TD
M2 MIAGE
PRM2
Big Data

Fichier : **TD_02_PRM2_Nov-2016**.docx

Date **Novembre 2016**

Rédacteur : DCN

Équipe d'étudiant(e)s

-
-
-
-
-
-

Diffusion

- Étudiants M2 Miage

Big Data

Désanonymisation des navigateurs à l'aide d'une DMP

Didier Certain Professeur associé
☎02 99 84 74 57 ☎06 08 63 63 28
e-mail : didier.certain@univ-rennes1.fr

1- Contexte

Profil numérique, empreinte numérique, identité numérique

Tout internaute qui utilise un navigateur web peut être caractérisé par le profil numérique de son navigateur web. On peut même envisager d'identifier, de manière unique, un internaute en se basant sur le profil numérique de son (ou de ses) navigateur(s).

Ainsi, d'après l'étude *Définition d'un profil par ses données de recherches* [TAMI2007], il est possible de déterminer un utilisateur de façon unique en fonction de ses centres d'intérêt, et donc, en fonction de ses recherches sur le Web et de leur évolution au cours du temps. Or ces informations sont fréquemment présentes sur les navigateurs web (notamment parce que les navigateurs mémorisent de nombreux paramètres liés au comportement de navigation). Et, quand ces informations ne sont pas présentes, on peut les reconstituer à partir d'autres données présentes sur les navigateurs web.

De même, une autre étude [HAYE2014] montre qu'une personne peut être identifiée par la liste des sites qu'elle a visités. Or ces informations sont fréquemment présentes sur les navigateurs web. Ces derniers gèrent, en effet, un historique qui n'est nettoyé qu'à la demande exprès de l'internaute.

Si l'on suit ce raisonnement, on devrait pouvoir identifier un internaute à partir de son environnement et donc de ses relations ou des modifications de ses relations au sens large. Ainsi, on pourrait donc identifier un internaute par son empreinte numérique laissée sur Internet. Le mot « empreinte » signifie ici le résumé concis de ses traces numériques. C'est, en quelque sorte, la signature qui valide l'identité de l'internaute. L'identité est définie comme un ensemble de caractéristiques et la signature est le résumé concis de cet ensemble de caractéristiques.

Aux États-Unis, par exemple, il existe des services d'analyse comportementale tels que DQE SOFTWARE[1] ou d'analyse de compilation de donnés comme SIGNIFYD[2].

Actuellement, en France, il n'existe pas de service de vérification de l'identité d'un internaute. C'est une bonne chose pour les libertés individuelles. Mais c'est également un problème lorsqu'il s'agit de lutter contre les tentatives de fraude.

Pourtant, quand on étudie des services existants, comme celui de Kount [3] aux USA, on découvre un service qui enrichit les données avec des métadonnées, soit pour une future analyse, soit pour une analyse en temps réel. Et on perçoit vite l'utilité de ce type de service pour les commerces en ligne.

Empreinte du navigateur (*browser fingerprint*)

Plusieurs études sur la traçabilité des internautes montrent que l'on peut suivre les traces des internautes non seulement en utilisant des « *cookies* »[4], mais aussi de cookies cachés ou « *evercookies* » [NYTI2010], [KAMK2010].

Aujourd'hui, la technique de traçabilité la plus répandue est l'empreinte du navigateur ou « *browser fingerprint* », Cette technique est décrite par *Electronic Frontier Foundation* dans un article de vulgarisation pour le grand public [ECKE2010].

La technique du *browser fingerprint* permet d'identifier un utilisateur unique à travers les caractéristiques de son navigateur web parce qu'**on suppose qu'il y a égalité entre internaute et navigateur**. C'est évidemment une approximation puisque, d'une part, il y peut y avoir plusieurs internautes qui utilisent le même navigateur d'un même microordinateur et, d'autre part, un internaute particulier peut utiliser plusieurs navigateurs distincts sur un microordinateur (ou sur plusieurs microordinateurs).

[1] DQE Software est un éditeur de solutions spécialisées dans l'optimisation de la qualité des données : http://www.dqe-software.com/

[2] SIGNIFYD aide les e-commerçants à vendre leurs produits et services en toute confiance en les protégeant contre la fraude : http://www.signifyd.com/

[3] Kount offre des solutions pour la gestion de la fraude et du risque : http://www.kount.com/

[4] Le *cookie* est l'équivalent d'un petit fichier texte stocké sur le navigateur de l'internaute.

