Mémoire de Stage de 4e année

FL-Minifer

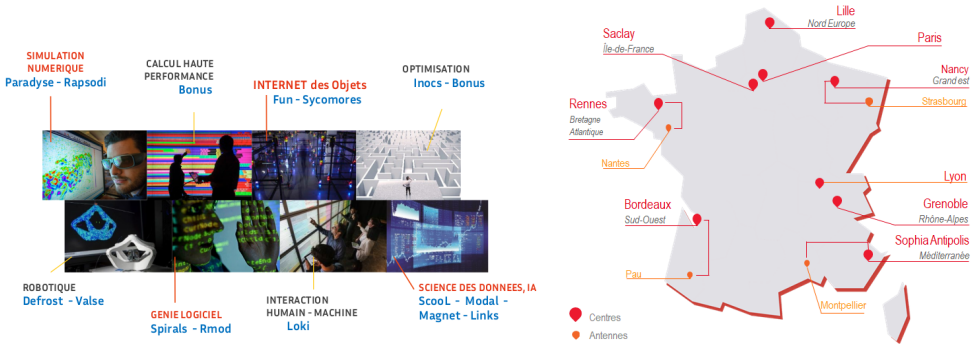
A Tool To Minify and Unify AdBlocker's Filter Lists

Maxence NEUS

Polytech Lille	**Inria**
Secrétariat - Bureau Boulevard Paul Langevin - Cité Scientifique 59655 VILLENEUVE D'ASCQ CEDEX 03-28-76-73-60 03-28-76-73-61	40 Av. Halley 59650 Villeneuve d'Ascq Tuteur Entreprise : **Walter RUDAMETKIN** Tuteur Ecole
	**Walter RUDAMETKIN**

2022

Contents



1 Présentation de l'entreprise

1.1 Inria

Le centre de recherche Inria fondé en 1967 est un établissement publique de recherche se spécialisant dans le domaine des mathématiques et de l'informatique. Le centre de Lille Nord Europe fait partie des 9 centres autour de la France et comporte deux complexes: à la Haute Borne et à Euratechnologie. Au sein du centre dirigé par Mireille RÉGNIER operent 15 équipes dont mon équipe d'acceuil: SPIRALS.

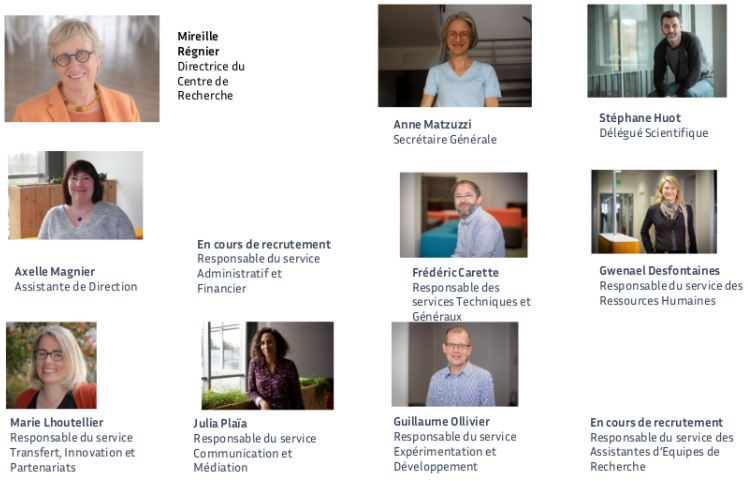


Figure 1: Équipe de direction

1.2 SPIRALS

est une équipe jointe de l'INRIA et de CRISTAL qui se focalise sur les systèmes distribués et l'ingéniérie système. les projets de l'équipe comportent l'étude de solutions autonomes, efficaces et adaptatives pour la récupération et le traitement de données. Dirigée par Lionel Seinturier, l'équipe est actuellement composée de 11 chercheurs permanents, 7 postdocs, 17 doctorants et 7 ingénieurs.

2 Contexte du projet

2.1 Introduction aux adBlockers

Les cookies: Le modèle économique du web est centré autour des données utilisateur et de la façon dont elles sont utilisées, majoritairement pour cibler au mieux des publicités qui pourront alors faire un maximum de convertions (nombres d'achats par publicité imprimée sur une page). Evidement pour cibler efficacement un utilisateur il est nécessaire de pouvoir associer un ensemble de données à un utilisateur, historiquement les cookies ont été l'outil le plus répendu pour remplir cette fonction: en incorporant un identifiant unique à chaque utilisateur dans ses requêtes HTML.

