Campagne de collecte de données et vie privée

Nicolas Haderer[1], Miguel Núñez, del Prado Cortez[2,3], Romain Rouvoy[1], Marc-Olivier Killijian[2], and Matthieu Roy[2,3]

[1] INRIA Lille – Nord Europe, Project-team ADAM
University Lille 1, LIFL – CNRS UMR 8022, France
{nicolas.haderer,romain.rouvoy}@inria.fr
[2] LAAS-CNRS, France
[3] Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse, France
{mnunezde,mkilliji,mroy}@laas.fr

Résumé Les communautés scientifiques ont souvent recours à la simulation dans le but de valider leurs théories. Cependant, la pertinence des résultats obtenus est fortement dépendante de la qualité des traces générées par les simulateurs. Ce phénomène est particulièrement vrai lorsque l'on considère les traces de mobilité humaine qui sont difficilement prévisibles. Dans ce contexte, la popularité des nouvelles générations de smartphones, équipés d'une grande variété de capteurs (GPS, bluetooth, accéléromètre, etc.), offre de nouvelles perspectives pour la collecte de données réalistes au sein d'une population. Cependant, la nature sensible, du point de vue de la vie privée, des informations collectées représente un des principaux obstacles aux déploiement généralisé d'une application de collecte de données et à son adoption auprès des utilisateurs.

C'est pourquoi nous présentons UBILAB , une nouvelle plate-forme permettant aux scientifiques de mettre en place facilement des campagnes de collecte de données et d'inférer automatiquement différentes attaques sur les données partagées par les utilisateurs mobiles afin de les avertir d'un risque potentiel d'atteinte à leurs informations privées.

1 Introduction

La nouvelle génération de smartphones (Android, iPhone), maintenant équipée d'une grande variété de capteurs (GPS, bluetooth, accéléromètre, etc.), offre de nouvelles perspectives à diverses communautés scientifiques afin de réaliser différentes campagnes de collectes de données massives d'une population et de son environnement. Ces données peuvent ainsi être exploitées pour mieux comprendre les mouvements d'une population, de mettre au point de nouveaux protocoles de communication, d'analyser les interactions sociales des utilisateurs, etc. La nature sensible des données collectées, généralement couplant des informations temporelles et géographiques, peuvent révéler des informations critiques sur la vie privée d'un utilisateur (résidence privée, opinion politique ou réseau social), même si celles-ci ont été préalablement anonymisées. Ce risque potentiel représente un des principaux obstacles aux déploiement généralisé d'une application de collecte de données et à son adoption auprès des utilisateurs.

Dans ce contexte, nous présentons UBILAB , une plate-forme dédié à la gestion de campagnes de collecte de données auprès d'utilisateurs de téléphones mobiles. UBILAB est le résultat de la l'association de deux plate-formes : ANTDROID [5] et GEPETO (GeoPrivacy Enhanced TOolkit)[1]. Ce système profite ainsi de l'architecture de ANTDROID pour rapidement mettre en place une campagne de collecte de données, et

des algorithmes de GEPETO pour inférer automatiquement différentes attaques sur les données partagées par les utilisateurs mobiles afin de les avertir d'un risque potentiel d'atteinte à leurs vies privées.

2 La plateforme AntDroid

La plate-forme AntDroid est composée de deux parties — chacune est destinée aux différents acteurs évoluant dans la plate-forme : les scientifiques et les cobayes. Le serveur d'application dédié, destiné aux scientifiques, repose sur le style architectural REST (*REpresentational State Transfer*) fournissant l'ensemble des services pour la définition, la diffusion et l'exploitation d'une expérience de collecte de données. L'application cliente pour smartphone est destinée aux utilisateurs voulant participer à une campagne de collecte de données.