En fonction de plusieurs données techniques transmises par le navigateur web (données initialement prévues pour le bon fonctionnement des sites web mais qui sont, ici, détournées de leur usage initial) et grâce à quelques scripts non intrusifs, il est possible de déterminer, de façon relativement simple, l'identité numérique d'un internaute.

Comme le montre le graphique ci-après, tiré de l'étude d'*Electronic Frontier Foundation* [ECKE2010], sur 409 296 *browser fingerprints*, 83,6 % sont uniques et 8,2 % sont présents entre 2 et 9 fois.

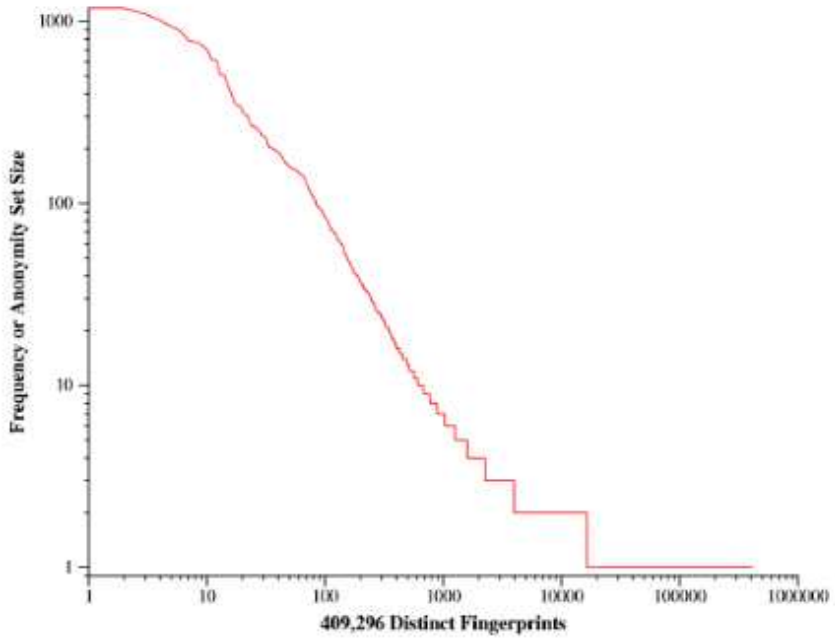

Fig. 1 – Répartition des fingerprints multiples (à gauche) et uniques (à droite)

Les *browsers fingerprints* peuvent varier dans le temps. Mais cette variation dépend du style des internautes. Un internaute technophile aura des *browsers fingerprints* qui varient plus rapidement que les *browsers fingerprints* des internautes technophobes. Néanmoins, malgré les variations possibles, l'empreinte du navigateur permet de suivre la trace d'un internaute dans le temps, pendant un certain temps.

L'étude [ECKE2010] estime la volatilité du *device fingerprint*, En 5 jours, 50 % des utilisateurs ont changé d'empreinte. À plus long terme, la tendance montre que 75 % des utilisateurs changent d'empreinte en 15 jours.

Cependant, nous pouvons également noter que, dans 65 % des cas, il est possible grâce à un algorithme assez simple de retrouver les *fingerprints* avec un taux de succès de 99,1 %.

Pour aller plus loin que le simple *browser fingerprint*, plusieurs chercheurs [ACAR2014] ont évalué récemment, l'ensemble des techniques existantes et leurs utilisations : *browser fingerprint*, *respawning* (résurrection de *cookies* à l'aide du langage *Flash* et de *cookies* dédiés) et *evercookies* (technique consistant à stocker de 13 façons différentes les *cookies* afin qu'ils ne soient jamais effacés).

Comme le démontrent ces chercheurs, ces techniques sont répandues sur les sites web destinés au grand public. Par ailleurs, une autre étude [HYLL2013] a démontré qu'il est difficile pour un utilisateur néophyte de passer à travers les mailles du *browser fingerprint*. En revanche, cela reste possible pour un utilisateur averti (disons un *geek*) comme le montre les études [FELC2011], [BIRI2011] et [LAPE2015].

