Protection de données personnelles pour la sécurité sur Internet

Denis Migdal, Christophe Rosenberger

Protection de données personnelles pour la sécurité sur Internet

D. Migdal[1] C. Rosenberger[1]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

{denis.migdal, christophe.rosenberger}@ensicaen.fr

Résumé

De nombreuses applications sur Internet nécessitent d'avoir des informations sur l'internaute, pour vérifier qu'il a bien le droit d'accéder à un service numérique (vérification d'une preuve d'identité comme un mot de passe), pour éviter des attaques (pédopornographie, usurpation de profil,...) ou pour donner de la confiance aux autres utilisateurs (réseaux sociaux). Après avoir proposé une méthode de génération d'une signature non-cryptographique basée sur l'identité de l'individu tout en protégeant sa vie privée [5], nous lançons aujourd'hui une campagne de collecte sur `https://trust.greyc.fr` *dans le but d'améliorer et de valider notre méthode de génération.*

1 Introduction

La consommation de services numériques sur Internet est de nos jours importante que ce soit pour les réseaux sociaux, le commerce électronique ou les jeux en ligne. À titre d'exemple, en 2016, 96% des français ayant demandé un extrait du casier judiciaire l'ont fait sur Internet [1]. Néanmoins, plusieurs données personnelles peuvent être récupérées lors de l'usage d'un service numérique sur Internet soit fournies par l'internaute (notamment sur les réseaux sociaux) soit collectées automatiquement.

Les services numériques sur Internet collectent de plus en plus des données personnelles liées à l'internaute parfois à des fins légitimes (détection de fraude, examen à distance, ...) mais aussi à des fins non conformes aux conditions de collecte (vente à d'autres services, consolidation d'identités, ...). Ces données personnelles peuvent être liées à l'individu (donnée biométrique, nom, âge, ...), au navigateur (version, type, ...), à la machine de l'internaute (système d'exploitation, matériel, résolution de l'écran). Toutes ces informations parfois collectées dans un contexte légitime peuvent aller jusqu'à identifier l'individu ce qui pose un problème majeur de respect de la vie privée.

Notre principale contribution est de proposer une méthode de génération d'une signature non-cryptographique sous la forme d'un code binaire lié à l'identité numérique d'un individu. Ce code ne permet pas de remonter aux informations utilisées pour le calculer et permet également de réaliser des comparaisons avec d'autres codes. Les informations utilisées vont du navigateur utilisé, à la machine, jusqu'à l'individu. Des pré-traitements sont réalisés sur ces données afin de calculer le code dans la dernière étape. L'identification d'attaques comme la détection de plusieurs comptes associés à une identité, ainsi que l'authentification, constituent des exemples d'applications de ce code. De plus amples détails sont aussi disponibles dans nos précédentes publications [5].

1. D'après les chiffres 2017 du SGMAP, `https://goo.gl/wNh3kH`

2 Travaux antérieurs

Le *Browser Fingerprinting* permet de suivre les utilisateurs dans leur navigation Internet grâce aux données discriminantes qu'un service donné peut récolter, souvent dans l'objectif de proposer des "services personnalisés" correspondant au profil-type de l'utilisateur. Les sites Panopticlick [2], IAmUnique [4], et UniqueMachine [1] permettent de calculer son *Browser Fingerprint* à partir des données collectées par le site, afin de déterminer le degré d'unicité de l'empreinte calculée parmi celles déjà collectées. Plus le *browser fingerprint* est unique, plus un service aura capacité à le discriminer.

Cependant, le *browser fingerprint* peut varier, e.g. par le changement du navigateur, de sa configuration [6], ou tout simplement de machine. Le but n'est pas d'identifier l'utilisateur de façon certaine, mais d'identifier un ensemble de sessions de navigations appartenant à un même utilisateur. Les données utilisées pour cela peuvent être liées, e.g., au matériel (e.g. carte graphique [1], écran), au système d'exploitation, au navigateur utilisé, à sa configuration, ou aux polices installées [2, 4].

3 Méthode proposée

L'objectif de notre méthode est de calculer un code binaire lié à une personne à partir d'informations personnelles (techniques et biométriques). Ce code répond à différentes exigences :

- *Non inversible* : le code ne doit pas donner d'informations sur les données personnelles collectées.
- *Confidentialité* : la valeur des attributs ne peut être connue, ni déduite, par le service.
- *Non-usurpation* : un tiers ne peut forger un code lui permettant d'usurper un utilisateur légitime.
- *Révocation* : l'utilisateur légitime doit pouvoir révoquer un code existant.
- *Conservation de la similarité* : Si les données d'un individu sont similaires alors les codes résultant doivent l'être.

La figure 1 présente le principe général de la méthode proposée. Un simple mot de passe est utilisé comme clé secrète [3]. Dans ce cas, Alice par la saisie du mot de passe consent à donner ce code binaire au service.

Dans le cadre de nos travaux, les codes sont utilisés via leur distance de Hamming, donnant ainsi la similarité des données desquelles ils sont issues.

