Barcelona School of Informatics

Bachelor's degree in Informatics Engineering

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Bachelor's Thesis

Is Your VPN Trustworthy

Author

Ismael de la Garcia Cobos

Advisors

Pere Barlet Ros

Ismael Castell

Q1 2020/2021

Abstract

Nowadays people are more and more concerned about their privacy and safety on the internet. Also, being public as it is, some content is limited based on geolocation. On the other hand, remote work is also becoming a matter of fact. These things have empowered the use of Virtual Private Networks. However, it is well known that VPNs, if malicious, can steal all your personal information. In this paper we want to build a tool that can analyze and detect any malicious traffic in your VPN.

Resumen

Cada día la privacidad y la seguridad en internet es un tema que preocupa a más personas. Además, y a pesar de ser público, hay contenido que se limita dependiendo de la localización. Por otro lado, el trabajo en remoto está cada vez más a la orden del día. Todo esto ha hecho que crezca el uso de las *Virtual Private Networks.* Es bien sabido que una VPN con malas intenciones puede llegar a robar toda tu información personal y privarte de la privacidad. En este trabajo se pretende crear una herramienta que permita detectar trafico malicioso en un cliente VPN.

Resum

Cada dia la privacitat i la seguretat a internet és un tema que preocupa a més persones. A més, i tot i ser públic, hi ha contingut que es limita depenent de la localització. D'altra banda, el treball en remot està cada vegada més a l'ordre del dia. Tot això ha fet que creixi l'ús de les *Virtual Private Networks.* És ben sabut que una VPN amb males intencions pot arribar a robar tota la teva informació personal i privar-te de la privacitat. En aquest treball es pretén crear una eina que permeti detectar trànsit maliciós en un client VPN.

Acknowledgments

I would first like to thank my thesis advisors Pere Barlet and Ismael Castell at Politechnics University of Catalonia, due to his instant e-mail answers when I ran into a trouble spot or had a question about my research or writing. They consistently allowed this thesis to be my own work, but guided me in the right direction whenever They thought I needed it. Both of them arranged periodic meetings with me and other students to share our findings and find a solution to our problems together. Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you all who have helped in one way or another to make this project possible.

Revision history and approval record

Version	**Date**	**Description**
0	28/09/2020	Document creation
1	10/10/2020	Update. Corrections suggested by the teacher in delivery 1 - Project Scope.
2	12/10/2020	Update. Corrections suggested by the teacher in delivery 2 - Project Planning.
3	16/10/2020	Update. Corrections suggested by the teacher in delivery 3 - Economic Analysis.
4	12/01/2021	Update. Added planning adjustements.
5	14/04/2022	Update. State of the art and project description created. Initial results commented.
6	22/04/2022	Update. Adjusted figures, tables and conclusions to the last findings.

Table 1: Record of modifications made to this document

List of Figures

List of Tables

Table of contents

Glossary

The fields that end with a * are further explained in the appendix.

API: Application Programming Interface

DNS: Domain Name Server

DOM: Document Object Model

GC: General Costs

IP: Internet Protocol

IPv6: Internet Protocol Version 6

IR: Incidental Risks

ISP: Internet Service Provider

IT: Information Technology

MD5: Message-Digest Algorithm 5

OS: Operative System

P2P: Peer to Peer

TLS: Transport Layer Security

URL: Universal Resource Locator

VPN: Virtual Private Network

XSS: Cross Site Scripting

1. Introduction

A VPN (Virtual Private Network) extends a private network across a public network and enables users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network [40]. For security, the private network connection may be established using an encrypted layer , and users may be required to pass various authentication methods to gain access to the VPN. Figure 1.1 shows a simplified diagram of this structure. Nowadays, OpenVPN is the one of the most used protocols available.

VPNs are often considered the key to safety and privacy while browsing the internet. However, there are some important things that we have to take into consideration when choosing a VPN if we want to enjoy safety and privacy.

VPNs are supported in many OS. Some of them are Windows, Mac OS X, iOS, Android, and Linux (Ubuntu) and Chrome book. They also work on mobile devices, like Android, IPhone and IPad. Some routers also support VPNs.

While most VPN providers claim they do not keep logs, it is true that most companies keep basic connection logs(User IP address and VPN timestamp). It is very important to check the companies privacy policy and the location of their headquarters (in terms of jurisdiction).

It is also important to take into account the encryption standard used. Many VPN clients allow you to choose different standards. While the securer standards may slow down the internet connection, they will make sure your data is safer.

Typically, your ISP would have to pay special attention to your activity in order to determine whether you are using a VPN or not. However, if well protected, they will not be able to learn the nature of your activity. They are only able to see you are connected to a certain endpoint and the bandwidth you are using. If they want to learn more about you, they may request logs to your VPN provider (they probably need a court order to do that).

Strictly speaking, using a personal VPN does not prevent others from tracking you fully. Some tracking techniques can go through a VPNs layer of privacy and still identify you while

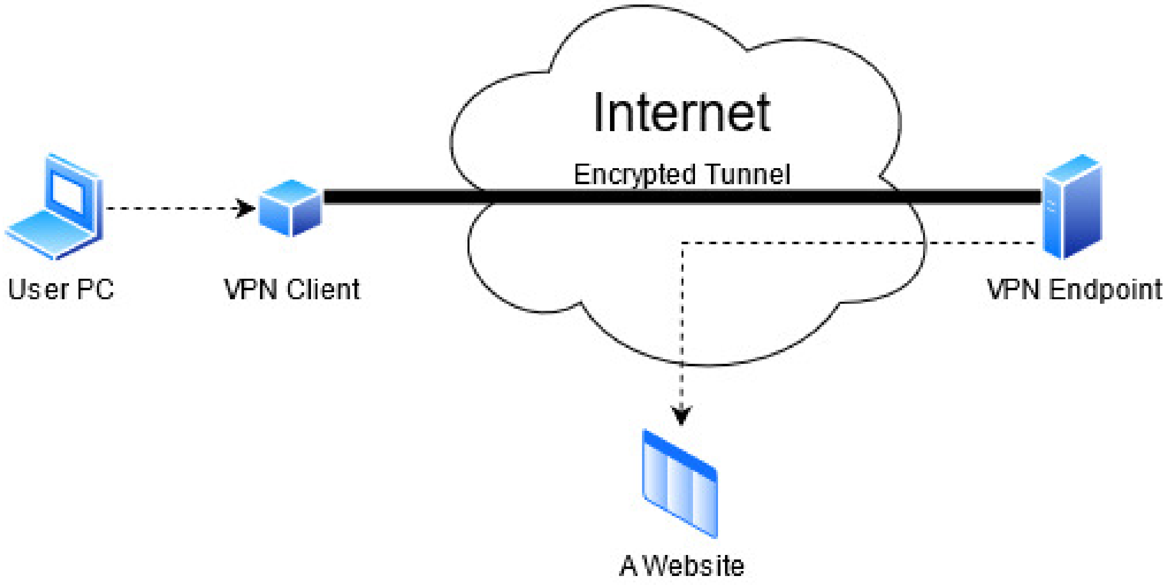


Figure 1.1: VPN diagram. Source: Original

browsing the internet. Some of this practices are simple, some others are more complicated. See appendix for more information, on section 11.2.

Notice that using a VPN does not protect you from malicious software. While some VPN providers offer additional protection, most do not. Viruses, spyware and malware can still harm you when you are using a VPN.

1.1 Justification

VPNs are widely used in today's world. Particular users normally use them to increase their privacy while using the internet. Others want to access content that is blocked in their country or location. Companies often use VPNs to access their resources remotely in a way these resources can not be stolen or seen by other users on the internet. VPNs are often linked with security and privacy. On the other hand, many people suspect that some VPN providers are selling their online activity to advertisement companies, but can not prove it true. Huge stirs have been raised when researchers prove a certain provider is violating the users privacy, but most of the times they have to prove it by analyzing the code of the VPN client [18]. In this thesis, I want to create a tool that can distinguish the legitimate and malicious traffic a VPN may generate. I identify as malicious traffic all the traffic or activity a VPN may generate that may compromise the users anonymity on the internet. Therefore, I do not want to create an antivirus software, but an anti-tracking tool.

Virtual Private Networks pretend to fight against privacy-abusive practices, or at least that is what they sell to their customers. When a user decides to use the services of a VPN provider, most of the times they have to read carefully their privacy policy and place their trust on the company that will provide them these services.

As of the date of begining of this project there is no tool that directly aims to detect malicious traffic VPNs may generate. Therefore the need of an automated tool that helps detect malicious traffic a VPN client might generate.

I have identified as worthy of our interest the following tracking methods:

- JavaScript Injection: It is a process that allows inserting custom JavaScript code in a page, either by entering the code into the address bar, or by finding an XSS vulnerability in a website. Note that changes can only be seen by the client and are not permanent. This is because JavaScript is a "client-side" language [17].
- TLS Interception: This is performed by software on middleboxes located in between the client and HTTPS website or on the clients machine, in the case of malware, anti-virus software, and ad injectors. Middlebox proxy software relies on the client having previously installed a root certificate onto their operating system. Any outgoing SSL/TLS connections from the client are terminated and re-established by the proxy to the server, which acts as an in-the-middle attacker [29].
- Traffic redirection: It consists in redirecting traffic to a malicious page using URLs embedded in website code, an .htaccess file, or a phishing email [30].

All this tracking methods can easily be used in the context of a VPN, specially when confined to browser add-on VPNs. We also contemplated DNS/IPv6 leaks as a dangerous practice that can lead to the users' privacy being compromised, but since this is not result of a malicious practice but rather comes from wrong programming, we do not consider these as important as the other three.

2. Project scope

A new way of working is approaching due to the recent changes in the world. Remote work has become a necessity in countries which were not ready for it. On the other hand, our privacy is threatened every day to a point most people can not understand. That is why many companies take advantage of this situation to make money out of it. It is widely spread that VPNs, specially the free versions, offer less than paid services. Many people state that free services may also sell the users privacy online to third party companies while saying they offer increased privacy and security.

Given this situation, I found the need of a tool that can protect users from malicious traffic by their VPN providers on a daily basis. Honesty should be the key to any agreement, even when using free software. Therefore, in this paper I want to propose a tool that can detect malicious traffic in a VPN client, specially the kind of traffic that may compromise the user's privacy.

2.1 Objectives

The main objectives of this project are:

- Learn tracking methods VPN clients can use.
- Learn to detect the VPN-related tracking methods.
- Create a tool that can detect VPNs that use this tracking methods.

From these, the following sub-objectives arise:

- Learn about online privacy and how VPN clients can help their users improve it.
- Learn about the most common tracking methods, specially those affected by VPNs.

- Learn about networking and traffic analysis.
- Learn how the different OS treat traffic.

2.2 Requirements and specifications

The requirements of the project come from the need of portability of this tool, for users of VPNs may use very different environments.

2.2.1 Functional Requirements

- It must not depend on other tools, like web browsers.
- It must never modify the users traffic in any means.
- It must be easily auditable.

2.2.2 Non-Functional Requirements

- The tool must be user-friendly and every user should be able to use it.
- It must be open-source.
- It must not slow down the connection more than a 10%.
- It must be easy to update.
- The tool must work in the main OS (Windows, Linux, Mac).
- It must work with, at least, the top 5 VPN clients according to number of users.

2.3 Actors

In this section I will list the stakeholders of this thesis, be them people who will benefit, be affected partake of it.

- **Developer:** The project will have a single developer, Ismael de la Gracia Cobos, as part of his bachelor's degree. He will research on the topics required, develop, test and document the project.

- **Advisors:** They will guide the developer in the project. The main advisors are Pere Barlet and Ismael Castell, a teacher in charge of many projects related to internet privacy and a doctorate student on internet privacy respectively.

- **VPN Providers:** VPN providers will definitely be affected by this project, for their services may become compromised after the development of the proposed tool. Any VPN provider conducting privacy-abusing practices may be exposed by this tool. However, all those who do not perform wrong practices will receive the support of a third party tool in their claims for privacy and security.

- **Advertising companies:** Every VPN that implements tracking in their clients later sells that information to an advertising company. The development of a tool that detects this interaction will surely punish this type of relations.

- **IT Departments:** Most companies now a days have an IT department that takes care of the installation of the necessary infrastructure for the company to work. These departments may benefit out this tool in auditing their systems for security, for they will be able to know if they are using real private networks.

- **Tech-savvy users:** This kind of users usually care about their privacy when connected to the internet. Most at times, they use VPNs to access restricted content or to navigate the internet privately. They may also wonder to what extent they are doing so. This tool will surely be interesting to them.

2.4 Obstacles and risks

Throughout the project, we may encounter unexpected obstacles and risks that we will need to affront. Here we put some examples:

- **Implementation errors:** Sadly, it is very common that developers make mistakes, for we are all humans. Debugging the type of tool we want to create may be hard, for it works at a very low level and the developer, Ismael de la Gracia, has no previous experience in this field.

- **Delays:** No room for distraction. The project has to progress carefully to arrive on time to the deadline. If we delay on something, we can cause the delay of the other projects.

- **Misconceptions:** Since the developer has no experience in this field and all the knowledge has just been acquired by self investigation, there might be some misconceptions that may lead to poor decisions. Therefore, it is important to investigate properly and to trust in the guidance of the advisors.

- **Bad documentation:** Trying to understand how a project is done is a waste of time if it does not have good documentation. It is important to document every progress and change we make to the project.

3. Methodology

The development of the project will be done using the Agile methodology, specifically Scrum [32]. Scrum is an iterative and incremental methodology, in such a way that a number of short iterations (sprints) of between 1 and 3 weeks are defined, and in each iteration value is added to the final product. This methodology divides the team into different roles. However, since there is only one developer in this project, he will carry all the tasks. In a Scrum project, a board must initially be filled (backlog) with the user stories punctuated according to the difficulty of each, and each sprint is assigned some of these user stories that will be divided between the team of developers, and must be completed in the end of the sprint. To make it easier to develop following this Scrum methodology, several tools will be used.

3.1 Tools

3.1.1 Git

Git [14] is a free and open source version control system. It is widely used in software development in cooperative projects to perform version management, through repositories and branches.

3.1.2 Github

GitHub [1] is a free software Git repository manager that also has other features such as error control, and continuous integration.

3.1.3 Git-flow

The Git-flow methodology [11] consists of a always stable Master Master branch, a Development branch, and Release and Feature branches. The git-flow methodology proposes that for each user history that is implemented, a new branch of Feature is created with origin in the

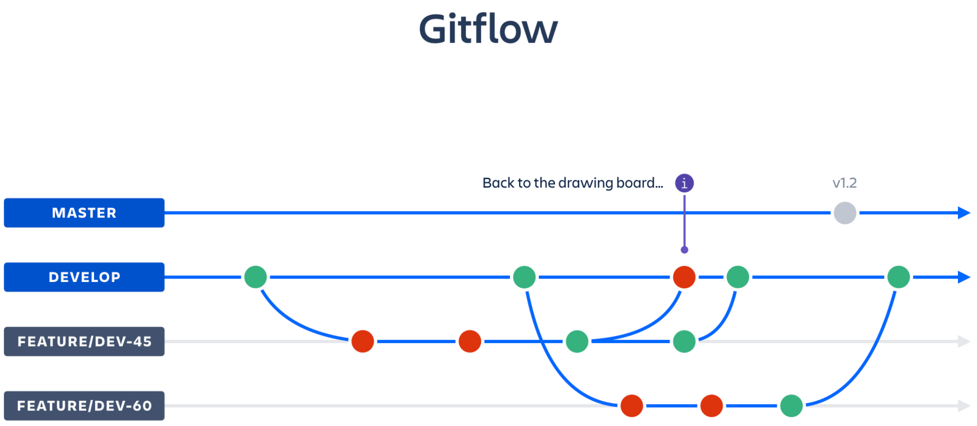


Figure 3.1: Git-flow example. Source: Atlassian.com

branch of Development, and that once this user history was complete the header is developed in Development. Figure 3.1 shows an example.

3.1.4 Taiga

Taiga [20] is a free and open source project management system. It has features to create and rate user stories, to manage the backlog, and to track issues.

3.1.5 Slack

"Connect your conversations with the tools and services that you use to get the job done. With over 2,000 apps and a robust API, the Slack [36] platform team works with partners and developers globally to build apps and integrations that streamline your work, automate mundane tasks and bring context into your conversations in Slack". We will use this tool to comunicate with the rest of the students and with the advisors.

3.1.6 Beautifier Online

"Beautify, unpack or deobfuscate JavaScript and HTML, make JSON/JSONP readable, etc.

All of the source code is completely free and open, available on GitHub under MIT licence, and we have a command-line version, python library and a node package as well." We use this tool to make legible all the resources downloaded by the crawler [5].

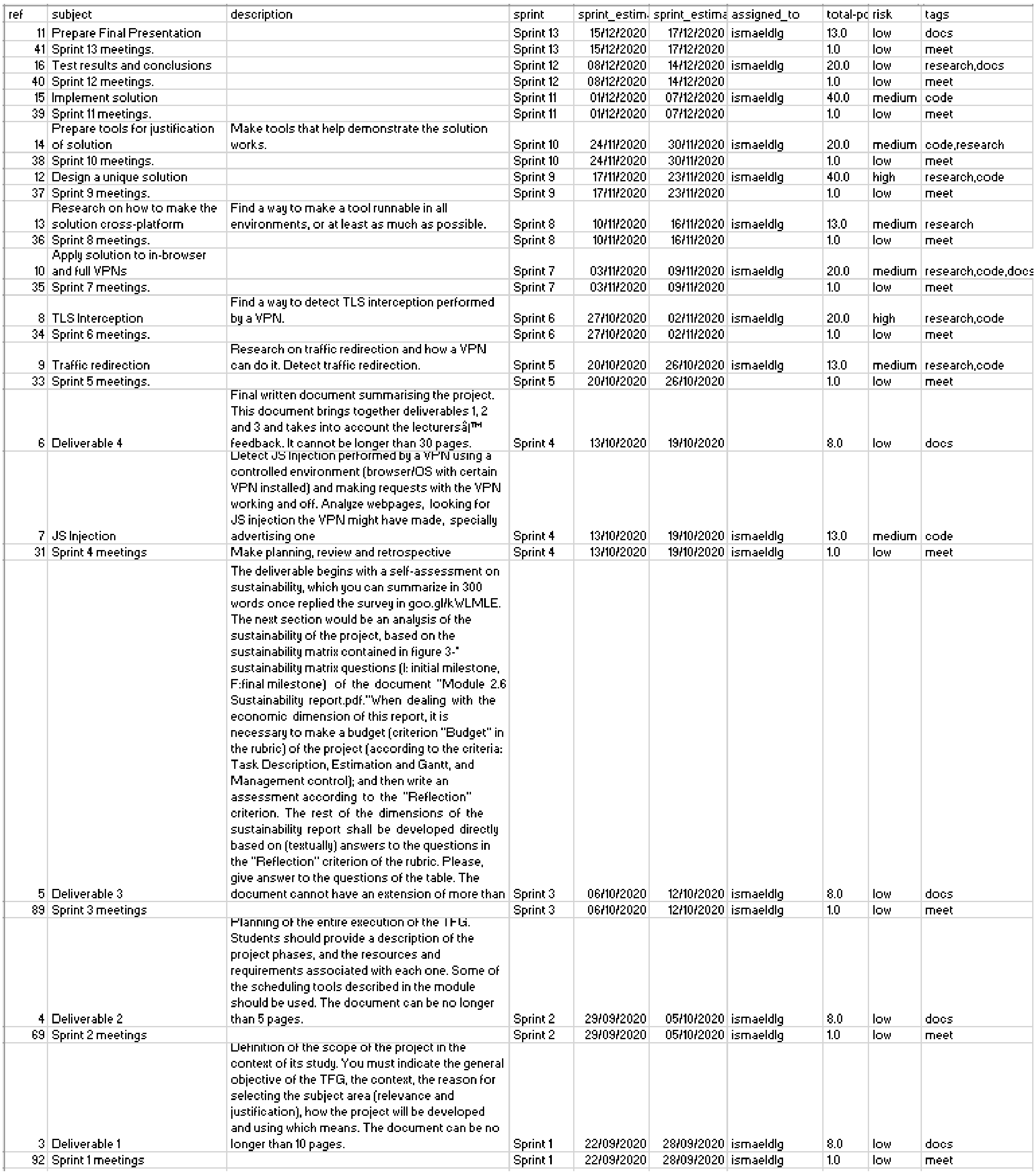
4. Project planning

As mentioned in the methodology of this project, the planning of the later has been performed and monitored using a tool called Taiga (refer to 3.1.4). You may find all the planning in the following sections.

