SOMMAIRE DU Nº 15

Techniques d'anonymisation tabulaire : concepts et mise en œuvre

Benjamin Nguyen [1] et Claude Castelluccia [2]

Dans ce document [3], nous présentons l'état de l'art des techniques d'anonymisation pour des bases de données classiques (i.e. des tables), à destination d'un public technique ayant une formation universitaire de base en mathématiques et informatique, mais non spécialiste. L'objectif de ce document est d'expliquer les concepts permettant de réaliser une anonymisation de données tabulaires, et de calculer les risques de réidentification. Le document est largement composé d'exemples permettant au lecteur de comprendre comment mettre en œuvre les calculs.

Introduction : le secret statistique

En France, il existe depuis 1951 la loi sur *l'obligation, la coordination, et le secret en matière de statistiques* [1]. L'histoire permet aisément de comprendre pourquoi ces questions ont vu le jour à la sortie de la seconde guerre mondiale. L'idée fondamentale défendue dans cette loi est que « *les renseignements individuels (...) et ayant trait à la vie personnelle et familiale et, d'une manière générale, aux faits et comportements d'ordre privé, ne peuvent faire l'objet d'aucune communication* ». De même en ce qui concerne l'Europe, ce principe est affirmé par l'article 338 du traité de l'Union [4] : « *L'établissement des statistiques se fait dans le respect de l'impartialité, de la fiabilité, de l'objectivité, de l'indépendance scientifique, de l'efficacité au regard du coût et de la confidentialité des informations statistiques* ». L'idée fondamentale qui existe depuis près de 70 ans est donc qu'il faut toujours s'assurer de

1. INSA Centre Val de Loire, Laboratoire d'informatique fondamentale d'Orléans.
2. Inria Grenoble Rhône-Alpes.
3. Également publié sur arxiv `https://arxiv.org/abs/2001.02650`.

la confidentialité des données privées lors d'opérations de traitement de données. L'Insee a publié un guide du secret statistique [24], précisant que « *pour les tableaux fournissant des données agrégées sur les personnes physiques, le secret statistique impose qu'on ne puisse pas connaître ou déduire des informations les concernant* ». Il est intéressant de noter que ce guide n'explique pas comment atteindre le secret statistique, mais montre des cas simples où le secret statistique n'est pas réalisé : « *Par exemple, si un tableau donne pour une commune la répartition par âge et situation matrimoniale, et que toutes les personnes âgées de 50 à 59 ans ont toutes pour état matrimonial « divorcé », le secret statistique n'est plus respecté dans ce tableau. En effet, si l'on sait par ailleurs que quelqu'un a entre 50 et 59 ans, le tableau nous informe que cette personne est divorcée* ». Nous verrons que ce problème persiste pour certaines techniques d'anonymisation.

Anonymisation : les principes

L'anonymisation dans la législation

Depuis 1978 et la loi « Informatique et libertés » [2], le principe de protection affirmé par la loi a évolué. En effet, les lois relatives au traitement de données personnelles précisent ce qu'est une telle donnée, ce qui permet de savoir par contraposée ce qu'est une donnée anonymisée : « *Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable,* ***il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne*** ». La législation française était donc, à l'origine, très rigide concernant une donnée anonyme, puisque la définition avait trait à une obligation de résultat (impossibilité que quiconque puisse remonter à la donnée originale). Nous verrons qu'une telle contrainte n'est pas compatible avec les méthodes d'anonymisation proposées, qui conservent toujours un risque de réidentification. Par conséquent, avec cette définition, toute donnée devrait être considérée comme une donnée personnelle, ce qui n'est clairement pas l'objectif recherché. En effet, nous allons voir qu'il est possible de quantifier la probabilité du risque de réidentification, et on pourra ainsi estimer qu'une donnée sera anonyme si ce risque est acceptable.

Exemple 1 (Données anonymes pré-RGPD)**.** *Un enregistrement, même utilisant un algorithme de chiffrement réputé sûr comme AES, ayant produit le n-uplet suivant (Nom : sEujSEsWzHioxae70aKE6w==, Âge : 27, Salaire : 36800K, Département :*

75) n'est pas anonyme *dans la mesure où la personne ayant chiffré le n-uplet est capable de le déchiffrer, puisqu'elle connaît la clé de chiffrement utilisée*[4].

Depuis mai 2018 et l'entrée en vigueur du règlement général sur la protection des données [3] (RGPD ou GDPR en anglais), cette obligation de résultat s'est transformée en obligation de moyens (cf. RGPD, considérant 26), éclairant l'interprétation de la signification de ce qu'est une donnée identifiante : « *Il y a lieu d'appliquer les principes relatifs à la protection des données à toute information concernant une personne physique identifiée ou identifiable. (...) Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que le ciblage. Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci* ». La définition d'une donnée identifiante dans la loi française est désormais celle référencée dans l'article 4 du RGPD. Il semble donc raisonnable de considérer la définition *relaxée* à l'obligation de moyens. Faute de cette interprétation, chercher à anonymiser des données tout en conservant une certaine utilité ne fait pas sens. Il faut donc accepter une part de risque de réidentification lorsqu'un processus d'anonymisation a lieu, et chercher à quantifier la probabilité de désanonymisation, l'impact, mais également le bénéfice qui pourra être obtenu suite au traitement des données. La définition de l'anonymisation se fait donc par rapport à la question de la réidentification.

Exemple 2 (Données anonymes post-RGPD)**.** *Considérons la table 2, générée en anonymisant une partie des données*[5] *de la table 1 (on a gardé uniquement les joueurs du PSG). La première ligne est une valeur hachée du nom utilisant une clé de chiffrement qui a été détruite (effacée). Dans ce cas, même si un attaquant connaît l'âge de la personne et son club, elle ne sera pas capable de déduire son salaire avec une probabilité supérieure à 50 %. Par exemple, Cavani, âgé de 32 ans et jouant au PSG peut avoir un salaire de 1,16 million ou 1,5 million.*

4. Pour ceux qui s'y intéressent, la clé de chiffrement utilisée ici est une clé basique : « ZZZZZZZZZZZZZZZZ ».

5. Ces salaires sont disponibles en accès libre dû au règlement sur le *fair play financier* de l'UEFA (Union européenne des associations de football). On pourrait s'interroger sur l'impact de ce règlement sur le droit à la vie privée des joueurs. Les données ont été extraites de `https://www.linternaute.com/sport/foot/1433459-salaires-de-ligue-1-le-nom-du-joueur-le-mieux-paye-revele-vendredi/`.