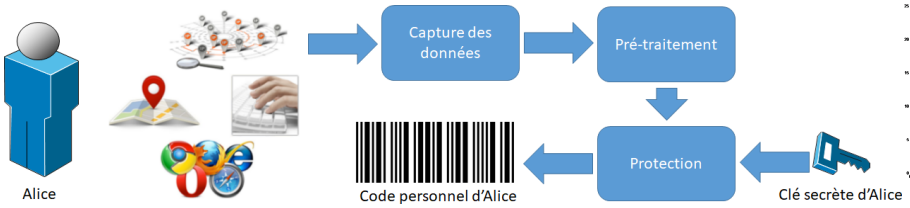

Figure 1 – *Principe de la méthode proposée*

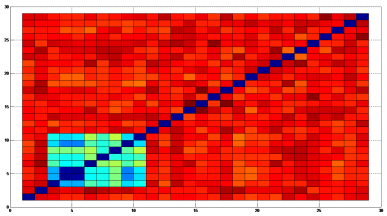

Figure 2 – *Matrice des distances entre les signatures (bleu, si similarité élevée, rouge si faible).*

3.1 Collecte de données personnelles

Lors du calcul de notre signature, nous utilisons plusieurs sources de données que nous présentons ci-dessous.

Navigateur : Le navigateur est authentifié à partir d'une clé générée aléatoirement au premier usage du navigateur, puis stockée sur ce dernier.

Localisation : Les adresses IP sont distribuées par plages, permettant ainsi, même en cas d'IP dynamiques, d'en déduire la localisation de l'utilisateur.

Données réseau : De manière analogue au *browser fingerprinting*, les données reçues des protocoles de communication servent à l'authentification de l'utilisateur.

Données biométriques : La biométrie comportementale de l'utilisateur peut être analysée à partir de ses actions claviers et souris.

3.2 Protection des données

Afin de protéger les données personnelles, le service numérique doit pouvoir exploiter notre signature sans connaître les informations utilisées pour la générer. Pour ce faire, nous utilisons l'algorithme Biohashing qui transforme des données à valeurs dans $\mathbb{R}^n$, en un modèle binaire (BioCode).

Cette transformation est non inversible et permet de conserver la similarité (distance de Hamming) des données en entrée. Cet algorithme a été initialement proposé pour le visage et les empreintes digitales par Teoh *et al.* [7] et nécessite un secret liée à l'utilisateur, ici, un mot de passe saisi par l'utilisateur [3].

4 Expérimentations

Une première campagne de collecte a été organisée en mars 2017. Avec 22 participants recrutés localement, l'échantillon n'est pas représentatif, mais a permis une première expérimentation de notre signature. Il en a résulté 29 signatures montrant une forte similarité lorsque issues d'une même personne, plus ou moins contrastée en fonction des variations de contexte (wifi, navigateur, ...). La figure 2 montre la matrice de distance des signatures, où seules les signatures 3 à 10 sont issues de la même personne (et donc issues de la même clé secrète).

Cette première expérimentation a permis de démontrer la capacité de la méthode proposée à produire un code exploitable pour des calculs de similarité d'informations personnelles. Cependant, nous visons désormais une collecte à plus grande échelle afin d'améliorer et quantifier les performances de notre méthode.

5 Conclusion et perspectives

Nous avons proposé une méthode permettant de calculer un code personnel lié à un internaute respectueux de la vie privée [5]. Ce code intègre différentes informations liées à son navigateur, sa façon de taper au clavier, ou sa localisation. Nous avons montré lors d'une première expérimentation qu'il était possible d'obtenir un code binaire proche pour la même personne malgré des différences de contexte. Nous visons désormais une expérimentation à plus grande échelle afin d'améliorer et quantifier les performances de notre méthode sur `https://trust.greyc.fr`.

Plusieurs applications de notre signature sont envisageables, dont l'authentification d'un internaute, l'usage pour identifier des comptes multiples par un service (similarité de codes calculés avec une clé unique). Ces applications constituent les perspectives de notre travail.

Remerciements

Les auteurs tiennent à remercier la région Normandie pour son soutien financier.

Références

[1] SL Yinzhi Cao and E Wijmans. Browser fingerprinting via os and hardware level features. *Network & Distributed System Security Symposium, NDSS*, 17, 2017.

[2] Peter Eckersley. How unique is your web browser ? *Privacy Enhancing Technologies*, 6205 :1–18, 2010.

[3] Patrick Lacharme and Aude Plateaux. Pin-based cancelable biometrics. *International Journal of Automated Identification Technology (IJAIT)*, 3(2) :75–79, 2011.

[4] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. Beauty and the beast : Diverting modern web browsers to build unique browser fingerprints. *Security and Privacy (SP)*, pages 878–894, 2016.

[5] Denis Migdal and Christophe Rosenberger. Towards a Personal Identity Code Respecting Privacy. In *International Conference on Information Systems Security and Privacy (ICISSP)*, Madeira, Portugal, January 2018.

[6] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator : Deceiving fingerprinters with little white lies. *Proceedings of the 24th International Conference on World Wide Web*, pages 820–830, 2015.

[7] A.B.J. Teoh, D. Ngo, and A. Goh. Biohashing : two factor authentication featuring fingerprint data and tokenised random number. *Pattern recognition*, 40, 2004.