The Torprinter project - GSoC'16 proposal

Pierre Laperdrix
https://plaperdr.github.io/
2nd year PhD Student
INRIA - Rennes, France

Summary - The Torprinter project: a browser fingerprinting website to improve Tor fingerprinting defences

The capabilities of browser fingerprinting as a tool to track users online has been demonstrated by Panopticlick and other research papers since 2010. The Tor community is fully aware of the problem and the Tor browser has been modified to follow the "one fingerprint for all" approach. Spoofing HTTP headers, removing plugins, including bundled fonts, preventing canvas image extraction: these are a few examples of the progress made by Tor developers to protect their users against such threat. However, due to the constant evolution of the web and its underlying technologies, it has become a true challenge to always stay ahead of the latest fingerprinting techniques.

I'm deeply interested in privacy and I've been studying browser fingerprinting for the past 2 years. I've launched 18 months ago the AmIUnique.org website to investigate the latest fingerprinting techniques. Collecting data on thousands of devices is one of the keys to understand and counter the fingerprinting problem.

For this Google Summer of Code project, I propose to develop the Torprinter website that will run a fingerprinting test suite and collect data from Tor browsers to help developers design and test new defences against browser fingerprinting. The website will be similar to AmIUnique or Panopticlick for users where they will get a complete summary with statistics after the test suite has been executed. It can be used to test new fingerprinting protection as well as making sure that fingerprinting-related bugs were correctly fixed with specific regression tests. The expected long-term impact of this project is to reduce the differences between Tor users and reinforce their privacy and anonymity online. In a second step, the website could open its doors to more browsers like Firefox or Chrome so that it could become a platform where vendors can implement significant changes in their browsers with regards to privacy and see the impact first-hand on the website. As one of its developer told me, Tails would also benefit from it to make sure that there are no differences between the fingerprints of Tor Browser inside and outside of Tails since extra protection are added to the Tor browser in Tails (link n°1 and n°2 on the issue with one example of a difference).

With the strong expertise I have acquired on the fingerprinting subject and the experience I have gained by developing the AmIUnique website, I believe I'm fully qualified to see such a project through to completion.

Website features

The main feature of the website is to collect a set of fingerprintable attributes on the client and calculate the distribution of values for each attribute like Panopticlick or AmIUnique. The set of tests would not only include known fingerprinting techniques but also ones developed specifically for the Tor browser.

The second main feature of the website would be for Tor users to check how close their current fingerprint is from “acceptable” fingerprints that most users should share. A list of actions should be added to help users configure their browser to reach one of these fingerprints.

The third main feature would be an API for automated tests as detailed by this page. This would enable automatic verification of Tor protection features with regard to fingerprinting. When a new version is released, the output of specific tests will be verified to check for any evolution/changes/regressions from previous versions.

The fourth main feature I'd like to include is a complete stats page where the user can go through every attribute and filter by OS, browser version and more.

The inclusion of additional features that go beyond the core functionalities of the site should be driven by the needs of the developers and the Tor community.

After discussion with the mentors on the “tor-dev” mailing list, additional questions have been answered on the project:

- The website should indicate the exact list of tests in the test suite. Every single Tor user should know in advance what test would be running in their browser.
- Anyone should have access to aggregate statistics but the exact values of every single fingerprint should be kept to a specific list of authorized users.
- When a new version is released, data collected from previous versions will be kept to study the impact of a proposed defense. Moreover, it could give an insight into how fast users are updating their browsers.
- New tests must be reviewed before being added on the website.
- The website should be accessible to every browser and not only to Tor users since we plan on opening it for every browser vendors.

Technical choices

In my opinion, the website must be accessible and modular. It should have the ability to cope with an important number of connections and be accessible to as many Tor developers as possible. With this in mind, I plan on using Django with a MongoDB database. Developing

the website in Python opens the door to many Tor developers and will prove to be better in the long run for its maintenance since a lot of scripting in the Tor community is done in this language. On the storage and statistics side, MongoDB is a good fit because it is now a mature technology that can scale well with an important number of data and connections. Moreover, the use of SQL databases for AmIUnique proved to be really powerful but the maintenance after the website was launched became a tedious task, especially when modifying the underlying model of a fingerprint to collect new attributes. The choice of a more flexible and modular database seems a better choice for maintenance and for adding/removing tests.
Finally, the use of Django and MongoDB complies with Tor guidelines so that the Torprinter website can be run on the Tor infrastructure (both Django and Mongo have stable Debian packages).

Estimated timeline
You will find below a rough estimate of the timeline for the three months of the GSoC.

Community bonding period : Discuss with the mentors and the community the set of features that should be included in the very first version of the website and clarify remaining open questions.

23 May - 27 June : Development of the first version of the website with the core features
Week 1 - Development of the first version of the fingerprinting script with the core set of attributes. Special attention will be given so that it is fully compatible with the most recent version of the Tor browser (and older ones too).
Week 2 - Start developing the front-end and the back-end to store fingerprints with a page containing data on your current fingerprint (try adding a view to see how close/far you are from an acceptable fingerprint).
Week 3 - Start developing the statistics page with the necessary visualization for the users. Modification of the back-end to improve statistics computation to lessen the server load.
Week 4 - Finishing the front-end development and refining the statistics page to get back the most relevant information.
Adding and testing an API to support automated tests.
Week 5 - Finishing the first version so that it is ready for deployment.
Start developing additional features requested by the community (rest API? account management?)

27 June - Mid July :

Deployment of the first version online for a beta-test with bug fixing.
Finishing development of additional features requested by the mentors/community.
Defining the list of new features for the second version.