The planning in Taiga and the one explained here may differ, for the tool gives me the ability to change every part of the planning whenever needed. Since the project will follow an agile methodology, that is the best approach. However, in this section I will describe the planning of the project when it started. Before the first sprint of the project, a backlog was created with the following user stories:

We consider that each story point translates to 1 hour of work. User stories have a risk potential, that is a certain degree of not going as planned, depending on: difficulty level, experience on the topic of the assigned user and unexpected events. I have classified user stories risk level using three levels: low, medium, high. I will consider that user stories may have an increase in their points of the following percentages according to their risk: 10, 50 and 100.

Each user story of the backlog was assigned to a different sprint. Sprints last 1 week. Therefore, the Gantt graph of the project looks like the following:



ref	subject	description	sprint	sprint_estim	sprint_estima	assigned_to	total-pc	risk	tags
11	Prepare Final Presentation		Sprint 13	15/12/2020	17/12/2020	ismaeldlg	13.0	low	docs
41	Sprint 13 meetings.		Sprint 13	15/12/2020	17/12/2020		1.0	low	meet
16	Test results and conclusions		Sprint 12	08/12/2020	14/12/2020	ismaeldlg	20.0	low	research,docs
40	Sprint 12 meetings.		Sprint 12	08/12/2020	14/12/2020		1.0	low	meet
15	Implement solution		Sprint 11	01/12/2020	07/12/2020	ismaeldlg	40.0	medium	code
39	Sprint 11 meetings.		Sprint 11	01/12/2020	07/12/2020		1.0	low	meet
14	Prepare tools for justification of solution	Make tools that help demonstrate the solution works.	Sprint 10	24/11/2020	30/11/2020	ismaeldlg	20.0	medium	code,research
38	Sprint 10 meetings.		Sprint 10	24/11/2020	30/11/2020		1.0	low	meet
12	Design a unique solution		Sprint 9	17/11/2020	23/11/2020	ismaeldlg	40.0	high	research,code
37	Sprint 9 meetings.		Sprint 9	17/11/2020	23/11/2020		1.0	low	meet
13	Research on how to make the solution cross-platform	Find a way to make a tool runnable in all environments, or at least as much as possible.	Sprint 8	10/11/2020	16/11/2020	ismaeldlg	13.0	medium	research
36	Sprint 8 meetings.		Sprint 8	10/11/2020	16/11/2020		1.0	low	meet
10	Apply solution to in-browser and full VPNs		Sprint 7	03/11/2020	09/11/2020	ismaeldlg	20.0	medium	research,code,docs
35	Sprint 7 meetings.		Sprint 7	03/11/2020	09/11/2020		1.0	low	meet
8	TLS Interception	Find a way to detect TLS interception performed by a VPN.	Sprint 6	27/10/2020	02/11/2020	ismaeldlg	20.0	high	research,code
34	Sprint 6 meetings.		Sprint 6	27/10/2020	02/11/2020		1.0	low	meet
9	Traffic redirection	Research on traffic redirection and how a VPN can do it. Detect traffic redirection.	Sprint 5	20/10/2020	26/10/2020	ismaeldlg	13.0	medium	research,code
33	Sprint 5 meetings.		Sprint 5	20/10/2020	26/10/2020		1.0	low	meet
6	Deliverable 4	Final written document summarising the project. This document brings together deliverables 1, 2 and 3 and takes into account the lecturersâ|™ feedback. It cannot be longer than 30 pages.	Sprint 4	13/10/2020	19/10/2020		8.0	low	docs
7	JS Injection	Detect JS Injection performed by a VPN using a controlled environment (browser/OS with certain VPN installed) and making requests with the VPN working and off. Analyze webpages, looking for JS injection the VPN might have made, specially advertising one	Sprint 4	13/10/2020	19/10/2020	ismaeldlg	13.0	medium	code
31	Sprint 4 meetings	Make planning, review and retrospective	Sprint 4	13/10/2020	19/10/2020	ismaeldlg	1.0	low	meet
5	Deliverable 3	The deliverable begins with a self-assessment on sustainability, which you can summarize in 300 words once replied the survey in goo.gl/kWLMLE. The next section would be an analysis of the sustainability of the project, based on the sustainability matrix contained in figure 3-" sustainability matrix questions (I: initial milestone, F:final milestone) of the document "Module 2.6 Sustainability report.pdf."When dealing with the economic dimension of this report, it is necessary to make a budget (criterion "Budget" in the rubric) of the project (according to the criteria: Task Description, Estimation and Gantt, and Management control); and then write an assessment according to the "Reflection" criterion. The rest of the dimensions of the sustainability report shall be developed directly based on (textually) answers to the questions in the "Reflection" criterion of the rubric. Please, give answer to the questions of the table. The document cannot have an extension of more than	Sprint 3	06/10/2020	12/10/2020	ismaeldlg	8.0	low	docs
89	Sprint 3 meetings		Sprint 3	06/10/2020	12/10/2020	ismaeldlg	1.0	low	meet
4	Deliverable 2	Planning of the entire execution of the TFG. Students should provide a description of the project phases, and the resources and requirements associated with each one. Some of the scheduling tools described in the module should be used. The document can be no longer than 5 pages.	Sprint 2	29/09/2020	05/10/2020	ismaeldlg	8.0	low	docs
69	Sprint 2 meetings		Sprint 2	29/09/2020	05/10/2020	ismaeldlg	1.0	low	meet
3	Deliverable 1	Definition of the scope of the project in the context of its study. You must indicate the general objective of the TFG, the context, the reason for selecting the subject area (relevance and justification), how the project will be developed and using which means. The document can be no longer than 10 pages.	Sprint 1	22/09/2020	28/09/2020	ismaeldlg	8.0	low	docs
92	Sprint 1 meetings		Sprint 1	22/09/2020	28/09/2020	ismaeldlg	1.0	low	meet

Figure 4.1: Backlog of the user stories of the project. Source: Original.

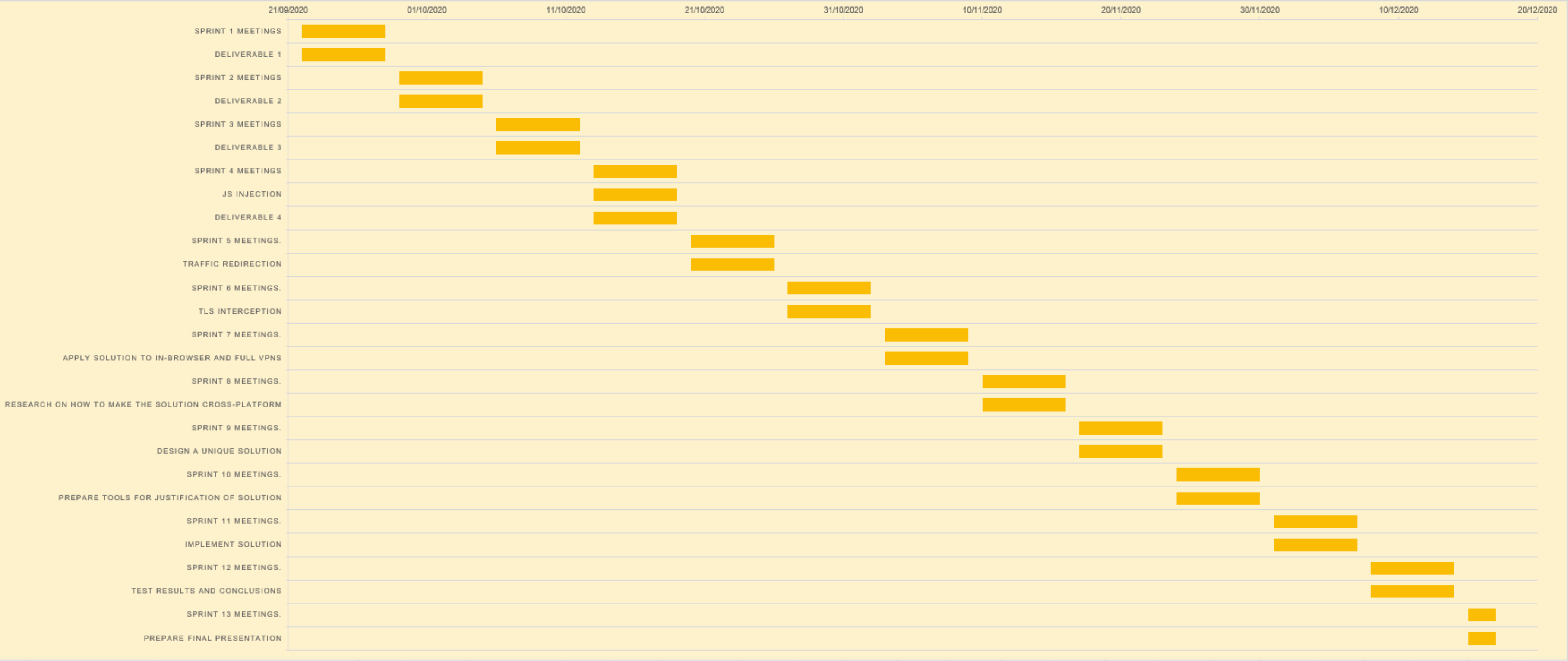


Figure 4.2: Gantt chart of the user stories. Source: Original.

Each user story has one or more tasks associated to it. When all tasks of a certain user story are closed, we consider the user story finished. Each task holds a percentage of the user story points. Find below the list of tasks and the user stories they are associated with:

4.1 Risk management

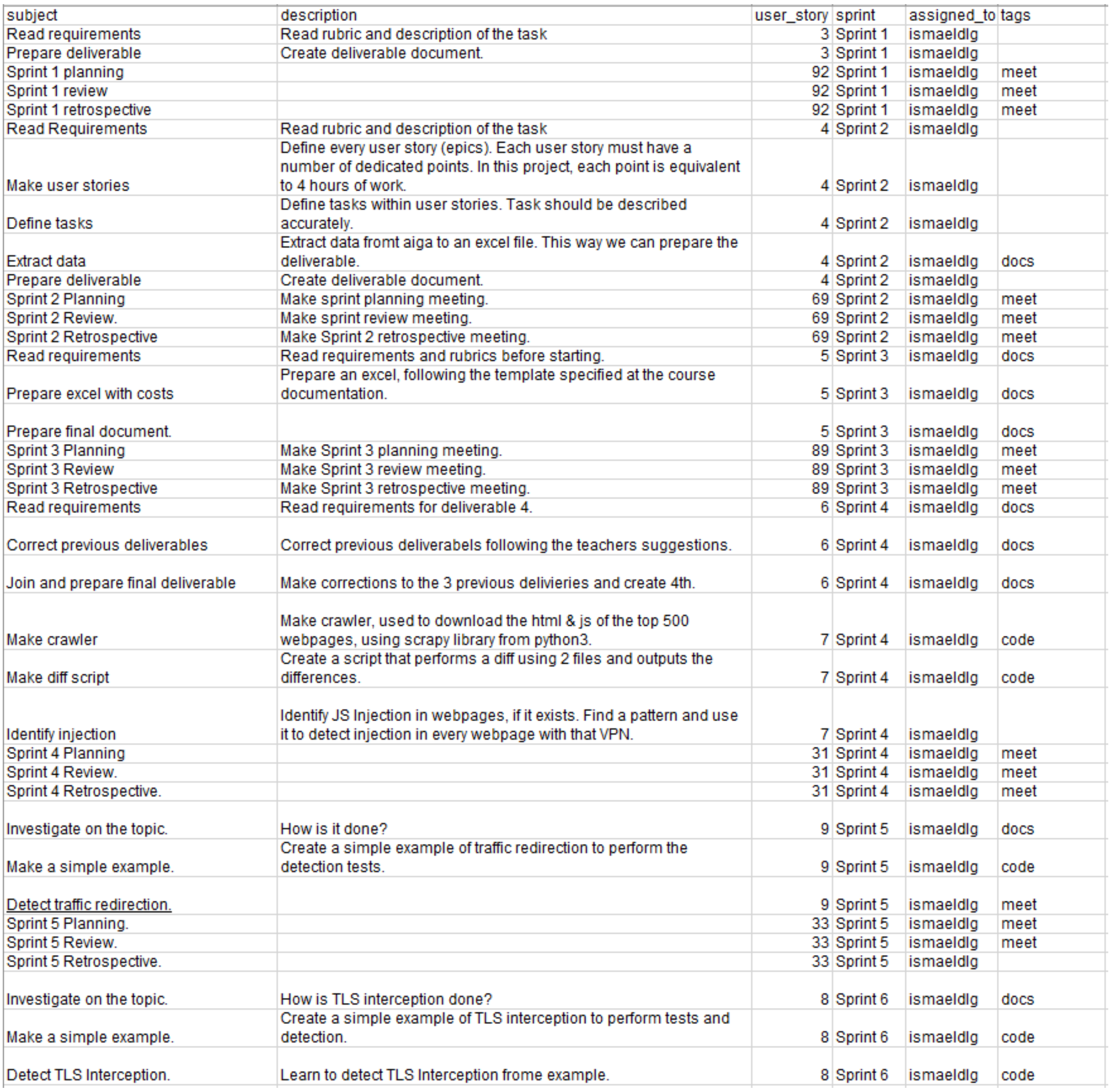
Most of the tasks have a low level of risk in the project tracking. This means their level of complexity may vary to the one assigned by the manager, but the task doesn't present any major threat. However, the tasks that have a medium or high level of complexity may create important challenges in the development of this project. Therefore, a backup plan has been studied in case these user stories do not develop as planned.

4.1.1 JS Injection

This task presents a threat to the development of the project because not the project manager nor the developer have experience in this field. Therefore, the user story can be more challenging than expected. This part of the challenge is contemplated by the 50% risk points assigned to the task.

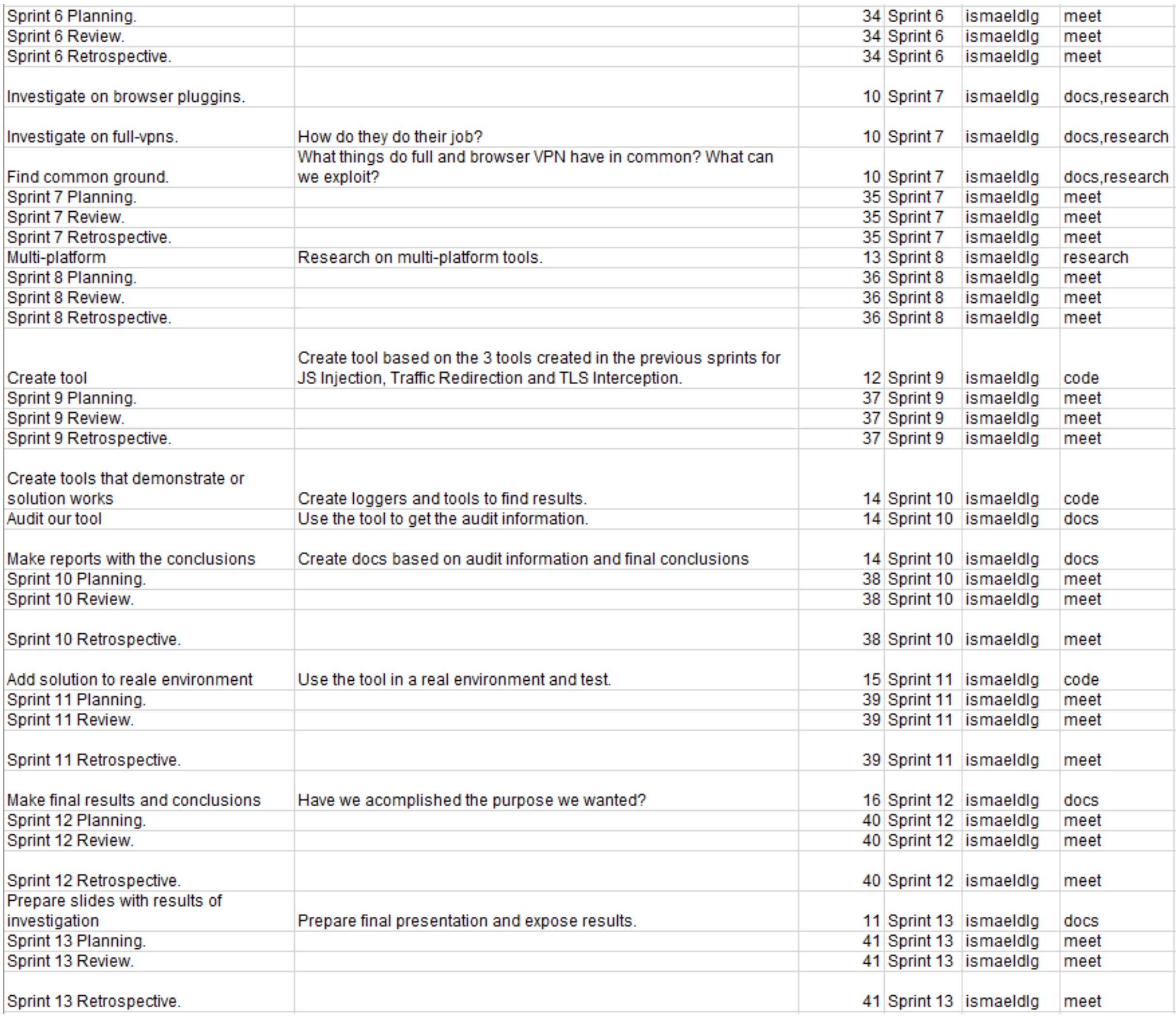
However, another two challenges are present. Since the solution the project wants to implement must work at network level, the challenge of analyzing a web page at network level arises. Most of the flows we will analyze are encrypted. Therefore, this task may not be possible at network level. In this case, it will be necessary to study the possibility of the tool working at two different levels, network level and application level. This would mean an increase of the risk of the task 15 Implement solution to high, for difficulty reasons, which means another 20 user-story points would be added to the project. This does not represent a big problem to the development of the project for tasks have been planned according to the risks planned, and, specially the last part of the project has an inferior workload and risk, therefore this task would be performed through sprints 10 and 11 instead of 11 only.

The later challenge is the performance of the tool. Analyzing every web page looking for JS Injection may be a resource consuming task, and may threat the maximum 10% slow-down allowed mentioned in the project requirements. If this is the case, and the algorithm can not be further optimized, an option will be added to the tool so that this functionality can be enabled and disabled at will, with the proper warning regarding slow-down issues. This is done this way because this project values privacy over other things, and I believe users interested in it will do



subject	description	user_story	sprint	assigned_to	tags
Read requirements	Read rubric and description of the task	3	Sprint 1	ismaeldlg	
Prepare deliverable	Create deliverable document.	3	Sprint 1	ismaeldlg	
Sprint 1 planning		92	Sprint 1	ismaeldlg	meet
Sprint 1 review		92	Sprint 1	ismaeldlg	meet
Sprint 1 retrospective		92	Sprint 1	ismaeldlg	meet
Read Requirements	Read rubric and description of the task	4	Sprint 2	ismaeldlg	
Make user stories	Define every user story (epics). Each user story must have a number of dedicated points. In this project, each point is equivalent to 4 hours of work.	4	Sprint 2	ismaeldlg	
Define tasks	Define tasks within user stories. Task should be described accurately.	4	Sprint 2	ismaeldlg	
Extract data	Extract data fromt aiga to an excel file. This way we can prepare the deliverable.	4	Sprint 2	ismaeldlg	docs
Prepare deliverable	Create deliverable document.	4	Sprint 2	ismaeldlg	
Sprint 2 Planning	Make sprint planning meeting.	69	Sprint 2	ismaeldlg	meet
Sprint 2 Review.	Make sprint review meeting.	69	Sprint 2	ismaeldlg	meet
Sprint 2 Retrospective	Make Sprint 2 retrospective meeting.	69	Sprint 2	ismaeldlg	meet
Read requirements	Read requirements and rubrics before starting.	5	Sprint 3	ismaeldlg	docs
Prepare excel with costs	Prepare an excel, following the template specified at the course documentation.	5	Sprint 3	ismaeldlg	docs
Prepare final document.		5	Sprint 3	ismaeldlg	docs
Sprint 3 Planning	Make Sprint 3 planning meeting.	89	Sprint 3	ismaeldlg	meet
Sprint 3 Review	Make Sprint 3 review meeting.	89	Sprint 3	ismaeldlg	meet
Sprint 3 Retrospective	Make Sprint 3 retrospective meeting.	89	Sprint 3	ismaeldlg	meet
Read requirements	Read requirements for deliverable 4.	6	Sprint 4	ismaeldlg	docs
Correct previous deliverables	Correct previous deliverabels following the teachers suggestions.	6	Sprint 4	ismaeldlg	docs
Join and prepare final deliverable	Make corrections to the 3 previous delivieries and create 4th.	6	Sprint 4	ismaeldlg	docs
Make crawler	Make crawler, used to download the html & js of the top 500 webpages, using scrapy library from python3.	7	Sprint 4	ismaeldlg	code
Make diff script	Create a script that performs a diff using 2 files and outputs the differences.	7	Sprint 4	ismaeldlg	code
Identify injection	Identify JS Injection in webpages, if it exists. Find a pattern and use it to detect injection in every webpage with that VPN.	7	Sprint 4	ismaeldlg	
Sprint 4 Planning		31	Sprint 4	ismaeldlg	meet
Sprint 4 Review.		31	Sprint 4	ismaeldlg	meet
Sprint 4 Retrospective.		31	Sprint 4	ismaeldlg	meet
Investigate on the topic.	How is it done?	9	Sprint 5	ismaeldlg	docs
Make a simple example.	Create a simple example of traffic redirection to perform the detection tests.	9	Sprint 5	ismaeldlg	code
Detect traffic redirection.		9	Sprint 5	ismaeldlg	meet
Sprint 5 Planning.		33	Sprint 5	ismaeldlg	meet
Sprint 5 Review.		33	Sprint 5	ismaeldlg	meet
Sprint 5 Retrospective.		33	Sprint 5	ismaeldlg	
Investigate on the topic.	How is TLS interception done?	8	Sprint 6	ismaeldlg	docs
Make a simple example.	Create a simple example of TLS interception to perform tests and detection.	8	Sprint 6	ismaeldlg	code
Detect TLS Interception.	Learn to detect TLS Interception frome example.	8	Sprint 6	ismaeldlg	code

Figure 4.3: Tasks of the project (One). Source: Original.