ID	*Âge*	*Club*	*Salaire*
Thiago Silva	*35*	*PSG*	*1160K*
Edison Cavani	*32*	*PSG*	*1500K*
Kylian Mbappé	*20*	*PSG*	*1730K*
Neymar Jr.	*27*	*PSG*	*3060K*
Dimitri Payet	*32*	*OM*	*500K*
Luiz Gustavo	*32*	*OM*	*500K*

TABLE 1. Salaires 2019

ID	*Âge*	*Club*	*Salaire*
dH7SdankcIhHDE1ATvErkg	*[30;39]*	*PSG*	*1160K*
fTRVz9bY2mHguqsmuPHtvw	*[30;39]*	*PSG*	*1500K*
x4TUcj1FQZkfSfnIeL05NA	*[20;29]*	*PSG*	*1730K*
jtILvRsZLVUETwlExAzpvw	*[20;29]*	*PSG*	*3060K*

TABLE 2. Salaires 2019 du PSG, table anonymisée

Objectif d'un processus d'anonymisation

Nous notons $\mathscr{D}$ une base de données, et $\mathscr{A}_X(\mathscr{D})$ sa version anonymisée selon un mécanisme $\mathscr{A}_X$.

Definition 1 (Réidentification)**.** *La réidentification est un processus (ou algorithme) prenant en entrée un jeu de données (anonymes), des connaissances annexes et cherchant à apparier des données anonymes avec des individus du monde réel.*

Exemple 3 (Réidentification)**.** *Considérons de nouveau la table 2. Supposons que la connaissance annexe d'un attaquant soit de savoir qu'il n'y a que trois joueurs du PSG qui gagnent plus de 1,5 million d'euros par mois (Neymar Jr., Mbappé et Cavani), et que l'âge de Cavani est de 32 ans, alors il est possible à cet attaquant de déduire le salaire de Cavani de la table 2 (1,5 million).*

Un processus d'anonymisation doit idéalement protéger contre toute réidentification, et doit *a minima* quantifier le risque de réidentification de la base de données. Cette réidentification peut être quantifiée par rapport à différents modèles [19] : le modèle du journaliste, du procureur, et du marketeur. Le modèle du journaliste considère que l'attaque est réussie si l'attaquant (le journaliste) arrive à désanonymiser n'importe quelle personne du jeu de données, et à la retrouver dans le monde réel. Le modèle du procureur considère que l'individu auquel on s'intéresse est dans la base de données, qu'on connaît toutes les informations sur lui, et que l'attaque est réussie si on réussi à le désanonymiser. Le modèle du marketeur cherche à apparier un maximum d'enregistrements pour lesquels on ne connaît que le quasi-identifiant (voir plus bas).

Âge	*Salaire*
35	*1160K*
32	*1500K*
20	*1730K*
27	*3060K*

TABLE 3. Résultats exacts : $Q1$

Exemple 4 (Risque du journaliste)**.** *Toujours dans l'exemple de la table 2, le modèle du journaliste fera comme hypothèse que le journaliste connaît toutes les données publiques relatives aux individus présents dans la base. Il connaîtra donc l'âge et l'équipe des individus : (Mbappé, 20, PSG), (Neymar Jr., 27, PSG), (Cavani, 32, PSG), (Silva, 35, PSG). Ici, il est incapable de différencier Cavani de Silva et Mbappé de Neymar Jr. soit une chance sur deux à chaque fois. Le risque de retrouver au moins l'un des quatre individus est donc égal à :* $R_J = 1 - (\frac{1}{2})^2 = 0.75$.

Exemple 5 (Risque du procureur)**.** *Le modèle du procureur fera comme hypothèse que le procureur connaît toutes les données publiques relatives à l'individu qui l'intéresse présent dans la base. Il connaîtra donc l'âge et l'équipe de : (Mbappé, 20, PSG). Ici, il est incapable de différencier Mbappé de Neymar Jr. Le risque de retrouver le salaire de Mbappé est donc égal à :* $R_P('Mbappe') = \frac{1}{2} = 0.5$. *On voit que le risque du procureur dépend de l'individu qu'il cherche à réidentifier. On peut donc définir le risque maximal du procureur comme étant* $R_P = \max_{i \in \mathscr{D}}(R_P(i))$.

Exemple 6 (Risque du marketeur)**.** *Le modèle du marketeur fait la même hypothèse que celui du journaliste, toutefois il cherche à calculer le nombre d'appariements réussis. On peut le voir comme une espérance normalisée du nombre de réidentifications réussies.* $R_M = \Sigma_{i \in \mathscr{D}} p(i)/|\mathscr{D}| = 0.5$.

Definition 2 (Utilité)**.** *L'utilité* $U(r(\mathscr{D}))$ *d'une requête* $r(\mathscr{D})$ *d'analyse de données se mesure par une métrique* `M` *qui indique la différence entre la valeur de* $U(r(\mathscr{D}))$ *et la valeur de* $U(r'(\mathscr{A}_X(\mathscr{D})))$*, où* r' *peut être* r *ou une version modifiée de* r *pour prendre en compte le processus d'anonymisation.*

Exemple 7 (Utilité du calcul du salaire par âge)**.** *Supposons que nous souhaitons calculer le salaire par âge sur la base de données constituée des quatre joueurs de notre exemple. Le résultat est indiqué dans la table 3.*

Si nous lançons la même requête sur les données anonymisées nous obtiendrons le résultat présenté dans la table 4. Nous pouvons évaluer l'utilité du deuxième calcul grâce à la fonction d'utilité suivante, qui calcule la moyenne de l'erreur normalisée (toute autre fonction d'évaluation d'erreur pourrait être pertinente) : $\mathtt{M}(U(Q_1(\mathscr{D}), U(Q'_1(\mathscr{A}_X(\mathscr{D}))))) = 1/4 \times (|\frac{1330-1160}{1160}| + |\frac{1330-1500}{1500}| + |\frac{2395-1730}{1730}| + |\frac{2395-3060}{3060}|) = 0.215$

Âge	*Salaire*
[30;39]	*1330K*
[20;29]	*2395K*

TABLE 4. Résultats sur données anonymes : Q'_1

Il est entendu que l'objectif de la publication d'une base de données est d'effectuer des analyses de données la concernant. Il s'agit donc pour le responsable de l'anonymisation de proposer un *compromis* acceptable (et le meilleur possible) entre d'une part la sécurité (anonymat) de la base de données, c'est-à-dire la difficulté de la réidentification, et d'autre part l'utilité (estimée ou calculée) de la base anonymisée.