Bibliographie - webographie :

[TAMI2007] Lynda Tamine, Nesrine Zemirli, Wahiba Bahsoun « Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information ». Information - Interaction -Intelligence, Éditions CEPADUES (2007) 7 (1), pp.5-25

[ECKE2010] Peter Eckersley "A Primer on Information Theory and Privacy" Electronic Frontier Fondation (janvier 2010)
https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

[NYTI2010] Article du New York Times sur la vulgarisation du « evercookie » : http://www.nytimes.com/2010/10/11/business/media/11privacy.html?hp&_r=0

[KAMK2010] + Article de synthèse sur le site Samy Kamkar : http://samy.pl/evercookie

[FELC2011] Guy de Felcourt « L'usurpation d'identité ou l'art de la fraude sur les données personnelles» - Collection Arès – Edition du CNRS (sept. 2011) - ISBN : 978-2-271-07243-6

[BIRI2011] Aroua Biri « Proposition de nouveaux mécanismes de protection contre l'usurpation d'identité pour les fournisseurs de services Internet » Thèse soutenue à l'INT (Institut National des Télécommunications) d'Évry, France (février 2011)

[HYLL2013] Corey Hyllested and Deb Linton "Device Fingerprinting: Opportunities for FTC Involvement" (Décembre 2013) http://people.ischool.berkeley.edu/~deb/portfolio/img/FTC-Final-Hyllested-Linton-Fall13.pdf

[ACAR2014] Gunes Acar, Christian Eubank, Steven Englehardt, , Marc Juarez, Arvind Narayanan, Claudia Diaz "The Web Never Forgets:Persistent Tracking Mechanisms in the Wild" https://securehomes.esat.kuleuven.be/~gacar/persistent/the_web_never_forgets.pdf

[HAYE2014] Brian Hayes "Uniquely Me! How much information does it take to single out one person among billions?" American Scientist (2014) http://www.americanscientist.org/libraries/documents/20142614253010209-2014-03CompSciHayes.pdf

[LAPE2015] Pierre Laperdrix (INSA-Rennes & INRIA Renne, France), Walter Rudametkin (University of Lille & INRIA Lille, France), Benoit Baudry (INRIA Rennes, France) « Mitigating browser fingerprint tracking: multi-level reconfiguration and diversification » (2015) http://diversify-project.eu/papers/laperdrix15.pdf

2- Problème à traiter

Shop4gift est une entreprise de distribution B2C. Shop4gift a l'ambition de conquérir la population mondiale. Et vous travaillez dans cet objectif chez Shop4gift. À cet effet, vous avez décidé de développer et d'administrer une base de données Relation-Clients qui sera capable de reconnaître les navigateurs web qui viennent butiner sur le site web www.shop4gift.com.

Le problème à traiter est le suivant. Vous allez construire une DMP (*Data Management Platform*) afin de tenter de désanonymiser les navigateurs qui vendront naviguer sur un site web *ad hoc*. Ensuite, vous remplirez cette DMP avec tous les *browser fingerprints* des navigateurs web qui viendront successivement butiner sur le site web *ad hoc*.

Dans un premier temps, vous ne travaillerez pas sur un vrai site web d'e-commerce B2C mais sur un prototype de site web permettant de tester le principe de la désanoymisation des *browser fingerprints*.

3- Analyse du problème

Q1 *R1*	Soit **T** la taille de la population mondiale. À un milliard près, par excès, quelle est la grandeur de **T** ?
Q2 *R2*	Soit **NB** le nombre minimum de bits pour coder **T** la taille de la population mondiale. Calculer **NB**.
Q3 *R3*	Soit **NO** le nombre minimum d'octets pour coder **T** la taille de la population mondiale. Calculer **NO**.

On prend, comme **Code-Client**, un tableau d'octets de longueur **NO**. Chaque **Code-Client** étant unique, ce code est une clé primaire possible de la base de données **Relation-Clients**. La table **TCC** listant tous les **Code-Client** possibles est composée de **T** lignes de longueur **NO** octets.

Q4 *R4*	Soit **ZCC** la taille en octets de la table **TCC** de toutes les clés **Code-Client** possibles. Calculer **ZCC**.

Quand un visiteur (prospect inconnu ou client déjà connu) navigue sur le site web de Shop4gift, le serveur consulte le navigateur web de façon à connaître les données numériques présentes sur ce navigateur web.

Pour chaque visiteur du site web www.shop4gift.com**,** on définit un champ **Browser-Fingerprint** de 4 096 octets qui rassemble toutes les données numériques présentes dans le navigateur. 4 096 octets est une taille minimale qui doit suffire dans la plupart des cas où les navigateurs web sont nettoyés régulièrement. En revanche, pour certains navigateurs web qui ne sont pas nettoyés, 4 096 octets peut être insuffisant pour stocker toutes les polices de caractères et tous les *plug-ins*.