Sprint 6 Planning.		34	Sprint 6	ismaeldlg	meet
Sprint 6 Review.		34	Sprint 6	ismaeldlg	meet
Sprint 6 Retrospective.		34	Sprint 6	ismaeldlg	meet
Investigate on browser pluggins.		10	Sprint 7	ismaeldlg	docs,research
Investigate on full-vpns.	How do they do their job?	10	Sprint 7	ismaeldlg	docs,research
Find common ground.	What things do full and browser VPN have in common? What can we exploit?	10	Sprint 7	ismaeldlg	docs,research
Sprint 7 Planning.		35	Sprint 7	ismaeldlg	meet
Sprint 7 Review.		35	Sprint 7	ismaeldlg	meet
Sprint 7 Retrospective.		35	Sprint 7	ismaeldlg	meet
Multi-platform	Research on multi-platform tools.	13	Sprint 8	ismaeldlg	research
Sprint 8 Planning.		36	Sprint 8	ismaeldlg	meet
Sprint 8 Review.		36	Sprint 8	ismaeldlg	meet
Sprint 8 Retrospective.		36	Sprint 8	ismaeldlg	meet
Create tool	Create tool based on the 3 tools created in the previous sprints for JS Injection, Traffic Redirection and TLS Interception.	12	Sprint 9	ismaeldlg	code
Sprint 9 Planning.		37	Sprint 9	ismaeldlg	meet
Sprint 9 Review.		37	Sprint 9	ismaeldlg	meet
Sprint 9 Retrospective.		37	Sprint 9	ismaeldlg	meet
Create tools that demonstrate or solution works	Create loggers and tools to find results.	14	Sprint 10	ismaeldlg	code
Audit our tool	Use the tool to get the audit information.	14	Sprint 10	ismaeldlg	docs
Make reports with the conclusions	Create docs based on audit information and final conclusions	14	Sprint 10	ismaeldlg	docs
Sprint 10 Planning.		38	Sprint 10	ismaeldlg	meet
Sprint 10 Review.		38	Sprint 10	ismaeldlg	meet
Sprint 10 Retrospective.		38	Sprint 10	ismaeldlg	meet
Add solution to reale environment	Use the tool in a real environment and test.	15	Sprint 11	ismaeldlg	code
Sprint 11 Planning.		39	Sprint 11	ismaeldlg	meet
Sprint 11 Review.		39	Sprint 11	ismaeldlg	meet
Sprint 11 Retrospective.		39	Sprint 11	ismaeldlg	meet
Make final results and conclusions	Have we acomplished the purpose we wanted?	16	Sprint 12	ismaeldlg	docs
Sprint 12 Planning.		40	Sprint 12	ismaeldlg	meet
Sprint 12 Review.		40	Sprint 12	ismaeldlg	meet
Sprint 12 Retrospective.		40	Sprint 12	ismaeldlg	meet
Prepare slides with results of investigation	Prepare final presentation and expose results.	11	Sprint 13	ismaeldlg	docs
Sprint 13 Planning.		41	Sprint 13	ismaeldlg	meet
Sprint 13 Review.		41	Sprint 13	ismaeldlg	meet
Sprint 13 Retrospective.		41	Sprint 13	ismaeldlg	meet

Figure 4.4: Tasks of the project (Two). Source: Original.

the same. Therefore, and if JS Injection has a high enough impact in the VPN tracking market to be considered in this project, it must be detected by the proposed tool.

4.1.2 TLS Interception

This task presents a threat to the development of the project because not the project manager nor the developer have experience in this field nor any previous knowledge. Therefore, the user story can be more challenging than expected. This part of the challenge is contemplated by the 100% risk points assigned to the task.

If it is not possible to add this solution to the tool proposed, other possibilities will be studied. This will be done in the task Investigate on the topic, related to the user story 8 - TLS Interception. This study and the development of the other solution are considered in the amount of points assigned to the user story and the amount of risk points it has (20 + 20).

Other user stories

The rest of the tasks with a risk level of medium may present a threat to the development of the project because not the project manager nor the developer have experience in the field nor any previous knowledge. Therefore, they can be more challenging than expected. This part of the challenge is contemplated by the 50% risk points assigned to each user story. There are no further challenges contemplated in these user stories as for the project planning.

5. Sustainability and budget

Before doing this part of my thesis I did a self-assessment quiz [38]. This survey has been designed by the members of the EDINSOST project, financed by the Ministry of Economy, Industry and Competitiveness. The main objectives of the EDINSOST project are to analyze the level of training in sustainability of teachers and students of the Spanish university system and define training proposals for both groups.

The objective of this survey is to obtain information about your knowledge and skills in sustainability. In this survey we use a broad definition of sustainability that includes its three dimensions: social, environmental and economic. The solution to a problem is sustainable when these three dimensions are considered simultaneously (holistically).

This analysis is based on the sustainability matrix proposed at Module 2.6 [39]. Since we are at an early state of the project, only the project put into production (PPP) will be considered.

5.1 Environmental

A project like the one described in this thesis requires a big amount of investigation and prior analysis of the tracking methods used by the different VPN providers and how each one of them is applied and can be fought. Therefore, the amount of time (and therefore electricity) consumed in the investigation process has been considered something to be conscious of. In order to update the project with the latest techniques, web crawlers [41] have been used. These crawlers have been deployed so that the time and resources consumed by the analytical part of the project are limited and automated as much as possible.

Furthermore, these spiders have been deployed and used in raspberries [28], a very lightweight computer. According to the investigations performed by other users [26, 27], the model used (raspberry pi zero) will consume approximately 0.51 Watts/hour which is little compared to the 0.86 Kilowatts/hour a normal server can consume [19, 7]. In the future, the project may use these to update itself automatically without consuming as much resources.

Regarding current solutions, and as stated in previous sections of the thesis, no solutions for the project described in this report have been found, and therefore no environmental analysis has been performed on them, so this project can not be compared to previous solutions.

5.2 Economic

The economic analysis is based on the defined user stories in the project planning 4.1 and in the 4.1 Risk management section.

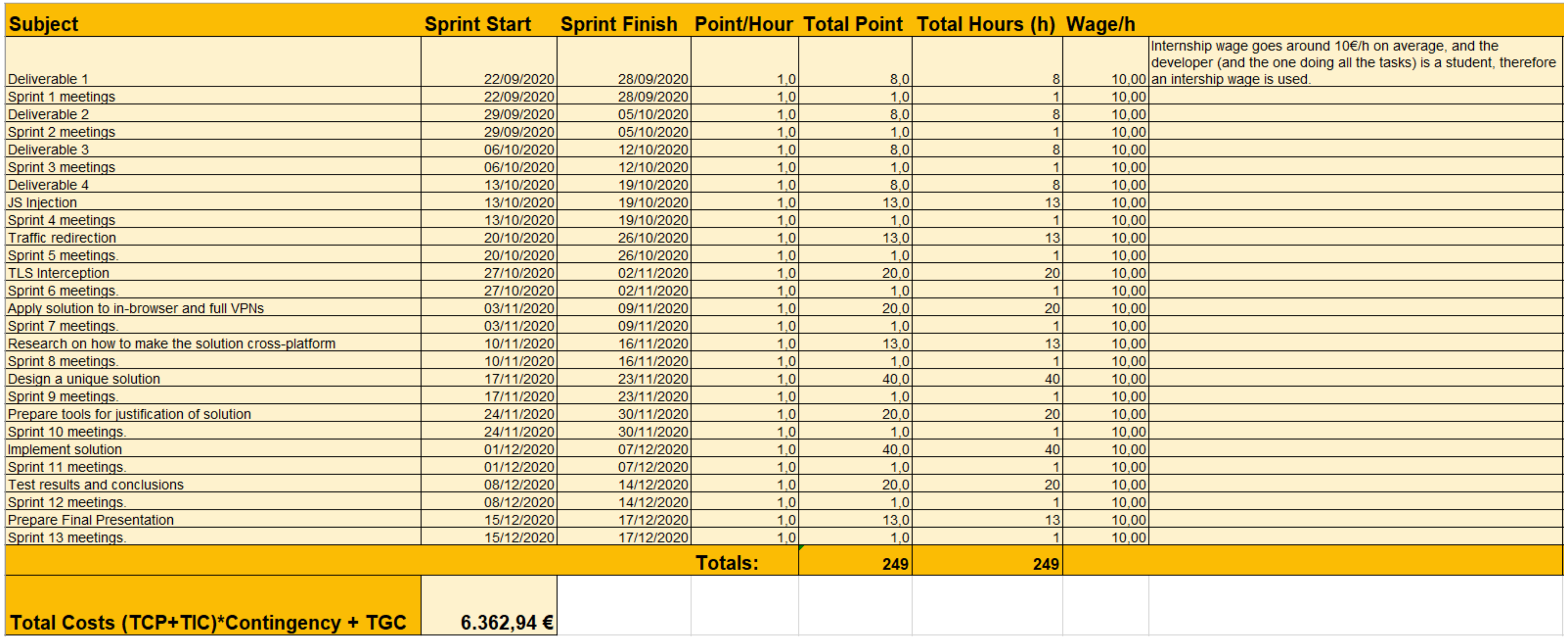
The costs considered in this section take into account different requirements established by the context of the project. Since this project is based on a Bachelor's Thesis, the one and only developer of all the tasks will be the author of this project. Therefore, and to make this analysis more consistent, the wages of all the workers in this project are considered to be around 10/h, which is on average the wage of an internship in the field of this project.

Also, given that this project has a high risk, specially regarding the planning, due to the fact that the author has no prior knowledge nor experience on the field, contingencies and margins are set high. All the costs of the project will include a 10% contingency and a risk according to the authors knowledge on that topic. IR (Incidental Risks) and IC (Incidental Costs) represent in most cases the lack of experience and therefore misappreciation of the complexity of the given tasks.

GC (General Costs) are costs associated with the expenses of developing the project that are not related to specific tasks. Since the objective of the project is to develop a software tool and the developer already has a computer able to do that job, this has not been included in the budget. However, the need of a raspberry will arise during the project and the electricity costs will still be there. Therefore, both of these things have been considered in the budget.

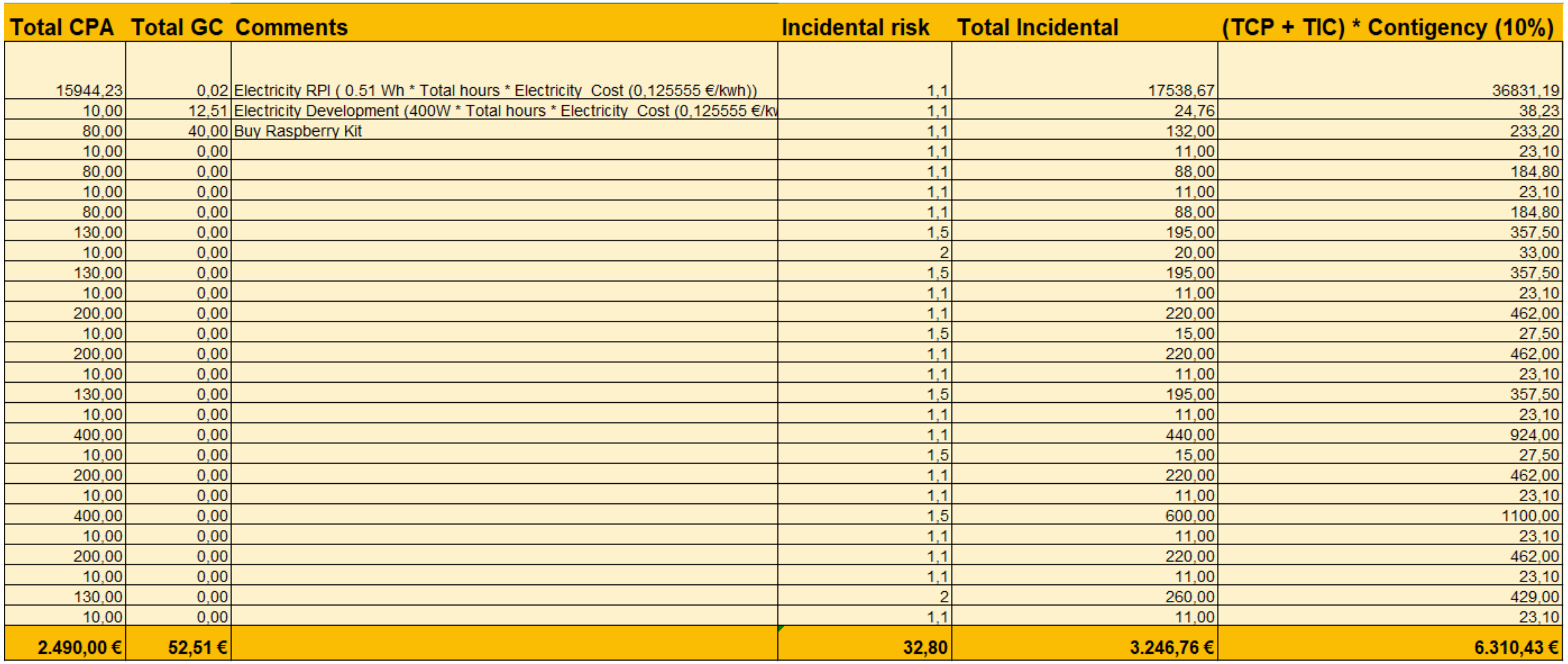
In any case, the budget of this project follows a flexible approach using the control system specified below and it is subject to changes during the development of the project.

Regarding the final cost, if everything in the project went as planned, it would be 2.490,00 (Total CPA) + 52,51 (Total GC), that is a total of 2.542,51. How ever, including the contingency costs and risks, the budget arises to the price of 6.362,94. As mentioned above, this is due to the lack of experience and the high risk of the project.



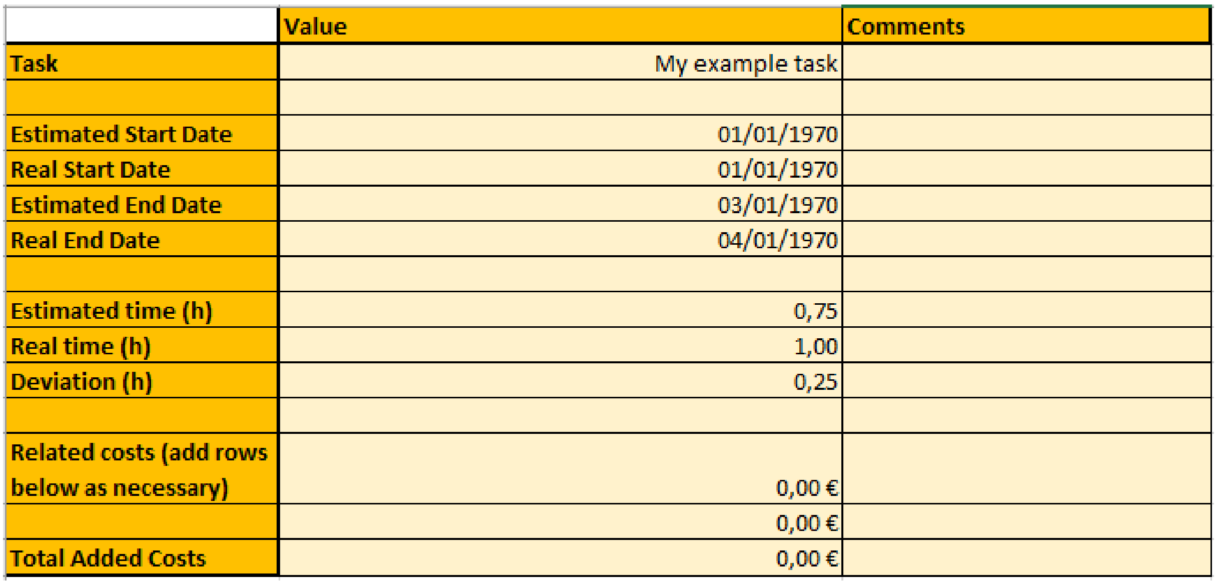
Subject	Sprint Start	Sprint Finish	Point/Hour	Total Point	Total Hours (h)	Wage/h	
Deliverable 1	22/09/2020	28/09/2020	1,0	8,0	8	10,00	Internship wage goes around 10€/h on average, and the developer (and the one doing all the tasks) is a student, therefore an intership wage is used.
Sprint 1 meetings	22/09/2020	28/09/2020	1,0	1,0	1	10,00	
Deliverable 2	29/09/2020	05/10/2020	1,0	8,0	8	10,00	
Sprint 2 meetings	29/09/2020	05/10/2020	1,0	1,0	1	10,00	
Deliverable 3	06/10/2020	12/10/2020	1,0	8,0	8	10,00	
Sprint 3 meetings	06/10/2020	12/10/2020	1,0	1,0	1	10,00	
Deliverable 4	13/10/2020	19/10/2020	1,0	8,0	8	10,00	
JS Injection	13/10/2020	19/10/2020	1,0	13,0	13	10,00	
Sprint 4 meetings	13/10/2020	19/10/2020	1,0	1,0	1	10,00	
Traffic redirection	20/10/2020	26/10/2020	1,0	13,0	13	10,00	
Sprint 5 meetings.	20/10/2020	26/10/2020	1,0	1,0	1	10,00	
TLS Interception	27/10/2020	02/11/2020	1,0	20,0	20	10,00	
Sprint 6 meetings.	27/10/2020	02/11/2020	1,0	1,0	1	10,00	
Apply solution to in-browser and full VPNs	03/11/2020	09/11/2020	1,0	20,0	20	10,00	
Sprint 7 meetings.	03/11/2020	09/11/2020	1,0	1,0	1	10,00	
Research on how to make the solution cross-platform	10/11/2020	16/11/2020	1,0	13,0	13	10,00	
Sprint 8 meetings.	10/11/2020	16/11/2020	1,0	1,0	1	10,00	
Design a unique solution	17/11/2020	23/11/2020	1,0	40,0	40	10,00	
Sprint 9 meetings.	17/11/2020	23/11/2020	1,0	1,0	1	10,00	
Prepare tools for justification of solution	24/11/2020	30/11/2020	1,0	20,0	20	10,00	
Sprint 10 meetings.	24/11/2020	30/11/2020	1,0	1,0	1	10,00	
Implement solution	01/12/2020	07/12/2020	1,0	40,0	40	10,00	
Sprint 11 meetings.	01/12/2020	07/12/2020	1,0	1,0	1	10,00	
Test results and conclusions	08/12/2020	14/12/2020	1,0	20,0	20	10,00	
Sprint 12 meetings.	08/12/2020	14/12/2020	1,0	1,0	1	10,00	
Prepare Final Presentation	15/12/2020	17/12/2020	1,0	13,0	13	10,00	
Sprint 13 meetings.	15/12/2020	17/12/2020	1,0	1,0	1	10,00	
			Totals:	**249**	**249**		
Total Costs (TCP+TIC)*Contingency + TGC	**6.362,94 €**						

Figure 5.1: Cost table. Part 1. Source: Original.