Approche Pratique

Du cas par cas. Il n'existe malheureusement pas de solution d'anonymisation universelle qui s'appliquerait à tous les types d'applications et de données. Une solution d'anonymisation est souvent le résultat d'un (long) travail d'optimisation entre les garanties en terme de sécurité qu'on souhaite fournir et l'utilité des données. Une solution d'anonymisation doit donc être développée au cas par cas et adaptée aux usages prévus et aux données traitées. Par exemple, des données sensibles, comme des données de santé, nécessiteront probablement des solutions d'anonymisation plus robustes que des données moins sensibles, comme par exemple des données de mobilité dans un musée. Par ailleurs, le type de publication envisagée est aussi important à considérer : des données publiées sur le Web, en libre accès, nécessiteront des garanties plus fortes que des données partagées avec un partenaire industriel avec qui un contrat juridique pourra être éventuellement signé.

Anonymisation et analyse de risques. Il est donc souvent souhaitable de combiner le travail de conception de solution d'anonymisation avec une analyse de risques. Cette tâche consiste à décrire précisément la nature des données utilisées, les traitements mis en œuvre, les différents acteurs, en considérant tant les aspects techniques qu'opérationnels. Il conviendra ensuite d'évaluer les risques sur la sécurité des données (confidentialité, intégrité et disponibilité) ainsi que leurs impacts potentiels sur la vie privée. Ce travail d'analyse permettra de développer la solution la mieux adaptée aux besoins et aux contraintes existants [10].

Évaluation. Afin d'évaluer les solutions d'anonymisation, le G29 (le groupement des autorités de protection des données européennes) propose trois critères : l'individualisation, la corrélation et l'inférence [10]. Ainsi, pour « démontrer » qu'une solution est correcte et conforme au RGPD, il faut démontrer que les données anonymisées ne permettent plus d'isoler les données qui appartiennent à un individu, ne permettent plus de relier entre eux des ensembles de données distincts concernant un même individu et ne permettent pas non plus de déduire de l'information sur un

individu. Dans le cas où un des trois critères n'est pas respecté, le G29 stipule que les données ne pourront être considérées comme anonymisées uniquement si une analyse détaillée démontre que les risques de réidentification sont maîtrisés ! Par ailleurs, étant donné que les techniques de réidentification s'améliorent, il est indispensable de réévaluer régulièrement le caractère anonyme des données produites.

La pseudonymisation

Une erreur courante consiste à considérer la pseudonymisation comme une solution d'anonymisation. La pseudonymisation est une technique simple qui consiste à remplacer la valeur d'un attribut « identifiant » (par exemple un nom) par une autre valeur, un « pseudo » (comme le cas de l'exemple 1). Ce « pseudo » peut être généré indépendamment de la valeur d'origine ou en être dérivé (par exemple, en appliquant une fonction de hachage). Clairement, le pseudonymisation permet de réduire le risque de mise en corrélation d'un ensemble de données avec l'identité originale d'un sujet, mais une réidentification indirecte est possible en utilisant, par exemple, les autres attributs [14]. Par conséquent, il est important de rappeler que la pseudonymisation est une mesure de sécurité utile, et qu'il faut encourager, mais ne constitue pas une solution d'anonymisation à part entière.

Techniques d'anonymisation

L'anonymisation moderne : naissance du k-anonymat

On peut dater la problématique moderne de l'anonymisation de la fin des années 1990 avec la publication par Sweeney de plusieurs articles où elle propose le concept de k-anonymat (*k-anonymity*), dont nous citons le plus connu [32] publié en 2002. En effet, à l'époque, lorsqu'on parlait d'anonymisation, on faisait référence au concept désormais connu sous le nom de *pseudonymisation*. Comme indiqué précédemment, la pseudonymisation est le remplacement de toutes les données directement identifiantes (comme le numéro de sécurité sociale) par une valeur aléatoire (pseudonyme). Sweeney a montré dans [32] qu'il était possible de réidentifier une base de données relationnelle pseudonymisée (SQL ou plus généralement un fichier tabulaire) en utilisant ses *quasi-identifiants*.

Definition 3 (Quasi-identifiant – QID)**.** *Soit $T(A_1,\ldots,A_n)$ une table composée de n attributs A_i. On appelle quasi-identifiant Q de T un ensemble d'attributs $Q = \{A_i,\ldots,A_j\} \subseteq \{A_1,\ldots,A_n\}$ dont la publication doit être contrôlée de manière suivante : Pour être un quasi-identifiant, Q doit être tel que*

```
SELECT A1,...,Aj
FROM T
GROUP BY A1,...,Aj
HAVING COUNT = 1
```

retourne un résultat non vide. Étant donné un ensemble de j attributs (e.g. DateNaissance, CP, Sexe), il existe dans l'instance de la table au moins une ligne qui ne partage pas ses valeurs avec une autre ligne de la table (e.g. le triplet (DateNaissance : 28/2/1980, CP : 18000, Sexe : F) n'apparaît qu'une seule fois).

On pourra noter, en terme de conception de base de données, que si Q est une clé alors Q est un QID mais pas forcément l'inverse. Si Q était une clé alors la requête précédente donnerait très exactement la projection de T sur Q.

