Visualizing Trackers Found in Browsing Data Using Privacy Badger

Aryan Srivastava
Brown University

Ragna Agerup
Brown University

William Kuenne
Brown University

Abstract

It has become increasingly difficult for an average person to understand how their data is being used on the internet. On average, 79% of websites globally are secretly tracking you [5]. At the same time, Google is tracking you on 86% of the 50,000 most common websites across the globe [7]. Both these examples paint a picture most people are not aware of because trackers seem invisible to the user on the pages they appear. Furthermore, this has contributed to a misconception of what amount of personal information is exposed online, because it is more than the eye can see. Ghostery, a browser extension that aims to tackle the problem of showing and blocking invisible trackers, have reported a total of about 2,000 trackers on the web [6]. We provide a way for users to collect data about the trackers and in addition, we have created a website to visualize the number of trackers found, on what websites they appear, and who they are. Lastly, we have provided a way for users to permanently opt-out of being tracked by the most notorious trackers found on user's profiles, in just one click.

1 Introduction

1.1 Goals of the Project

Privacy online is almost non-existent, and transparency in data collecting of private information by companies is equally so. There exists trackers everywhere, that track the user's activity and without being extremely observant or interested in privacy online, these trackers are invisible. However, these same trackers are being used by advertisers and analytics companies to track you, profile you, make inferences about your interests and retarget you with ads wherever you go online [8]. Furthermore, it is impossible to know how complete and invasive our profiles are online because we can't easily access them, or know who these trackers are.

Transparency for the users are lacking, and an evening searching the web leads to hundreds of trackers that are now collecting information about you, without your knowledge. Thus, our goals are as follows:

1. Provide transparency and educate users in how tracking online works, even for those who are not confident tech-savvy people.
2. Break down information of how each tracker works, who they are and what websites we found them on.
3. Providing users with the option to opt-out from tracking of each individual tracker.

1.2 The Current Situation

There are currently several tools that help identify and block trackers on the web, some of these being Privacy Badger, Ghostery and more recently browsers such as Firefox and Chrome. These tools seek to solve some of the issues we mention such as blocking unwanted trackers and showing which trackers appear on different websites. However, they don't provide a mechanism for the user to understand the bigger picture of where these trackers come from, how often they appear and how to opt out of them tracking your online browsing.

Privacy Badger [3] allows users to download the data it stores about their browsing history and trackers, which is collected when a user has the extension activated in their browser. This can help the user understand what trackers are blocked, and who the trackers are. However, the data is stored as a json file. A json file is almost impossible to read, and only ever convenient in scenarios when the person using it knows how to use it. In other words, it is not meant for reading by regular people. In our case, it has proved to be a good attribute to tackle our problem, and show the users the answers they are really looking for.

We aim to provide a mechanism for users, and potentially researchers, to use a modified version of Privacy Badger to collect data about their browsing habits, and provide a simple and user-friendly website to display visualizations summarizing the presence of trackers in the user's activity online.

2 Overview

When a user clicks on a website, let's say nytimes.com, there are a bunch of cookies on that page. To be able to read and access material, you accept the cookies. This is visible to you. What you don't visually accept, is trackers, which are still installed on that website to track your behaviour. In the case of New York Times, Privacy Badger found 5 trackers, tracking me (3 of which were blocked and one that was deemed necessary). In contrast, Fox News showed 22 trackers. Without Privacy Badger as an extension, I would not have know about their existence.

Now each tracker stores pieces of information about you. Many companies even do cross-web-tracking, which means that they are tracking your actions across several websites, to build a more in-depth picture about what type of user you are. All to be able to find more ads that can target your profile, and match what they consider your online "needs".

It is a common misconception that trackers don't know that much about you, because how can they possibly? The reality is that it is a lot more than you think. Most likely they know your gender, age, relationship status, education, income, diet, fitness routine, exact location and so much more [8]. We have broken down who these notorious trackers are for you in our product.

3 Design

3.1 Using the modified Privacy Badger Extension to collect data

To be able to get the proper results from our visualization, there needs to me made some small modifications to the Privacy Badger extension that we use to collect data.

First of all, the user needs to install our un-packaged modified Privacy Badger extension. There are instructions to do so on the GitHub that are simple and straightforward. Once a user has installed the modified extension on the browser of their choice (Chrome or Firefox).

Secondly, the user nees to delete the pre-trained seed data that Privacy Badger uses from the settings window. Privacy Badger has a pre-trained seed data set which exists to be better suited to block all trackers found on the internet. However, for our purpose of only finding each specific user's trackers, we do not want to include this data.