Total CPA	Total GC	Comments	Incidental risk	Total Incidental	(TCP + TIC) * Contigency (10%)
15944,23	0,02	Electricity RPI (0.51 Wh * Total hours * Electricity Cost (0,125555 €/kwh))	1,1	17538,67	36831,19
10,00	12,51	Electricity Development (400W * Total hours * Electricity Cost (0,125555 €/kv	1,1	24,76	38,23
80,00	40,00	Buy Raspberry Kit	1,1	132,00	233,20
10,00	0,00		1,1	11,00	23,10
80,00	0,00		1,1	88,00	184,80
10,00	0,00		1,1	11,00	23,10
80,00	0,00		1,1	88,00	184,80
130,00	0,00		1,5	195,00	357,50
10,00	0,00		2	20,00	33,00
130,00	0,00		1,5	195,00	357,50
10,00	0,00		1,1	11,00	23,10
200,00	0,00		1,1	220,00	462,00
10,00	0,00		1,5	15,00	27,50
200,00	0,00		1,1	220,00	462,00
10,00	0,00		1,1	11,00	23,10
130,00	0,00		1,5	195,00	357,50
10,00	0,00		1,1	11,00	23,10
400,00	0,00		1,1	440,00	924,00
10,00	0,00		1,5	15,00	27,50
200,00	0,00		1,1	220,00	462,00
10,00	0,00		1,1	11,00	23,10
400,00	0,00		1,5	600,00	1100,00
10,00	0,00		1,1	11,00	23,10
200,00	0,00		1,1	220,00	462,00
10,00	0,00		1,1	11,00	23,10
130,00	0,00		2	260,00	429,00
10,00	0,00		1,1	11,00	23,10
2.490,00 €	**52,51 €**		**32,80**	**3.246,76 €**	**6.310,43 €**

Figure 5.2: Cost table. Part 2. Source: Original.



	Value	Comments
Task	My example task	
Estimated Start Date	01/01/1970	
Real Start Date	01/01/1970	
Estimated End Date	03/01/1970	
Real End Date	04/01/1970	
Estimated time (h)	0,75	
Real time (h)	1,00	
Deviation (h)	0,25	
Related costs (add rows below as necessary)	0,00 €	
	0,00 €	
Total Added Costs	0,00 €	

Figure 5.3: Control template for each developer. Source: Original.

5.2.1 Control

Each user involved in the development of the project will be asked to fill table like the one in 5.3 for each task he develops. This will greatly help to control the expected budget and the real expenses on a task basis (less than a week). It will also help to know which members are doing a great job and which need some readjustments, be them in terms of tasks assigned, workspace or any other kind. The comments column is meant to justify the differences between the expected values and the real ones. With that, the project manager will be able to discern what to do in each specific situation, if any action is required.

All of this tables will be delivered in a sheet to the project manager at the end of the task. He will include them in another sheet that, based on the price per hour of the worker assigned, will calculate the total costs and deviations. When required, he will check the specific comments and justification in each task to know how to proceed.

5.3 Social

Personally, I think this project will contribute to me and anyone who partakes or uses it by giving better insight and understanding of what internet privacy means. Until now, there has been a lot of interest into that topic, and most users know well the dangers behind IP addresses and cookies. Lately, VPNs have been considered some of the master columns behind internet privacy, but sometimes they are not the key. Even the more so when people are concerned about their VPN providers, specially on free VPNs.

Until now, users had to read the terms and conditions of their VPN providers offered them, read their privacy policy and trust that these providers would follow good practices and attach to their policies. However, with a tool like the one proposed, this should not be the only reason to trust them.

On the other hand, this may have a huge impact on VPN providers and advertising companies, for one of their sources of income might disappear is the tool proposed is accomplished. As mentioned in the definition of the project, studies on VPN compromising the users privacy have been performed in different environments and using different strategies [18]. These type of studies are definitely useful to users who want to choose a VPN, however, they can not reach every solution and keep it up to date. A tool like the one proposed is not specific to a certain environment, for it works on an inferior level, the network layer. It can also be updated and maintained regularly, and therefore improves the other approaches.

6. Planning adjustments

Regarding the initial tasks, there have been some changes during the project. The high risks associated to the tasks resulted in unexpected extensions in the time planning that required adjusting it. However, the most important parts of the project were successfully adapted to this changes. However, some needed to change.

Due to time limitations, I considered TLS interception related tasks are now out of the scope of the project. The tasks related to identifying and isolating JavaScript Injection took way longer than expected becuase of their high complexity I have to remove TLS Interception tasks from the board, for they are not the main aim of the project. Therefore, I removed the task 9 TLS Interception.

Due to the new necessities perceived, task 7 JavaScript Injection has a bigger scope now. The crawler not only has to analyze the contents of the web page, but also the resources it loads. This is done because I realized it could happen that VPN clients tracked clients not only by modifying the web page, but also by loading external resources. Therefore, the tool analyzes performance logs of the web page to check all the requests performed when the page is loaded. Then, if the resource is a JavaScript file, or any type of application related file, it downloads it and stores it as something to be analyzed. The analyzing part has also changed for now the JavaScript Injection crawler only crawls web pages, and the data is analyzed afterwards due to the high amount of it to analyze.

The time planning has also changed. Task 7 JavaScript Injection took three more weeks than expected and 16 Test and results has been performed alongside the last. Therefore, since the delivery date has been postponed for three months, the rest of the dates have not changed, just been delayed.

7. State of the Art

In this section further information regarding VPNs and add-ons will be provided regarding the current state of these.

7.1 VPNs

A virtual private network (VPN) gives you online privacy and anonymity by creating a private network from a public internet connection.VPNs mask your internet protocol (IP) address so your online actions are virtually untraceable. Most important, VPN services establish secure and encrypted connections to provide greater privacy than even a secured Wi-Fi hotspot [24].

Essentially, they create a protected data tunnel between your local data network and an exit node that can be in another location far from the original. This tunnel is protected with encryption. However, there are different levels of encryption as well. By default OpenVPN includes AES-128 [4], SHA-1 [34], and RSA-2048 [31], while maximum encryption includes AES-256 [34], SHA-256 [35], and RSA-4096 [31]. An encryption key tells your computer how to decrypt or encrypt data. Data authentication is part of the encryption process and refers to the message authentication algorithm with which user data is authenticated. This is used to protect users from active security attacks [3].

7.1.1 Where can VPNs be installed

VPNs are supported in many OS. Some of them are Windows, Mac OS X, iOS, Android, and Linux (Ubuntu) and Chrome book. They also work on mobile devices, like Android, IPhone and IPad.

Some routers also support VPNs. Find some of the firmware that support VPNs below.

Firmware package	**Cost**	**Developer**
DD-WRT	Free	NewMedia-NET GmbH
Gargoyle	Free	Eric Bishop
OpenWrt	Free	Community driven development
OPNsense	Free	Deciso BV
Tomato	Free	Keith Moyer

Table 7.1: Firmware that support VPNs

7.1.2 Are they anonymous?

While most VPN providers claim they do not keep logs, it is true that most companies keep basic connection logs(User IP address and VPN timestamp). It is very important to check the companies privacy policy and the location of their headquarters (in terms of jurisdiction).

It is also important to take into account the encryption standard used. Many VPN clients allow you to choose different standards. While the securer standards may slow down the internet connection, they will make sure your data is safer.

Typically, your ISP would have to pay special attention to your activity in order to determine whether you are using a VPN or not. However, if well protected, they won't be able to learn the nature of your activity. They are only able to see you are connected to a certain endpoint and the bandwidth you are using. If they want to learn more about you, they may request logs to your VPN provider (they probably need a court order to do that). See more information on this topics in the following subheadings.

Personal information

It is well known that companies keep personal information when we use their products. Now a days, it is very common (if not obligatory) that web pages ask you to accept their privacy policy and the use they give to your data (be by using cookies or any other means). When choosing a VPN service, it is very important to read carefully the privacy policy of that company in order to be aware of the data they store about us. It would make no sense that we used a VPN to hide our personal data (IP, billing information, name, . . .) but the VPN itself stores and uses that information!

Different levels of logging

Companies may have different logging levels. These are the three main categories:

1. **Usage (browsing) logs** – These logs basically include online activity: browsing history, connection times, IP addresses, metadata, etc. From a privacy standpoint, you should avoid any VPN that collects usage data. Most of the VPN services that are collecting usage logs are free VPN apps, which are basically spyware. The data they collect is then sold to third parties, thereby monitoring the "free VPN" service.

2. **Connection logs** – Connection logs typically include dates, times, connection data, and sometimes IP addresses. Typically this data is used for optimizing the VPN network and potentially dealing with user problems or terms of use issues (torrenting, illegal activities, . . .).

3. **No logs** – No logs simply means the VPN service is not keeping any logs whatsoever. Having a truly no logs policy can be difficult to implement while at the same time enforcing restrictions, such as device connections or bandwidth. This is especially the case when VPNs need to enforce restrictions such as bandwidth or the number of devices being used per subscription.

While basic connection logs are not necessarily a problem, there is an increasing number of VPNs that keep connection logs, while falsely claiming to be a "no logs" service. Examples of this are Betternet, PureVPN, Windscribe, and TunnelBear. That is why it is very important to be aware of our VPN provider's privacy policy and the jurisdiction of their headquarters [2].

Data encryption

Essentially, data encryption provides users with a secure VPN tunnel whereby data is protected. However, there are different levels of encryption as well. By default OpenVPN includes AES-128 [4], SHA-1 [34], and RSA-2048 [31], while maximum encryption includes AES-256 [4], SHA-256 [35], and RSA-4096 [31]. An encryption key tells your computer how to decrypt or encrypt data.

Data authentication is part of the encryption process and refers to the message authentication algorithm with which user data is authenticated. This is used to protect users from active security attacks.

Further considerations

VPN servers [33] may use static or dynamic IP addresses. The later are the most common used. There are pros and cons for both approaches, and they can be important to the user depending on the use given to the VPN (and the budget available).

Dynamic IP address

- Pros - They provide greater anonymity for public WiFi hotspots, for they implement NAT and offer a wider range of IPs.
- Cons - IPs can be abused by other users of the network [3], be it by spamming or hacking. This can lead to being banned from certain websites, or being suspicious of illegal activity.

Static IP address

- Pros - Unique IP for user. Harder to block by ISP or Network Administrators, since they use their own ports.
- Cons - Expensive.

Some VPN servers [33] offer built-in DNS protection. This happens because there is a well-know flaw in some operative systems that, despite having routed the traffic through a VPN server [33], keep contacting websites using their original IP and DNS. This is a treat to your online security, anonymity and usefulness of your VPN. Al tough these DNS leaks can be prevented, it can be worth looking for a VPN with built-in protection.

Notice that using a VPN does not protect you from malicious software. While some VPN providers offer additional protection, most do not. Viruses, spyware and malware can still harm you when you are using a VPN.

Many VPN offer anonymous payment options, like Bitcoin, Altcoins, Alipay, CashU, PaySafeCard, Gift Cards and others. These can be an extra anonymity layer to take into account.

Deep packet inspection (DPI) [10] or packet sniffing is a type of data processing that inspects in detail the data being sent over a computer network [8], and usually takes action by blocking, re-routing, or logging it accordingly. It is used by ISPs, government agencies, and hackers to monitor and retain all of the data transmitted to and from your computer, including confidential and private information. To help customers bypass this, VPN providers are actively implementing modified OpenVPN protocols with added obfuscation layers, which masks VPN traffic away from view of DPI crawls. Check your VPN provider on this topic.

7.1.3 Free or Paid VPNs

Free VPN services normally offer slower bandwidth and a smaller server [33] range. It is also important to consider that everyone wants to make a living, so while they claim to be free, they may make money using other means, which, in some cases, compromise the users confidentiality (see the case of Hola in the appendix).

A legitimate and reliable paid VPN service offers several advantages, including a much more private environment, authentic guarantees, faster speeds, more servers [33], more extra features and add-on services, as well as being less likely to log traffic or connection logs. Normally, and given that you pay for them, they are more likely to be safe

Name	**Connections**	**Servers**	**Speed**	**Logs**	**Type**
ProtonVPN	1	3	Medium	None	Full VPN
Hotspot Shield	1	1	?	None	Full VPN
Windscribe	1	11	Low	Connection	Full VPN
TunnelBear	1	20	Medium	Browsing	Full VPN

Table 7.2: Free VPN examples.

Name	**Connections**	**Servers**	**Speed**	**Logs**	**Type**
ProtonVPN	5	55	High	None	Full VPN
Hotspot Shield	5	3200	High	None	Full VPN
Windscribe	1	11	High	Connection	Full VPN
TunnelBear	5	120	High	Browsing	Full VPN

Table 7.3: Paid VPN examples.

7.1.4 What does a VPN hide?

There is a lot of information that can be protected by means of using a VPN. Among it, we can find the following:

- **Your browsing history:** Since they create an encrypted data tunnel, VPNs can hide your browsing history from your ISP and other unwated overseers by not showing them straight away the data running through your network.

- **Your IP address:** IP addres, even though not the best, still is a common tracking technique. By hiding your personal IP address, VPNs can help mask your real identity.
- **Your Location:** Your physical location is often discerned by external sources by means of the IP address. By masking your IP address, VPNs can also hide your location.
- **Your device:** A VPN can help protect your devices, specially mobile ones by protecting your traffic wherever you go. Connecting to public Wi-Fi networks can put in danger your device's data, but a VPN will help mitigating that.

7.2 Browser add-on

Browser extensions main purpose is to extend your browser with additional features. To do this, they may modify webpages, use the browser storage or connect to external services, for example. However, they can be used to remove unwanted features or functionalities, like pop-up ads and other aspects of a websites core behavior that a user wishes to opt out of. Even though browser companies make some of these extensions, most of them are made by third-party developers. Finding out which ones to trust isnt always easy, although each browser will often offer you a selection depending on what you are trying to do with your machine.

7.2.1 What can they do?

Depending on the browser, the permissions a extension can have differ. You may find quoted the permissions the extensions can use in Firefox browser [25] and in Chrome browser [9]. The truth is that extensions specify the permissions they require, and if users want to use them, they have to accept those permissions.

However, to give some general ideas, browser extensions can normally access to the following data if they require to do so.

- Clipboard.
- Browser History.
- Cookies.

- Downloads.
- Webpage document.
- Browser local storage.
- System data.

Name	**Location**	**User Activity**	**Website Content**	**Personal Information**
HTTPS Everywhere	No	No	No	No
Google Translate	Yes	Yes	Yes	No
Zoom Scheduler	No	Yes	No	No
BattleTabs	Yes	Yes	No	Yes

Table 7.4: Browser add-on examples and the data they collect.

7.3 VPN Extensions

A browser extension is a simpler, more lightweight proxy solution (usually a http proxy) that is often used by people who desire faster connection speeds than those provided by their primary VPN client. However, it is very important to understand that VPN extensions do not provide a full VPN tunnel. They normally are a browser-based proxy service [37].

7.3.1 What are the advantages of a VPN extension?

A VPN browser extension is great for just flipping on and off when you need it. Proxy browser extensions are most useful for quickly accessing regional web pages when you can't be bothered to turn off your VPN client - or change the server you are currently connected to.

When using a VPN proxy extension, you can stay connected to a VPN server in the US (for example) but change the browser extension to the UK (for example) to quickly access a website in that specific region.

7.3.2 Kinds of proxy extensions

- HTTP Proxy: Only useful for geolocation purposes. Not secure.
- HTTPS Proxy: Encrypts HTTP traffic, but does not necessarily proxy DNS requests.
- SSL Proxy: Basically the same as HTTPS Proxy.
- SOCKS Proxy: Relatively new protocol that permits consumers to relay a broader selection of data using TCP or UDP. It is a proxy that can be set up securely with encryption.

Type	**Name**	**URL**
HTTP Proxy	HolaVPN	https://hola.org/
HTTPS Proxy	TouchVPN	https://touchvpn.net/
SOCKS Proxy	VPN Master	https://github.com/Emano-Waldeck/vpn-master

Table 7.5: Examples of Proxy extensions.

8. Description of the project

This project focuses in detecting JavaScript injection performed by VPN extensions with the objective to obtain user data. Note that the term "resource" is often used in this section. I use resource to refer to any script or iframe present in a certain page or loaded in it.

The main hypothesis of the project are:

1. A VPN extension that injects JavaScript code into web pages to track users will inject the code in all pages.

2. Common files found in the crawled list should not be found if performed without that said VPN extension.

Name
vimeo.com
theguardian.com
ebay.com
buydomains.com
usatoday.com
lefigaro.fr
bing.com
4shared.com
elpais.com
mediafire.com
abc.es
elmundo.es

Table 8.1: Web pages visited with crawler.

Table 8.1 shows the list of websites crawled. This list is a proof of concept. Depending on the results obtained with it, a large-scale analysis may be developed. This selection responds to websites that are likely to present tracking code or advertising mechanisms, for this would make it less obvious for an extensions to inject ads in those.

The extensions tried to use servers in Spain or close to it, to reduce location-related changes in the websites. Also note that each website has been visited a minimum of 100 times with each extension. All the websites have been visited more than 500 times without any VPN extension. Websites are visited several times because I noticed resources found in them can vary depending on the visit.

The VPNs tested in this project meet two requirements. First one is that they are free, for these are the ones most likely to look for alternative ways of making profit out of their users and because they are easily accessible. Second requirement is that these VPNs are browser extensions. This second requirement comes from the focus of this project, for browser extensions have to capability of altering the website, and this implies they have the possibility of injecting malicious JavaScript code.

The extensions tested in this project are:

All the modules follow a script-like design. They are run with options and argument in the command line. In the following sections you may find a more detailed explanation of the structure of the project and its behaviour.

8.1 Data Crawler

The data crawler module is in charge of extracting the data that will later be analyzed by the Data Processor.

8.1.1 How does it work?

Figure 8.1 shows a diagram of the crawler's behaviour. When crawlers are launched, a thread is created for each VPN extensions provided. The crawler launches a Chromium browser instance with all the data cleared and loads the given extension on it, along with our Page Downloader custom extension. From this point, it waits for user confirmation to start crawling the list of websites. This works in this way because selenium does not have the capability to interact with browser extensions in any way, and it is required that VPN extensions are activated manually before each execution.

Name	**Open Source**	**Chrome Rating**
1clickVPN	No	4.5
AdGuard VPN	Yes	4.5
AStar VPN	No	4.7
Betternet VPN	No	4.4
BrowSec VPN	No	4.5
Daily VPN	No	4.0
DotVPN	No	3.7
EarthVPN	No	4.6
Free VPN	No	3.5
Hola VPN	No	4.9
Hotspot VPN	Yes	4.4
Hoxx VPN Proxy	No	4.7
IP Unblock	No	4.1
PP VPN	No	4.2
Prime VPN	No	4.5
RUSVPN	No	4.6
Surf VPN	No	4.1
Touch VPN	No	4.6
Urban VPN	No	4.6
VeePN	No	4.8
VPN Proxy Master	No	4.6

Table 8.2: Extensions analyzed in the project

Once manual confirmation has been performed, the crawl is started. After each website, all data from the browser is cleared using the chrome menu and selenium utilities. This is done to avoid any interference between websites because of cookies or cache data. Every website can be visited multiple times by a certain crawler if options are specified.

The crawler (using the Page Downloader extension mentioned above) downloads all the <script> and <iframe> tags found in the page visited, and also uses the performance logs of the browser to detect every network requests performed while in the page. If the requests is of mime-types text/*, application/javascript or application/x-javascript, it downloads its contents.

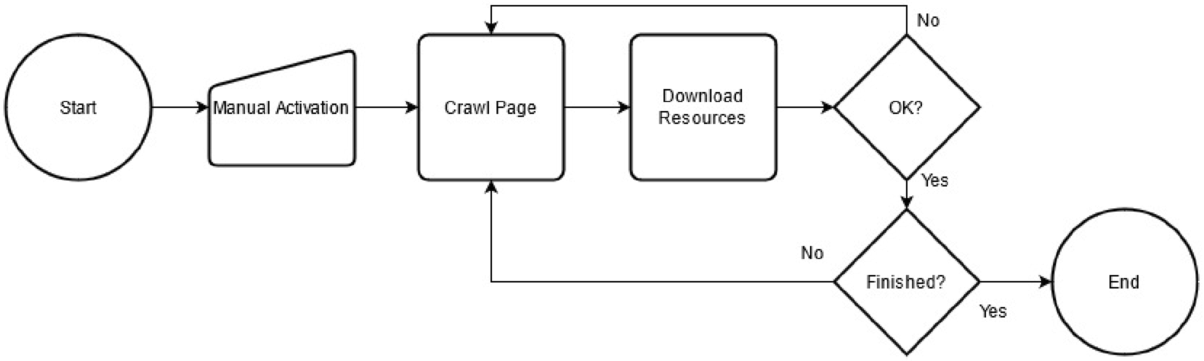


Figure 8.1: Crawler behaviour diagram. Source: Original.