Selon le type de clientèle (audacieuse, conservatrice, technophile ou technophobe, confiante ou parano), on pourra ajuster la taille du champ **Browser-Fingerprint.**

Le tableau ci-dessous donne un exemple des données présentes usuellement dans *le browser fingerprint.* Voir également le site web *amiunique.org.*

Attribute	Value
User agent	Mozilla/5.0 (X11; Linux i686) Gecko/20100101 Firefox/25.0
HTTP accept	text/html,application/xhtml+xml,application/xml; q=0.9,*/*;q=0.8 gzip, deflate en-US,en;q=0.5
Plugins	Plugin 0: IcedTea-Web 1.4.1; Plugin 1: Shockwave Flash 11.2 r202
Fonts	Century Schoolbook, DejaVu Sans Mono, Bitstream Vera Serif, URW Palladio L, ...
HTTP DoNotTrack	1
Cookies enabled	Yes
Platform	Linux i686
OS (via Flash)	Linux 3.14.3-200.fc20.x86 32-bit
Screen resolution	1920x1080x24
Timezone	-480
DOM session storage	Yes
DOM local storage	Yes
I.E. User data	No

Tab. 2 – Exemple de brower fingerprint (source : [LAPE2015])

La table **TAD** de tous les champs **Browser-Fingerprint** possibles est composée de **T** lignes de deux colonnes (clé, valeur).

Chaque ligne **Browser-Fingerprint** de la table **TAD** doit être indexée par une clé afin de la retrouver plus rapidement.

La clé de la table **TAD** est composée de NO octets et la valeur contenant le champ **Browser-Fingerprint** est composée de 4 096 octets.

Q5 Soit **ZAD** la taille en octets de la table **TAD** composée de lignes (clé, valeur) **Browser-Fingerprint** possibles. Calculer **ZAD**.

R5

On définit une fonction de hachage telle que : **Clé = H (Valeur)**

La fonction **H** doit avoir les propriétés suivantes :

1- La clé doit être de longueur **NO** octets
2- La valeur doit être de longueur **4 096** octets.
3- Les collisions doivent être aussi rares que possibles.

Q6 *R6*	Proposer une fonction de hachage **H** ayant les propriétés requises.

Quand un visiteur arrive sur le site web, Shop4gift calcule son **Browser-Fingerprint** et tente de le comparer aux champs **Browser-Fingerprint** déjà présents dans la table **TAD**.

Pour cela, le système calcule la clé = H(**Browser-Fingerprint)** et parcourt la table **TAD** afin d'écrire **Browser-Fingerprint** dans ce champ s'il est libre.

S'il y a déjà un **Browser-Fingerprint** dans le champ dont on a calculé la clé, le système compare le nouveau **Browser-Fingerprint** et l'ancien **Browser-Fingerprint.**

S'ils sont identiques, cela signifie que ce navigateur est déjà passé sur le site web. On a ainsi désanonymisé le visiteur grâce à l'empreinte de son navigateur.

En revanche, si le nouveau **Browser-Fingerprint** et l'ancien **Browser-Fingerprint** sont différents, le système doit mesurer leur différence. Pour cela, on utilise la distance de Levenshtein.

Q7 *R7*	Développez un prototype de site web permettant de tester le remplissage de la table **TAD** avec les visites successives de navigateurs web.

Q8 *R8*	Fournissez les statistiques d'utilisation de la table **TAD** : nombre de visites, nombre de collisions sur la fonction H, nombre de visiteurs différents, nombre de fois où des collisions ont concerné des **Browser-Fingerprint** différents, distribution statistique des distances de Levenshtein.

4- Résultat attendu

Vous fournirez un rapport expliquant ce que vous avez compris de la problématique et vous répondrez aux questions Q1 à Q6.

Vous livrerez une URL (de type www.shop4gift.com) me permettant de tester sur votre prototype plusieurs navigateurs web. Cf. Q7

Vous livrerez une autre URL permettant d'accéder aux statistiques de fonctionnement de votre prototype Cf. Q8.

À noter : La complexité des données relève du *« big data ».* Le volume des données est de l'ordre de grandeur du téraoctet. En conséquence, il est hors de question de travailler « à la main ». Il faut entièrement automatiser les processus afin de limiter les temps de production (et donc les coûts de production) de ces statistiques.

Quoi qu'il en soit, vous êtes entièrement libres des méthodes, des outils ou des langages que vous utiliserez pour implémenter votre prototype.