On peut décider, indépendamment de toute connaissance annexe, que Q est un QID, puisqu'il s'agit d'une propriété de l'instance de la table. Par contre, le risque de réidentification est différent et dépend de la connaissance annexe de l'attaquant : si on identifie un QID mais qu'on est *sûr* que l'attaquant ne possède aucune base avec ce même QID, on pourra estimer que le risque de réidentification via ce QID est faible. À l'inverse, identifier un QID sur une table contenant des données sensibles, ainsi qu'une base de données annexe contenant ce même QID et des données identifiantes présente un risque réel.

Exemple 8 (QID des footballeurs)**.** *On peut considérer que le couple (âge, club) forme un quasi-identifiant pour la table 1. Ainsi, Thiago Silva est le seul joueur du PSG à être âgé de 35 ans. Mbappé est le seul joueur du PSG à être âgé de 20 ans. Toutefois il y a plusieurs joueurs âgés de 27 ans (Neymar, Kurzawa, Sarabia).*

Rappelons que la logique de la définition repose sur le fait qu'il est possible de retrouver dans une *autre* base de données à la fois Q mais aussi un attribut (ou un ensemble d'attributs comme le couple nom/prénom) identifiant. Notons que si Q_1 est un quasi-identifiant et $Q_1 \subseteq Q_2$, alors Q_2 est un quasi-identifiant. Lors de l'étude des quasi-identifiants, il convient donc de s'intéresser au(x) quasi-identifiant(s) minimal(aux) par rapport à l'inclusion.

Sweeney définit alors le critère de k-anonymat, et propose une garantie de sécurité par rapport à la réidentification.

Definition 4 (k-anonymat)**.** *La publication d'une version anonyme $P_{\mathscr{D}}$ d'une base de données $\mathscr{D}$ respecte le critère de k-anonymat par rapport à un quasi-identifiant Q si et seulement si chaque valeur de $q \in Q$ dans $P_{\mathscr{D}}$ apparaît au moins k fois. On parle alors de* classe d'équivalence *pour tous les n-uplets qui ont la même valeur q.*

Exemple 9 (2-anonymat des footballeurs)**.** *La table 5 est 2-anonyme par rapport au QID (**âge**, **club**). On voit toutefois qu'on a été obligé de modifier le domaine de définition de l'attribut **âge** de l'année à la décennie.*

Le k-anonymat peut être vu comme une *contrainte* que doit respecter une version publiée du jeu de données. Plusieurs algorithmes permettant de respecter une telle contrainte existent : des algorithmes basés sur de la suppression (effacement de la valeur d'un attribut ou effacement d'un n-uplet), des algorithmes basés sur de

Âge	*Club*	*Salaire*
[30;39]	*PSG*	*1160K*
[30;39]	*PSG*	*1500K*
[20;29]	*PSG*	*1730K*
[20;29]	*PSG*	*3060K*

TABLE 5. Données 2-anonymes

la généralisation (modification de la valeur d'un attribut pour le généraliser, tout en conservant la même signification, par exemple généralisation de la commune vers le département ou la région), ou des algorithmes combinant ces deux techniques, comme le préconise Sweeney dans [32]. Une fois qu'on a procédé à cette transformation pour respecter la contrainte sur les quasi-identifiants, on peut publier les données, en leur associant les autres informations (considérées comme des données sensibles).

Optimalité des algorithmes. L'une des questions débattues au sein de la communauté était celle de *l'optimalité* d'une telle anonymisation, c'est-à-dire une transformation qui ferait perdre le moins d'informations, tout en respectant la contrainte [27]. Meyerson et Williams ont montré que trouver la valeur optimale est difficile (NP-difficile), même s'il existe des $O(k \log k)$-approximations calculables en temps polynomial.

Risque de réidentification. La protection *revendiquée* par une base de données anonymisée selon la contrainte de k-anonymat est que chaque n-uplet étant confondu avec $k-1$ autres, la probabilité de retrouver le n-uplet correct si on connaît les valeurs exactes du quasi-identifiant est de $1/k$. Il est important de souligner que la garantie proposée, comme toutes les garanties en anonymisation, est une garantie *probabiliste*. Il convient donc au responsable de traitement mettant en œuvre l'anonymisation de décider du risque de réidentification qu'il est prêt à accepter, et de choisir le paramètre de k en conséquence.

Faiblesse du modèle du k-anonymat

L'intérêt principal du k-anonymat est qu'il est facile à comprendre. Les algorithmes permettant de l'implémenter sont également assez rapides (il ne faut guère plus de quelques secondes pour anonymiser une base de données de plusieurs dizaines ou centaines de milliers de lignes). Toutefois, le modèle n'est pas robuste par rapport aux *attaques d'homogénéité*, comme proposées par Machanavajjhala *et al.* [26]. Une telle attaque a lieu lorsque les valeurs sensibles associées à une valeur donnée de quasi-identifiant sont toutes identiques. Dans ce cas, on peut déduire que toutes les personnes ayant cette valeur de quasi-identifiant ont la même donnée sensible, qu'on est capable de déduire.

Âge	*Club*	*Salaire*
[30;39]	*PSG*	*1160K*
[30;39]	*PSG*	*1500K*
[20;29]	*PSG*	*1730K*
[20;29]	*PSG*	*3060K*
[32]	*OM*	*500K*
[32]	*OM*	*500K*

TABLE 6. Données 2-anonymes avec un risque de réidentification

Il est donc impossible de donner *a priori* une garantie sur le risque de réidentification, ce qui fait qu'une utilisation du *k*-anonymat seul ne présente pas de garanties d'anonymat raisonnables.

Exemple 10 (Les footballeurs marseillais)**.** *Considérons la base de données contenant des joueurs d'autres clubs, construite à partir de l'intégralité de la table 1, et présentée de manière 2-anonyme dans la table 6. Si on sait que seuls les joueurs Payet et Gustavo ont 32 ans et jouent à Marseille, on sera capable de déduire que leur salaire est de 500, puisqu'ils ont tous les deux le même salaire.*

Extensions du modèle du k-anonymat

De multiples modèles ont été proposés afin de se prémunir contre les attaques d'homogénéité, en particulier la ℓ-diversité [26] (ℓ-*diversity*), la t-proximité [25] (t-*closeness*), la δ-divulgation [11] (δ-*disclosure*), la β-ressemblance [12] (β-*likeness*). Tous ces modèles rajoutent des contraintes sur les valeurs sensibles des classes d'équivalence. Considérons la plus simple, la ℓ-diversité.