Each of this is named with the MD5 hash of its contents.

When an error occurs (lost internet connection, website unreachable, etc.) another crawler is launched that skips the website that caused the error. However, manual activation of the extension is required again.

8.1.2 Structure

Figure 8.2 shows UML diagram of the crawler. The main script manages the different threads generated. Each thread is composed of a Manager, which makes sure the Crawler is goes through the website list given.

8.2 Data Processor

The data processor module is in charge of storing resources in the database and using said data to find suspicious files.

8.2.1 How does it work?

Figure 8.3 show a diagram of its behaviour.The data processor can do two tasks:

- Store resources in the DB: The data processor can store resources in the database to be later analyzed. It uses the folder structure created by the data crawler to deduce all the necessary data from the resource. Note that resources are categorized according, mainly to the following characteristics: when it was obtained, from which website it came, which extension was used and its contents. Therefore, I can organize Resources in specific

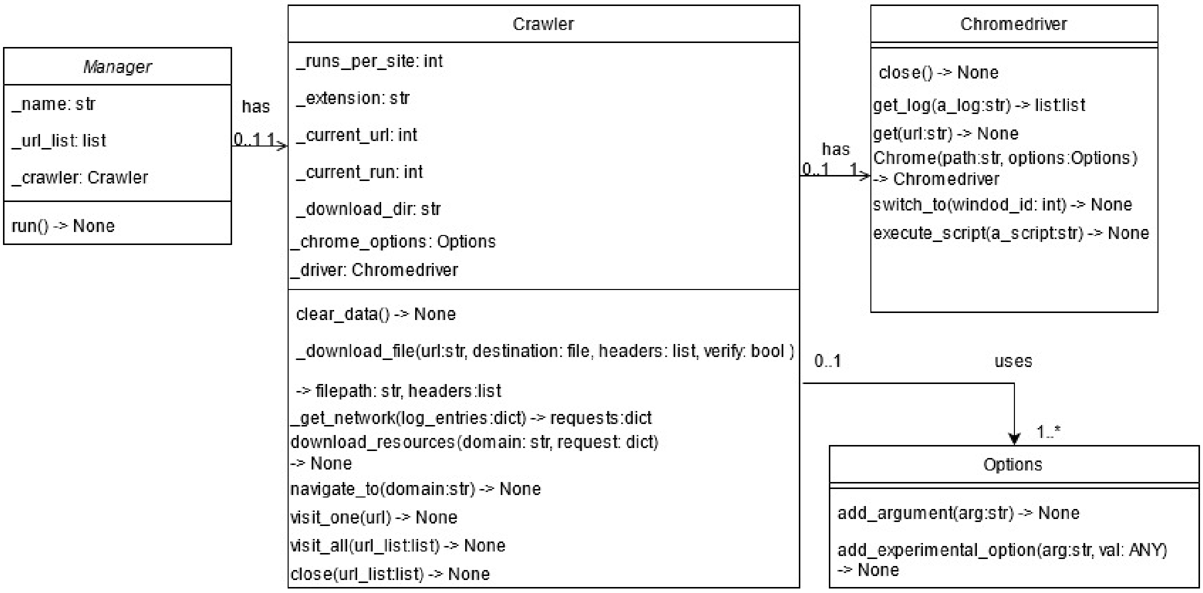

Manager
_name: str
_url_list: list
_crawler: Crawler
run() -> None
has
0..1 1
Crawler
_runs_per_site: int
_extension: str
_current_url: int
_current_run: int
_download_dir: str
_chrome_options: Options
_driver: Chromedriver
clear_data() -> None
_download_file(url:str, destination: file, headers: list, verify: bool)
-> filepath: str, headers:list
_get_network(log_entries:dict) -> requests:dict
download_resources(domain: str, request: dict)
-> None
navigate_to(domain:str) -> None
visit_one(url) -> None
visit_all(url_list:list) -> None
close(url_list:list) -> None
has
0..1 1
Chromedriver
close() -> None
get_log(a_log:str) -> list:list
get(url:str) -> None
Chrome(path:str, options:Options)
-> Chromedriver
switch_to(windod_id: int) -> None
execute_script(a_script:str) -> None
0..1
uses
1..*
Options
add_argument(arg:str) -> None
add_experimental_option(arg:str, val: ANY)
-> None

Figure 8.2: Crawler Domain UML. Source: Original.

Runs (when it was obtained), and Runs in RunCollections, according to their extension. However, to make resources themselves meaningful, they store the extension and webpage they belong to. Resources also store a Variety, that is, an example of that resource (content of the resource and MD5 hash).

- Process data: The data processor can also use the resources stored in the database. The analyzing process is fairly simple: it consists on comparing all runs performed by a certain VPN extension to find common files in all of them. This sits on the hypothesis one. Then these common files are compared to the common files found in the list with no VPN extensions installed. The difference of the last two are the "suspicious" files, that are later analyzed manually.

  Resources are compared by similarity of their contents. When two resources reach a certain similarity threshold, I can say they are the same. This is because some scripts can be dynamically generated by the extensions and therefore contain small changes (like a URLs or a token).

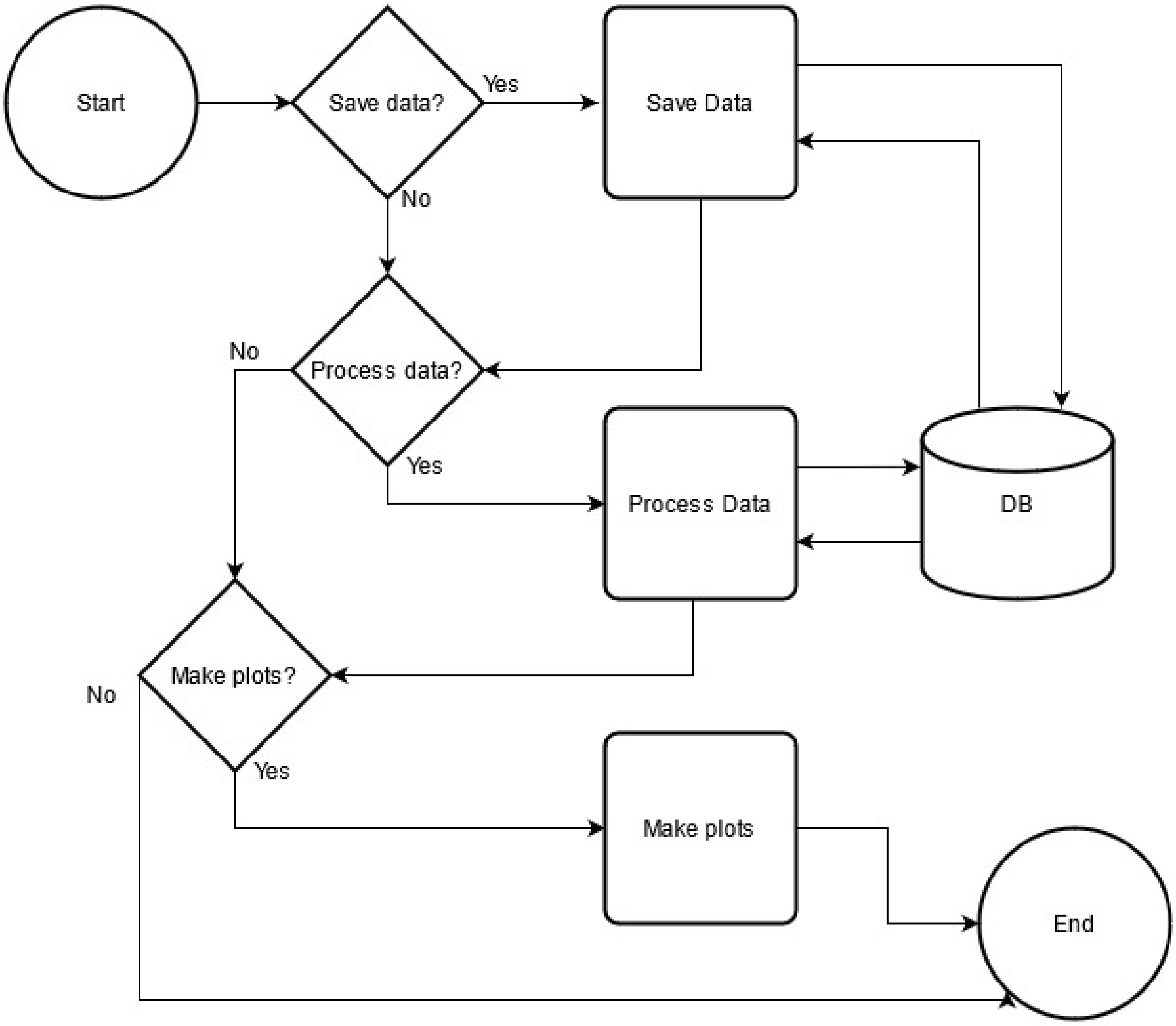


Figure 8.3: Data Processor behaviour diagram. Source: Original.

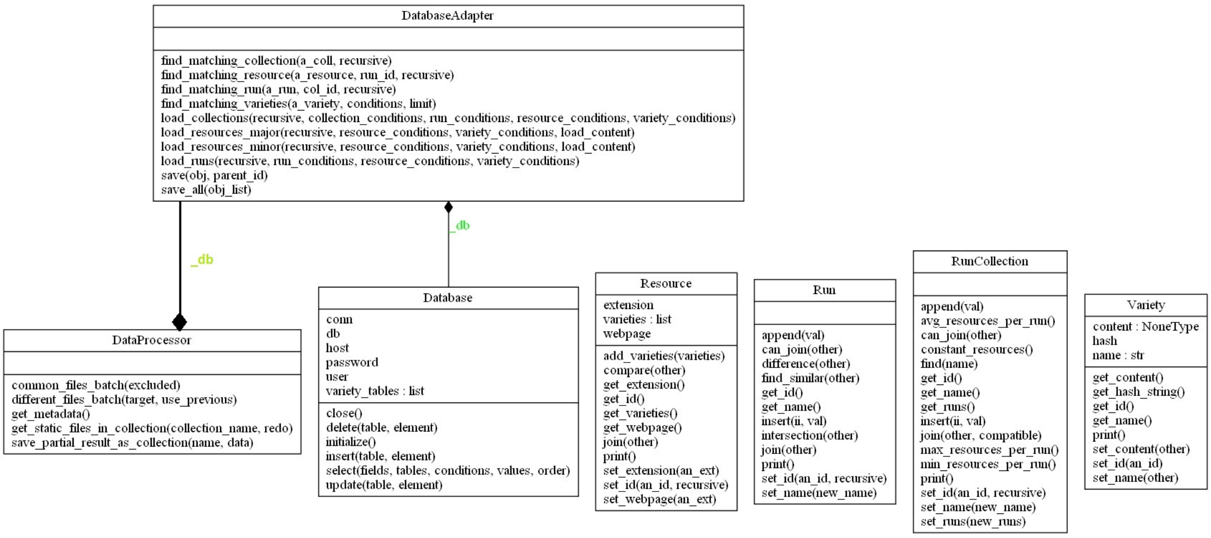


Figure 8.4: Data Processor Domain UML. Source: Original.

8.2.2 Structure

Figure 8.4 shows a UML diagram of the Processor. The main script chooses weather to store new resources in the DB or to start an analysis of what is present in it.

8.3 Technologies

The original idea of the project was an extension-like software that was able to discern weather a specific web page had been injected JavaScript code by the extension. However, after analyzing more deeply the implications of this approach and the technologies available, I reached the conclusion that something like this would not work for two main reasons:

1. Our project compares resources found in a page with and without VPN extension. An extensions-like approach would make the user experience significantly slower, for each page would be visited a minimum of two times.

2. To be sure a resource does not belong to a site, I need to do as many visits to that site as possible. This would also make a live version of this product unfeasible, and therefore the extension approach does not fit our needs

After discarding this approach, two more came to our minds. The first one was to create a service that would check for JavaScript injection in extensions indefinitely. The second, a script-like approach that would be executed when needed in a remote machine (like a computing server). For complexity reasons, the second approach was chosen.

The other two issues that raised where how would I do the crawling of the web pages and compare the similarity of resources? After doing some research on these topics, I found the following options:

8.3.1 Crawling

A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically operated by search engines for the purpose of Web indexing (web spidering) [42].

- Octoparse: Desktop application made for automatic data extraction from web pages. No coding required.
- Scrapy: Python library created to make the creation of web crawlers easy.
- Selenium: Framework for automated software testing of web applications capable of interacting with browsers.

In our case, Selenium [22] was the best choice, for it allows us to interact with a web browser, load self-made extensions and load dynamic JavaScript by interacting with the browser. Being Chrome one of the most used browsers, I used it in our tests. Note however, most of the extensions available in Chrome also exist for other browsers.

8.3.2 Data processing

Given that Selenium was chosen for the data crawling, I needed a technology compatible with Selenium for the data processing. Python is a well-known programming language very popular because that lets you work quicker and integrate your systems more effectively, and of course can use the Selenium framework along with many other libraries. Therefore, at this point of the project I focused in finding answers in Python regarding how to compare two files to tell how similar they are.

There are a lot of studies regarding NLP [23] and similarity of sets [12]. In this project, the datasketch 1.5.3 module from Python was used, more specifically their MinHash implementation. it lets you estimate the Jaccard similarity (resemblance) between sets of arbitrary sizes in linear time using a small and fixed memory space [21]. MinHash and Jaccard similarity is useful to detect document resemblance and repeated documents in a large dataset, as proved my Andrei Z. Broder in his paper [6].

9. Final results

In this section you will find an analysis of the results obtained by the project described in the two previous sections. That is, the result of analyzing the data obtained with the Crawler module using the Data Processor module.

To sum up the behaviour of the Crawler module, it launched automated crawlers that navigated through a certain list of web pages several times. Each of those visits to each of those web pages started with a clean storage, so that previous web pages visited or previous visits to that page do not alter the contents of that page in that visit. When a page was visited by the crawler, it downloaded every script and iframe that page had, along with every external resource loaded that belongs to the types "text/*", "application/*" according to the request used to load it.

Regarding the Data Processor module, it stored the files of the crawler in the database if they were new and then analysed them by two means. Firstly, it searched for files common in all usages of a certain VPN extension (to discard web page-specific resources) and also compared those common files to the common files of that list of web pages visited by a browser with no VPN extension at all (to discard common resources loaded by all those web pages).

In the following sections I will discuss the results found for all extensions tested. Furthermore, the results of the project will be compared with what the Privacy Policy of each of these products says regarding user data collection and usage.

9.1 Metadata Analysis

Before proceeding to look for files injected by the VPN extensions, a general data analysis was performed. In this section you will find the results obtained from said analysis. The figures mentioned in this section are of two types: bar charts and grouped bar charts. The first type of chart shows the number of resources loaded in a specific web page of the list on average by all extensions. The later show the minimum, maximum and the average number of resources loaded on a specific web page. Note that the normal bar charts contain a dotted line. It is there only to help visualize the number of resources the page has without a VPN active compared to the others. In all charts, NO_VPN refers to the results obtained without any VPN extension loaded.

The data shown in figures 9.1 and 9.2 indicates that, in general, when a web page is visited with a VPN extension of the ones mentioned on table1.1 it loads more resources of types text/* and application/*javascript than usual. And this same thing may indicate that the extensions is injecting code onto the page. It could also be that the extension is using external resources during its normal behaviour. This rises the need of finding what this additional loaded resources are.

This is specially noticeable in the charts of abc, elmundo, elpais and lefigaro. These sites normally contain ads of their own. That is why I think they are more likely to be injected ads by the VPNs, for they will most likely remain unnoticed.

After executing the data processor, three examples of this additional resources have been found. The following sections analyze these resources found and give a bit of context on the VPN extension that loaded them and its privacy policy.

9.2 Hola VPN

Hola, based out of Israel, is mainly a free VPN service. It also offers paid subscriptions. Hola was the first P2P based VPN service created. But this VPN services did not become famous for its improvements, but for its bad practices. Their privacy policy [15] states that:

- **Location of their headquarters** - They are located in Israel, outside the 14-Eyes Countries and Jurisdiction [13]. However, Israel is cooperative with this alliance, where they share all their espionage information.
- **Logging level** - They log the following data: browser type, web pages visited, time spent on this pages, access times and dates. We could say they know nearly everything about your internet activity. They store usage logs.
- **Personal data** - They store your IP address, name, email and they may share it to "subsidiaries and affiliated companies".

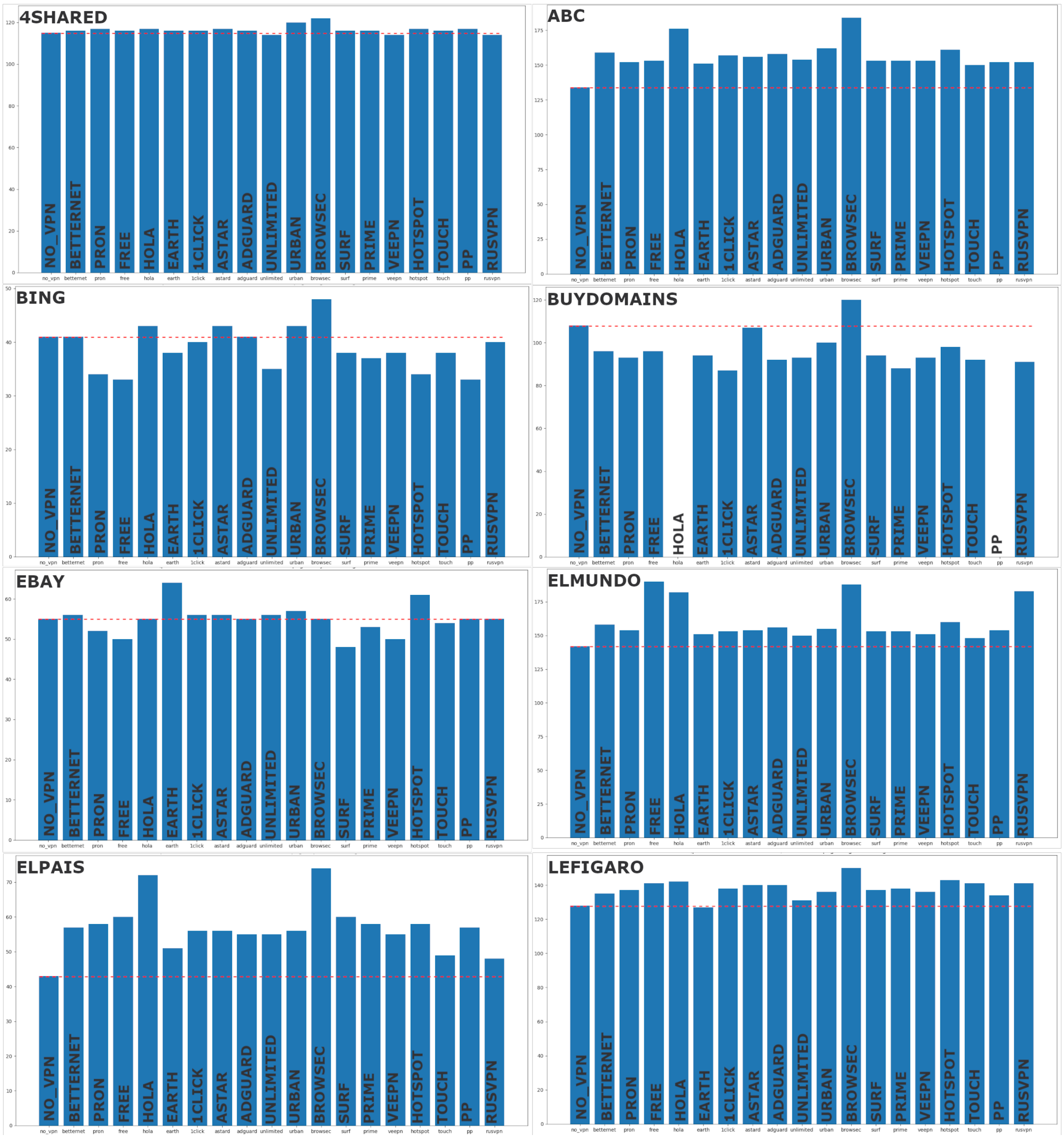
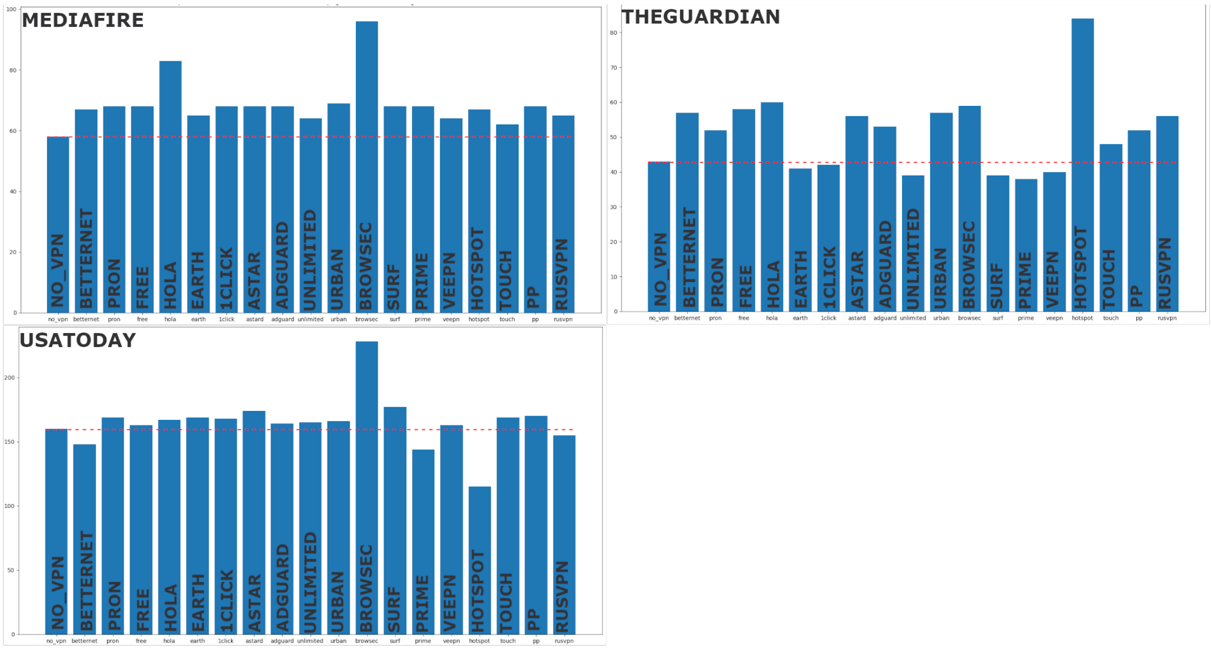


Figure 9.1: Common resources found in all runs of pages (One). Source: Original.



