Données de mobilité : protection de la vie privée vs. utilité des données

Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Sara Bouchenak, Nicolas Marchand, Bogdan Robu

Compas'2017 : Parallélisme/ Architecture / Système
Sophia Antipolis, France, du 27 au 30 juin 2017

Données de mobilité : protection de la vie privée vs. utilité des données

Sophie Cerf[1], Vincent Primault[2], Antoine Boutet[2], Sonia Ben Mokhtar[2],
Sara Bouchenak[2], Nicolas Marchand[1], Bogdan Robu[1]

[1] Univ. Grenoble-Alpes, CNRS, GIPSA-lab, F-38000 Grenoble, France
[2] INSA-Lyon, LIRIS, CNRS, Distributed Systems Research Group, Lyon, France
{firstname.lastname}@gipsa-lab.fr, {firstname.lastname}@insa-lyon.fr

Geo-Indistinguishability

Résumé
La généralisation des appareils mobiles a facilité l'apparition de bases de données de mobilité. Ces dernières peuvent poser des problèmes de divulgation de données sensibles lors de leur publication. Des Mécanismes de Protection de la vie Privée pour les données de Mobilité (LPPM) ont été développés pour garantir formellement les besoins de protection de vie privée. Cependant, cela ne se fait pas sans dégradation de l'utilité de la base de données résultante. La configuration des ces LPPM permet de jouer sur ce compromis entre vie privée et utilité. Nous proposons PULP, un mécanisme réalisant cette configuration de manière automatique en fonction d'objectifs de vie privée et d'utilité, en se basant sur la modélisation de l'impact de la configuration d'un LPPM sur la vie privée et l'utilité. Notre approche a été évaluée sur un LPPM de l'état de l'art et quatre bases de données, les résultats montrent l'efficacité de notre solution pour garantir les objectifs.

Mots-clés : protection de vie privée, utilité de données, base de donnée de mobilité, Mécanisme de Protection de la vie Privée pour les données de Mobilité LPPM

1. Introduction

L'utilisation de plus en plus répandue de services utilisant la localisation des utilisateurs pour améliorer leur qualité a multiplié l'apparition de bases de données de mobilité. Ces bases de données de mobilité peuvent se révéler très utiles en post-analyse, mais peuvent cependant engendrer des violations de la vie privée des utilisateurs. Pour pallier ce problème, des mécanismes de protection (nommés LPPM) ont été proposés dans la littérature. Leur principe est de transformer un ensemble de données de mobilité en un ensemble brouillé de données de mobilité, les transformations apportées pouvant être de toutes sortes : modification, ajout, retrait, etc. Les différents LPPM peuvent être classifiés suivant les garanties de vie privée qu'ils apportent, parmi lesquels la k-anonymité [16], qui cache un utilisateur parmi k-1 autres ou bien ϵ-confidentialité différentielle [7] qui borne la quantité de données dévoilée par le retrait d'une donnée. Cependant, tous ces LPPM doivent être configurés (choix de k, ϵ, etc.), ce qui est loin d'être une tâche évidente à réaliser et à évaluer. Primault [14] et Agir [3] présentent deux mécanismes de l'état de l'art qui proposent des solutions heuristiques qui explorent itérativement différentes configurations de LPPM pour en choisir la meilleure. Cependant, aucune des solutions de l'état de l'art ne permet de garantir des objectifs en terme de préservation de vie privée et d'utilité des données.

Il n'existe aucune méthode standard pour évaluer dans quelle mesure la vie privée et l'utilité de données de mobilité sont préservées, pas plus que de quantification du compromis entre protection de la vie privée et utilité. De plus, une solution de configuration de LPPM capable de garantir des objectifs de vie privée et d'utilité est manquante. Dans ce papier nous proposons PULP, une approche de modélisation adaptative des LPPM qui lie leur configuration à la protection de la vie privée (en terme de protection des points d'intérêt) et à l'utilité (en terme de précision spatiale), capable de s'adapter aux différentes bases de données et LPPM. Ensuite nous utilisons cette modélisation pour en déduire des lois de configuration basées sur la garantie d'objectifs, exprimés comme niveaux minimums à garantir pour les métriques ou bien par le compromis protection de la vie privée vs. utilité à atteindre. PULP est évalué avec différentes bases de données : la modélisation est précise, rapide et adaptative, et les lois de configuration ont la même fiabilité que la modélisation.
La suite du papier est organisée comme suit : le contexte de notre travail (données, LPPM, métriques) est fixé et une illustration de motivation est montrée, puis PULP est présenté (modélisation et configuration). Une évaluation est réalisée avant de conclure le papier.