Definition 5 (ℓ-diversité)**.** *Une classe d'équivalence respecte la contrainte de ℓ-diversité si elle contient au moins ℓ valeurs « représentatives » pour la donnée sensible. Une base de données (ou table) est dite ℓ-diverse si toutes ses classes d'équivalence respectent la contrainte de ℓ-diversité.*

Exemple 11 (ℓ-diversité des footballeurs)**.** *La table 5 est* 2*-anonyme et* 2*-diverse, car chaque classe d'équivalence est composée d'au moins deux n-uplets et chaque classe d'équivalence est associée à au moins deux valeurs sensibles différentes. À noter qu'il y a ici deux classes d'équivalence : ([30;39], PSG) et ([20;29], PSG).*

La table 6 est certes 2*-anonyme, mais n'est que* 1*-diverse puisque pour la classe d'équivalence ([32], OM), il n'y a qu'une seule valeur sensible :* 500*K*.

Il est ensuite possible de discuter de ce que signifie précisément « représentatives » ou combien de fois ces valeurs représentatives doivent apparaître dans une classe d'équivalence pour que le critère de ℓ-diversité soit atteint. Si on comprend

bien l'objectif que cherche à remplir ce modèle, il faut prendre garde à traiter les données sensibles en prenant en compte leur sémantique (d'où la question de données « représentatives »). En effet, il ne s'agit pas seulement d'avoir ℓ valeurs sensibles *syntaxiquement* différentes, encore faut-il qu'on ne puisse pas déduire des informations sensibles concernant les utilisateurs, comme par exemple que tous les utilisateurs d'une classe d'équivalence sont tous atteints d'une pathologie grave ou chronique.

Definition 6 (t-proximité)**.** *Une classe d'équivalence respecte la contrainte de t-proximité si la distance entre la distribution de chaque attribut sensible de cette classe et la distribution de chaque attribut sensible de la table complète ne dépasse pas un seuil t. Une base de données (ou table) respecte la contrainte de t-proximité si toutes ses classes d'équivalence respectent la contrainte de t-proximité.*

Exemple 12 (t-proximité)**.** *Le développement d'un exemple de t-proximité étant un peu long, nous référons le lecteur à l'article de Li* et al. [25] *pour un exemple sur des données médicales.*

La t-proximité précise la définition de « représentativité » des valeurs, en obligeant la distribution des données sensibles de chaque classe d'équivalence à ressembler, à un facteur t près, à la distribution générale de cette même donnée sensible. Commence à se poser alors la question de l'utilité des données. Sous contrainte de t-proximité, les données ne paraissent pas forcément directement exploitables. Toutefois, il reste possible de dégager des tendances, ou d'effectuer des calculs généraux ou corrélations sur l'ensemble de la table.

Il peut être délicat de savoir comment paramétrer la valeur du t de ce modèle. C'est l'objectif du modèle de δ-divulgation : quantifier le gain d'information d'un attaquant qui observe les classes d'équivalence, et qui connaît aussi la distribution des valeurs sensibles.

Definition 7 (δ-divulgation)**.** *Soit une valeur sensible v_i avec une fréquence p_i dans la base de données originale, et une fréquence $q_{i,j}$ de cette valeur dans une classe d'équivalence Ec_j. La classe d'équivalence Ec_j est dite δ-divulgation-privée si et seulement si* $\forall v_i, |\log(q_{i,j}/p_i)| < \delta$.

Exemple 13 (Les limites de la δ-divulgation)**.** *Malheureusement, si p_i est grand et même pour un δ petit, il n'y a pas de borne maximale réelle sur $q_{i,j}$. Par exemple, on peut choisir $p_i = 0.5$ et $q_{i,j} = 1$ et $\delta = 0.5$ et on aura bien* $\log(1/0.5) = log(2) = 0.3$*, or si $q_{i,j} = 1$, cela signifie qu'on connaît avec certitude la valeur de la donnée sensible.*

D'autres modèles existent permettant, par exemple, de mesurer certains critères comme la probabilité d'appartenance d'un individu au jeu de données échantillonné par rapport à un ensemble de personnes. C'est le cas de la δ-présence [29] (δ-*presence*).

Publications successives

La publication successive d'un jeu de données (par exemple la liste des malades d'un hôpital tous les mois) donne lieu à un problème difficile d'anonymisation. En effet, la publication de deux jeux de données peut donner lieu à une attaque par *différence*, qui consiste a faire la différence entre les deux jeux de données. Si l'attaquant connaît les individus qui sont arrivés ou sont partis de l'hôpital entre-temps, il sera capable de remonter à l'ensemble de leurs données sensibles. Certains modèles comme la *m*-invariance [34] ont été proposés, mais leur utilité est assez réduite, puisque les publications multiples accumulent des données qui ont déjà été publiées, afin de se protéger. La meilleure manière de faire des publications multiples sera d'utiliser la technique de confidentialité différentielle que nous détaillons dans la suite.

Differential Privacy [6]

Intérêt et définition

La DP, introduite par Dwork [17, 18], n'est pas un modèle d'anonymisation, mais la caractéristique d'une opération (ou exécution d'un algorithme) sur des données qui présentent certaines garanties de confidentialité. Il est donc tout à fait possible d'avoir des algorithmes cherchant à atteindre un modèle spécifique d'anonymat, tout en proposant des garanties de DP. Par exemple, Domingo-Ferrer et Soria-Comas montrent qu'il est possible d'atteindre des garanties de DP avec le modèle de la *t-closeness* [16].

La DP est très en vogue, car elle permet de quantifier un risque de réidentification « absolu ». Toutefois, le paramétrage de ce risque de réidentification n'est pas simple, en particulier lorsque l'ensemble des données sensibles est grand. En effet, la DP fonctionne au mieux lorsqu'il n'y a qu'un faible nombre de valeurs pour les données sensibles (par exemple vrai/faux, masculin/féminin...).