MEDIAFIRE

THEGUARDIAN

USATODAY

Figure 9.2: Common resources found in all runs of pages (Two). Source: Original.

- **Encryption level** - Non-existent. Hola VPN uses a HTTP proxy tunnel, transmitting information in plain text.

Hola is a special type of VPN that uses a P2P network to provide its services, as mentioned above. This means that users themselves are used as proxy servers, with all the security risks this implies. Take into consideration that, when using this VPN, other users can use your computer bandwidth for illegal purposes. Furthermore, they are aware that malicious users may bypass their security checkups and gain access to your computer and they are warning you about that in their security policies, saying they will try their best to fix this issues.

Furthermore, the results obtained by the Data Processor reveal that Hola VPN is injecting file 11.1.1 in every web page visited. This script is most likely related with tracking, as the element "pageAnalytics" indicates in line 5. It also interacts with an object called ga_data in lines 7-8, which is most likely referring to Google Analytics data.

Considering their privacy policy mentions they log pages visited and time spent in them (browsing data), it makes sense to think they use the previous script, maybe among other methods, to track this parameters.

I have also compared the pages visited with no VPN and with Hola VPN active. Figures 9.3 and 9.4 show clear examples of injected ads in websites. In fact, Example 2 does not have an ad injected, but thew page was modified to insert the add adding a blank space and the ad failed loading.

9.3 Hotspot Shield

Hotspot Shield, based in the United States is mainly a paid VPN service. However, it also offers a free subscription that implies receiving personalized advertisements. The following analysis will be centered on the free version of the VPN, for it is the one I have analyzed.

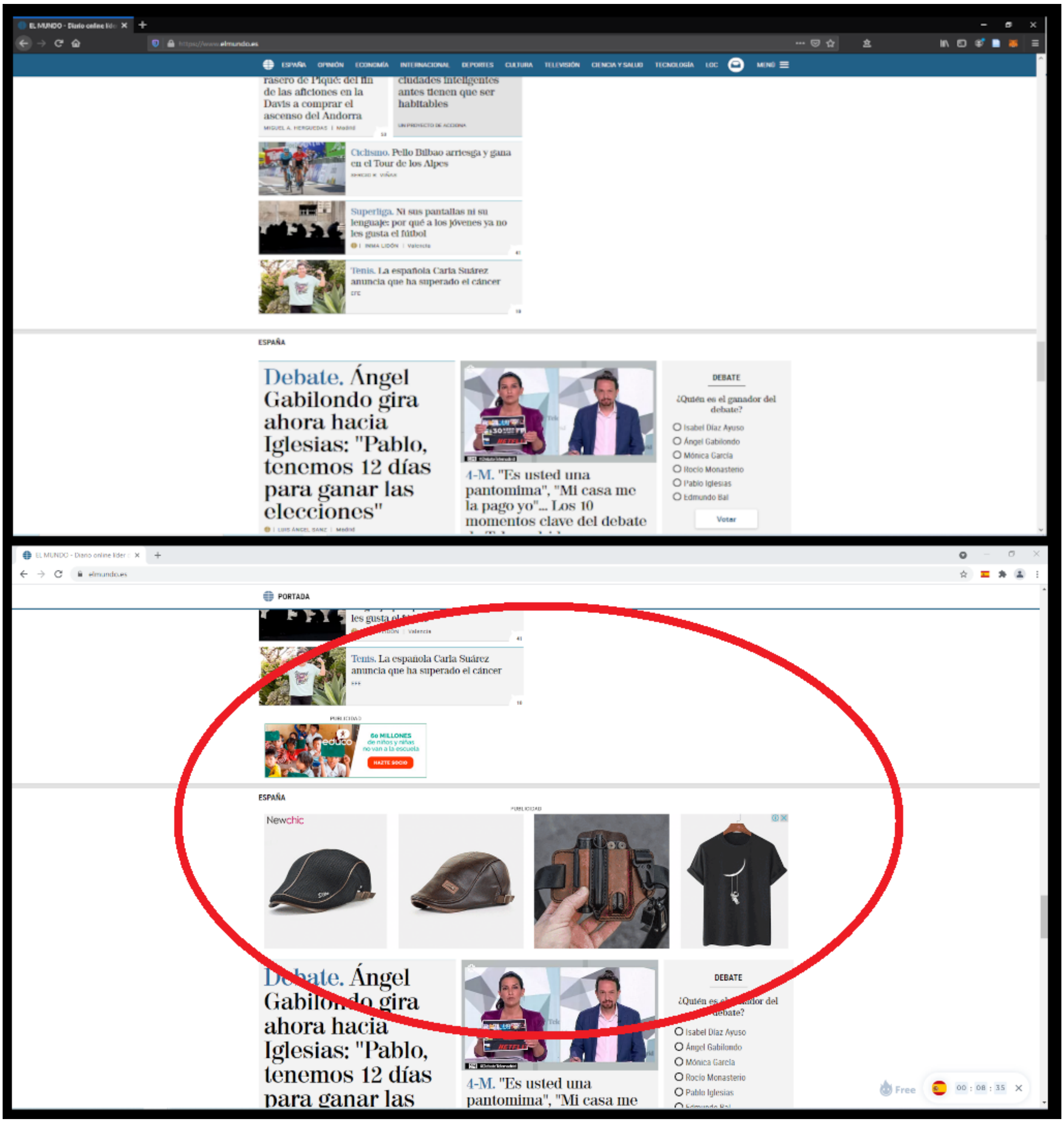
- **Location of their headquarters** - Hotspot Shield is headquartered in the US, one of the founding members of the 5/9/14-Eyes Alliance [13].
- **Logging level** - They say they only store connection logs, but the information in their logs is abundant. They log the following data: Email, Username, Unique mobile ID, Hardware model, OS version, language, "Network information" (Not specified), Location, Advertising ID, MAC Address.
- **Personal data** - Free users personal information is shared with third-party companies. That is: location, Advertising ID, Mobile ID, MAC Address and Wireless carrier.
- **Encryption level** - It uses 256-bit AES.

Their privacy policy states that they may place specific advertisements and share personal information with third parties. The data processor has found the two files 11.1.2 and 11.1.2 repeated throughout the web page list.

Due to the length of the isolated file found for this VPN, the appendix only contains the relevant parts of it for this project. The file 11.1.2 uses a scraper in line 6 to obtain the data specified in the string contained in the variable "userKeys". Among other parameters, it tries to obtain the user gender and age, which are not specified in the Privacy Policy of said VPN.

File 11.1.2 has to do with Amazon Advertisement service. Even though the script is too big to fully comprehend all the actions behind it, references to the AdServer, iframes and ads in general are found in multiple places in the script. The amazon ad server is referenced in multiple lines, along with some of its domains. Therefore, and in conjunction with what is found in the privacy policy and the comparisons below, this apparently is a script used by the VPN to inject advertisements in web pages.

I have also compared the pages visited with no VPN and with Hotspot VPN active. Figures 9.5 and 9.6 show clear examples of injected ads in websites.


EL MUNDO - Diario online líde
https://www.elmundo.es
ESPAÑA OPINIÓN ECONOMÍA INTERNACIONAL DEPORTES CULTURA TELEVISIÓN CIENCIA Y SALUD TECNOLOGÍA LOC MENÚ
UN PROYECTO DE ACCIONA
ESPAÑA
DEBATE
Votar
EL MUNDO - Diario online líder
elmundo.es
PORTADA
PUBLICIDAD
educo
60 MILLONES de niños y niñas no van a la escuela
HAZTE SOCIO
ESPAÑA
PUBLICIDAD
Newchic
Free
00 : 08 : 35

rasero de Piqué: del fin de las aficiones en la Davis a comprar el ascenso del Andorra

MIGUEL A. HERGUEDAS | Madrid

ciudades inteligentes antes tienen que ser habitables

Ciclismo. Pello Bilbao arriesga y gana en el Tour de los Alpes

Superliga. Ni sus pantallas ni su lenguaje: por qué a los jóvenes ya no les gusta el fútbol

INMA LIDÓN | Valencia

Tenis. La española Carla Suárez anuncia que ha superado el cáncer

EFE

Debate. Ángel Gabilondo gira ahora hacia Iglesias: "Pablo, tenemos 12 días para ganar las elecciones"

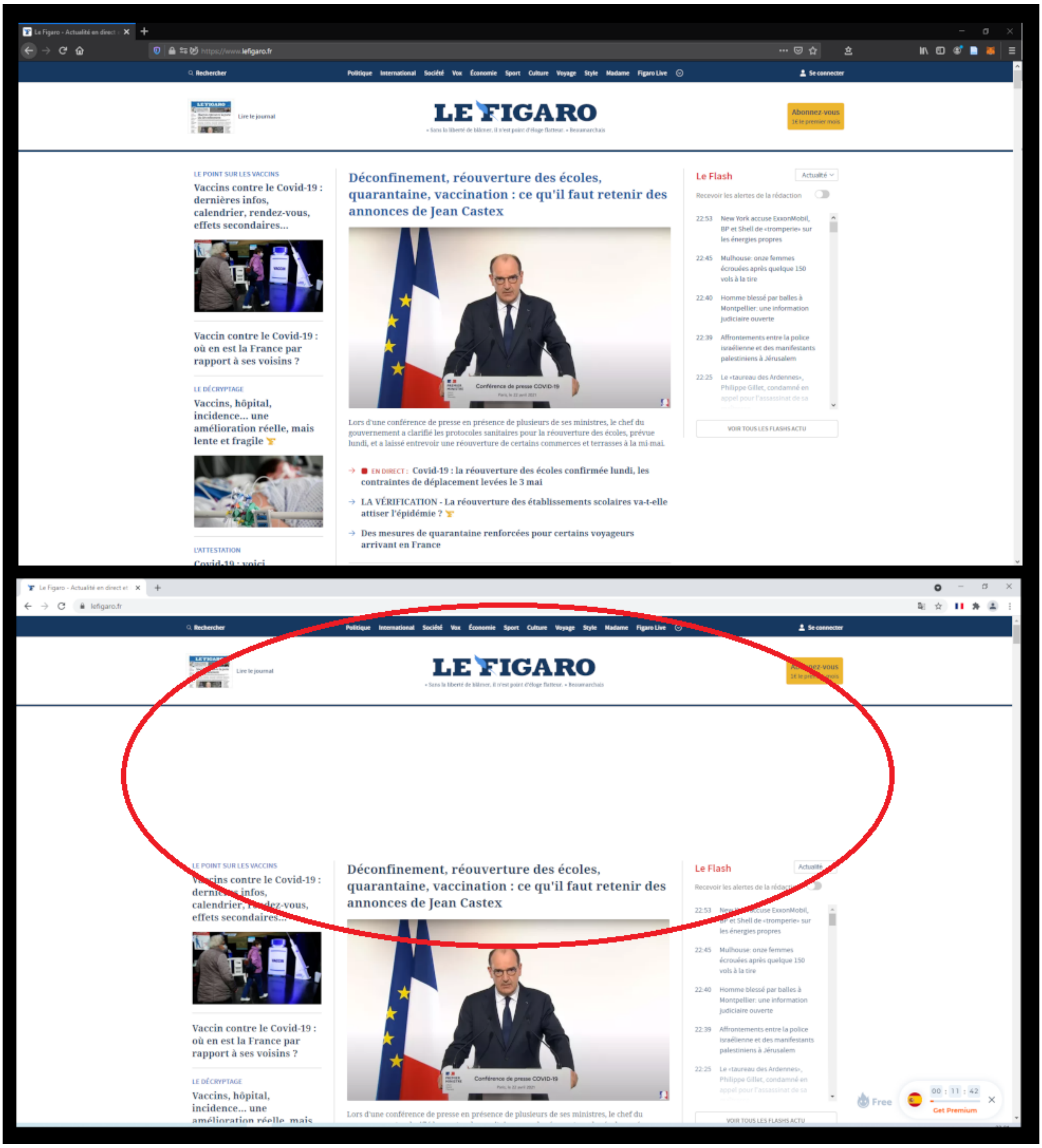
LUIS ÁNGEL SANZ | Madrid

4-M. "Es usted una pantomima", "Mi casa me la pago yo"... Los 10 momentos clave del debate

¿Quién es el ganador del debate?

○ Isabel Díaz Ayuso
○ Ángel Gabilondo
○ Mónica García
○ Rocío Monasterio
○ Pablo Iglesias
○ Edmundo Bal

Figure 9.3: Example 1. On top, page visited with no VPN extension. At the bottom, page visited with Hola VPN. Source: Original.



Le Figaro - Actualité en direct

https://www.lefigaro.fr

LE POINT SUR LES VACCINS

Vaccins contre le Covid-19 : dernières infos, calendrier, rendez-vous, effets secondaires...

Vaccin contre le Covid-19 : où en est la France par rapport à ses voisins ?

LE DÉCRYPTAGE

Vaccins, hôpital, incidence... une amélioration réelle, mais lente et fragile

L'ATTESTATION

Covid-19 : voici

Déconfinement, réouverture des écoles, quarantaine, vaccination : ce qu'il faut retenir des annonces de Jean Castex

Lors d'une conférence de presse en présence de plusieurs de ses ministres, le chef du gouvernement a clarifié les protocoles sanitaires pour la réouverture des écoles, prévue lundi, et a laissé entrevoir une réouverture de certains commerces et terrasses à la mi-mai.

→ EN DIRECT : Covid-19 : la réouverture des écoles confirmée lundi, les contraintes de déplacement levées le 3 mai

→ LA VÉRIFICATION - La réouverture des établissements scolaires va-t-elle attiser l'épidémie ?

→ Des mesures de quarantaine renforcées pour certains voyageurs arrivant en France

Le Flash

Recevoir les alertes de la rédaction

22:53	New York accuse ExxonMobil, BP et Shell de «tromperie» sur les énergies propres
22:45	Mulhouse: onze femmes écrouées après quelque 150 vols à la tire
22:40	Homme blessé par balles à Montpellier: une information judiciaire ouverte
22:39	Affrontements entre la police israélienne et des manifestants palestiniens à Jérusalem
22:25	Le «taureau des Ardennes», Philippe Gillet, condamné en appel pour l'assassinat de sa

Le Figaro - Actualité en direct

lefigaro.fr

LE POINT SUR LES VACCINS

Vaccins contre le Covid-19 : dernières infos, calendrier, rendez-vous, effets secondaires...

Vaccin contre le Covid-19 : où en est la France par rapport à ses voisins ?

LE DÉCRYPTAGE

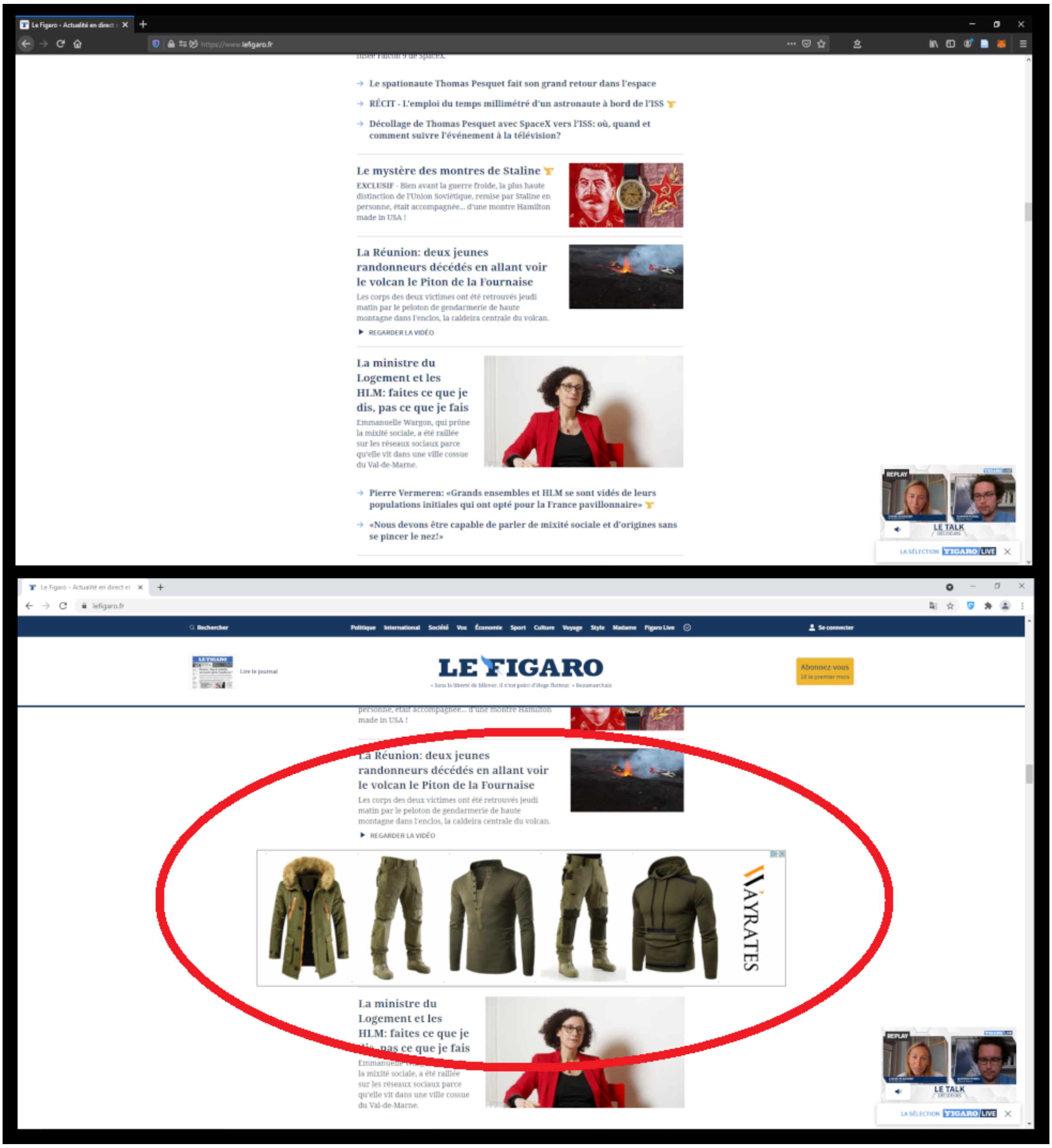
Vaccins, hôpital, incidence... une amélioration réelle, mais

Déconfinement, réouverture des écoles, quarantaine, vaccination : ce qu'il faut retenir des annonces de Jean Castex

Lors d'une conférence de presse en présence de plusieurs de ses ministres, le chef du

Le Flash

Figure 9.4: Example 2. On top, page visited with no VPN extension. At the bottom, page visited with Hola VPN. Source: Original.