La définition d'une campagne est décrite avec un langage de script dédié AntDroid Scripting Language (cf. Listing 1.1), permettant de spécifier les données qui doivent être collectées par le téléphone mobile (lignes 6-11) et l'événement qui déclenche la collecte (ligne 5). Ce choix permet aux scientifiques de bénéficier d'une grande flexibilité pour la définition du schéma de leurs données n'imposant aucune structure particulière. Les données collectées sont ensuite stockées au format XML et peuvent être facilement manipulées par le langage XQuery pour analyse et extraction. AntDroid est une plate-forme évoluant dans le *cloud*, proposant une interface web pour créer et gérer une campagne de collecte de données sans requérir l'installation complexe de logiciels.

```
var gsm = new Experiment("GSM_Signal_Strength")                          1
// Update of GPS position every 120 seconds                              2
gsm.configure("Location", { strategy:"fine", period:120 })               3
// Event triggered when location changes                                 4
gsm.onLocationStateChange(event) {                                       5
    return { trace: {                                                    6
        lat:  event.getLatitude(),                                       7
        lon:  event.getLongitude(),                                      8
        time: event.getTime(),                                           9
        ss:   event.getSignalStrengthLevel()                             10
}   } }                                                                  11
// Start collecting activity traces                                      12
gsm.start()                                                              13
```

Listing 1.1. GSM Signal Strength Experiment

L'application cliente est une application Android téléchargeable, utilisée par une communauté d'utilisateurs pour partager leurs traces d'activités. Pour ce faire, l'utilisateur s'abonne à une ou plusieurs campagnes publiées par des scientifiques. Ces campagnes correspondent à des scripts de collecte automatique des informations requises par le scientifique. Ces scripts sont téléchargés via le serveur d'application dédié puis interprétés par un moteur de script intégré dans l'application cliente. Afin de maîtriser toute diffusion d'information, l'application cliente dispose de différents contrôles permettant aux utilisateurs d'autoriser ou non la collecte de certaines informations jugées trop sensibles (par ex., sa position). L'application intègre également un certain nombre d'optimisations permettant de limiter sa consommation énergétique afin de ne pas perturber les usages des utilisateurs.

3 Analyses des traces de mobilité avec GEPETO

Les données collectées par le téléphone mobile peuvent ensuite être envoyées manuellement par l'utilisateur ou automatiquement lorsque le téléphone est alimenté en

courant pour limiter sa consommation énergétique. Avant d'être disponible pour les scientifiques, les données sont stockées dans une base de données temporaire ou un ensemble d'analyses sera effectué par la plate-forme GEPETO (GeoPrivacy Enhanced TOolkit) permettant de détecter si les données partagées peuvent comporter des risques vis-à-vis de la vie privée de l'utilisateur.

3.1 Le processus d'analyse

Le processus d'analyse commence par l'élaboration d'un modèle de mobilité basé sur les chaines de Markov (MMC) d'un individu [4]. Ce modèle repose sur un automate probabiliste où les états représentent des points d'intérêt (*POI*) et les transitions, c'est-à-dire les déplacement d'un *POI* à un autre. Le modèle produit peut être à gros grain (représentation globale du mouvement à travers plusieurs villes) ou à grain fin (représentation plus spécifique dans une seule ville avec plus de sémantique). La Figure 1 montre les modèles de Bob : à gauche, un modèle générique des déplacements et à droite, un modèle plus spécifique pour la ville (Toulouse) où il passe le plus de temps.

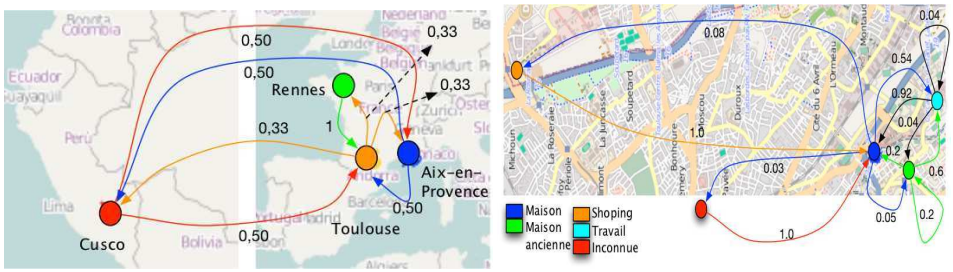

FIGURE 1. Exemple de *MMC* globale et spécifique de *Bob*.