Lastly, the last step for it to work is for the user to enable learning based on browsing, and disable the sending of "Global Privacy Control" [9] and "Do Not Track" [2] signals. Upon the completion of these simple steps, Privacy Badger will start collecting data about the user's browsing and the trackers it found.

When the user is ready to download their data (for example, after they have visited the specific websites they wish to get data for), they can simply export user data from the "Manage Data" tab in the settings of the extension. This will download a json file including a list mapping every tracker that was found during the user's browsing to a list of websites it was found on.

4 Visualization of Website

The visualization aims to break down the data in a clear way so that the user has a complete understanding of what websites contain what trackers, and who the trackers are. We created a web page in React.js that allows the user to upload their downloaded Privacy Badger file, which then parses the data and displays it.

We have implemented several features that make it easy to break down the data and show the user how the tracking works. Our first component is a bar chart that sorts the websites combined with the number of trackers found on each website. The bar chart displays the top 10 websites with the most trackers on them. This will give the user a clear picture of what websites track a lot and how many trackers you can find on these pages.

The next component is an interactive button menu, which contains the same top 10 websites from the bar chart. Upon clicking on either of the websites, a list of each tracker that appeared on that websites is shown. Displayed is a list of trackers, sorted in alphabetical order. The websites are displayed from highest to lowest, with the website with the highest number of trackers first.

Lastly, we have implemented a bubble chart that maps the number of occurrences of each tracker on the websites visited by the user. This shows what tracker does the most tracking on you. In addition, we have created a database with the most common trackers and their opt out links, and thus, clicking one of the bubbles will direct you to that tracker's opt out page. There you can not only opt out of tracking, but most often also request to delete the data that each company has stored on you.

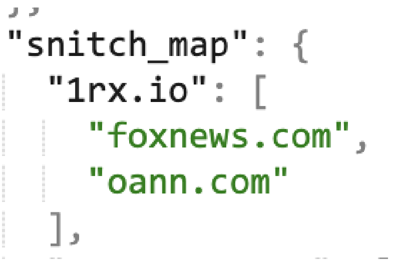
5 Implementation

Our implementation consists of two components. It is the modified Privacy Badger extension that collects the user's data, and the visualization website that breaks down and shows that data.

Our project has been implemented in React.js using JavaScript, HTML and CSS. On the back-end, we created a server where the file uploaded by the user is posted. On the front-end, the file is fetched and displayed to the user through various visualizations. We use express to create the server, and multer to handle the file that is uploaded. We used Nodemon to monitor changes that auto-reloads when changes are made. This makes it so that we don't have to restart the server manually each time.

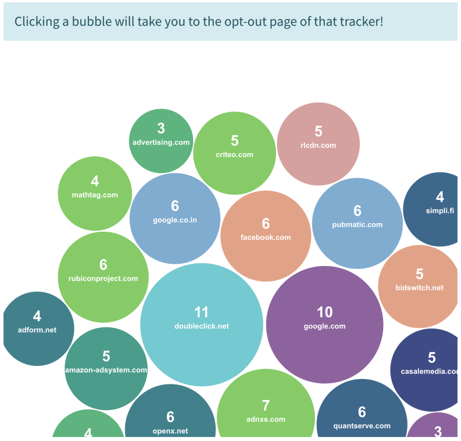
On the front-end side, we have used react bootstrap for styling and react-router-dom to connect the html pages for a smooth transition. Additionally, for the visualization, we have used D3, a JavaScript library which contributes in producing dynamic, interactive data visualizations in web browsers. D3 helped us create the different visualization components, such as the bubble chart displayed below in Figure 2 and the bar chart discussed earlier.

We wrote data processing functions that obtained useful lists and statistics from the 'snitch_map' stored in the json user data file collected from the modifier Privacy Badger. Figure 1. Shows an example of an entry in the snitch_map. The key stores the tracker and the value stores a list of all the websites it was found on.



```
"snitch_map": {
  "1rx.io": [
    "foxnews.com",
    "oann.com"
  ],
```

Figure 1: This figure shows an example entry in the snitch_map of the .json file containing the user data. "1rx.io" represents the tracker which is then mapped to the two websites it was found on.



Clicking a bubble will take you to the opt-out page of that tracker!