→ Le spationaute Thomas Pesquet fait son grand retour dans l'espace

→ RÉCIT - L'emploi du temps millimétré d'un astronaute à bord de l'ISS

→ Décollage de Thomas Pesquet avec SpaceX vers l'ISS: où, quand et comment suivre l'événement à la télévision?

Le mystère des montres de Staline

EXCLUSIF - Bien avant la guerre froide, la plus haute distinction de l'Union Soviétique, remise par Staline en personne, était accompagnée… d'une montre Hamilton made in USA !

La Réunion: deux jeunes randonneurs décédés en allant voir le volcan le Piton de la Fournaise

Les corps des deux victimes ont été retrouvés jeudi matin par le peloton de gendarmerie de haute montagne dans l'enclos, la caldeira centrale du volcan.

▶ REGARDER LA VIDÉO

La ministre du Logement et les HLM: faites ce que je dis, pas ce que je fais

Emmanuelle Wargon, qui prône la mixité sociale, a été raillée sur les réseaux sociaux parce qu'elle vit dans une ville cossue du Val-de-Marne.

→ Pierre Vermeren: «Grands ensembles et HLM se sont vidés de leurs populations initiales qui ont opté pour la France pavillonnaire»

→ «Nous devons être capable de parler de mixité sociale et d'origines sans se pincer le nez!»

personne, était accompagnée… d'une montre Hamilton made in USA !

La Réunion: deux jeunes randonneurs décédés en allant voir le volcan le Piton de la Fournaise

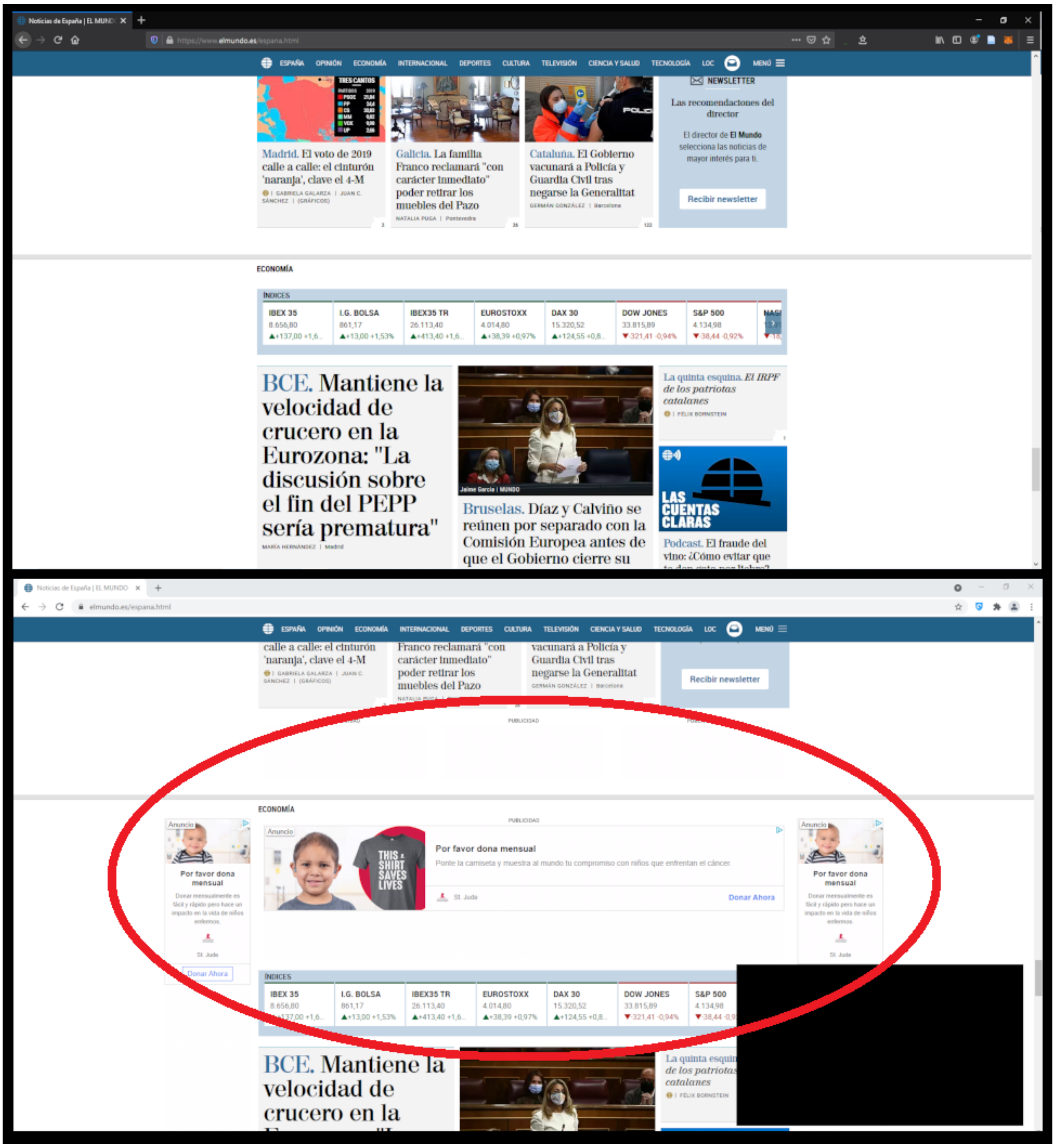
Les corps des deux victimes ont été retrouvés jeudi matin par le peloton de gendarmerie de haute montagne dans l'enclos, la caldeira centrale du volcan.

▶ REGARDER LA VIDÉO

La ministre du Logement et les HLM: faites ce que je dis, pas ce que je fais

Emmanuelle Wargon, qui prône la mixité sociale, a été raillée sur les réseaux sociaux parce qu'elle vit dans une ville cossue du Val-de-Marne.

Figure 9.5: Example 1. On top, page visited with no VPN extension. At the bottom, page visited with Hotspot VPN. Source: Original.



Madrid. El voto de 2019 calle a calle: el cinturón 'naranja', clave el 4-M

GABRIELA GALARZA | JUAN C. SÁNCHEZ | (GRÁFICOS)

Galicia. La familia Franco reclamará "con carácter inmediato" poder retirar los muebles del Pazo

NATALIA PUGA | Pontevedra

Cataluña. El Gobierno vacunará a Policía y Guardia Civil tras negarse la Generalitat

GERMÁN GONZÁLEZ | Barcelona

Las recomendaciones del director

El director de El Mundo selecciona las noticias de mayor interés para ti.

ÍNDICES							
IBEX 35	I.G. BOLSA	IBEX35 TR	EUROSTOXX	DAX 30	DOW JONES	S&P 500	NASI
8.656,80	861,17	26.113,40	4.014,80	15.320,52	33.815,89	4.134,98	
▲+137,00 +1,6..	▲+13,00 +1,53%	▲+413,40 +1,6..	▲+38,39 +0,97%	▲+124,55 +0,8..	▼-321,41 -0,94%	▼-38,44 -0,92%	▼-18

BCE. Mantiene la velocidad de crucero en la Eurozona: "La discusión sobre el fin del PEPP sería prematura"

MARÍA HERNÁNDEZ | Madrid

Bruselas. Díaz y Calviño se reúnen por separado con la Comisión Europea antes de que el Gobierno cierre su

La quinta esquina. *El IRPF de los patriotas catalanes*

FÉLIX BORNSTEIN

Podcast. El fraude del vino: ¿Cómo evitar que

Por favor dona mensual

Ponte la camiseta y muestra al mundo tu compromiso con niños que enfrentan el cáncer

Por favor dona mensual

Donar mensualmente es fácil y rápido pero hace un impacto en la vida de niños enfermos.

[redacted]

Figure 9.6: Example 2. On top, page visited with no VPN extension. At the bottom, page visited with Hotspot VPN. Source: Original.

9.4 Final remarks

To conclude with the result analysis I want to highlight the following points.

- A webpages resources vary to some degree on each visit by themselves. But the charts work on an average of all the visits, that are more than 100 per website.
- The data processor only isolates scripts present in every visit of a page. Scripts that are missing in at least one run are directly discarded.
- Scripts found must be present in the page or in the performance logs of the web browser. This means they are directly in the web page or loaded when it is visited.

The processing of the data was aiming for a strict approach to reduce the set of results as much as possible. During the testing of this project, I realized some pages do not load all their scripts in every visit. However, most of them appear in a great number of the visits you perform to them, even when you are not using a VPN. VPN extensions may use this behaviour to hide some of their tracking too, for example, not using the JavaScript injection in every site they visit, or not using it in every visit to a site. If the approach was not so strict, maybe more injections would appear. However, since the intention of this project was to prove a concept, this approach was good enough to find two extensions doing this type of tracking.

Some files isolated by the data processor were not useful. For example, a error response message to a request was collected in almost all pages. This happened because Google was not allowing access to a certain page or resource because the crawler was getting blacklisted in their robots.txt file. Situations like this made a requirement to check each file after they are selected by the data processor.

10. Conclusion

To conclude, I must say this project has been very challenging from beginning to end. Since the objective of the project was something new that had not been done before, finding answers to each question that raised was a true challenge. Many adjustments in focus have been made, but time and effort have put this project in the right direction.

The original objectives of the project where much more ambitious, but the underlying complexity of those made me focus on JavaScript Injection, that was in my opinion the most interesting part of this research work. To do this, I have had to study how domains track users by means of all the available tracking methods and learn to differentiate those from the ones a VPN extension could use.

Learning to compare files was also a great challenge. Thanks to the articles related to MinHash I could find a way to solve this problem, and using Python to develop this project made possible having a library that implemented this solution without excessive effort from my side.

Analyzing the results was by far the most complicated part, because it was at this point that all the data was put together and the errors were clearly shown. Some of the data was not healthy and was put a part. Adjusting the number of permutations the MinHash used for files and the comparison threshold the files used to determine similarity took a lot of tests.

However, I have to say that all the main objectives established at the beginning have been finally accomplished. I have developed a tool able to detect JavaScript injection, as it has been proved in the cases of Hola VPN and Hotspot VPN. The parameters of the tool could be further adjusted using more data and analyzing the results deeper, but the proof of concept works as imagined.

This project could be further expanded in different ways. Some of them are:

- **Expand web sites and VPNs:** As mentioned in the project description, this was a proof of concept. Now that this project has been finalized, it could be good to expand further the amount of sites visited and VPNs analyzed for wider and more trustworthy results.

- **Test with full VPNs:** Originally this project was meant to cover both browser VPN extensions and system-wide VPNs, but the later could not be tested due to time limitations. A good future project could be to apply the idea described in this paper to system-wide VPNs.

- **Improve the system:** The designed tool consist of two pieces that work separately. Each of them has certain limitations. For example, the Crawler can not resume broken crawls by himself. It requires human intervention to turn on the VPN extensions. The Data processor uses batch processing to get the desired results, but this process could be greatly enhanced. And the two modules could be put to work together.

- **Tweak the values used:** The tool implemented was using some configurable parameters, like the similarity threshold used to tell whether two files are the same or not and the number of permutations used to create MinHashes. This values could be further tested to obtain different results. The strict approach could also be adapted for a more permissive one, opening the path for a lot more resources to appear as isolated files by the data processor.

11. Appendix

11.1 Source code

11.1.1 hola files

hola_file01

- MD5 hash: 00308765366b9ea57d09ef5c8a6b0074
- Beautified Content:

```
<
script >
    function firePromoAnalytics(e) {
        try {
            var analytics = document.getElementById("pageAnalytics");
            if (analytics) {
                var bpt = window.ga_data.route.basePageType,
                    ssts = window.ga_data.route.ssts.replace('/',
                    ':'),
                    parent = getParentByTagName(e.target, 'A'),
                    index = e.target.dataset.index || parent.dataset
                    .index || 0,
                    localName = e.target.dataset.moduleName || parent
                    .dataset.moduleName || '';

                if (bpt === 'homefront') {
                    ssts = 'home'
                }

                if (index && parseInt(index)) {
                    index = parseInt(index);
                }

                var eventString = bpt + ':' + ssts + ':' + localName +
                    ':' + index;
                analytics.fireEvent(eventString, {}, true);
            } else {
                if (window.newrelic) window.newrelic.noticeError(
                    'page analytics tag not found');
            }
        } catch (ex) {
            if (window.newrelic) window.newrelic.noticeError(ex);
        }
    }

function getParentByTagName(el, tag) {
    while ((el = el.parentElement) && el.tagName !== tag);
    return el;
} <
/script>
```

11.1.2 hotspot files

hotspot_file01

- MD5 hash: 00308765366b9ea57d09ef5c8a6b0074
- Beautified Content:

```
< script >
    (function() {
        var attr, config, custDelimit, dataObj, libUtil, omitKeys,
            prefix, userKeys;
        libUtil = Krux('require:util.library-tag');
        dataObj = Krux('scrape.javascript', 'gciKruxData');
        userKeys =
            'userGender,userMeterState,userPersonalizedLinkId,userPersonalizedLinkSessionId,userStatus,userYe
        omitKeys = 'none';
        custDelimit = 'none';
        prefix = libUtil.resolvePrefix('text', 'web',
            'undefined');
        config = {
            'userKeys': userKeys ? userKeys.split(',') :
                void 0,
            'omitKeys': (omitKeys ? omitKeys.split(',') : [])
                .concat([/gtm\./,
                    /sessionid/i, /\.phpsessid$/i,
                    /\.sid$/i, /\.zenid$/i,
                    /\.requestid$/i
                ]),
            'omitValues': [/.*@.*(?:\..*)+/, /gtm\./,
                /^(https?:)?\/\/[^\/]+/,
                /^\/[^\/]+/, /.{255}/
            ],
            'customDelimited': custDelimit ? custDelimit
                .split(',') : void 0,
            'caseSensitive': 'web' === 'true',
            'useFullPath': 'false' === 'true',
            'useLastValue': 'false' === 'true',
            'optimizeNames': 'false' === 'true',
            'convertAttrNames': []
        };
        if (prefix) {
            config.convertAttrNames.push({
                pattern: /((?:page|user)_attr_)/,
                replacement: '$1' + prefix
            });
        }
        attr = Krux('ingestDataLayer', dataObj, config);
    }).call(); <
/script>
```

hotspot_file02

This file is not full in thsi document for length reasons. However, you may find the relevant parts of the file below. When code is removed from the file, the following string is used to indicate it: /* [...] */.

- MD5 hash: bd8f224a043d5da15665d4e3a88b97e2
- Beautified Content:

```
/*! amazon-dtb-javascript-api - apstag - v7.59.00 - 2021-01-25 10:26:47 */ !
/* [...] */
  function f() {
    var t = "amzn_lsTest";
    try {
      return window.localStorage.setItem(t, t), window.localStorage
        .removeItem(t), !0
    } catch (t) {
      return !1
    }
  }

  function b() {
    return document.cookie.split("; ").map((function(t) {
      return t.split("=")
    }))
  }

/* [...] */

  function m(t) {
    try {
      var e = t.innerWidth || t.document.documentElement.clientWidth || t
        .document.body.clientWidth,
        n = t.innerHeight || t.document.documentElement.clientHeight || t
        .document.body.clientHeight;
      return "".concat(e, "x").concat(n)
    } catch (t) {}
    return "x"
  }
```

```
/* [...] */

      displayAdServer: function(t, e) {
        switch (e.type) {
          case "SLOT_RENDER_ENDED_SET":
            return s(s({}, t), {}, {
              slotRenderEndedSet: !0
            });
          case "NO_BID_ON_ADSERVER_SLOTS":
            return s(s({}, t), {}, {
              noBidSlotIDs: t.noBidSlotIDs.concat(e.slotIDs)
            });
          case "REQUESTED_BID_FOR_ADSERVER_SLOTS":
            return s(s({}, t), {}, {
              noBidSlotIDs: t.noBidSlotIDs.filter((function(t) {
                return !Object(o.j)(e.slotIDs, t)
              }))
            });
          case "SHOULD_SAMPLE_SLOT_RENDER":
            return s(s({}, t), {}, {
              shouldSampleRender: e.value
            });
          default:
            return s(s({}, t), {}, {
              noBidSlotIDs: d(t.noBidSlotIDs)
            })
        }

/* [...] */

  var r, i, o = ["amznbid", "amzniid", "amznsz", "amznp"],
    c = ["amznbid", "amzniid", "amznp", "r_amznbid", "r_amzniid",
    "r_amznp"];
  (i = r = r || {}).new = "NEW", i.exposed = "EXPOSED", i.set = "SET", i
    .rendered = "RENDERED";
  var a, s, u, d = "apstagDebug",
    f = ["redux", "fake_bids", "verbose", "console", "console_v2",
    "errors"],
    l = "apstagDebugHeight",
    b = "apstagDEBUG",
    p = "apstagCfg",
    m = 0,
    g = 0;
  (s = a = a || {}).amznbid = "testBid", s.amzniid = "testImpression", s
    .amznp = "testP", s.crid =
    "testCrid", (u || (u = {})).video = "v";
  var y, h, O, j, v = ["amznbid", "amznp"];
  (h = y = y || {}).__apsid = "ck", h.__aps_id_p = "ckp", h.aps_ext_917 =
    "st", (j = O = O || {})
    .noRequest = "0", j.bidInFlight = "1", j.noBid = "2";
  var S = "600",
    _ = "7.59.00",
    w = "https://",
    D = "function" == typeof XMLHttpRequest && void 0 !== (
      new XMLHttpRequest).withCredentials,
    E = "apstagLOADED",
    T = 13,
    I = 1e4
},

/* [...] */

  function l(t) {
    var e = new Image;
    return e.src = t, f.push(e), e
  }!0 === Object(o.c)("exposePixels") && (window.apstagPixelQueue = u,
    window.apstagPixelsSent = f);
  var b, p = {
      adServer: [],
      ampAdContext: [],
      appended: [],
      AaxSlotSizes: [],
      bidRender: [],
      bidRenderState: [],
      bidType: [],
      "blockedBidders-fetchBids": [],
      "blockedBidders-init": [],
      ccpa: [],
      cmpVar: [],
      creativeSize: [],
      deals: [],
      fetchBids: [],
      fifFlow: [],
      gdpr: [],
      idRemap: [],
      iframe: [],
      renderFootprint: [],
      resizeIframe: [],
      schain: [],
      simplerGpt: [],
      slots: [],
      slotType: [],
      targeting: [],
      tcfVar: [],
      unusedDeal: [],
      useSafeFrames: []
    },
    m = [],
    g = !1;

/* [...] */

  function f(t) {
    var e = function() {
      if ("undefined" == typeof Reflect || !Reflect.construct) return !1;
      if (Reflect.construct.sham) return !1;
      if ("function" == typeof Proxy) return !0;
      try {
        return Date.prototype.toString.call(Reflect.construct(Date, [], (
          function() {}))), !0
      } catch (t) {
        return !1
      }
    }();
    return function() {
      var n, r, i, o = b(t);
      if (e) {
        var a = b(this).constructor;
        n = Reflect.construct(o, arguments, a)
      } else n = o.apply(this, arguments);
      return r = this, !(i = n) || "object" !== c(i) && "function" !=
        typeof i ? l(r) : i
    }
  }


  var m = function() {
      d(e, i.c);
      var t = f(e);

      function e(n) {
        var i;
        return a(this, e), p(l(i = t.call(this, n.targetId, Object(r.m)(n,
              "invCode") ? n.invCode :
            Object(r.m)(n, "tagId") ? n.tagId : n.targetId)), "rawSlot",
          void 0), p(l(i), "mediaType",
          "display"), i.rawSlot = n, i
      }

/* [...] */

      return u(e, [{
        key: "reportError",
        value: function(t, e) {
          Object(o.b)(t, "AppNexusAdServer-".concat(e))
        }
      },

/* [...] */

      }, {
        key: "setTargeting",
        value: function(t, e) {
          try {
            if (!Object(r.m)(window, "apntag") || !Object(r.m)(
                window.apntag, "requests")) return;
            Object(r.m)(window.apntag.requests, "keywords") || (
                window.apntag.requests
                .keywords = {}), window.apntag.requests.keywords[
              t] = e
          } catch (t) {
```

```
      this.reportError(t, "setTargeting")
     }
    }
   }, {
    key: "getTargeting",
    value: function(t) {
     try {
      if (!Object(r.m)(window, "apntag") || !Object(r.m)(
        window.apntag, "requests"))
       return [];
      Object(r.m)(window.apntag.requests, "keywords") || (
       window.apntag.requests
       .keywords = {});
      var e = window.apntag.requests.keywords[t];
      return void 0 === e ? [] : [e]
     } catch (t) {
      return this.reportError(t, "getTargeting"), []
     }
    }
   },

/* [...] */

   {
    key: "hasAdServerObjectLoaded",
    value: function() {
     try {
      return Object(r.m)(window, "apntag") && Object(r.m)(
        window.apntag, "loaded") && !0 ===
       window.apntag.loaded
     } catch (t) {
      return this.reportError(t, "hasAdServerObjectLoaded"), !
       1
     }
    }
   },

/* [...] */


   return t = c, (e = [{
    key: "reportError",
    value: function(t, e) {
     Object(o.b)(t, "SmartAdServer-".concat(e))
    }
   }, {
    key: "cmdQueuePush",
    value: function(t) {
     try {
      window.sas.cmd.push(t)
     } catch (t) {
      this.reportError(t, "cmdQueuePush")
     }
    }
   }, {
    key: "hasAdServerObjectLoaded",
    value: function() {
     try {
      return Object(r.m)(window, "sas") && Object(r.m)(window
        .sas, "__smartLoaded") && !
       0 === window.sas.__smartLoaded
     } catch (t) {
      return this.reportError(t, "hasAdServerObjectLoaded"), !
       1
     }
    }
   },

/* [...] */

 function o(t) {
  var e = [];
  try {
   t.hasAdServerObjectLoaded() && (e = t.getSlots())
  } catch (t) {
   Object(r.b)(t, "getDisplayAdServerSlots")
  }
  return e
 }

/* [...] */
```

```

        if (Object(y.m)(e.data, "renderData")) {
          var d = e.data.renderData,
            f = d.id;
          if (Object(y.m)(d, "renderStart") || Object(y.m)(d,
              "renderEnd")) {
            var l = d.renderStart,
              b = d.renderEnd;
            if (l) t.renderTimes[f] = l;
            else if (b && 0 !== t.renderTimes[f]) {
              var p = b - t.renderTimes[f],
                m = {
                  _type: "iframeRender",
                  c: "dtb",
                  pid: O.b,
                  crt: p
                };
              Object(j.b)(m)
            }
          }
        }
        if (Object(y.m)(e.data, "blockData")) {
          var g = e.data.blockData.blockInfo,
            h = {
              _type: "malwareBlock",
              c: "dtb",
              tpbr: 1,
              pid: O.b,
              info: g
            };
          Object(j.b)(h)
        }
      }
    }), !0)
  }
},

/* [...] */

        function D(t, e) {
          try {
            return Object(qt.k)(Lt.a.getState().targetingKeys[t]) ? e ? [
                "amzniid_sp"
              ] : Lt.a.getState()
              .targetingKeys[t].filter((function(t) {
                return -1 < t.indexOf("amzniid") && t.indexOf(
                  "amzniid_sp") < 0
              })) : ["amzniid"]
          } catch (t) {
            return Object(Xt.b)(t, "_getAllBidIdKeys"), []
          }
        }

          try {
            if (Object(Gt.c)("iframe", "friendly"), void 0 === t.html)
              throw new Error(
                "No HTML available for ad, most likely the creative has expired"
                );
            t = ne({
              hasRendered: !1,
              hasTimedOut: !1
            }, t), (e = N(t)).id = "apstag-f-iframe-".concat(Object(qt
              .e)()), B(t.doc, e, (
              function() {
                var t = r.bind(null, !1);
                try {
                  n = null !== e.contentDocument && Object(qt.j)([
                        "complete", "interactive"
                      ], e
                      .contentDocument.readyState) ? (t(),
                      "doc-ready") : null !== e
                    .contentDocument && "uninitialized" !== e
                    .contentDocument.readyState ? (e
                      .contentDocument.addEventListener(
                        "DOMContentLoaded", t), "dom-listener") : (e
                      .addEventListener("load", t), "iframe-listener"
                      ), Object(Gt.c)("fifFlow", n),
                    setTimeout(r, 1e3, !0)
                } catch (t) {
                  Object(Xt.b)(t,
                    "_loadAdIntoFriendlyIframe-setAttributes")
                }
              }))
          } catch (t) {
            Object(Xt.b)(t, "_loadAdIntoFriendlyIframe", !0)
          }
        }

/* [...] */
```

