# Rapport intermédiaire n° 1 (entre 5-10 pages)

| | | | |
|---|---|---|---|
| **Acronyme** | PRIVA'MOV | | |
| **Titre du projet** | Mobilité et vie privée : de la collecte à l'analyse des données | | |
| **Coordinateur scientifique** | Nom | Prénom | Fonction |
| | Ben Mokhtar | Sonia | CR CNRS |
| | Laboratoire | Téléphone | Mail |
| | LIRIS | 04 72 43 63 07 | sonia.ben-mokhtar@liris.cnrs.fr |

| **Liste Partenaires** | N° | Laboratoire / Equipe | Correspondant scientifique |
|---|---|---|---|
| **Partenaires académiques** | 1 | LIRIS-DRIM | Sonia Ben Mokhtar, Omar Hasan, Lionel Brunie |
| | 2 | EVS-ITUS | Olivier Brette |
| | 3 | CITI-Privatics | Cédric Lauradoux, Mathieu Cunche |
| | 4 | CITI-Urbanet | Hervé Rivano, Razvan Stanica |
| | 5 | LET-ENTPE | Patrick Bonnel |
| **Partenaires praticiens** | 6 | | |

## I. Rappel des objectifs de la période concernée

- Travail sur des mécanismes de protection de la vie privée des usagers en situation de mobilité. Les contributions pour cette partie sont décrites dans les sections : II.1, II.2 et II.3.
- Achat de dispositifs de collecte de traces de mobilité. Cette partie est décrite dans la section II.4.
- Développement d'une plateforme de collecte de traces de mobilité. Cette partie est décrite dans la section II.5.
- Déploiement des dispositifs sur des usagers et collecte de données. Cette partie est décrite dans la section II.6.

## II. Avancées scientifiques

### *II.1. État de l'art : mécanismes de protection de la privée*

La littérature propose de nombreux mécanismes de protection de la vie privée en situation de mobilité. On différencie deux scénarios majeurs : la protection de traces statiques avant publication, telles que celles qui vont être collectées par le projet Priva'Mov, et la protection des utilisateurs de services géolicalisés (ou LBS, pour *Location-Based Services*) en temps réel, qui doit leur permettre de profiter des services géolocalisés sans craindre pour leur vie privée. Dans un premier temps, nous nous sommes concentrés sur les mécanismes de protection en temps réel. Nous avons élaboré une classification en six catégories des mécanismes existants.

La première famille de mécanismes utilise une pseudonymisation des données. Cela consiste à remplacer une information identifiant l'utilisateur (son nom, prénom, date de naissance) par un identifiant ne comportant pas d'information personnelle (par exemple un identifiant numérique généré aléatoirement). La plupart des mécanismes des autres familles comment par appliquer des pseudonymes avant d'aller plus loin. Il est aussi possible de mélanger les pseudonymies dans certaines régions appelées mix-zones afin d'empêcher de tracer un utilisateur par son pseudonyme, comme proposé dans [11].

Une deuxième famille de mécanismes a été regroupée sous le terme de *spatial cloaking*. Le principe est d'englober la position des utilisateurs dans une zone géographique plus large. Cela a pour double effet de réduire la précision de la position de l'utilisateur (on ne sait pas où il se trouve exactement dans la zone) et d'ajouter de l'incertitude sur l'utilisateur qui utilise un LBS (on ne sait pas quel utilisateur parmi ceux qui se trouvent dans la zone envoie la requête). Des solutions plus anciennes comme [1] fonctionnent sur une architecture très centralisée, où un serveur de confiance procède à la génération des zones qui seront utilisées. Des solutions décentralisées où les utilisateurs communiquent directement entre eux ont également été proposées comme [2].

Une troisième famille de mécanismes a recours à la génération de faux utilisateurs, appelés *dummies*, pour déboussoler le LBS. Lors que chaque requête, les utilisateurs envoient en plus de leur vraie position $n$ autres positions venant de faux utilisateurs dont les trajectoires sont générées localement. Le LBS ne sait ainsi pas où se trouve réellement l'utilisateur parmi les $n + 1$ positions qu'il reçoit. Tout l'enjeu est de générer des trajectoires réalistes qui ne trahissent pas l'utilisateur. Des solutions simples reposant essentiellement ont d'abord été proposées comme [3] avant que des solutions plus élaborées faisant appel à des données extérieures (recensement, statistiques de trafic, etc.) ne soit conçues comme [4].

Une quatrième famille de mécanismes repose sur l'ajout de bruit à la position des utilisateurs. La conséquence est de décaler spatialement la position réelle de l'utilisateur. Le LBS ne sait ainsi pas où se trouve précisément l'utilisateur, ce qui peut permettre de masquer des lieux sensibles (lieux de culte, partis politiques, soins médicaux, etc.). Différentes méthodes existent pour ajouter du bruit offrant des garanties différentes. Une proposition récente reprend la notion populaire de *differential privacy* [10] et l'étend pour la protection de données spatiales [5].

Une cinquième famille de mécanismes exploite des techniques cryptographiques pour protéger l'information et obtenir une réponse. Chaque mécanisme répond à un problème bien précis (par exemple identifier des lieux autour de soi ou trouver des amis aux alentours) et propose un protocole adapté à ce problème. Par exemple les auteurs de [7] proposent trois protocoles pour repérer ses amis tandis que [8] propose un moyen d'obtenir des statistiques sur les utilisateurs se trouvant dans une zone donnée.

