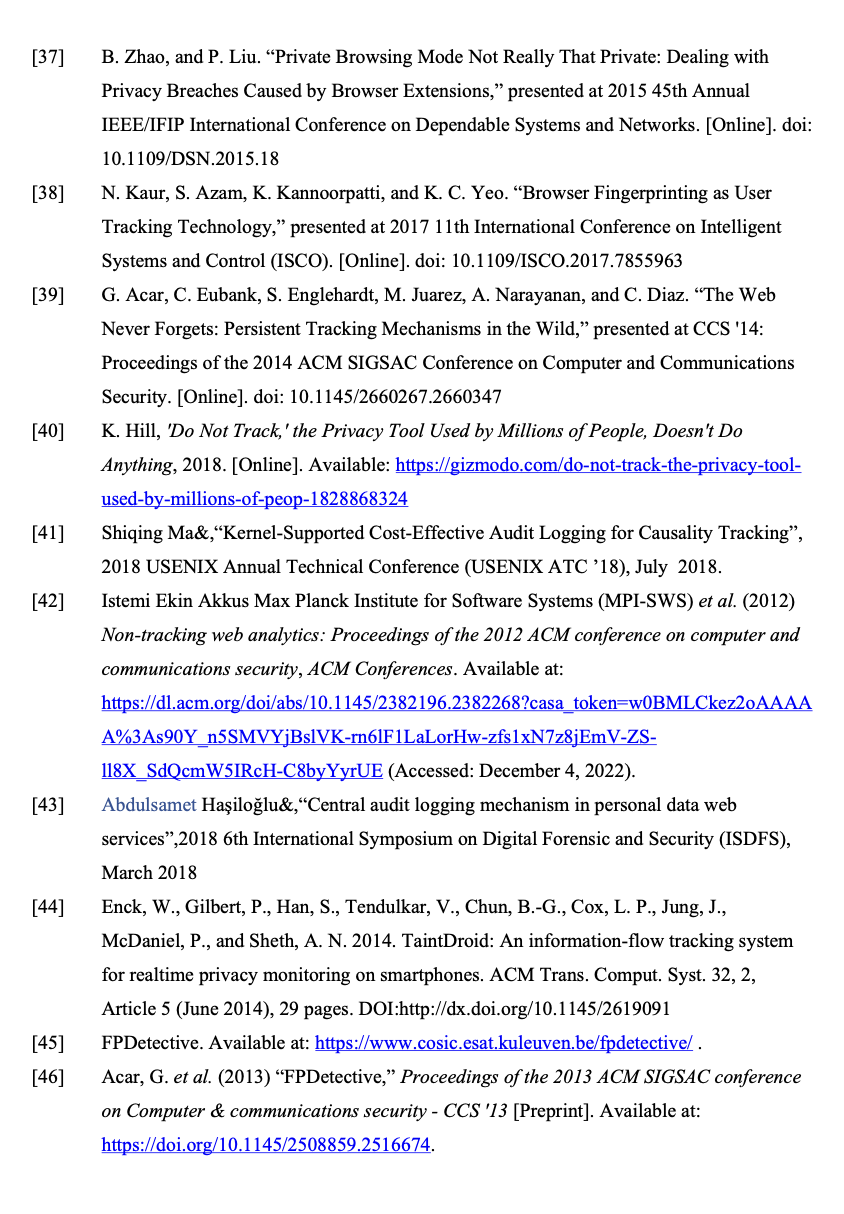
In this paper, we offered a comprehensive view of the current issues of the Internet, including the most significant third-party tracking and identity theft, which may result in business obtains our privacy data to make money or even leak our social security number to online thefts. Nine solutions from three aspects including tracking defense tools, privacy enhancement technique and tracking audit tools have been introduced with its pros and cons. None of the solution is perfect to protect or identify all tracking, some of them have limited scope like Microsoft tracking protection list or private browsing only protects the user while using web browser and some of them like DNT even have questionable performance. As tracking method are evolving and the user data are important on business in the big data era, the investment and study of defensing method will be even significant in the future to protect users from collecting the valuable data unnoticeably.

**References**









Graphical user interface

Description automatically generated with low confidence