Avec ce modèle (MMC), il est d'abord possible de calculer la prédictibilité potentielle en utilisant l'équation proposée par Gambs et *al.* [3] rappelée dans l'équation 1 ci-dessous. Par exemple, supposons que le vecteur stationnaire pour le MMC spécifique de Bob est $VS = \{0,68, 0, 24, 0, 06, 0, 02, 0, 01\}$, alors, selon le résultat sus-citée, la prédictibilité est de 64%. Il est également possible d'améliorer la prédictibilité, c'est-à-dire avoir un modèle plus intrusif, en utilisant un modèle qui mémorise les n derniers *POIs* où l'individu était (tel que le modèle *n*-MMC[3]) ou un modèle qui introduit le temps comme le *t*-MMC [2].

$$Pred = \sum_{k=1}^{l} (\pi(k) \times P_{max\_out}(k, *)) \quad (1)$$

Ensuite, en observant la configuration du MMC, il est possible d'en déduire des sémantiques des POIs basées sur la densité des états et les transitions. On peut alors attacher des labels aux points d'intérêts, tels que maison, travail, loisir, etc., qui sont, en eux-même, des « quasi identificateurs » et permettent à un adversaire de trouver l'identité d'un utilisateur anonyme [1]. Par la suite, il est possible d'associer une adresse physique à un état quelconque en utilisant les cordonnées GPS du medioid, grâce aux services de géocodage inversé. Dans la table 1, nous présentons, comme illustration, le résumé des informations mentionnées ci-dessus concernant les POIs de Bob.

Densité	Latitude	Longitude	Label	Adresse
390	43.57291	1.46875	Maison	4-6 Allée des sciences Appliquées
70	43.563	1.4774	Travail	L.A.A.S. - C.N.R.S.
61	43.56743	1.46617	Ancienne maison	Résidence Universitaire Clément Ader
57	43.57664	1.46806	Inconnue	Chemin des Herbettes
20	43.62799	1.48265	Shopping	Rue Saint-Jean Balma

TABLE 1. Résumé des informations concernant les POIs de Bob.

Le résultat de ces analyses est ensuite renvoyée à l'utilisateur afin qu'il puisse évaluer si les données collectées peuvent compromettre sa vie privée. L'utilisateur peut alors ensuite décider de valider les données pour les rendre directement disponible pour les scientifiques, de les supprimer, ou d'appliquer des algorithmes d'assainissement (distorsion aléatoire, sous échantillonnage, etc..) fourni par GEPETO pour ajouter du bruit sur les traces de mobilité.

4 Conclusion

Nous avons présenté UBILAB , une plate-forme pour la gestion de campagnes de collecte de données. Nous avons montré comment cette plate-forme peut être utilisée par les scientifiques pour diffuser et exploiter une expérience de collecte de données et les différents mécanismes pour avertir les participants si leurs données partagées peuvent comporter des risques vis-à-vis de leurs informations privées. Dans ce cas, les données peuvent être supprimées ou dégradées en appliquant différents algorithmes d'assainissement fournis par la plate-forme. Cependant, une trop grande dégradation des données peuvent les rendre inutiles pour les scientifiques. Nous envisageons donc dans de futur travaux d'élaborer un mécanisme permettant d'avoir un compromis entre la protection des informations privées des utilisateurs et la qualité des données offertes aux scientifiques.

Références

1. S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Gepeto : a geoprivacy-enhancing toolkit. *AINA'09 Workshop on Advances in Mobile Computing and Applications : Security, Privacy and Trust, Perth, Australia.*, August 2009.
2. S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Towards temporal mobility markov chains. In *DYNAM, an OPODIS workshop*, Toulouse, France., December 2011.
3. S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Next place prediction using mobility markov chains. In *Mobility, Privacy and Measurement.*, Bern, Switzerland, April 2012.
4. S. Gambs, M.-O. Killijian, and M. N. n. del Prado Cortez. Show me how you move and i will tell you who you are. In *Transactions on Data Privacy*, volume 2, pages 103–126, Catalonia, Spain, August 2011.
5. N. Haderer, R. Rouvoy, and L. Seinturier. AntDroid : A distributed platform for mobile sensing. Rapport de recherche RR-7885, INRIA, Lille, France., February 2012.