Considérons un algorithme *ALG* et deux bases de données $\mathscr{D}_1$ et $\mathscr{D}_2$, telles que $\mathscr{D}_1 = \mathscr{D}_2 \cup d$ et $d \notin \mathscr{D}_2$. La garantie que cherche à fournir la DP est qu'en observant $\Omega = ALG(\mathscr{D})$, le résultat de l'exécution de *ALG* sur un jeu de données $\mathscr{D}$, il sera *très difficile* de savoir si $\mathscr{D} = \mathscr{D}_1$ ou $\mathscr{D} = \mathscr{D}_2$. Dit autrement, Ω ne doit pas changer beaucoup, selon que d est présent ou absent de la base de données utilisée en entrée de *ALG*. Un seul individu ne doit donc pas foncièrement changer la valeur de l'exécution de *ALG*.

6. Nous utilisons ici le terme anglais, ou son abréviation *DP*. La traduction française communément admise est *confidentialité différentielle.*

Definition 8 (ε, δ*-differential privacy*)**.** *Soient* $\varepsilon \in \mathbb{R}^+$ *et* $\delta \in \mathbb{R}^{*+}$*. On dit qu'un mécanisme ALG respecte la contrainte de* ε, δ*-differential privacy si et seulement si* $Pr[ALG(\mathscr{D}_1) = \Omega] \leqslant \exp(\varepsilon) \times Pr[ALG(\mathscr{D}_2) = \Omega] + \delta$*. Si* $\delta = 0$*, on parle alors simplement d'*ε*-differential privacy.*

Note : $Pr[ALG(\mathscr{D}_1) = \Omega]$ signifie « la probabilité d'observer Ω comme résultat de l'exécution de *ALG* sur la base de données $\mathscr{D}_1$ ».

Il découle de la définition de la DP que l'algorithme *ALG* doit être un algorithme *probabiliste et non déterministe*, c'est-à-dire que plusieurs exécutions successives de l'algorithme sur la même entrée peuvent produire des résultats différents. En effet, si *ALG* était déterministe, on ne pourrait pas choisir ε aussi petit qu'on veut.

Il est également très important de souligner que les garanties de DP s'appliquent également à tous les *post-traitements*. C'est-à-dire que tous les algorithmes d'analyse de données exécutés sur une base de données qui est passée par un processus de ε*-differential privacy* produiront des résultats ayant cette même garantie, avec le même ε.

La DP en pratique

La garantie proposée par la DP est que la probabilité d'observer une valeur plutôt qu'une autre ne doit pas être sensiblement différente (i.e. à $\exp(\varepsilon)$ près) selon qu'un individu est présent ou pas. Nous allons voir comment cela peut être appliqué en pratique.

Exemple 14 (Un exemple d'algorithme DP : la réponse aléatoire à un sondage)**.** *Considérons l'algorithme classique de* réponse aléatoire à un sondage *(ou* randomized response, RR*). Son objectif est de protéger les réponses des individus à une question de type* `vrai/faux` *selon le processus suivant :*

> *(1) La personne lance (secrètement) une pièce de monnaie, si elle tombe sur face, elle donne la vraie réponse. (On considère ici une probabilité égale pile/face).*
>
> *(2) Si la pièce de monnaie tombe sur pile, alors la personne relance une pièce. Si cette fois c'est face, elle répond* `vrai`*, si c'est pile, elle répond* `faux`*.*

Une rapide analyse montre que si une personne a répondu `vrai` à la question, alors la probabilité que ce soit la vérité est de 0.75, alors que la probabilité qu'en vérité ce soit `faux` est de 0.25. Si ce qui nous intéresse est la proportion totale dans un ensemble de personnes de `vrai/faux` alors on voit qu'on est en mesure d'estimer cette valeur par rapport aux réponses observées : si x est le nombre de personnes ayant la propriété `vrai`, x_O le nombre de personnes observées ayant répondu `vrai` et n le nombre total de personnes, alors on est capable d'estimer $x \approx x_{est} = 2 \times x_O - n/2$.

Par rapport au ε, il faut comparer les valeurs possibles d'espérance et de variance pour les jeux de données contenant un individu particulier, et un jeu de données ne le

comprenant pas. Considérons l'objectif qui est de calculer la fonction x_{est}, que l'on atteint en mesurant x_O. Toutes choses étant égales par ailleurs, la réalisation de x_O pour une base de données de $\mathscr{D}_1$ contenant un individu ayant la caractéristique `vrai` et un jeu de données $\mathscr{D}_2$ où il n'est pas présent est supérieure dans 75 % des cas et identique dans 25 % des cas. Ainsi, avec les paramètres de notre exemple, on aura trois fois plus de chances de tomber juste en choisissant `vrai` si la réalisation de l'algorithme est supérieure à l'espérance. En d'autres termes, $\varepsilon = \ln(3) \approx 1.09$.

Réciproquement, on peut aussi calculer la probabilité P qu'il faut appliquer au premier jet avant de relancer pour un ε donné. On pourrait montrer que pour des valeurs de ε très proches de 0, il faut choisir $P = 2\varepsilon$.

Évaluer la protection statistique à partir de ε

Il est intéressant de noter, par rapport à cet exemple, que nous avons discuté de la valeur de ε mais que ce qui nous intéresse véritablement est le risque de retrouver la valeur de la donnée sensible, soit ici la valeur $P = 0.75$. P peut se calculer à partir de ε, sachant que dans notre exemple $P + \bar{P} = 1$ et $P = exp(\varepsilon) \times \bar{P}$, soit $P = 3 \times (1 - P) = 0.75$.

D'une manière plus générale, il est également possible de remonter à cette probabilité dans le cas où le nombre de valeurs de la donnée sensible est fini, et vaut n.

