Avancement mémoire

Antoine Marchal

20 mars 2014

Plan

Présentation brève

La première étape de ce mémoire était d'identifier les moyens qui sont utilisés pour nous tracer sur Internet.
Il se focalise sur le traçage au niveau du client (navigateur web) via notamment :

- Les cookies web
- Les scripts JavaScript
- Les images chargées depuis des sites tiers
- Flash
- ...

Quels sites nous tracent ?

Afin de se rendre compte de l'étendue de ce traçage, l'élaboration d'un outil était nécessaire.

- Il est réalisé en Java
- Il se compose de deux modules principaux et indépendants :
  - Crawler
  - Parser

Crawler

- Il utilise Selenium afin de lancer une instance de Firefox
- Cette instance de Firefox est munie de deux extensions :
  - Firebug : récupère tous les éléments chargés sur la page
  - NetExport (extension de Firebug) : exporte au format HAR

Crawler : problèmes rencontrés

- Selenium ne permettait pas de récupérer tous les éléments
  => J'ai alors mis en place un proxy
- Cependant, le proxy ne permettait pas d'exécuter du JavaScript et beaucoup de sites n'étaient alors pas traités correctement (entre 20 et 30% des sites sur le TOP 100)
  => J'ai finalement décidé de supprimer le proxy et d'utiliser les extensions Firefox citées précédemment

Crawler : en pratique

J'ai lancé le crawler sur le TOP 1000 et cela a pris environ 5h

- Sur une machine de la salle Intel
- Avec Xvfb (émulation de X)
- Une seule remarque : le stockage des fichiers occupe beaucoup d'espace disque

Parser

Le parser permet de traiter les fichiers HAR et d'identifier les sites qui utilisent (potentiellement ou pas) des trackers.
Il y a des aspects différents :

- Identifier les éléments chargés par le site et comparer avec une base de données de trackers (Ghostery*) => fiable
- Déterminer si les éléments chargés par le site sont stockés sur des serveurs tiers (utilisation des DNS) => moins fiable
- Identifier (provenance) les cookies créés sur le machine

*J'ai demandé à Ghostery la permission d'utiliser leur base de données de trackers.

Parser : problèmes rencontrés

- Certains sites hébergent leur contenu sur un autre domaine
  => La solution est alors d'utiliser les DNS mais cela peut entraîner des faux positifs

Parser : en pratique

J'ai lancé le parser sur le TOP 200 et cela a pris moins de 5 min

- Sur laptop personnel
- Au niveau de l'implémentation, seule la vérification par rapport à la base de données de trackers est terminée

Suite...

- Installer des extensions censées protéger la vie privée et relancer le programme sur cette instance particulière de Firefox
- Regarder si le nombre de trackers varie par pays (TOP 100 de pays différents)
- Déterminer le pourcentage de sites se conformant au DNT
- Récupérer et identifier les cookies Flash
- ...

Démonstration