2. Contexte et motivation

Dans cette section nous posons le contexte de PULP : les données de mobilité sont présentées ainsi que les mécanismes de protection de l'état de l'art. Puis nous définissons les métriques qui nous servirons à quantifier la protection de vie privée et l'utilité des données. Enfin nous illustrons le problème de configuration d'un LPPM avec des données réelles.

2.1. Données réelles de mobilité

Une base de données de géolocalisation est un ensemble de traces d'utilisateurs T_{u_i}. Chaque trace est constituée d'ensembles d'événements, des quadruplets $\{u_i, lat, lng, t\}$ donnant la position sur la surface de la terre avec (lat, lng) à un instant t d'un utilisateur u_i.
Dans la suite nous utiliserons quatre jeux de données de mobilité : Cabspotting [12], Geolife [17], Privamov [5], et le Mobility Data Challenge (MDC) [10, 9]. Toutes ces bases de données sont des enregistrements de données réelles d'utilisateurs dans différentes villes du monde. Afin de pouvoir les comparer, nous avons réduit leur taille à celle du plus petit jeu de données, à savoir Cabspotting qui dure 30 jours, en sélectionnant les périodes les plus actives.

2.2. Mécanismes de protection de vie privée pour les données de mobilité (LPPM)

Un LPPM est un mécanisme permettant de protéger une base de données de mobilité. Comme protéger ne se résume pas à l'anonymiser, de nombreux LPPM existent, chacun assurant une protection différente. Parmi les plus connus sont Hermers++ [11] (génération de fausses données), Promesse [13] (ajout de perturbation) ou encore Never Walk Alone [2] (aggrégation des utilisateurs). Dans la suite de cet article nous présenterons un LPPM en particulier : Geo-Indistinguishability [4], même si d'autres ont été étudiés mais dont les résultats ne figureront pas dans l'évaluation faute de place.
Geo-Indistinguishability (Geo-I) est un LPPM qui agit selon le principe de confidentialité différentielle [7] qui vise à brouiller une base de données tout en s'assurant que son utilité n'en est pas trop dégradée. Geo-I bruite spatialement les données, en ajoutant aux coordonnées un bruit Laplacien paramétré par ϵ (en m^{-1}). La quantité de bruit est inversement proportionnelle à ϵ : plus on veut bruiter nos données, plus il faudra appliquer un ϵ petit.

2.3. Mesurer la protection de vie privée et l'utilité des données

Les notions de protection de vie privée et d'utilité des données n'ont pas de définitions quantificatrice communément acceptées dans la littérature. Or ces quantifications sont indispensables

quand on cherche à protéger un jeu de données de manière fiable. Les métriques que nous avons choisies sont détaillées dans les paragraphes suivants. Il aurait été possible d'utiliser d'autres indicateurs, cependant notre choix a été motivé par une étude de l'état de l'art : nous considérerons qu'un utilisateur cherche à cacher ses arrêts [8] (i.e. la mesure de la protection de la vie privée se base sur une attaque d'un adversaire sans connaissance *a priori*), et que l'utilité est liée à la justesse spatiale de la localisation divulguée [6].

Protection de la vie privée Afin de quantifier la protection de vie privée, nous introduisons la notion de Point d'Intérêt (POI), comme étant un lieu où un utilisateur s'est arrêté pendant suffisamment longtemps. Notre métrique de protection de la vie privée ρ évalue le recouvrement des POI d'un utilisateur u_i entre la trace originale T_{u_i} et celle protégée T'_{u_i}.
Plus précisément, un POI est le centroïde d'une zone circulaire de diamètre variable dans laquelle l'utilisateur a passé un temps significatif. Pour définir ρ, nous calculons la précision du recouvrement (proportion des POI retrouvés par rapport aux POI de la trace protégée) ainsi que son rappel (proportion des POI retrouvés par rapport aux POI de la trace originale). Pour la phase de recouvrement, deux POI sont considérés comme identiques s'il sont suffisamment proches (poi_{lim} paramétrant la distance limite maximale). Afin de tenir compte à la fois de la précision et du rappel du recouvrement, nous définissons la métrique de vie privée ρ comme étant la moyenne harmonique (F-score) de ces deux mesures moyennées pour chaque utilisateur. Enfin nous en prenons l'opposé et ramenons la mesure entre 0 et 1 :

$$\rho = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{2 \times precision_{poi}(T_{u_i}, T'_{u_i}) \times rappel_{poi}(T_{u_i}, T'_{u_i})}{precision_{poi}(T_{u_i}, T'_{u_i}) + rappel_{poi}(T_{u_i}, T'_{u_i})} \quad (1)$$

