PhD proposal

Improving online privacy through content blocking and information restriction

1 Supervisors

- Romain Rouvoy (Professor, University of Lille, romain.rouvoy@univ-lille.fr)
- Pierre Laperdrix (Full-time researcher, CNRS, pierre.laperdrix@univ-lille.fr)

2 Research team

The PhD student will join the Spirals (https://team.inria.fr/spirals) project-team between the University of Lille and Inria, within the UMR CRIStAL Laboratory.

Université de Lille
UMR CRIStAL
Bâtiment M3, Université de Lille 1, 59655 Villeneuve d’Ascq – FRANCE

Inria Lille - Nord Europe
Parc Scientifique de la Haute Borne
40, avenue Halley - Bat. B, Park Plaza
59650 Villeneuve d’Ascq – FRANCE

3 Scientific Context

Since its inception, the web has grown substantially and websites have turned into rich client-side experience customized for the user where third parties supply a considerable amount of content. The increasing reliance on third parties has brought a number of privacy issues to the web with web tracking being at the top of that list [7]. With cookies and browser fingerprinting, users are on the losing side of privacy as they can be tracked across the domains they are visiting. To regain control, browser vendors like Mozilla and Apple have added in their own browsers a tracking protection mechanism (called Enhanced Tracking Protection
for Firefox [3] and Intelligent Tracking Protection for Safari [5]) aimed at preventing tracking on the web. Yet, essential functionality of a website is sometimes so intertwined with tracking code that using these protective mechanisms can transmissively “break” a webpage. We define “page breakage” as an undesirable behavior on a webpage and it includes, but is not limited to, page slowdowns, page freezes, page crashes, page errors and page display issues. In order to push online privacy forward, there is a real need today to identify and block properly tracking entities on the web without the current usability costs associated with it.

4 PhD project

Positioned in the context of web security and privacy, this PhD project will focus on the three following topics.

Blocking non essential content of a page Most approaches to block content on the web are based on blacklists. When an element matches a URL or a regular expression present in a blacklist, this element is not loaded no matter its intended functionality. For a lot of scripts, this approach is fine as their only purpose is to put a cookie in the browser’s storage. For other scripts, the result can range from display issues to a complete page freeze [1,2,4].

In this PhD, the student will investigate how to design a system that will take any third-party script found online and provide a detailed set of information about it. This would go beyond recent studies [6,7] as it would not be limited to tracking scripts and would use techniques like information flow control, machine learning and deobfuscation for finer-grained analysis. The ultimate goal would be to differentiate elements that are essential to web browsing from elements that are not. This distinction can help in the design of advanced defense mechanisms that could go much further than the tracking protection we see today.

Automatic detection of page breakage Following directly on the previous topic, one big challenge on the web is to identify page breakage. A page that does not load may be easy to identify but one where one crucial component does not trigger the right action is a very different problem.

In this PhD, the student will investigate how to automatically detect page breakage. This includes detecting the type of breakage (e.g., errors or freezers) and its severity (e.g., is the whole page unresponsive or only one button?).

Restricting information sent to a webserver To improve user experience on the web, browsers send information like the version of the browser or the OS being used. The problem is that the diversity of modern devices is so great today that it opened the door to a technique called browser fingerprinting. By collecting a set of information related to a user’s device, any third-party can build a fingerprint of the device and use it to track an individual online [8]. This technique is particularly dangerous as it is completely stateless and requires no permission to be executed. One way to mitigate the effects of fingerprinting is to send less information or, at least, less precise information so that not a single user has a unique fingerprint.
In this PhD, the student will assess the impact of restricting information sent to a server on usability and page breakage. This includes blocking information from APIs that are used for fingerprinting and reducing the quantity of information in legacy headers. This topic goes hand in hand with the previous ones as identifying scripts that use some APIs only to collect device information can be blocked without any impact on the rendered webpage.

5 Skills summary

The PhD candidate should have a background in computer science. Knowledge in Web programming is a strong plus. As is a common practice in the Spirals research teams, all source code is expected to be open sourced. The student should publish high-level academic papers, as well as participate in related open source communities. This should assist in the technological transfer from academic prototypes to industry-ready tools.

References