11.2 Tracking Techniques

In this section you may find some information regarding non-VPN related tracking techniques.

11.2.1 Tracking cookies

While using a VPN can hide yourIP address from the internet, cookies can still track your activity. Tracking cookies are not blocked by your VPN when browsing the internet, and they are not removed from your system afterwards either. However, your web browser can do that for you. Every web browser can block third-party cookies and remove all of them after the browsing session has finished (when you exit the browser), and if yours does not, you should think about changing your web browser. Sometimes this is as simple as browsing in incognito mode.

A VPN, however, can help in the sense that it will hide from cookies your real IP address, since the VPN server is the one connecting to the internet. Therefore, IP based identification will not work on you. But all the other information will be there: browsing history, queries performed, preferences...

11.2.2 Referrer URLs

A referrer URL is the web address of the previous website where you clicked a link to get to the current website. These can be used for several reasons, and recording your browser history is one of them. They are often used alongside with cookies, but these are not required in order to use referrer URLs. If someone clicks on a link to `example.org` at `example.com/links.htm`, then `example.org's` visitor log will show `example.com/links.htm` as his referral URL [16].

Some web browsers have an option not to send referral URL information to websites. Generally speaking, VPNs do not do that. Users may also do this manually by copying the link address and pasting it into the browser's address bar.

11.2.3 Web beacons

A web beacon is a technique used on web pages and email to unobtrusively (usually invisibly) allow checking that a user has accessed some content . The first web beacons were small digital image files that were embedded in a web page or email. The image could be as small as a single pixel, and could be of the same color as the background, or completely transparent (thus the name "tracking pixel"). When a user opens the page or email where such an image was embedded, they might not see the image, but their web browser or email reader would automatically download the image, requiring the user's computer to send a request to the host company's server, where the source image was stored. This request would provide identifying information about the computer, allowing the host to keep track of the user.

The identifying information provided by the user's computer typically includes its IP address, the time the request was made, the type of web browser or email reader that made the request, and the existence of cookies previously sent by the host server. The host server can store all of this information, and associate it with a session identifier or tracking token that uniquely marks the interaction.

11.2.4 Browser fingerprinting

A device fingerprint, machine fingerprint or browser fingerprint is information collected about a remote computing device for the purpose of identification. Fingerprints can be used to fully or partially identify individual users or devices even when cookies are turned off.

The attributes of the fingerprint depend on the web application, for each of them may use a different set of attributes. For reference, `amiunique.org` uses the following data to create their fingerprint:

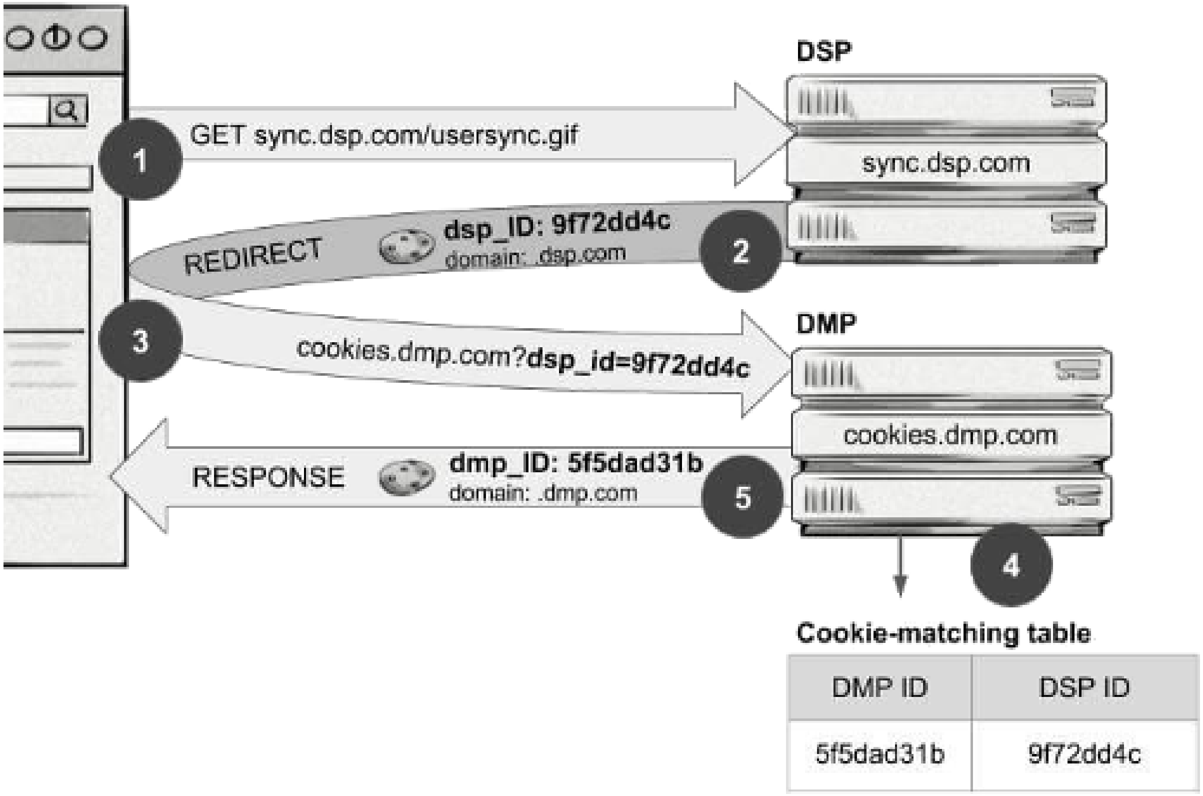
- the User agent header
- the Accept header
- the Connection header
- the Encoding header
- the Language header
- the Upgrade Insecure Requests header

- the Referer header
- the Cache-Control header
- the BuildId of the browser
- the list of plugins
- the platform
- the cookies preferences (allowed or not)
- the Do Not Track preferences (yes, no or not communicated)
- the timezone
- the screen resolution and its color depth
- the use of local storage
- the use of session storage
- a picture rendered with the HTML Canvas element
- a picture rendered with WebGL
- Supported Audio formats
- Supported Video formats
- the presence of AdBlock
- the list of fonts

Nor VPNs, nor most ad-blockers can protect users from this type of tracking. Some ad-blockers are working on blocking this type tracking by removing the HTML canvas element from a web page when certain conditions are met, for this is usually used for tracking purposes.

11.2.5 Cookie syncing

The biggest problem cookies face is that they can only be read on the domain that created them. This means that AdTech companies can't read cookies created by other AdTech platforms or by the website itself, essentially limiting their effectiveness for advertising purposes on other



DMP ID	DSP ID
5f5dad31b	9f72dd4c

Figure 11.1: Cookie syncing diagram. Source: Amazon.com

websites. This is faced by creating a unique identifier (UID) for each user and sharing it with other companies, like a translation table that gathers each companies ID of that user and translates it to a UID.

11.2.6 Supercookies

Supercookies are not stored in your computer. Instead, an ISP inserts a piece of information unique to a user's connection into the HTTP header. They are injected at the network level as Unique Identifier Headers (UIDH). This means clearing your browser data will not delete that cookie. Blockers can't block it either.

A zombie cookie remains intact as it hides outside of your browser's regular cookie storage. Zombie cookies target local storage, HTML5 storage, RGB color code values, Silverlight storage, and more. That's why they're known as zombie cookies. An advertiser must only find an existing cookie in one of those locations to resurrect the rest.

If an ISP decides to track you with supercookies, there is not really much you can do, except encrypting your traffic (so that at least, they do not know what you are doing). This can be done by using HTTPS-only websites, and by using a VPN. However, the last if the safer option.

Bibliography

[1] Inc 2020 GitHub. *Build software better, together.* en. Feb. 2008. URL: https://github.com (visited on 09/29/2020).

[2] *6 Best No Log VPNs for 2019 that Take Your Privacy Seriously.* en-US. Mar. 2019. URL: https://pixelprivacy.com/vpn/no-log-vpn/ (visited on 04/15/2021).

[3] *Active and Passive attacks in Information Security.* en-us. Section: Computer Networks. Sept. 2018. URL: https://www.geeksforgeeks.org/active-and-passive-attacks-in-information-security/ (visited on 04/15/2021).

[4] *Advanced Encryption Standard.* en. Page Version ID: 980596722. Sept. 2020. URL: https://en.wikipedia.org/w/index.php?title=Advanced_Encryption_Standard&oldid=980596722 (visited on 09/29/2020).

[5] beautifier. *Online JavaScript beautifier.* URL: https://beautifier.io/ (visited on 04/17/2021).

[6] Andrei Z. Broder. "Identifying and Filtering Near-Duplicate Documents". en. In: *Combinatorial Pattern Matching.* Ed. by Gerhard Goos et al. Vol. 1848. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–10. ISBN: 978-3-540-67633-1 978-3-540-45123-5. DOI: 10.1007/3-540-45123-4_1. URL: http://link.springer.com/10.1007/3-540-45123-4_1 (visited on 04/18/2021).

[7] Power Consumption Calculator. *Server Rack Power Consumption Calculator.* en-US. Section: blog. July 2019. URL: https://www.racksolutions.com/news/blog/server-rack-power-consumption-calculator/ (visited on 10/12/2020).

[8] *Computer network.* en. Page Version ID: 1017506648. Apr. 2021. URL: https://en.wikipedia.org/w/index.php?title=Computer_network&oldid=1017506648 (visited on 04/15/2021).

[9] *Declare permissions.* en. URL: https://developer.chrome.com/docs/extensions/mv2/declare_permissions/ (visited on 04/15/2021).

[10] *Deep packet inspection.* en. Page Version ID: 1015609991. Apr. 2021. URL: https://en.wikipedia.org/w/index.php?title=Deep_packet_inspection&oldid=1015609991 (visited on 04/15/2021).

[11] Vincent Driessen. *Using git-flow to automate your git branching workflow.* Aug. 2010. URL: https://jeffkreeftmeijer.com/git-flow/ (visited on 09/29/2020).

[12] *Estimating Similarity of Two or More Sets Snowflake Documentation.* URL: https://docs.snowflake.com/en/user-guide/querying-approximate-similarity.html (visited on 04/18/2021).

[13] *Five Eyes, Nine Eyes and 14 Eyes: VPNs & Jurisdiction.* en-US. Mar. 2018. URL: https://thebestvpn.com/5-9-14-eyes-countries/ (visited on 04/21/2021).

[14] git-scm.org. *Git.* Apr. 2005. URL: https://git-scm.com/ (visited on 09/29/2020).

[15] *Hola Better Internet Access censored sites.* en. URL: https://hola.org/ (visited on 04/17/2021).

[16] *How are Cookies Affected by a VPN?* en-US. Oct. 2018. URL: https://vpnpros.com/blog/how-cookies-affected-by-vpn/ (visited on 04/20/2021).

[17] *How to Use JavaScript Injections.* en. Oct. 2020. URL: https://www.wikihow.com/Use-JavaScript-Injections (visited on 04/15/2021).

[18] Muhammad Ikram et al. "An Analysis of the Privacy and Security Risks of Android VPN Permission-enabled Apps". en. In: *Proceedings of the 2016 ACM on Internet Measurement Conference - IMC '16.* Santa Monica, California, USA: ACM Press, 2016, pp. 349–364. ISBN: 978-1-4503-4526-2. DOI: 10.1145/2987443.2987471. URL: http://dl.acm.org/citation.cfm?doid=2987443.2987471 (visited on 09/29/2020).

[19] Spiceworks Inc. *Server power consumption.* en. URL: https://community.spiceworks.com/topic/293347-server-power-consumption (visited on 10/12/2020).

[20] Taiga Agile LLC. *Taiga.io.* Sept. 2016. URL: https://taiga.io/ (visited on 09/29/2020).

[21] *MinHash datasketch 1.0.0 documentation.* URL: http://ekzhu.com/datasketch/minhash.html (visited on 04/18/2021).

[22] Baiju Muthukadan. *Selenium with Python Selenium Python Bindings 2 documentation.* URL: https://selenium-python.readthedocs.io/ (visited on 04/17/2021).

[23] *Natural Language Processing Essay - Language on Study Boss.* en-us. Apr. 2019. URL: https://studyboss.com/essays/natural-language-processing.html (visited on 04/18/2021).

[24] Norton. *What is a VPN?* en. URL: https://us.norton.com/internetsecurity-privacy-what-is-a-vpn.html (visited on 04/15/2021).

[25] *permissions - Mozilla | MDN.* URL: https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/manifest.json/permissions (visited on 04/15/2021).

[26] Raspberry Pi. *power supply - How much energy does the Raspberry Pi consume in a day?* 2020. URL: https://raspberrypi.stackexchange.com/questions/5033/how-much-energy-does-the-raspberry-pi-consume-in-a-day (visited on 10/12/2020).

[27] Raspberry Pi. *RasPi power usage measurements ALL Models - Page 5 - Raspberry Pi Forums.* 2020. URL: https://www.raspberrypi.org/forums/viewtopic.php?f=63&t=6050&p=291334&hilit=watts+power#p425381 (visited on 10/12/2020).

[28] Raspberry Pi. *Teach, Learn, and Make with Raspberry Pi Raspberry Pi.* 2020. URL: https://www.raspberrypi.org/ (visited on 10/12/2020).

[29] Team Poppyseed. *TLS Interception and SSL Inspection ů TLSeminar.* Mar. 2017. URL: https://tlseminar.github.io/tls-interception/ (visited on 04/15/2021).

[30] *Prevent A URL Redirect Attack | Malicious Redirects | SiteLock.* en-US. Jan. 2020. URL: https://www.sitelock.com/blog/prevent-url-redirect-attacks/ (visited on 04/15/2021).

[31] *RSA (cryptosystem.* en. URL: https://en.wikipedia.org/wiki/RSA_(cryptosystem (visited on 09/29/2020).

[32] Scrum.org. *What is Scrum?* en. Mar. 2010. URL: https://www.scrum.org/resources/what-is-scrum (visited on 09/29/2020).

[33] *Server (computing).* en. Page Version ID: 1012771846. Mar. 2021. URL: https://en.wikipedia.org/w/index.php?title=Server_(computing)&oldid=1012771846 (visited on 04/15/2021).

[34] *SHA-1.* en. Page Version ID: 979702000. Sept. 2020. URL: https://en.wikipedia.org/w/index.php?title=SHA-1&oldid=979702000 (visited on 09/29/2020).

[35] *SHA-2*. en. Page Version ID: 977780884. Sept. 2020. URL: https://en.wikipedia.org/w/index.php?title=SHA-2&oldid=977780884 (visited on 09/29/2020).

[36] Slack. *El motor de tu trabajo*. es-ES. 2020. URL: https://slack.com/intl/es-es/ (visited on 09/29/2020).

[37] *The best and worst VPN extensions | Are VPN extensions secure?* en. URL: https://proprivacy.com/vpn/guides/vpn-extension (visited on 04/15/2021).

[38] EDINSOST UPC. *Cuestionario de Estudiantes de Ingeniería Informatica*. es. 2020. URL: goo.gl/kWLMLE (visited on 10/12/2020).

[39] FIB UPC. *Module 2.6 - SUSTAINABILITY ANALYSIS FOR THE BACHELORS THESIS*. en. 2020.

[40] *Virtual private network*. en. Page Version ID: 979378541. Sept. 2020. URL: https://en.wikipedia.org/w/index.php?title=Virtual_private_network&oldid=979378541 (visited on 09/29/2020).

[41] *Web crawler*. en. Page Version ID: 981476806. Oct. 2020. URL: https://en.wikipedia.org/w/index.php?title=Web_crawler&oldid=981476806 (visited on 10/12/2020).

[42] *Web crawler*. en. Page Version ID: 1018112513. Apr. 2021. URL: https://en.wikipedia.org/w/index.php?title=Web_crawler&oldid=1018112513 (visited on 04/18/2021).