Théorème 1 (Probabilité de réidentification)**.** *Soit n valeurs pour une donnée sensible, et soit x_O une observation d'un algorithme respectant la contrainte de DP avec une valeur ε. La probabilité de pouvoir déduire la valeur de la donnée sensible à partir de cette observation est de $P \leqslant \frac{exp(\varepsilon)}{exp(\varepsilon)+n-1}$*

Démonstration. Soit P_1 la probabilité de retrouver la valeur correcte. Soient $P_2, \ldots, P_n$ les autre probabilités. On a $\Sigma_{i \in [1;n]} P_i = 1$ et $\forall j > 1, P_1 \leqslant exp(\varepsilon) \times P_j$. Si on considère que $\forall j > 1, k > 1, P_j = P_k$ alors on a $P_1 \leqslant \frac{exp(\varepsilon)}{exp(\varepsilon)+n-1}$. □

On voit que dans le cas $n = 2$ et $\varepsilon = log(3)$ on retrouve bien $P_1 \leqslant 3/4$. On voit aussi, que pour des jeux de données où il y a un grand nombre de valeurs de données sensibles, un ε de l'ordre de 10 peut fournir une bonne protection. Par exemple si on a $n = 10^6$ et $\varepsilon = 10$ alors $P \approx 0.0216$.

Théorème de composition

La DP permet de résoudre le problème de la publication successive.

Théorème 2 (Composition de la DP)**.** *Un processus $ALG'(\mathscr{D})$ composé de l'exécution successive de k fois le même processus sur le même jeu de données $ALG(\mathscr{D})$ respectant la contrainte d'ε-DP respectera une contrainte de $(k \times \varepsilon)$-DP.*

Note : si les processus sont exécutés sur des jeux de données indépendants, alors on prend simplement la valeur maximale des ε mis en jeu.

Exemple 15 (Réponses aléatoires multiples à un sondage)**.** *Reprenons l'exemple précédent, et appliquons exactement le même processus aléatoire, mais deux fois de suite. Supposons que la réponse véritable soit* `vrai`*. Nous répondons donc la vérité la première fois avec une probabilité 0.75 et de même la seconde fois. La valeur de* ε *devient donc* $2\log(3)$*. On comprend assez naturellement que* ε *augmente à chaque fois qu'on observe de nouveau le résultat de l'algorithme, et en effet, si on exécute l'algorithme un grand nombre de fois, il paraît assez naturel de penser que la valeur apparaissant le plus fréquemment sera la réponse véritable de l'individu.*

Grâce à la DP, on est donc en mesure de quantifier très exactement le coût (en termes d'augmentation du ε global) de la publication d'une donnée. Toutefois, le théorème de composition de la DP ne doit pas faire croire qu'il est possible de publier à l'infini des données en respectant des garanties fortes. Il est donc possible d'exploiter le théorème de deux manières :

(1) On se donne à l'avance un certain « budget » (sous la forme d'un ε), et on s'autorise k publications. Dans ce cas, chaque publication p_i devra utiliser un $\varepsilon_i = \varepsilon/k$.

(2) On effectue k publications avec divers ε_i. Alors, au final, on peut estimer le « risque » encouru $\varepsilon = \Sigma_i \varepsilon_i$.

Quelques exemples d'utilisation de la DP

Suite à des attaques sur son recensement de 2000 [15], le bureau du recensement américain expérimente depuis 2018 l'utilisation de garanties de DP qu'elle souhaite mettre en œuvre pour son recensement de 2020 [5]. L'un de ses problèmes principaux est de bien régler le compromis entre protection et utilité des données.

En 2016, Google a proposé un framework nommé RAPPOR [20] pour récupérer des données des utilisateurs (afin d'effectuer des calculs de statistiques) avec des garanties de DP. La valeur choisie dans ce cas de figure était de $\ln(3)$, avec une approche inspirée de l'algorithme de *réponse aléatoire*.

Apple a également été l'un des précurseurs à utiliser la DP dans ses algorithmes d'apprentissage sur les données des utilisateurs des iPhones [33]. Ils ont appliqué leurs algorithmes d'IA au calcul des emojis les plus fréquents par langue, aux profils de consommation énergétiques de leur navigateur, ou encore à la découverte de nouveaux mots (comme des noms propres) pour le correcteur orthographique. Dans ce cas de figure, les valeurs de ε utilisées sont entre 2 et 8, pour une utilité d'apprentissage donnée.

Enfin, on pourra mettre en garde contre les approches qui cherchent à maximiser l'utilité, sans toutefois utiliser la DP, mais qui peuvent néanmoins être jugées conformes par rapport à la législation, comme Diffix [21], et pour lesquelles des attaques de désanonymisation ont été démontrées possibles [23].

Logiciel de mise en œuvre des techniques d'anonymisation

Un outil *open source* développé par l'Université Technologique de Munich, ARX[7] [31], permet de réaliser des anonymisations selon de nombreux modèles à partir de données originales en format tabulaire. L'outil permet également d'estimer les risques de désanonymisation selon les modèles présentés plus haut, et même de manière plus fine en calculant la distribution des probabilités de désanonymisation et non simplement les valeurs maximales.

Conclusion

Cet article est introductif et ne décrit pas toutes les techniques d'anonymisation qui existent. Une description plus exhaustive peut être trouvée dans [22, 13, 10].

Nous avons essentiellement discuté, dans cet article, la question de l'anonymisation appliquée aux données tabulaires. Il y a de nombreux autres domaines, en particulier le domaine de l'anonymisation des données de géolocalisation, qui ont donné lieu à un grand nombre de publications ces dernières années [14, 28, 8, 7]. Ces données sont, par nature, difficiles à anonymiser car elles sont très identifiantes. En effet, des études ont montré que la connaissance de trois ou quatre points spatiaux-temporels d'une trajectoire suffisait pour réidentifier, avec une probabilité élevée, une personne dans une population de plusieurs millions d'individus [14]. Différents types d'anonymisation ont été proposés dans la littérature. Certaines solutions proposent de publier uniquement des statistiques sur les différentes trajectoires, comme leur longueur moyenne ou les endroits les plus souvent visités. D'autres approches proposent de publier des données synthétiques, c'est-à-dire des trajectoires générées artificiellement à partir des caractéristiques statistiques des vraies trajectoires [9]. Finalement, d'autres solutions proposent de modifier les trajectoires avant de les publier, par exemple, en groupant les trajectoires similaires [8] ou en y ajoutant du bruit [6].

Lors du choix de la technique d'anonymisation à mettre en œuvre, il convient de rappeler que seule la DP permet d'avoir une garantie (probabiliste) indépendante des connaissances des attaquants. Les autres modèles doivent faire des hypothèses sur la connaissance des attaquants, et peuvent donc être parfois contournés si un attaquant dispose d'informations qu'on ne l'imaginait pas avoir (ou qui sont collectées postérieurement à l'anonymisation).

Une question difficile à trancher est la valeur du ε pour la DP. Nous constatons, dans les faits, que les entreprises mettant en œuvre la DP optent pour des valeurs autour de $\varepsilon = 1$ ou plus, mais il faut bien comprendre que la sécurité liée à cet ε dépend aussi du nombre de valeurs différentes possibles parmi lesquelles il faut choisir. Ce n'est pas la même chose d'avoir $\varepsilon = 1$ avec deux valeurs (auquel cas

7. ARX est disponible sur `https://arx.deidentifier.org/`

comme on l'a vu avec l'exemple 14 de la réponse aléatoire, une réponse pourra être très fréquente de l'ordre de 75 % de chances d'observer la bonne valeur), et $\varepsilon = 1$ avec 1000 valeurs où toutes les probabilités seront de l'ordre de 10^{-3} et avec une valeur légèrement supérieure (3.10^{-3}) pour la bonne valeur. Quoi qu'il en soit, il faut choisir le plus petit ε possible permettant de continuer à obtenir des résultats exploitables pour l'algorithme d'analyse de données.

Finalement, il faut rappeler que l'anonymisation ne constitue qu'un élément dans une stratégie globale de gouvernance des données qui doit aussi considérer, par exemple, l'audit des systèmes, la gestion du stockage des données ou la gestion des droits d'accès [30].

Remerciements

Les auteurs souhaitent remercier Christine Froideveaux pour sa suggestion de l'écriture de cet article, pour sa relecture attentive et ses commentaires, et Cédric Eichler pour des discussions autour de l'applicabilité de la DP.

Références

[1] Loi n° 51-711 sur l'obligation, la coordination et le secret en matière de statistiques. *Journal Officiel de la République Française*, 6 juin 1951 :6013, 1951.

[2] Loi n° 78-17 relative à l'informatique, aux fichiers et aux libertés. *Journal Officiel de la République Française*, 6 janvier 1978.

[3] RÈGLEMENT (UE) 2016/679 DU PARLEMENT EUROPEN ET DU CONSEIL du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/ce (règlement général sur la protection des données)(texte présentant de l'intérêt pour l'eee). *Journal Officiel de l'Union Européenne*, 2016/679, 2016.

[4] Traité sur l'union européenne et du traité sur le fonctionnement de l'union européenne. *Journal Officiel de l'Union Européenne*, 2016/C 202/01, 2016.

[5] John M. Abowd. The U.S. census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, page 2867, 2018.

[6] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone : Uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, Mexico*, pages 376–385, 2008.

[7] G. Acs, G. Biczok, and C. Castelluccia. *Privacy-Preserving Release of Spatio-Temporal Density*. Handbook of Mobile Data Privacy, ISBN : 978-3-319-98161-1, Springer, 2018.

[8] Gergely Acs and Claude Castelluccia. A Case Study : Privacy Preserving Release of Spatio-temporal Density in Paris. In *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2014.

[9] Gergely Ács, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE Trans. Knowl. Data Eng.*, 31(6) :1109–1121, 2019.

[10] Article 29 Data Protection Working Party. Avis 05/2014 sur les techniques d'anonymisation, April 2014.

[11] Justin Brickell and Vitaly Shmatikov. The cost of privacy : Destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 70–78, New York, NY, USA, 2008. ACM.

[12] Jianneng Cao and Panagiotis Karras. Publishing microdata with a robust privacy guarantee. *Proc. VLDB Endow.*, 5(11) :1388–1399, July 2012.

[13] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2(1-2) :1–167, 2009.

[14] Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd : The privacy bounds of human mobility. *Scientific Reports, Nature*, March 2013.

[15] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210, 2003.

[16] Josep Domingo-Ferrer and Jordi Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization. *Knowl.-Based Syst.*, 74 :151–158, 2015.

[17] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, pages 1–12, 2006.

[18] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4) :211–407, 2014.

[19] Khaled El Emam and Fida Kamal Dankar. Research paper : Protecting privacy using k-anonymity. *JAMIA*, 15(5) :627–637, 2008.

[20] Giulia C. Fanti, Vasyl Pihur, and Úlfar Erlingsson. Building a RAPPOR with the unknown : Privacy-preserving learning of associations and data dictionaries. *PoPETs*, 2016(3) :41–61, 2016.

[21] Paul Francis, Sebastian Probst Eide, and Reinhard Munz. Diffix : High-utility database anonymization. In *Privacy Technologies and Policy - 5th Annual Privacy Forum, APF 2017, Vienna, Austria, June 7-8, 2017, Revised Selected Papers*, pages 141–158, 2017.

[22] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing : A survey of recent developments. *ACM Comput. Surv.*, 42(4) :14 :1–14 :53, June 2010.

[23] Andrea Gadotti, Florimond Houssiau, Luc Rocher, Benjamin Livshits, and Yves-Alexandre de Montjoye. When the signal is in the noise : Exploiting diffix's sticky noise. In *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019.*, pages 1081–1098, 2019.

[24] INSEE. Guide du secret statistique, 2018.

[25] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness : Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115, 2007.

[26] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *L*-diversity : Privacy beyond *k*-anonymity. *TKDD*, 1(1) :3, 2007.

[27] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. ACM.

[28] Darakhshan J. Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N. Wright. Dp-where : Differentially private modeling of human mobility. In *BigData Conference*, pages 580–588, 2013.

[29] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 665–676, New York, NY, USA, 2007. ACM.

[30] Institute of Medicine. *Sharing Clinical Trial Data : Maximizing Benefits, Minimizing Risk*. The National Academies Press, Washington, DC, 2015.

[31] Fabian Prasser and Florian Kohlmayer. Putting statistical disclosure control into practice : The ARX data anonymization tool. In *Medical Data Privacy Handbook*, pages 111–148. 2015.

[32] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5) :571–588, 2002.

[33] Differential Privacy Team. Learning with privacy at scale. 1(8), 2017.

[34] Xiaokui Xiao and Yufei Tao. M-invariance : Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 689–700, New York, NY, USA, 2007. ACM.