Mid July - 23th August :
Adding a system to make the website as flexible as possible to add/remove tests easily (A pull-request system? A test submission form where admins review tests before they are included in the test suite?)
Developing additional features for the website.
Making sure that the website can be opened to more browsers (work done at design time to support any browsers will be tested here)
Bug fixing

Code sample
In 2014, I developed the entire AmIUnique.org website from scratch. Its aim is to collect fingerprints to study the current diversity of fingerprints on the Internet while providing full details to users on this subject. It was the first time that I built a complete website from the design phase to its deployment online.
One of the first challenge that I encountered was to build a script that would not only use state-of-the-art techniques but that could simply work on the widest variety of browsers. Testing a script for a recent version of a major browser like Chrome and Firefox is an easy task since they implement the latest HTML and JavaScript technologies but making sure that the script runs correctly on older browsers like Internet Explorer is another story. Juggling with a dozen different virtual machines was necessary to obtain a bug-free and stable version of the script. A small beta-test was required to make sure that everything was good to go for what is now the foundations of the AmIUnique website. The totality of the source code for AmIUnique and my other projects can be found on GitHub.
A second challenge that I faced was to deal with the increasing load of users so that the server could return personalized statistics to visitors in a timely manner (less than 2/3s). By having a separate entity that updates statistics in real time on top of the database, I managed to drastically reduce the server load. With the number of Tor users around the world, the website needs from the get go to handle a high load of visitors and statistics computation and my previous experience on that specific task will prove useful.

For the very first version of Torprinter, I plan on testing well-known and widespread fingerprinting techniques to make sure that there is no variation among Tor users. These include HTTP headers and known JavaScript objects. There should be no need for any

Flash attributes since plugins are not present in the Tor browser (thus removing complex code in charge of correctly loading the Flash object).
For this proposal, I have also developed a special page with 7 different tests that are mainly targeted at the Tor browser to give an idea of what tests can be included that are more suited to the Tor users.
Tests n°5, n°6 and n°7 are broader and also concerns the Firefox browser.
You can found a working version of the script on a special webpage (need to scroll to make the results appear): https://plaperdr.github.io/torScript.html
The script can be found here: https://plaperdr.github.io/assets/tor/tor.js

Test n°1
Test the size of the current window - As reported by ticket n°14098
https://trac.torproject.org/projects/tor/ticket/14098
Test n°2
Test the support of emoji - As reported by ticket n°18172
https://trac.torproject.org/projects/tor/ticket/18172
Test n°3
Analysis of the "scroll" behavior of the window - As investigated by
http://jcarlosnorte.com/security/2016/03/06/advanced-tor-browser-fingerprinting.html
Test n°4
Test the size of current fallback font by using the canvas API to render some text (no need for user permission like canvas extraction) - Custom test
Test n°5
Test the difference between OS on the maximum font size - Custom test
Test n°6
Test the difference between OS on the Date API - As reported by ticket n°15473
https://trac.torproject.org/projects/tor/ticket/15473
Test n°7
Test the difference between OS on the Math class - As reported by ticket n° 13018
https://trac.torproject.org/projects/tor/ticket/13018

Why the Tor project?
I want to work with the Tor project because I firmly believe that everyone should be able to protect and control their online identity and have strong privacy online. Tools like Tor are really important steps towards that goal and building a website to improve the Tor browser fingerprinting defences is my own way of using my knowledge on the fingerprinting domain to contribute to the Tor community.

Experience working on projects
In my engineering school, we worked on projects in groups that would last between 2 to 10 months. More details on these projects can be found on my personal webpage.
Everything that I'm developing now is open-source and can be found on GitHub and all contributions are welcome.

Will you be working full-time on the project for the summer, or will you have other commitments too (a second job, classes, etc)?
I'll be working full-time during the summer. (I have a commitment at the very beginning of the coding period where I have to present my second publication on my findings from the AmIUnique dataset at the S&P conference in San Jose.)

Will your project need more work and/or maintenance after the summer ends? What are the chances you will stick around and help out with that and other related projects?
I expect the core set of features to be done before the end of the summer. Work after the summer ends would mainly be to add extra features, to facilitate maintenance or add new ways to visualize data. I'm also aiming to make the site as accessible as possible so that the community can contribute to it as much as possible.

What is your ideal approach to keeping everybody informed of your progress, problems, and questions over the course of the project? Said another way, how much of a "manager" will you need your mentor to be?
I plan to post on my blog a progress every week of my work and I'll keep the mentors updated every step of the way. The mentor would mainly be here as an advisor to push the website in directions that would be more suited for the Tor community and for the developers in terms of features (See open questions in the project section for details on that).

What school are you attending? What year are you, and what's your major/degree/focus? If you're part of a research group, which one?
I'm currently in my second year as a PhD student in the DiverSE team at INRIA Rennes. My subject is to use diversity and software engineering techniques to design solutions against browser fingerprinting.

How can we contact you to ask you further questions?
You can contact me by mail at: pierre.laperdrix@irisa.fr.
My nickname on IRC is “SuperOctupus”.

Are you applying to other projects for GSoC and, if so, what would be your preference if you're accepted to both?
It's the only GSoC project I'm applying for.

Is there anything else that we should know that will make us like your project more?
I truly believe that this project can help the Tor browser and Tor users around the world to be better protected against unwanted tracking on the Internet. I also believe that my strong expertise in browser fingerprinting and in web development make me an ideal candidate to provide the Tor community with a strong and relevant contribution
Ticket n°6119 has been opened for more than four years now. I think it is time to close it at the end of the summer.