Une sixième et dernière famille de mécanismes propose de concevoir depuis le départ les LBS en intégrant les aspects de vie privée dès le départ. En étant ainsi « *privacy by design* », les garanties offertes aux utilisateurs peuvent être très bonnes, tout en garantissant des résultats précis, au prix d'une intégration impossible dans les LBS existants. La solution proposée par les auteurs de [9] semble être prometteuse.

### *II.2. Étude détaillée d'un mécanisme de protection*

Dans un second temps, nous nous sommes attachés à étudier dans le détail un des mécanismes précédemment répertoriés pour mieux comprendre son fonctionnement, ses forces et ses faiblesses. Nous avons choisi un travail récent publié dans CCS en 2013 et nommé *geo-indistinguishability* [5].

Reposant sur l'ajout de bruit, ce mécanisme a l'avantage de fonctionner de façon très simple, sans calcul coûteux, et de s'intégrer très facilement avec des LBS existants. De plus, s'appuyant sur la notion de *differential privacy* [10], il fournit des garanties de vie privée prouvables.

Néanmoins, il souffre d'un défaut majeur qui apparaît dès lors qu'on veut protéger de multiples fois une même position d'un utilisateur. Nous avons tous des répétitions dans notre mobilité (par exemple le trajet domicile-travail) et donc de mêmes endroits sont à protéger jour après jour. L'ajout de bruit reposant sur une distribution probabiliste connue, le plus grand nombre de fois une position donnée va être protégée, le plus facilement elle va pouvoir être ré-identifiée en croisant les informations obtenues jour après jour. C'est en substance ce que nous avons montré dans un [6]. Même à un fort niveau de protection, nous arrivions selon nos expérimentations à retrouver 35 % des POIs des utilisateurs dans un rayon de 500 mètres du vrai point. À un faible niveau de protection cela montre à 63 % des POIs ainsi réidentifiés. De plus, un fort niveau de protection est susceptible d'entraîner une forte dégradation des performances, 90 % des résultats retournés par un LBS étant ainsi inutilisés à un fort niveau de protection. Cela entraîne ainsi une forte augmentation de la bande passante ainsi que du temps de calcul effectué sur le téléphone et donc potentiellement une perte de batterie.

### *II.3. Conception de nouveaux mécanismes de protection*

Après avoir étudié un mécanisme existant, la suite logique est de proposer notre propre solution en construisant sur ce que nous avons précédemment analysé. Ayant compris toutes les difficultés à protéger la position d'un utilisateur qui envoie des requêtes à un LBS en situation de mobilité, nous avons choisi de nous intéresser à la protection de traces statiques avant publication. Ces traces ont beaucoup de valeur, tant pour les industriels que pour les chercheurs, mais leur publication comporte de nombreux enjeux en termes de vie privée. Si les garanties ne sont pas suffisamment fortes, les utilisateurs risquent de ne pas participer à notre campagne de collecte. De plus, des directives européennes[1] obligent les organismes qui publient de telles données à les anonymiser, sans pour autant imposer une solution technique particulière.

L'état de l'art sur ce sujet précis est déjà bien rempli mais laisse encore une large place à de nouveaux mécanismes. Nous souhaitons nous inscrire dans la lignée de papiers récents comme [12] qui pensent qu'il n'y a pas un mécanisme de protection unique qui satisfait tous les scénarios mais que l'anonymisation doit se faire en fonction des données et de l'usage que l'on souhaite en faire. Notre idée-clé est d'aller à contre-courant de solutions comme [5] qui cherchent à protéger uniquement la composante spatiale des données. Or les données de mobilité sont des données spatio-temporelles qui contiennent un identifiant d'utilisateur, une position géographique et une pastille temporelle. Nous souhaitons conserver au maximum intacte la position géographique pour maximiser l'utilité des données et jouer sur chacune des deux autres composantes. Ce travail a donné lieu à deux nouveaux mécanismes, l'un agissant principalement sur la composante temporelle pour garantir une vitesse constante sur la trajectoire et l'autre mélangeant les trajectoires des utilisateurs pour empêcher des les tracer.

Nous avons montré dans le travail précédent que l'identification de ces points d'arrêt permet de trouver les points d'intérêt d'une personne, et que cette information est très sensible. Or, si l'on est incapable d'identifier les endroits où l'utilisateur s'arrête, il devient bien plus difficile d'extraire de la connaissance d'une trace. Cela peut être le cas si jamais l'utilisateur semble être en mouvement constant, c'est-à-dire que sa vitesse reste constamment élevée. C'est l'idée de base de notre premier mécanisme : forcer une vitesse constante sur une trace, en réalisant un lissage de la vitesse. Certes, la précision de l'information temporelle est mécaniquement réduite, mais reste suffisante pour

1 http://ec.europa.eu/justice/data-protection/

répondre à des requêtes faisant intervenir des fenêtres temporelles (e.g., combien de personnes étaient à cet endroit entre 8h et 11h ?), d'autant plus que l'information spatiale reste elle très précise.

Le mélange des trajectoires est une idée qui n'est pas nouvelle et a été proposée très tôt sous le nom de *mix-zones*, comme évoqué plus haut. Ces approches tentent usuellement d'optimiser le placement des mix-zones afin d'obtenir une utilité maximale. Nous nous travaillons sur un jeu de données statique, ce qui rend le problème différent. Nous pouvons donc mener une analyse sur les traces dont nous disposons afin de générer des mix-zones optimales. L'idée de base est que si deux utilisateurs sont dans la même zone dans la même fenêtre temporelle, alors nous avons une chance que leurs trajectoires soient échangées (il s'agit d'un mélange des identifiants de tous les utilisateurs contenus dans cette fenêtre spatio-temporelle). Nous introduisons ainsi un mécanisme de protection opportuniste, qui fournit une protection au prix d'une dégradation très réduite des données.

Notre approche opportuniste fournissant une protection plus faible que d'autres approches, nous préférons ainsi intégrer ce mécanisme avec celui de lissage de la vitesse, afin de former un nouveau mécanisme combinant deux techniques complémentaires. En agissant ainsi sur les identifiants des utilisateurs et sur la dimension temporelle, nous fournissons une protection complète sans déformer la dimension spatiale, contrairement à ce qui a été fait dans l'état de l'art. Nous avons réalisé une évaluation extensive, sur le mécanisme combiné mais aussi sur le lissage de la vitesse seul. Contrairement à certains articles utilisant des données synthétiques, nous avons fait le choix d'utiliser des jeux de données réels. Nous nous sommes comparés à deux mécanismes de l'état de l'art, un basé sur *k*-anonymity, l'autre sur *differential privacy*. L'évaluation a été faite en anonymisant plusieurs jeux de données et en comparant certaines propriétés avant et après anonymisation. Il s'agit à la fois de propriétés quantitative (taille des jeux de données, écarts spatiaux, etc.) et de propriétés qualitatives (erreur lors de l'exécution de requêtes de comptage). Les résultats montrent que notre solution surpasse notre compétiteur utilisant la *k*-anonymity et est comparable à celle utilisant de la *differential privacy* ; selon les usages, l'une ou l'autre peut être mieux adaptée. Si une précision temporelle est requise, un mécanisme à base de *differential privacy* pourra mieux se comporter (e.g., si on veut suivre l'utilisation des TCL en temps réel). Mais si une grande précision spatiale est requise (e.g., si on veut compter le nombre de personnes passées dans une rue bien définie), notre mécanisme se comportera de façon meilleure.

Nous souhaitons poursuivre le travail sur notre mécanisme, notamment en étudiant d'autres façons de l'implémenter. Nous pourrions par exemple nous servir de la connaissance du réseau de rues (librement accessible) pour pré-traiter les données afin de rattacher chaque point à une rue, un chemin, une ligne de tramway. Cela permettrait ainsi de mieux cacher les endroits où l'utilisateur se rend (principalement des bâtiments) tout en conservant une utilité pour un large panel de cas d'utilisations, notamment ceux liés aux transports en commun ou au trafic. Une seconde piste de travail pour la dernière année du projet concerne une étude plus détaillée des utilisateurs. Nous souhaitons en particulier nous intéresser à leur sensibilité. Tous les utilisateurs ne se comportent pas de la même façon : certains visitent des endroits plus sensibles que d'autres, certains sont très réguliers tandis que d'autres ont un comportement plus chaotique, etc. Il s'agit de choses observables dans les traces de mobilité et donc quantifiables. Savoir inférer le degré de sensibilité d'une trace et proposer à l'utilisateur d'utiliser et adapter un mécanisme de protection en fonction serait un résultat important.

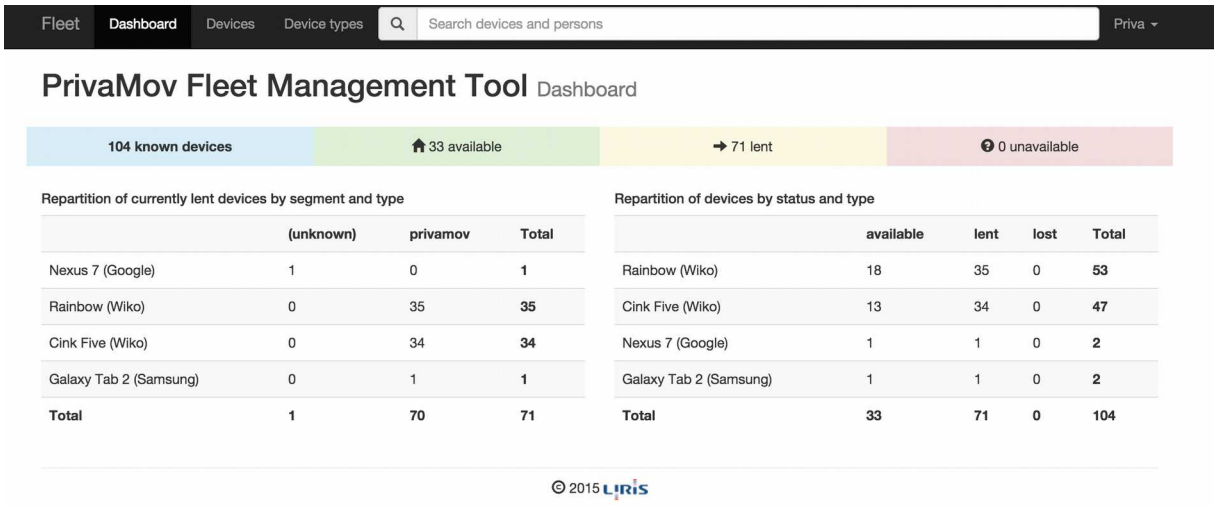
### *II.4. Plateforme de collecte*

Dans le cadre du projet Priva'Mov, nous avons acquis une centaine de dispositifs mobiles. L'objectif était de les distribuer à des utilisateurs afin qu'ils s'en servent au quotidien et que nous puissions collecter leurs usages, notamment en situation de mobilité. Tout d'abord, une étude à été mené afin de

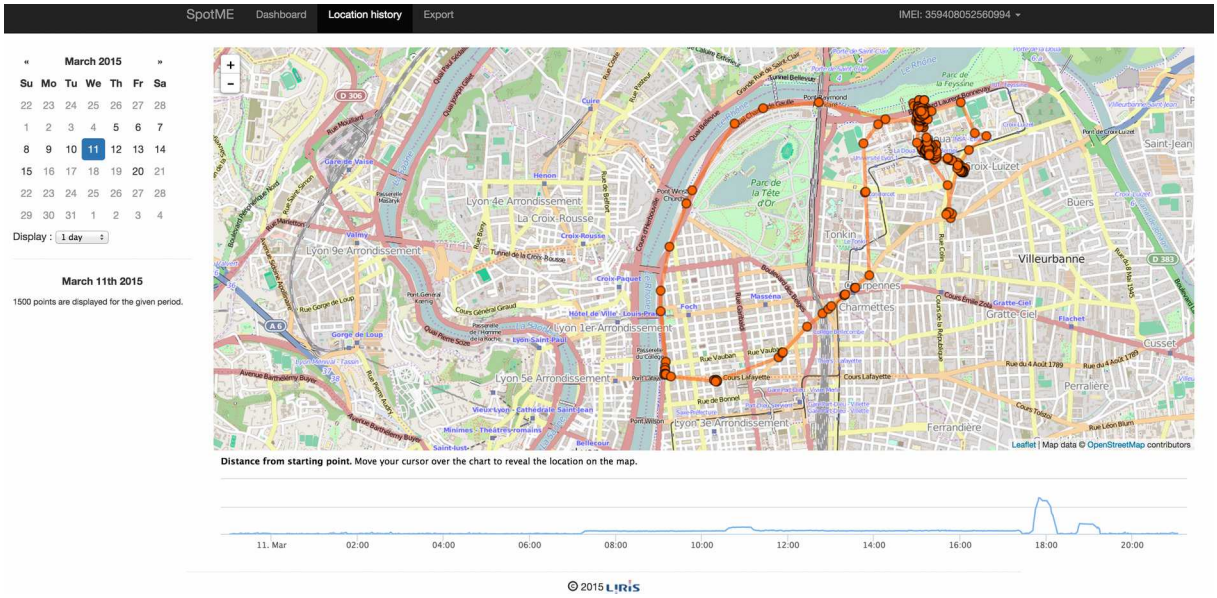
tester les capacités de plusieurs types de dispositifs, tablettes et téléphones. Il fallait notamment qu'on puisse y installer l'application procédant à la collecte, que celle-ci tourne sans heurt et que l'autonomie reste malgré tout suffisante pour permettre un usage normal du dispositif. À l'issue des tests et après avoir étudié les usages possibles, il a été décidé de privilégier l'achat de smartphones.
En particulier, nous avons au total :

- 2 téléphones Wiko Cink five (167€)
- 2 téléphones Samsung GALAXY Core GT-i8262 (183€)
- 2 tablettes ASUS NEXUS 2 (sans 3G 269€)
- 2 tablettes ASUS ME371MG (avec 3G 219€)
- 47 téléphones Wiko Cink 5 (150€)
- 52 téléphones Wiko Rainbow (125€)

Une application web a été développée afin de supporter la logistique autour des téléphones. Une capture d'écran de la page principale de cette application est représentée ci dessous :




Elle permet de procéder à leur inventaire et de suivre à qui ils ont été prêtés, durant quelle période de temps ainsi que les caractéristiques de l'expérimentation en cours sur ce téléphone. Cela nous permet lorsque nous recevons les données d'un téléphone de pouvoir ensuite savoir à quelle personne elles sont liées. Une seconde application web permet aux utilisateurs de pouvoir visualiser les données que nous collectons. Chacun dispose d'un lien unique qui lui permet de vérifier que la collecte se passe correctement, de visualiser les données sur une carte ainsi que de télécharger ses données s'il le souhaite. Une capture d'écran de ce site de visualisation pour un utilisateur donné est représentée ci dessous :




## II.5. Architecture de collecte

Afin de réaliser la collecte de données, il était nécessaire de développer une architecture de collecte. Cette architecture est constituée de deux entités : l'application déployée sur les téléphones, qui permet de recueillir les données d'utilisateurs participant à une expérience, et le serveur de collecte, qui reçoit les données transmises par l'application. Nous présentons dans un premier temps, les détails de l'application puis ceux du serveur.

### II.5.1. Application mobile

L'application mobile que nous avons déployée sur les smartphones Priva'mov se base sur le framework Funf que nous avons modifié pour les besoins du projet.

Ce framework fournit les outils nécessaires pour recueillir des données périodiquement. Chaque grandeur mesurée (position, accélération, signaux Wi-Fi, etc.) constitue une sonde. Chaque sonde peut donc être activée périodiquement pour une certaine durée durant laquelle, les données propres à cette sonde peuvent être recueillies. On peut également forcer, lors de l'activation de la sonde, une requête de données. Les données recueillies par une sonde sont converties au format JSON afin de faciliter la maintenance de l'application et de visualiser aisément si des données ne sont pas recueillies. Ces sondes sont interfacées avec une base de données sqlite locale au travers d'un pipeline. Ce pipeline permet de faire le lien entre le service, qui tourne en tâche de fond, l'activité, qui est l'interface utilisateur, et les sondes. On définit donc pour un pipeline une instance sqlite.

Nous avons dans un premier temps défini les sondes qui étaient importantes dans le cadre du projet. Les sondes retenues sont les suivantes :

- La sonde GPS qui fournit la position du téléphone et de l'utilisateur.

- La sonde accéléromètre qui permet de connaître les mouvements que subit le téléphone.

- La sonde Wi-Fi qui scanne les réseaux Wi-Fi visibles du téléphone.

- La sonde cellulaire qui scanne la cellule à laquelle le téléphone est relié et les cellules voisines.

- La sonde Bluetooth qui scanne les dispositifs Bluetooth à proximité du téléphone.

- La sonde SMS qui conserve les dates d'arrivée et d'envoi des SMS ainsi que le fait que le correspondant appartient au carnet d'adresse du téléphone.

- La sonde appels qui conserve les dates d'appels émis, reçus, manqués ainsi que le fait que le correspondant appartient au carnet d'adresse du téléphone.

- La sonde batterie qui recueille l'état de charge du téléphone.

- La sonde navigateur qui stocke les recherches web réalisées par l'utilisateur sur le navigateur web du téléphone.

- La sonde applications qui liste les applications dont un processus est actif sur le téléphone.

- La sonde réseau qui récupère le contenu d'un fichier du téléphone qui stocke le volume de données échangées pour chaque application du téléphone.

Durant la phase de test, nous avons rapidement constaté que l'activation simultanée de plusieurs sondes posaient des soucis sur l'accès à la base de données. Nous en avons donc déduit qu'il était préférable de modifier le comportement de l'application pour allouer une instance de sqlite à chaque sonde. Avec le comportement périodique permis par le framework Funf, nous avons réalisé que nous pouvions perdre des données sur différentes sondes. Pour certaines d'entre elles, un compromis peut être fait pour déterminer le seuil minimal d'activation d'une sonde en fonction du degré de précision que l'on souhaite obtenir. En revanche, pour des sondes comme les appels ou les SMS, on ne peut pas garantir de ne pas manquer un événement.durant la période d'inactivation. De plus, le choix d'une fréquence élevée d'activation/désactivation va augmenter la consommation énergétique. C'est par exemple le cas de la sonde GPS, où la puce GPS va consommer plus d'énergie sur les premières secondes d'activation lors de la phase de recherche des satellites. Nous avons donc également fait le choix de modifier le comportement des sondes pour les laisser actives en permanence à l'écoute d'événements du système. Les données qui sont recueillies sont donc celles captées par le téléphone dans un fonctionnement normal, sans forcer des scans des réseaux environnants par exemple. Cependant, si pour un déploiement particulier, on avait besoin de forcer des captures de données, ceci est tout à fait réalisable. Il faudrait modifier les paramètres d'activation des sondes en question et on pourrait avoir une application de collecte plus agressive et moins performante en énergie. De la même manière, on peut décider de n'utiliser que certaines sondes afin de gagner en performance. Durant la phase de tests, les téléphones avaient une autonomie d'une journée (du matin jusqu'au soir). Il fallait donc que les téléphones soient chargés le matin et les recharger le soir pour ne pas arriver à cours de batterie.

L'interface utilisateur est longtemps restée basique fournissant des informations sur la carte SIM et sur l'IMEI du téléphone avec deux boutons pour mettre à jour la configuration de l'application et transmettre les données stockées par l'application. Nous avons ajouté une couche de visualisation des données de localisation. Pour le moment, les données qui sont visualisées sont celles qui sont encore stockées sur le téléphone, mais nous sommes actuellement en train de développer côté application et serveur les procédures permettant à l'utilisateur de visualiser un historique de déplacements plus anciens.

Les fichiers transmis vers le serveur de collecte sont les fichiers sqlite contenant les données des sondes. Lors de la transmission de ces fichiers, un identifiant de déploiement ainsi qu'un identifiant du téléphone sont transmis. Ainsi lors de la réception côté serveur, on est en mesure d'associer les données au bon terminal et à la bonne collecte. Un téléphone pourra être utilisé pour plusieurs collectes successives et plusieurs collectes, avec des configurations différentes, peuvent exister en parallèle.

Après avoir détaillé les motivations des choix de conception de l'application, nous allons présenter comment le serveur de collecte recueille les données et les extrait pour les rendre exploitables pour analyse.

### *II.5.2. Serveur de collecte*

Le serveur de collecte stocke les données transmises par les téléphones. Pour la partie collecte, le serveur utilise deux bases de données. Une première base de données qui sert de zone tampon pour la seconde qui va décompresser les fichiers transmis et faire les contrôles nécessaires avant le stockage des données brutes. Cette décomposition en deux bases de données permet de décorréler la partie réception des données de la partie extraction et vérification des données transmises.

La seconde base de données vérifie que les fichiers transmis n'ont pas déjà été reçus. Cette situation pourrait se produire dans le cas où le téléphone n'aurait pas reçu la réponse du serveur après envoi d'un fichier et que le téléphone n'aurait donc pas supprimé ce fichier. Pour réaliser cette fonctionnalité, une opération de hachage est réalisée sur les fichiers et un index est placé sur les hashs. Un contrôle est également fait sur le fichier afin de ne pas conserver des fichiers qui auraient été corrompus lors de l'écriture sur le téléphone ou lors de la transmission. C'est cette base de données qui réalise la correspondance entre les données transmises, la sonde, le téléphone et la collecte auxquels ces données appartiennent.

Afin de traiter ces données, nous avons la possibilité de l'exporter sous différents formats (JSON, XML, CSV). Ce sont sur ces exports que des travaux de recherche ont été réalisés. Nous avons également créé une base de données SQL en utilisant postgreSQL pour faciliter la consultation et le traitement des données collectées. Nous utilisons postGIS afin d'améliorer les fonctions de traitement spatial des données. Nous avons défini des index sur le champ indiquant la date de la mesure et sur l'identifiant du téléphone afin de diminuer les temps de sélection de données sur une certaine plage temporelle ou pour un certain nombre d'utilisateurs.

### *II.6. Déploiements de la plateforme et collecte de données*

Après quelques tests par les membres du projet eux-mêmes, des déploiements réels de la plateforme ont été réalisés. Le premier a pris place lors de la conférence Middleware en décembre 2014 à Bordeaux, où était proposé aux participants d'emprunter un téléphone pour la durée de la conférence. Cela nous a permis de tester un déploiement à échelle réduite (une trentaine de téléphones ont été distribués) et avec des enjeux moindres, ainsi que de valider le bon comportement de notre dispositif de collecte, tant du point de vue des téléphones que de la réception des données.

Forts de cette expérience, nous avons lancé une expérience à plus grande échelle de mars à août 2015. Nous avons ciblé des membres de nos laboratoires ainsi que des étudiants de l'INSA. Au cours de ce second déploiement à plus grande échelle, nous avons distribué environ 80 téléphones en plus de ceux que portent les membres du projet. Ce second déploiement nous a permis de collecter des données à plus grande échelle, tout en faisant remonter les problèmes liés au traitement d'un volume de données plus important.

## III. Interactions entre les disciplines impliquées et la valeur ajoutée par cette pluridisciplinarité

Tous les partenaires du projet ont participé aux différentes phases du projet : allant de la définition des données à collecter aux type de dispositifs à acquérir en passant par les techniques d'obfuscation de données à privilégier. Ces décisions collectives ont été prises lors des réunions plénières du projets qui se sont tenues aux dates suivantes :

- 14/11/2013
- 15/04/2014
- 20/11/2014
- 11/06/2015

La base de donnée de traces de mobilité étant encore en cours de collecte, l'interaction entre les différentes disciplines s'est en particulier focalisée sur le raffinement des cas d'utilisation pour les différents partenaires. En particulier, les différents laboratoires porteurs de cas d'utilisation des traces (économie, transports, informatique) ont mentionné les données qu'ils souhaitent voir collectées par le projet Priva'Mov et les usages qu'ils souhaitent en faire. Cela a nécessité une bonne compréhension des besoins de chacun, le vocabulaire n'étant pas nécessairement le même, voire même certains mots (mobilité, trace, etc.) n'ayant pas le même sens dans chacune des disciplines.
Le défi est de collecter des données qui soient facilement exploitables par tous, des informaticiens aux sciences humaines et sociales. Les besoins récoltés sont bien différents, tous ne sont pas intéressés par les données, et cela ajoute donc un défi supplémentaire au projet.

Par ailleurs, un début de collaboration avec un autre projet porté par des collègues SHS de l'université de Franche-Comte ont été discuté à la dernière réunion plénière. En particulier, notre collègue Olivier Brette, membre du projet a invité à cette dernière réunion Thomas Buhler MCF en aménagement de l'espace et urbanisme à l'université de Franche-Comté) qui a obtenu un financement régional pour un projet (TELEM) qui vise à conduire une analyse longitudinale par panel des pratiques de mobilité et de consommation énergétique. Ce projet ayant des objectifs très similaires à ceux du projet Priva'Mov, des pistes de collaboration très prometteuses ont été discutées (e.x., utilisation dans le projet TELEM des techniques de protection des données développées dans Priva'Mov, partage de données collectées entre les deux projets, etc.).

## IV. Résultats obtenus, publications, valorisation et exploitation des résultats

### *IV.1. Site web du projet*

Un site web pour le projet a été mis en place, à l'adresse http://liris.cnrs.fr/privamov. Il récapitule les objectifs du projet, les différentes personnes impliquées ainsi que les résultats que nous avons obtenus.

### *IV.2. Sur les mécanismes de protection de la vie privée*

L'état de l'art des mécanismes de protection a été présenté lors de la réunion du collège doctoral MDPS[2] qui s'est tenue en décembre 2013 à Lyon. Cela a été l'occasion d'avoir du retour sur la classification et de discuter de possibles futures pistes de recherche. Un *survey* a également été rédigé et doit être finalisé afin de pouvoir être publié dans une conférence ou un journal.

Un article [14] portant sur les travaux introduits en II.2 a été publié au *workshop Mobile Security Technologies* (associé à la conférence *Security & Privacy, classée A* par CORE*[3]). L'article a ensuite été présenté à San Jose, CA, USA en mai 2014 devant une soixantaine de personnes. Le format moins strict des *workshops* a permis des échanges enrichissants lors de la présentation, mettant en valeur les points d'accroche et ceux qui nécessitent encore d'être davantage approfondis. Les

2 https://www.dimis.fim.uni-passau.de/MDPS/fr/

3 http://www.core.edu.au/

résultats de cet article ont été à nouveau présentés lors de la réunion du collège MDPS de juin 2014 à Milan.

Lors du déploiement de la plateforme à la conférence Middleware, un poster [13] avait été réalisé conjointement avec une équipe de recherche INRIA à Lille, qui travaille sur des problématiques de *crowd-sensing*. Nous disposions d'un bureau pour distribuer les téléphones et présenter le projet à côté du bureau d'inscription, et étions présent lors de la session posters pour échanger avec les autres participants.

Les résultats concernant le mécanisme de protection combiné introduit en II.3 ont été soumis à la conférence *International Conference on Distributed Computing Systems (ICDCS, classée A par CORE)* en décembre 2014. Ils ont ensuite été présentés lors de la réunion du collège MDPS en décembre 2014 à Besançon. L'article [15] soumis à été accepté comme papier court accompagné d'un poster, et a été présenté en juillet 2015 à Columbus, OH, USA. Cela a permis des échanges intéressants avec les autres participants à la conférence.

Une version longue du papier se focalisant uniquement sur le mécanisme de lissage de la vitesse a été soumis à la conférence *International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom, classée A par CORE)*. Les résultats qu'il présente ont été présentés lors de la réunion du collège MDPS en juin 2015 à Passau, Allemagne, puis l'article [16] ayant été accepté à Helsinki, Finlande en août 2015.

### *IV.3. Sur les risques d'atteinte à la vie privée dus aux technologies de surveillance*

Le contexte du projet PrivaMov a été l'occasion de mener des travaux sur de potentielles attaques sur la vie privée via des technologies de surveillance. La plateforme de collecte a été mise à profit dans le cadre d'une étude visant à évaluer la tracabilité des téléphones par un ensemble de points d'accès Wi-Fi malicieux. Cette étude considère un attaquant qui a réussi à prendre le controle d'un grand nombre de points d'accès Wi-Fi et qui les exploitent pour tracer les déplacements des utilisateurs d'appareils Wi-Fi passant à porté. En effet, les appareils équipés de WiFi révèlent leurs présence en emettant régulièrement des messages de recherche contenant un identifiant unique, et ceci même quand ils ne sont pas connectés à un point d'accès.

Grâce à la plateforme de collecte Privamov, il a été possible de constituer un jeu de données relatif aux points d'accès se trouvant à portée des smartphones. A partir de ce jeu de données, nous avons pu évaluer l'efficacité du traçage que pourrait effectuer un tel système.
Les résultats montrent qu'en zone urbaine, une faible fraction de points d'accès Wi-Fi compromis (2 pour cent) permet de reconstruire des traces de mobilité très précises. Ces travaux ont été présentés en Juillet 2015 au Workshop on Surveillance & Technology co-localisé avec PETS.

### *IV.4. Sur les données collectées*

Actuellement, trois déploiements ont été réalisés pour la première collecte dont l'identifiant est insa-2015. Les données recueillies dans le cadre de cette collecte l'ont été durant la phase de tests avec trois téléphones, durant un déploiement test lors de la conférence Middleware 2014 à Bordeaux pendant 3 jours sur une dizaine d'utilisateurs et auprès d'étudiants et de personnels de l'INSA Lyon à partir du mois de mars.

Nous fournissons ici des informations sur cette collecte. La figure 1 représente le volume de données capturé par téléphone et enregistré dans notre base. La figure 2 quant à elle

représente le nombre d'enregistrements par sonde et par téléphones pour les trois sondes principales : GPS, Wi-Fi et réseau cellulaire.

De manière synthétique la base de donnée collecté à ce jour a une taille d'environ 86GO. A la dernière vérification effectuée la semaine dernière, notre base contenait une totalité de

- 101451807 enregistrements GPS.

- 2524583 enregistrement pour le réseau cellulaire.

- 19432320 enregistrements Wi-Fi.

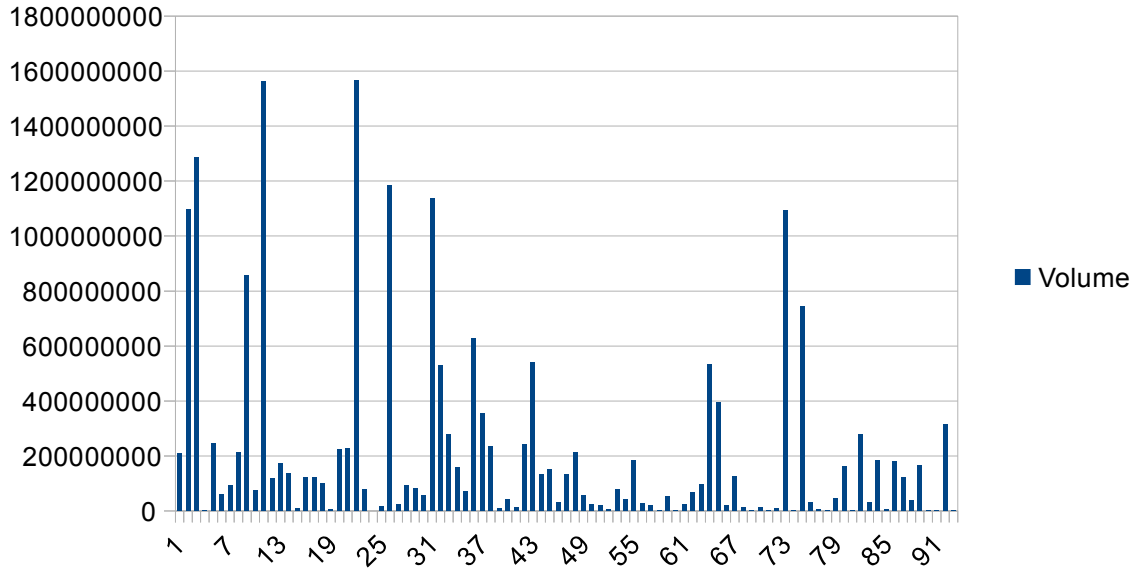



*Figure 1 : Volume de données stockées par téléphone*

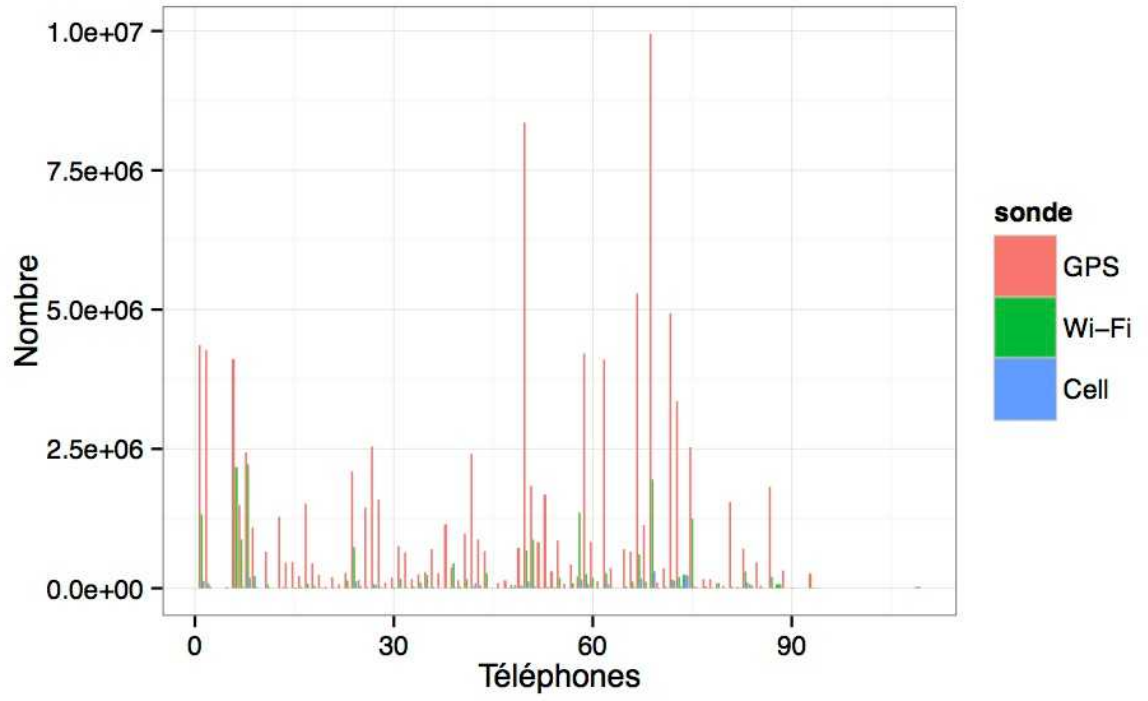



*Figure 2: Nombre d'enregistrements par téléphones pour différentes sondes*

## V. Prochaines étapes

Nos prochaines étapes consistent en :

- L'analyse des données collectées.
- La poursuite des travaux autour de la protection des données de mobilité.
- La réalisation de nouvelles campagnes de collecte de données afin de varier l'échantillon de traces collectées. Cet aspect est particulièrement important pour nos collègues SHS.

## VI. Références

[1] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The New Casper: Query Processing for Location Services Without Compromising Privacy," in *Proceedings of the 32nd International Conference on Very Large Data Bases*. VLDB Endowment, 2006, pp. 763–774.

[2] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems," in *Proceedings of the 16th International Conference on World Wide Web*. ACM, 2007, pp. 371–380.

[3] D. Quercia, I. Leontiadis, L. McNamara, C. Mascolo, and J. Crowcroft, "SpotME If You Can: Randomized Responses for Location Obfuscation on Mobile Phones," in *Proceedings of the 2011 31st International Conference on Distributed Computing Systems*. IEEE Computer Society, 2011, pp. 363–372.

[4] P. Shankar, V. Ganapathy, and L. Iftode, "Privately Querying Location-based Services with SybilQuery," in *Proceedings of the 11th Interna- tional Conference on Ubiquitous Computing*. ACM, 2009, pp. 31–40.

[5] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential Privacy for Location-based Sys- tems," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. ACM, 2013, pp. 901–914.

[6] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie, “Differentially Private Location Privacy in Practice,” in *Proceedings of the 2014 Mobile Security Technologies Conference*, May 2014.

[7] G. Zhong, I. Goldberg, and U. Hengartner, “Louis, Lester and Pierre: Three Protocols for Location Privacy,” in *Privacy Enhancing Tech- nologies*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4776, pp. 62–76.

[8] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li, “Privacy and accountability for location-based aggregate statistics,” in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 653–666.

[9] S. Guha, M. Jain, and V. N. Padmanabhan, “Koi: A Location-Privacy Platform for Smartphone Apps,” in *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX, 2012, pp. 183–196.

[10] C. Dwork, “Differential Privacy,” in *Automata, Languages and Programming*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, vol. 4052, pp. 1–12.

[11] B. Palanisamy and L. Liu, “MobiMix: Protecting location privacy with mix-zones over road networks,” in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE)*, Apr. 2011, pp. 494–505.

[12] G. Acs and C. Castelluccia, “A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris,” in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data.*

[13] N. Haderer, V. Primault, P. Raveneau, et al. "Towards a Practical Deployment of Privacy-preserving Crowd-sensing Tasks," in *Middleware Posters and Demos '14*, 2014.

[14] V. Primault, S. Ben Mokhtar, C. Lauradoux and L. Brunie. "Differentially Private Location Privacy in Practice," in *Proceedings of the Third Workshop on Mobile Security Technologies (MoST)*, 2014.

[15] V. Primault, S Ben Mokhtar, L. Brunie. "Privacy-preserving Publication of Mobility Data with High Utility," in *Proceedings of the 35th IEEE International Conference on Distributed Computed Systems (ICDCS)*, 2015.

[16] V. Primault, S. Ben Mokhtar, C. Lauradoux, L. Brunie. "Time Distortion Anonymization for the Publication of Mobility Data with High Utility," in *Proceedings of the 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2015.