Utilité des données Pour évaluer l'utilité d'un jeu de données, nous en regardons l'acuité spatiale. Plus précisément, nous considérons le recouvrement des zones spatiales visitées par les utilisateurs, dans la trace originale et dans celle protégée. Pour cela nous définissons une fonction $cell$ qui répertorie les cellules visitées par une trace, en se basant sur le découpage spatial de Google S2 [15]. Ce découpage est paramétré pour gérer différentes grandeurs de cellules. La précision et le rappel du recouvrement sont calculées avec la fonction $cell$, et notre mesure finale d'utilité est leur moyenne harmonique :

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \frac{2 \times precision_{cell}(T_{u_i}, T'_{u_i}) \times rappel_{cell}(T_{u_i}, T'_{u_i})}{precision_{cell}(T_{u_i}, T'_{u_i}) + rappel_{cell}(T_{u_i}, T'_{u_i})} \quad (2)$$

2.4. Motivation : impact d'un LPPM sur les métriques de vie privée et d'utilité

Pour illustrer l'importance d'une configuration adaptée d'un LPPM, nous avons réalisé plusieurs expériences dans lesquelles il est appliqué un LPPM sur une même base de données, chaque fois avec une configuration différente. Ensuite les métriques ρ et μ sont calculées.
La Fig. 1 montre l'évolution des métriques de protection de vie privée et d'utilité lorsque Geo-I est appliqué sur le jeu de données Geolife avec différentes valeurs de son paramètre ϵ. On remarque que lorsque ϵ est petit, c'est à dire quand on ajoute beaucoup de bruit au données, la vie privée est bien préservée (peu de POIs ont été correctement ré-identifiés) alors que l'utilité est faible. Inversement lorsque ϵ est grand on tend à la minimisation de l'impact du LPPM sur les données, donc à une mauvaise préservation de la vie privée mais à une bonne utilité. De plus, quand ϵ augmente l'utilité augmente mais la confidentialité diminue : cela correspond bien au compromis constaté entre ces deux métriques. On voit ainsi sur ces figures l'importance du choix de la configuration d'un LPPM pour modérer son action.

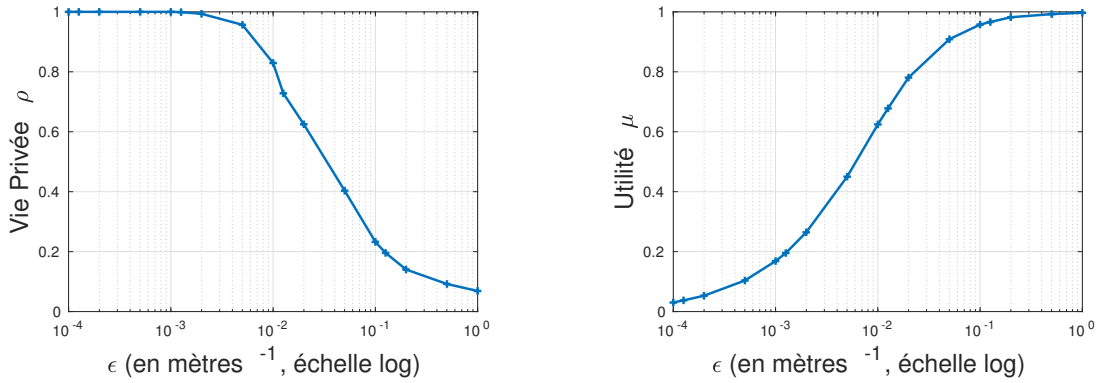


FIGURE 1 – Impact de la configuration d'un LPPM sur la protection de la vie privée et sur l'utilité des données (jeu de données : Geolife)

3. PULP : protection de vie privée et conservation d'utilité pour les données de mobilité

Nous proposons PULP, un outil de configuration automatique de LPPM permettant d'atteindre des objectifs de protection de vie privée et d'utilité. PULP prend en entrée une base de données de mobilité ainsi que les objectifs de protection de vie privée et d'utilité, définis pas l'utilisateur. PULP calcule en sortie la configuration du LPPM à appliquer. PULP agit en trois phases : d'abord une caractérisation de l'impact du LPPM sur la base de données vis-à-vis de nos métriques, puis une modélisation de cet impact et enfin une loi de configuration utilisant ces modèles. Ces trois parties de PULP sont détaillées ci-après et représentées dans le schéma de la Fig. 2

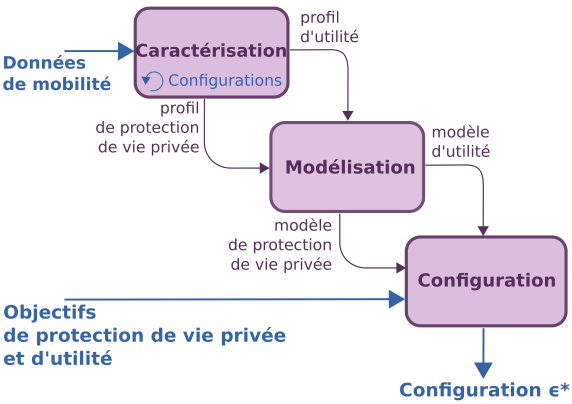


FIGURE 2 – Schéma de principe du fonctionnement de PULP

3.1. Caractérisation d'une base de données vis-à-vis d'un LPPM

Durant la phase de caractérisation, un LPPM avec différentes configurations est appliqué à la base de données de mobilité considérée. Les valeurs de protection de la vie privée et d'utilité des données brouillées sont ensuite mesurées. Un profil caratéristique de la base de données par rapport à un LPPM est ainsi créé, consitué d'une liste de valeurs des métriques. Le choix des valeurs du paramètre de configuration à appliquer dépend du LPPM considéré. Plus les configurations choisies seront représentatives de l'ensemble de la plage de fonctionnement du LPPM, plus la caractérisation de la base de donnée sera précise.

3.2. Modélisation du niveau de protection de vie privée et d'utilité des données

Nous proposons maintenant une modélisation permettant de capturer l'impact d'un LPPM sur l'utilité et la préservation de vie privée d'une base de données. Ce modèle prend donc en son entrée les profils de la base de données et prédit en sortie les valeurs des métriques ρ et μ de vie privée et d'utilité suivant le paramétrage du LPPM. Par la suite, on peut en déduire pour chaque configuration d'un LPPM la relation liant utilité et vie privée.
Pour la modélisation nous avons choisi une fonction qui représenterait au mieux le comportement observé en expérimentation (voir Fig. 1), qui simplifierait la conception de lois de confi-

guration et qui pourrait s'adapter largement pour d'autres LPPMs et d'autres jeux de données. La fonction arctan qui a un comportement de transition entre deux valeurs limites semble *a priori* adéquate, sa paramétrisation permettant de l'adapter aux différents cas. Il en résulte la formulation suivante :

$$\rho(\epsilon) = a_\rho \arctan(b_\rho(\ln(\epsilon) - c_\rho)) + d_\rho \quad (3)$$
$$\mu(\epsilon) = a_\mu \arctan(b_\mu(\ln(\epsilon) - c_\mu)) + d_\mu \quad (4)$$

Les paramètres a et d définissent la valeur des seuils de saturation, respectivement leur écart et leur décalage sur la valeur de la métrique. Le paramètre c définit la configuration du LPPM donnant une valeur moyenne de la métrique considérée. Enfin le paramètre b caractérise la vitesse de transition entre les deux valeurs seuils.

Afin de s'adapter au LPPM et au jeu de données utilisés, les huit paramètres des éq. (3) et (4) sont adaptés au moyen d'un algorithme d'optimisation. Nous utilisons la fonction *fminunc* de Matlab [1] qui permet de minimiser l'écart entre les données d'expériences et la modélisation.