Malheureusement pour les géants du web, l'idée que quelques entreprises puissent avoir accès à une quantité alarmante de leurs données simplement par leur navigation sur le web ne plait pas vraiment aux utilisateurs. Suite à des campagnes de sensibilisation de plus en plus d'utilisateurs du web se sont retrouvés confrontés à cette réalité et désirent autant que possible limiter l'accès à leurs données de navigation. Pour cela plusieurs compagnies se sont misent à créer des extensions pour navigateurs qui permettent de bloquer les dis "tracking cookies". En effet tous les cookies n'ont pas pour but de traquer l'utilisateur, la technologie a été créée pour garder des informations utilisateurs essentielles à la fonctionnalité du site comme rester connecter à son compte utilisateur ou sauvegarder l'état du panier sur les sites de shopping en ligne. Ces cookies dits "essentiels" doivent être filtrés pour garder les fonctionnalités du site, une méthode largement utilisée est de bloquer les cookies dits "tierces" qui proviennent de domaines différents de celui du site visité étant donné que ceux ci sont bien souvent utilisés uniquement comme tracking cookies et peuvent donc être bloqués sans crainte de nuire aux fonctionnalités du site.

La publicité sur le web: Comme dit précédement, le modèle économique du web consiste à montrer le plus de publicités intéressantes possible à l'utilisateur pour qu'il achète le plus de produits possible. Ce modèle entraîne logiquement un comportement où les sites utilisent chaque pixel d'espace libre de contenu sur leurs sites pour y servir des publicités grâce à un modèle d'enchères où les distributeurs de publicités enchèrissent le plus sur les utilisateurs avec un intérêt probable pour leurs produits.

Les utilisateurs veulent donc se débarasser un maximum des publicités à la fois pour leur confort mais aussi pour protéger les plus vulnérables des pratiques prédatoriales des entreprises. C'est là l'utilité des adBlockers qui

scannent les éléments de la page et bloquent ceux qui sont détectés comme étant des publicités.

Les rêgles de filtrage: Pour détecter les élements qui font parties de publicités, les adBlockers utilisent des listes de filtres qui contiennent des caractèristiques d'élements publicitaires rencontrés sur le web. Ces listes sont maintenus régulièrement par la communauté et par des entreprises d'adBlockers pour parer aux nouvelles tentatives de contournement des distributeurs de publicités. Les listes sont de simples fichiers texte dont la syntaxe de base est définie par AdBlockPlus (ABP) ici et elle est étendue par uBlockOrigin (uBO) ici. La syntaxe comporte deux grands types de rêgles:

- Les *Network Rules*
- Les *Cosmetic Rules*

Le cas le plus simple, les *Network Rules* correspondent à des rêgles qui bloquent les éléments selon l'url dont ils sont importés, ce grâce à une syntaxe similaire à des expressions réguliaires. C'est à dire que la rêgle décrits un modèle d'url qui correspond à un ensemble d'url qui devront être bloquées, par exemple "http://example.com/ads/banner*.gif" bloquera les élements venant de la source "http://example.com/ads/banner420.gif" (le caractère * correspondant à n'importe quelle suite de caractère) mais pas "http://example.fr/ads/ad123.png".

Plus puissantes, les *Cosmetic Rules* bloquent des élements dont le css correspond à une certaine description, par exemple la rêgle "###ad-boxes" bloque les élements sur la page qui ont une id css de "ad-boxes".

Il est également possible d'écrire des exceptions qui permettent de laisser passer des élements qui seraient bloqués par une rêgle trop générale et qui nuirait aux fonctionnalités du site.

2.2 Présentation AmIUnique

Le Fingerprinting sur le web Comme décrit en 2.1 les cookies ont longtemps été l'outil principal pour suivre les utilisateurs sur le web, mais les efforts des adBlockers et des navigateurs comme Firefox ou Brave ont rendus leurs utilisation plus difficile ou moins efficaces qu'avant, et avec la demande pour les données utilisateur encore grimpante, de nouveaux outils plus avancés sont développés pour continuer à suivre les utilisateurs sur le web.

Un outil qui remplis cette fonction peut-être mieux encore que les cookies le faisaient auparavant est le **Browser Fingerprinting** ([1]) le principe derrière celui-ci consiste à appliquer une approche inspirée de la *data science* et d'obtenir un maximum de bribes d'information sur l'utilisateur (ou ici plutôt son navigateur) afin de les agréger en un identificateur unique permettant de suivre l'utilisateur grâce à la configuration de son navigateur que les cookies soit acceptés ou non.

AmIUnique est un projet d'Inria Lille qui consiste à étudier les caractèristiques du navigateur qui permettent de l'identifier et de faire des statistiques sur les changements de l'identificateur au cours du temps grâce à une extension.

3 Corps du projet

4 Extras

References

[1] Benoit Baudry Pierre Laperdrix Walter Rudametkin. "Beauty and the Beast: Diverting modern web browsers to build unique browser fingerprints". In: *IEEE Symposium on Security and Privacy* (2016).

Résumé

Lors de ce stage blablabla

Mots clés:

Abstract

During this internship blablabla

Keywords: