Anonymisation et confidentialité différentielle appliquées à des données spatio-temporelles: cas d'usage portant sur la billettique

Vincent Thouvenot [1] & Thibaut Dubois [2] & Stephane Lorin [3]

[1] *Thales SIX & GTS France, 4 Avenue des Louvresses, vincent.thouvenot@thalesgroup.com*
[2] *Thales SIX & GTS France, 4 Avenue des Louvresses, thibo.dubois@hotmail.com*
[3] *Thales SIX & GTS France, 4 Avenue des Louvresses, stephane.lorin@thalesgroup.com*

Résumé. Smartphone, carte d'abonnement, compteur intelligent énergétique, etc., les sources de données personnelles sont nombreuses. Si ces données peuvent apporter des fortes valeurs ajoutées, que ce soit aux citoyens, aux collectivités ou aux entreprises, celles-ci doivent être protégées. La réglementation autour de la donnée personnelle évolue et se renforce (voire la RGPD). L'anonymisation des données peut être utilisée pour les protéger. Nous présentons ici deux méthodes d'anonymisation de données billettiques. La première méthode est fondée sur un algorithme de généralisation, alors que la seconde cherche à respecter la confidentialité différentielle.

Mots-clés. Anonymisation, Billettique, Confidentialité différentielle, Données spatio-temporelles, Etude de cas, Généralisation

Abstract. Smartphone, subscription card, smart meter in the world of energy, ..., there are many sources of personal data. If this data can provide strong added value, whether to citizens, communities or businesses, it must be protected. It is in this context that the regulation of personal data is evolving and strengthening (see the GDPR which came into force in May 2018). Anonymization of data can be used to protect it. We will present here two methods of anonymizing ticketing data. The first method is based on a generalization algorithm, while the second seeks to respect differential privacy.

Keywords. Anonymization, Differential Privacy, Generalization, Spatial-Temporal data, Ticketting, Use case study

1 Introduction

Les données de billettique sont liées aux activités et aux déplacements dans les transports en commun (métro, bus, tramway, etc.) par les usagers. Ces données peuvent notamment être collectées à partir de carte d'abonnement destiné au transport (par exemple, la carte SimpliCités en Lorraine, la carte Navigo en Ile de France, etc.). Typiquement, ces cartes permettent de collecter l'instant et le lieu de validation des usagers. Grâce à ces données,

les matrices Origine / Destination peuvent être construites et la planification du réseau plus facilement optimisée. Cependant, ces données sont par nature personnelles.

La RGPD, entrée en vigueur en mai 2018, est une réglementation Européenne portant sur le traitement des données personnelles. Cette réglementation vise à garantir un cadre précis à l'utilisation des données personnelles, respectant la transparence, une finalité déterminée, une adéquation des données collectées aux besoins, l'exactitude des données, le droit à l'oubli, à la portabilité et à l'accès, une limitation dans le temps de la conservation des données et la mise en place de mesure de sécurité de la donnée, ainsi que la responsabilisation de tous les acteurs concernés.

L'anonymisation de données personnelles permet de protéger les individus. Un procédé d'anonymisation est un processus irréversible qui cherche à protéger de trois risques:

- Individualisation: Peut-on isoler un individu ?
- Corrélation: Peut-on relier des ensembles de données distincts concernant un même individu ?
- Inférence: A l'aide de variables explicatives externes, peut-on déduire des informations sur un individu ?

L'anonymisation consiste en la réalisation d'un compromis entre la protection des individus et l'utilité future de celles-ci anonymisées. Il n'existe pas de procédé générique d'anonymisation car celle-ci dépend de trois paramètres:

- Typologie des données: des données structurées (par exemple une base de données médicales) n'est pas traitées comme des données non structurées (par exemple, un réseau social)
- Utilisation future des données: l'anonymisation dégrade l'information. Selon l'étude voulue par la suite, les informations nécessaires ne doivent pas avoir été détruites.
- Temps: les méthodes validées peuvent devenir obsolètes.

2 Etat de l'Art général sur l'anonymisation

Dans cette partie, les identifiants sont les éléments qui permettent de (quasiment) directement identifier les individus. Les quasi-identifiants sont des éléments qui permettent d'isoler des gens lorsqu'ils sont combinés. Les variables sensibles sont les variables qui ont de la valeur ajoutée.

Les techniques de permutation et de puzzling changent l'architecture des données (e.g. Agrawal et Srikant (2000), Zhang *et al.* (2007)). Les techniques d'addition de bruit sont populaires. Aggarwal et Yu (2008) propose un Etat de l'Art sur les méthodes de randomisation. Ces méthodes présentent l'avantage de ne pas dépendre du passé des

données et peuvent donc être appliquées durant la collection des données. Cependant, comme le montre Liu *et al.*, ces méthodes sont sensibles à des attaques. Par exemple, lors d'ajout de bruit, l'attaquant peut utiliser du filtrage ou des statistiques bayésiennes pour ré-identifier les individus. Lorsqu'un bruit multiplicatif est utilisé, l'attaquant peut ré-identifier les données s'il posséde un sous-échantillon d'entrée et de sortie .

La confidentialité différentielle (e.g. Dwork (2006), Dwork et A. Roth(2014)) porte sur des assurances théoriques d'un mécanisme de perturbation. L'idée derrière cette technique est d'ajouter un bruit aléatoire tel que le mécanisme de dé-identification produise la même sortie avec des probabilités similaires lorsque deux jeux de données sont adjacents, c'est-à-dire ne changeant que d'un individu. Soit $\varepsilon \in \mathbb{R}^+$, un mécanisme d'anonymisation M, et Im_M l'image de M. M est dit ε-différentiel si pour tout D et D' deux jeux de données adjacents et pour chaque $D^* \in Im_M$:

$$P\left(M(D) = D^*\right) \leq \exp(\varepsilon) P\left(M(D') = D^*\right).$$

La confidentialité différentielle est adaptée au cas où des requêtes sont effectuées sur le jeu de données personnelles. Dans ce cas, une méthode classique pour respecter la confidentialité différentielle est d'utiliser un mécanisme de Laplace.

1. Calcul de la l1-sensibilité: contribution de l'individu le plus influent sur la requête f (qui peut être une moyenne, une médiane, etc.):

$$\Delta_1 f = \max_{D,D'} ||f(D) - f(D')||_1, \text{ avec } D, D' \text{ deux jeux adjacents.}$$

2. Pour une requête f, le mécanisme de Laplace défini par

$$M(D, f, \varepsilon) = f(D) + (Y_1, \ldots, Y_k),$$

où (Y_i) variables aléatoires i.i.d. de loi de Laplace $Lap(\Delta_1 f/\varepsilon)$ est ε- différentiel.

Soit $(\varepsilon, \delta) \in (0,1)^2$, un mécanisme d'anonymisation M, et Im_M l'image de M. M est dit (ε, δ)-différentiel si pour tout D et D' deux jeux de données adjacents et pour chaque $D^* \in Im_M$:

$$P\left(M(D) = D^*\right) \leq \exp(\varepsilon) P\left(M(D') = D^*\right) + \delta.$$

Une méthode classique pour respecter la confidentialité est d'utiliser un mécanisme Gaussien.

1. Calcul de la l2-sensibilité: contribution de l'individu le plus influent sur la requête f (qui peut être une moyenne, une médiane, etc.):

$$\Delta_2 f = \max_{D,D'} ||f(D) - f(D')||_2, \text{ avec } D, D' \text{ deux jeux adjacents.}$$

2. Pour une requête f, le mécanisme Gaussien défini par

$$M(D, f, \varepsilon, \delta) = f(D) + (Y_1, \ldots, Y_k),$$

où (Y_i) variables aléatoires i.i.d. de loi Normale $N(0, \Delta_2^2 f \frac{\ln(1.25/\delta)}{\varepsilon^2})$ est (ε, δ)- différentiel.

Une autre manière de protéger les individus consiste à créer des agrégats d'individus. Sweeney (2002) propose la K-Anonymisation où des classes d'équivalence d'individus de classe K minimum sont crées en généralisant et supprimant des quasi-identifiants. La K-Anonymisation peut être inefficace si tous les individus d'une classe ont la même valeur pour une variable sensible. Pour faire face à ce problème, Machanavajjhala (2006) propose la l-diversité qui force à ce que chacune des classes d'équivalence de taille K minimum ait l valeurs différentes des variables sensibles. Dans la t-proximité (Li (2007)) impose que dans chaque classe d'équivalence, la variable sensible suit la même distribution que dans la population totale. La généralisation et la suppression est un problème NP compliqué. Domingo-Ferrer et Torra (2005) propose d'utiliser de la micro-agrégation pour notamment traiter les variables catégorielles ordinales et continues. L'idée est ici de partitionner les données dans des clusters de taille k minimum puis d'appliquer un opérateur d'agrégation (par exemple la moyenne pour les variables numériques) à chaque cluster et renvoyer la valeur agrégée. Domingo-Ferrer et Torra (2005) utilise l'algorithme de MDAV pour réaliser cette tâche.

3 Anonymisation spatio-temporelle: données simulées sur le réseau de transport de Lisbonne

3.1 Données et cas d'usage

Nous utilisons pour ce travail des données simulées de validation du transport public de Lisbonne. Celui-ci est composé de métro, de tramway et de bus. Des passagers au profil différent (femme/homme, age, profil professionnel) sont simulés pendant un mois. Une validation comprend un instant et lieu de validation ainsi que du profil du passager et d'un numéro d'identifiant associé à chaque usager. Juste chiffrer ou supprimer le numéro d'identifiant ne suffit pas à protéger les passagers. En effet, grâce à l'instant et le lieu de validation, nous pouvons très aisément isoler les individus. Les données, bien que pseudonymisées, ne sont pas anonymisées.

L'objectif est de fournir les données pour permettre de pouvoir obtenir des informations telles que le nombre d'étudiants qui sont venus dans une zone précise de Lisbonne sur une plage horaire fine. Juste agréger les données au niveau des stations sur une plage horaire fine ne suffit pas: de nouveau des individus peuvent être isolés (validation dans une petite gare, la nuit, etc.).

3.2 Description des algorithmes utilisés

Nous évaluons les méthodes en fonction de l'erreur qu'elles amènent par rapport aux vraies données, à leur temps de calculs et leur niveau de protection qu'elles assurent.

3.2.1 Généralisation

Pour cette méthode, nous travaillons jour par jour. Pour chaque journée, nous cherchons la grille géographique la plus fine possible qui permet d'assurer qu'il y ait au minimum k validations sur chacune des plages horaires de 15 minutes.

Pour chacune des stations, nous calculons la proportion de pas de temps pour lesquels l'hypothèse de k validations minimum n'est pas respectée. La station qui a la proportion la plus élevée est alors agrégée à la station la plus proche. Nous fixons un nombre maximal de groupes. Tant que ce nombre n'est pas atteint ou qu'il reste des (groupes de) stations ne respectant pas l'hypothèse de k validations minimum, nous répétons le procédé. Si à la fin du processus, certains groupes de stations comportent encore des pas de temps avec moins de k validations, le nombre de validations pour les observations en question est remplacé par k.

3.2.2 Confidentialité différentielle

Le Fourier Perturbation Algorithm est un algorithme de perturbation des séries temporelles qui respecte des hypothèses de confidentialité différentielle proposé par Rastogi et Nath(2010).

1. Calcul des coefficients de Fourier $F = \{F_1, \ldots, F_n\}$ d'une série de taille n
2. Supprimer les $n - k$ derniers coefficients de la décomposition
3. Ajouter un bruit de Laplace i.i.d. $Lap(\sqrt{k}/\varepsilon)$ sur les k coefficients restants
4. Inversion de la transformation de Fourier avec les coefficients perturbés

L'algorithme précédent réalise alors la ε- différentialité.

Acs et Castelluccia (2014) propose un mécanisme s'inspirant de cet algorithme pour l'utiliser sur des données de télécommunication. Pour cela, les auteurs proposent de remplacer la décomposition en série de Fourier par une transformée de cosinus dicrete, qui permet de réduire les composantes de haute fréquence. Ce faisant, ils conservent la ε- confidentialité différentielle (voir Asc *et al.* (2012)) tout en diminuant l'erreur. Pour diminuer la variance du bruit à ajouter, les auteurs proposent d'utiliser un mécanisme Gaussien, et non Laplacien, et respecte alors la (ε, δ)- différentialité, avec (ε, δ) des paramètres des bruits Gaussiens *i.i.d.* ajoutés.

Certaines stations ont peu d'observations: y ajouter du bruit directement risque de conduire à des résultats incohérents. Les stations sont préalablement agrégées entre stations voisines avant l'ajout du bruit. Le redimensionnement final des séries impliquent un autre mécanisme de Laplace.

4 Conclusion

Dans cette communication, nous présenterons un état de l'Art sur les techniques d'anonymisation, ainsi que les principes liés à la confidentialité différentielle et expliciterons certains des mécanismes permettant de la respecter. Nous illustrerons notre propos sur des données simulées du réseau de transport de Lisbonne. Une application Shiny sera également présentée.

Bibliographie

R. Agrawal and R. Srikant, "Privacy preserving data mining," in ACM SIGMOD Conference, 2000.

Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," in 2007 IEEE 23rd International Conference on Data Engineering, April 2007, pp. 116–125.

C. C. Aggarwal and P. S. Yu, "A survey of randomization methods for privacy-preserving data mining," in Privacy-Preserving Data Mining - Models and Algorithms, 2008, pp. 137–156.

K. Liu, C. Giannella, and H. Kargupta, "A survey of attack techniques on privacy-preserving data perturbation methods."

C. Dwork, Differential Privacy. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, pp. 211–407, 2014.

L. Sweeney, "K-anonimity: A model for protecting privacy,"International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in IN ICDE, 2006.

N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond kanonymity and l-diversity," in 2007 IEEE 23rd International Conference on Data Engineering, April 2007, pp. 106–115.

J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," Data Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195–212, 2005.

Acs, Castelluccia and Chen, 2012, Differentially Private Histogram Publishing through Lossy Compression

Acs and Castelluccia, 2014, A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris

Rastogi and S. Nath. Differentially private aggregationof distributed time-series with transformation and encryption. InSIGMOD, 2010