À partir des éq. (3) et (4) et en remplaçant la variable ϵ, on obtient le modèle de la confidentialité en fonction de l'utilité (parfaitement symétrique en ρ et μ) :

$$\rho(\epsilon) = a_\rho \arctan\left(\frac{b_\rho}{b_\mu} \tan\left(\frac{\mu(\epsilon) - d_\mu}{a_\mu}\right) + b_\rho c_\mu - b_\rho c_\rho\right) + d_\rho \quad (5)$$

3.3. Configuration automatique d'un LPPM

Nous souhaitons maintenant obtenir des lois de configuration qui donnent la valeur adéquate ϵ^* du paramètre de configuration du LPPM réalisant des objectifs définis par le détenteur de la base de données à brouiller. Ces objectifs peuvent être de quatre natures différentes, définissant quatre variantes de loi de configuration :

- garantir les niveaux minimaux d'utilité et de vie privée désirés : $\rho \geq \rho_{min}$ et $\mu \geq \mu_{min}$,
- garantir le niveau minimal d'utilité souhaité et maximiser autant que possible de la protection de vie privée : $\mu \geq \mu_{min}$ et ρ $maximal$,
- garantir le niveau minimal de protection de vie privée souhaité et maximiser autant que possible de l'utilité : $\rho \geq \rho_{min}$ et μ $maximal$,
- garantir le compromis entre utilité et vie privée défini tel que $\rho = w.\mu$, où l'on accorde w fois plus d'importance au critère de garantie de vie privée qu'à celui d'utilité.

Nous développerons uniquement le modèle correspondant à la quatrième formulation d'objectifs par souci de place.

Garantir un compromis vie privée/utilité Considérons un objectif de configuration tel que $\{\rho = w.\mu\}$, avec $w \in \mathbf{R}^+$ représenté sur la Fig. 3 pour $w_1 = 2$, $w_2 = 1$ et $w_3 = 1/3$. L'utilisation du modèle (éq. (5)) nous permet d'obtenir une équation donnant la valeur du paramètre de configuration réalisant les objectifs. Une solution analytique est cependant complexe à obtenir pour cette équation non-linéaire : nous avons choisi de la résoudre numériquement, par exemple en utilisant le solveur de matlab *fminunc* :

$$\epsilon^* = argmin_\epsilon |\rho(\epsilon) - w.\mu(\epsilon)| \quad (6)$$

La convergence de la solution est assurée par la convexité de l'équation à minimiser.

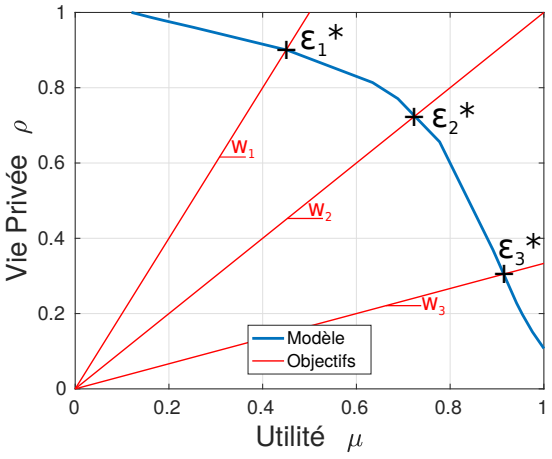


FIGURE 3 – Illustration d'une loi de configuration pour différents objectifs. Données : Geolife

4. Évaluation de PULP

L'évaluation de PULP est réalisée en utilisant les jeux de données décrits en section 2.1 (par souci de place les illustrations ne concerneront que deux d'entre eux) avec le LPPM Geo-I.

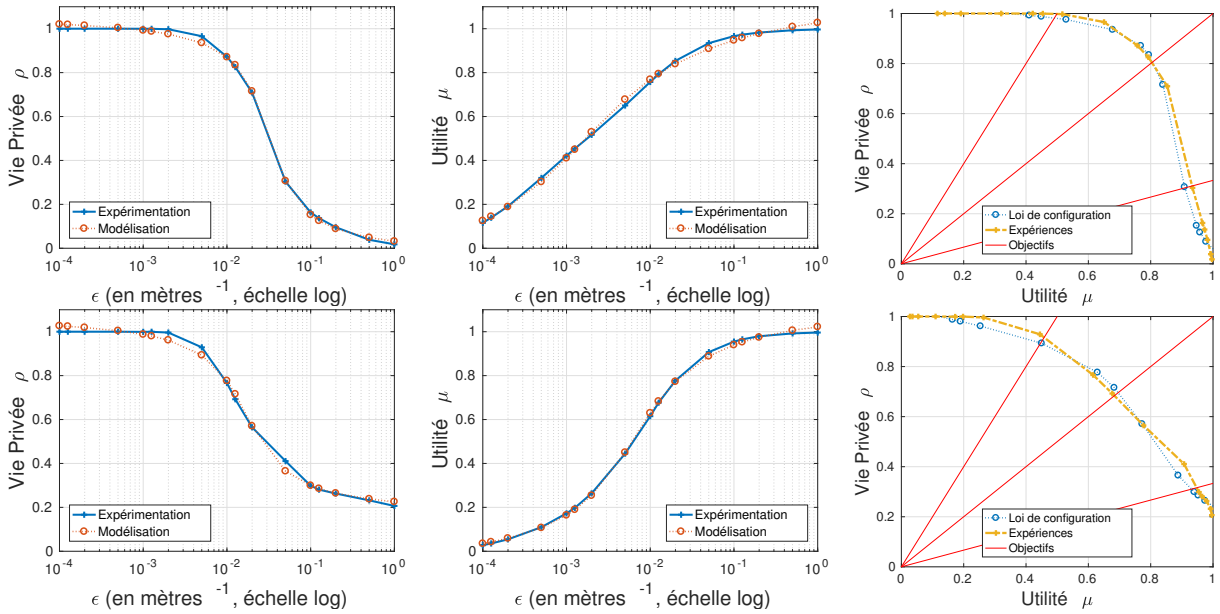
4.1. Dispositif expérimental

Les expérimentations de mesure d'utilité et de vie privée ont été réalisées grâce à l'outil de caractérisation. Cet outil est exécuté sur une machine sous Ubuntu 14.04, ayant 50Gb de RAM et 12 cœurs de 1.2GHz. Pour les calculs de modélisation configuration de PULP, nous avons utilisé le logiciel Matlab dans sa version R2016b, tournant sur une machine Ubuntu 14.04 ayant 3.7Gb de RAM et 4 cœurs de 2.5GHz.

Nous avons choisi de paramétrer les métriques de manière à correspondre à un scénario urbain dense, étant donné la nature des bases des données dont nous disposons. Ainsi, un point d'intérêt a un diamètre de 200m et une durée minimale de 15 minutes. Pour considérer deux POI comme identiques, il faut qu'ils soient à une distance de maximum $\mathrm{poi}_{\mathrm{lim}} = 100\mathrm{m}$. Enfin pour l'utilité, la taille des cellules a été choisie avec un paramètre de taille 15, qui correspond grossièrement à des losanges d'environ 300m à nos latitudes.

4.2. Évaluation de la modélisation dans PULP

La Fig. 4 montre la superposition des données d'expérimentation (obtenus avec l'outil de caractérisation) et de leur modélisation sur Cabspotting et MDC, pour la protection de vie privée (colonne 1) et pour l'utilité (colonne 2), pour 17 configurations différentes de Geo-I (4 par décade de l'intervalle de fonctionnement usuel de son paramètre de configuration). Le modèle est une très bonne estimation des données d'expérimentation puisque la variance de l'erreur de modélisation est de moins de 10^{-3} pour chacun des modèles. Notre modèle capture efficacement le comportement du système et les variations entre jeux de données. Le temps nécessaire pour obtenir les profils expérimentaux est de l'ordre de la minute pour les bases de données considérés. Le temps de modélisation est lui de l'ordre de la milli-seconde.



Modélisation de la vie privée | Modélisation de l'utilité | Configuration par compromis

FIGURE 4 – Évaluation de PULP sur Geo-I, pour les jeux de données Cabspotting (première ligne) et MDC (seconde ligne)

4.3. Évaluation de la loi de configuration dans PULP

Nous évaluons maintenant la fiabilité des lois de configuration grâce à la Fig. 4 (colonne 3). Pour tous les jeux de données considérés, la variance de l'écart entre la configuration trouvée par PULP et les données d'expérience est de moins de 1.5 10^{-3}. La loi de configuration de PULP considérée permet donc de garantir les objectifs désirés. De plus, le temps d'éxécution de l'algorithme de configuration est rapide, de l'ordre de la milli seconde.

5. Conclusion

Nous avons présenté PULP, un mécanisme qui vise à garantir la protection de la vie privée des utilisateurs dans une base de données de mobilité lors de sa divulgation, tout en maintenant de l'utilité dans les données brouillées dans le but d'un post traitement. Après avoir défini les métriques de vie privée et d'utilité que nous considérons, nous avons modélisé l'impact sur nos métriques du mécanisme de protection de vie privée Geo-I, en fonction de sa configuration ; ce qui nous a permis de quantifier le compromis entre vie privée et utilité. En nous basant sur cette modélisation, nous proposons des lois de configuration basées sur des objectifs de vie privée et utilité. L'évaluation de nos solutions avec quatre jeux de données montre de très bons résultats et démontre l'adaptabilité de PULP.

Par la suite, nous prévoyons d'investiguer le compromis vie privée/utilité de manière plus fine, par exemple en détaillant par utilisateur. Des travaux pour étendre notre solution à d'autres métriques de vie privée et d'utilité ainsi qu'à d'autres LPPM sont en cours.

Bibliographie

1. Find minimum of unconstrained multivariable function - MATLAB fminunc - MathWorks Australia.
2. Abul (O.), Bonchi (F.) et Nanni (M.). – Never walk alone : Uncertainty for anonymity in moving objects databases. – In *ICDE*, pp. 376–385, 2008.
3. Agir (B.), Papaioannou (T.), Narendula (R.), Aberer (K.) et Hubaux (J.-P.). – User-side adaptive protection of location privacy in participatory sensing. *GeoInformatica*, vol. 18, n1, 2014, pp. 165–191.
4. Andrés (M. E.), Bordenabe (N. E.), Chatzikokolakis (K.) et Palamidessi (C.). – Geo-indistinguishability : Differential Privacy for Location-based Systems. – In *CCS*, pp. 901–914, 2013.
5. Boutet (A.), Ben Mokhtar (S.) et Primault (V.). – *Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets*. – Research report, LIRIS UMR CNRS 5205, octobre 2016.
6. Chatzikokolakis (K.), Palamidessi (C.) et Stronati (M.). – Constructing elastic distinguishability metrics for location privacy. *Proceedings on Privacy Enhancing Technologies*, vol. 2015, n 2, 2015, pp. 156–170.
7. Dwork (C.). – Differential Privacy. *In : Automata, Languages and Programming*, pp. 1–12. – Springer Berlin Heidelberg, 2006.
8. Gambs (S.), Killijian (M.-O.) et del Prado Cortez (M. N.). – Show Me How You Move and I Will Tell You Who You Are. *Transactions on Data Privacy*, vol. 4, n2, août 2011, pp. 103–126.
9. Kiukkonen (N.), J. (B.), Dousse (O.), Gatica-Perez (D.) et J. (L.). – Towards rich mobile phone datasets : Lausanne data collection campaign. – In *ICPS*, 2010.
10. Laurila (J. K.), Gatica-Perez (D.), Aad (I.), Blom (J.), Bornet (O.), Do (T. M. T.), Dousse (O.), Eberle (J.) et Miettinen (M.). – From big smartphone data to worldwide research : The mobile data challenge. *Pervasive Mob. Comput.*, vol. 9, n6, décembre 2013, pp. 752–771.

11. Pelekis (N.), Gkoulalas-Divanis (A.), Vodas (M.), Kopanaki (D.) et Theodoridis (Y.). – Privacy-aware querying over sensitive trajectory data. – In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 895–904. ACM, 2011.
12. Piorkowski (M.), Sarafijanovic-Djukic (N.) et Grossglauser (M.). – CRAWDAD dataset epfl/mobility (v. 2009-02-24). – Downloaded from http ://crawdad.org/epfl/mobility/20090224, feb 2009.
13. Primault (V.), Ben Mokhtar (S.), Lauradoux (C.) et Brunie (L.). – Time distortion anonymization for the publication of mobility data with high utility. – In *TrustCom*, pp. 539–546, 2015.
14. Primault (V.), Boutet (A.), Ben Mokhtar (S.) et Brunie (L.). – Adaptive location privacy with alp. – In *SRDS*, 2016.
15. S2, a spherical geometry library. – Available online at https ://github.com/google/s2-geometry-library-java.
16. Sweeney (L.). – k-Anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, n5, 2002, pp. 557–570.
17. Zheng (Y.), Zhang (L.), Xie (X.) et Ma (W.-Y.). – Mining interesting locations and travel sequences from gps trajectories. – In *WWW*, pp. 791–800, 2009.