Figure 2: Above is a Bubble Chart representing the number of times a tracker was found across different websites. The bubbles are interactive, and on click you can opt-out. As shown, doubleclick.net and google.com are on top, indicating that they have the most tracking on average. Doubleclick.net is in fact a division within Google called the Digital Marketing platform that " makes its money from online advertisers and publishers". Thus, they make money on tracking you. [4]

6 Modifying Privacy Badger

For our visualization, we wanted to get a better sense of all of the trackers online. We learned that the current Privacy Badger model did not store a complete set of the trackers found on the websites we visited, and thus we ended up modifying Privacy Badger to work as we wanted it to. This required changing only one line of code! Moreover, the change was only changing the value of the "TRACKING_THRESHOLD" variable in the constants.js file from 3 to some high value; we changed it to 9999999.

This worked because Privacy Badger blocks a tracker that has been observed tracking on more than TRACKING_THRESHOLD websites. If a tracker is blocked, any new websites it is found on are not added to the collected data. A blocked tracker, if it is a dynamic tracker, also cannot load other trackers. Therefore, making sure Privacy Badger never blocks anything by making TRACKING_THRESHOLD a high value fixes both of these issues.

Lastly, we got a lot of help from one of the developers of Privacy Badger through their GitHub repository's issues page to clarify what we needed to change in order to get our desired result [1].

7 Evaluation

We collected data from a browsing session on multiple news websites. Visualizing the downloaded data from the modified Privacy Badger on our website yielded interesting and insightful results. We found out the most notorious websites and trackers and were able to give the user a way to opt-out of most of them.

Our product is user friendly and an easy, convenient and fun way to find out more about online trackers, in particular trackers specific to the user. In the future, we would like to further improve our final product by incorporating a feature that can opt the user out of all of the trackers that were found, as well as be able to offer the opportunity to request to have all of their data deleted from these notorious tracking companies databases.

8 Limitations

Our Privacy Badger modification requirement can be considered a small limitation to this project, however, it has not been problematic to the completion of our product. Since we depend on the user being able to use our version of Privacy Badger to get accurate results about their trackers, we do depend on them being able to download our version of the code and install it properly on their browser. This can be considered a small complication, however, and it only requires a few extra steps. To address this issue, we have added some very clear instructions on how to deal with that on the main page of our visualization page.

Another limitation is that our list of opt-out links for the trackers are static, and needs to be updated manually. This can be tedious, in particular when the number of trackers continue to grow. This can be addressed by potentially linking it to an already existing shared database, that contains a much larger data set than what we have collected so far.

9 Individual Contributions

Aryan - I worked on the data collection method by figuring out how to modify and use Privacy Badger for our purpose. I also worked on the visualization website, mostly writing the data processing functions and some styling.

Ragna - I worked on the visualization and implemented the back-end server for posting and loading the data. I also used the data processing functions that Aryan wrote to create components that rendered and displayed the visualizations. Lastly, I worked on the database for the trackers and their opt-out links.

You can visit our GitHub repository containing our final product by accessing this link.

References

[1] Privacy Badger Aryan Srivastva. Understanding the user data json file 2716. 2020. https://github.com/EFForg/privacybadger/issues/2716.

[2] Electronic Frontier Association. Do not track. 2020. https://www.eff.org/issues/do-not-track.

[3] Privacy Badger. How to get started with privacy badger. 2020. https://privacybadger.org/#How-does-Privacy-Badger-work.

[4] Joanna Geary. Doubleclick (google): What is it and what does it do? 2016. https://www.theguardian.com/technology/2012/apr/23/doubleclick-tracking-trackers-cookies-web-monitoring.

[5] Ghostery. Ghostery study: 79 percent of websites globally are secretly tracking your personal data. 2019. https://www.ghostery.com/press/ghostery-global-tracking-study/.

[6] Ghostery. Submit a tracker. 2020. www.ghostery.com/submit-a-tracker/.

[7] John Koetsier. Google is tracking you on 86% of the top 50,000 websites on the planet. 2020. https://www.forbes.com/sites/johnkoetsier/2020/03/11/google-is-tracking-you-on-86-of-the-top-50000-websites-on-the-planet/?sh=d10e6c2750fb.

[8] mozilla. What you need to know about online tracking (and how to stop it). 2020. https://blog.mozilla.org/firefox/how-to-stop-web-trackers/#:~:text=What%20are%20web%20trackers%3F,website%20you've%20never%20visited..

[9] Global Privacy. Take control of your privacy. 2020. https://globalprivacycontrol.org/.