CHAPITRE **12**

Publication de données : dispose-t-on de techniques d'anonymisation suffisamment robustes ?

Benjamin Nguyen

Laboratoire d'informatique fondamentale d'Orléans, (EA 4022),

INSA Centre Val de Loire et Université d'Orléans

I – De la pseudonymisation à l'anonymisation

Historiquement, la publication de documents contenant des données personnelles précède l'utilisation massive de l'informatique. On pourra penser notamment au caviardage de documents officiels lors de leur diffusion auprès du grand public, lorsque certaines parties sont jugées trop sensibles, souvent du fait qu'elles révèlent les identités de certains protagonistes. La suppression de certains mots – en particulier les noms des individus concernés, ou les lieux – dans un texte ou un document, a longtemps été l'unique approche de protection utilisée lors de la publication de données sensibles. Comme cette technique apparaît facile à adopter dans le cas de documents papier, le concept de caviardage (ou de suppression) a naturellement été le premier à être adopté pour la protection de documents « électroniques », dans le cadre d'une démarche de pseudonymisation.

En effet, la vaste majorité des données se trouvent désormais dans des bases, dont une des caractéristiques repose sur la structuration, dont le format le plus simple consiste en des tables composées de plusieurs colonnes (les « attributs ») et de nombreuses lignes (les « enregistrements »). Les attributs n'ont pas tous la même criticité : certains attributs peuvent être directement identifiants, comme des numéros de sécurité sociale, ou le couple nom/prénoms qui identifie les individus, hors homonymie. D'autres attributs sont sensibles, c'est-à-dire que la connaissance du lien entre leur valeur et celle d'un attribut identifiant révélerait une information confidentielle, par exemple les revenus de l'individu « Bernard Arnault » dans la base de données de l'impôt sur le revenu, ou encore le diagnostic du patient numéro 12345678 de la base de données de l'Institut Gustave-Roussy.

Pseudonymiser une base de données (ou un document) signifie remplacer les éléments directement identifiants par un pseudonyme, à l'instar du caviardage des noms ou de l'utilisation de faux noms dans des articles de presse. La pseudonymisation est une procédure non triviale, qui dépend du cas d'utilisation. On pourra se poser (de manière non exhaustive) les questions suivantes : s'agit-il d'utiliser un pseudonyme simplement pour améliorer la sécurité de données en interne, en cas de fuite de données ? S'agit-il de partager des données avec des sous-traitants ? Qui va procéder à la construction des pseudonymes ? Qui connaîtra la correspondance entre les pseudonymes et les vrais identifiants ? Nous référons le lecteur intéressé par cette question au rapport produit pour l'Europe, « Techniques et meilleures pratiques de pseudonymisation »[1], qui présente des recommandations techniques sur la mise en œuvre d'un processus de pseudonymisation.

Toutefois, pseudonymiser n'est pas anonymiser, au sens du Règlement européen sur la protection des données (RGPD). Les vulnérabilités de la pseudonymisation ont été montrées expérimentalement par Latanya Sweeney[2] à la fin des années 1990. Selon cette étude, beaucoup de personnes ont des valeurs uniques si on considère des (petits) ensembles d'attributs, qui au premier abord ne semblent ni sensibles, ni identifiants. En utilisant la triplette date de naissance-sexe-code postal, elle a été capable de croiser une base de données médicales censée être anonymisée, achetée à un hôpital, avec une base de données publiques (une liste électorale), et ainsi de réidentifier William Weld, à l'époque gouverneur du Massachusetts. Une fois reconnue dans la base de données de l'hôpital, elle a pu retrouver toutes ses données hospitalières. À la faveur d'une étude des données du

1. Jensen M., Lauradoux C. et Limniotis K., « Techniques et meilleures pratiques de pseudonymisation », European Union Agency for Cybersecurity (ENISA), nov. 2019.
2. Son article, *"A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002, vol. 10, nº 5, p. 557-570, est considéré comme fondateur de ce modèle très utilisé.

recensement de 1990, Latanya Sweeney a montré que les trois attributs de 87 % de la population des États-Unis possédaient une valeur unique. Ce résultat d'unicité est transposable à des degrés variables à tous les pays. Le résultat de Latanya Sweeney, montrant la vulnérabilité de la pseudonymisation par son incapacité à empêcher la réidentification, a marqué le début de travaux sur l'anonymisation moderne. Sa proposition, le *k-anonymat*, simple à comprendre et à mettre en œuvre, est notamment utilisée dans le domaine médical aux États-Unis, en adéquation avec le Health Insurance Portability and Accountability Act (HIPAA) de 1996[3].

Les attributs utilisés pour la réidentification sont appelés des quasi-identifiants. Il en existe plusieurs définitions, la plupart informelles, dont la définition historique de Latanya Sweeney. Nous adoptons la définition informelle suivante[4].

Définition 1 [quasi-identifiant] : un ensemble de n attributs (A_1, …, A_n) d'une base de données est un quasi-identifiant si (i) il est raisonnable que ces attributs puissent être connus par un attaquant[5] et (ii) une proportion raisonnable des individus de la base de données possède une valeur unique de ces attributs.

Néanmoins, si le *k-anonymat* résout certains problèmes liés aux quasi-identifiants simples, son analyse de sécurité est complexe, puisqu'il faut considérer pour chaque attaquant le type de données qu'il est susceptible de connaître.

II – L'anonymisation au sens du RGPD

En Europe, le RGPD propose de mettre en place un ensemble de règles ayant trait à la protection des données personnelles. Un seul considérant mentionne l'anonymisation, mais son impact est colossal : en effet, si une donnée est anonyme, alors elle n'est plus personnelle, et elle échappe donc aux obligations du RGPD. Il est donc capital (du point de vue de la réglementation) de savoir si une donnée est anonyme ou non.

Citons ici le considérant 26 :

> « Il y a lieu d'appliquer les principes relatifs à la protection des données à toute information concernant une personne physique identifiée ou identifiable. Les données à caractère

3. Le HIPAA (*www.hhs.gov/hipaa/index.html)* est une loi fédérale des États-Unis qui réglemente la diffusion de données personnelles par les acteurs de la santé.
4. On pourra se référer à Nguyen B. et Castelluccia C., « Techniques d'anonymisation tabulaire : concepts et mise en œuvre », *Bulletin de la Société informatique de France* 2020, vol. 1024, nº 15, p. 23-41,pour une définition plus formelle.
5. Un attaquant est une personne ou une entité contre laquelle on souhaite se protéger.

> personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique 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. Il n'y a dès lors pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche. »

On voit déjà que le considérant indique très clairement que des données pseudonymisées doivent être considérées comme identifiantes, puisqu'elles peuvent être utilisées dans le cas d'une attaque de réidentification en utilisant des « informations supplémentaires ». Comme démontré par Latanya Sweeney, la pseudonymisation ne peut donc pas être considérée comme une technique d'anonymisation. Si elle peut être utilisée comme une mesure de protection complémentaire de données personnelles, elle ne suffit pas pour rendre une donnée « non personnelle ».

Notons qu'aucune technique spécifique n'est citée par le texte qui limite ses exigences à la prise en compte de l'état de l'art des connaissances en termes d'anonymisation pour juger si une technique est suffisamment protectrice. Un principe « d'effectivité », c'est-à-dire la prise en compte du coût et du temps de la réidentification, est adopté pour décider si des données sont anonymes ou non. Il est intéressant de noter qu'une des particularités des techniques d'anonymisation est d'être incapable de fournir des garanties de protection absolues. Elles peuvent en revanche fournir des garanties probabilistes, par exemple sous forme de probabilité maximale de réidentification d'un individu dans une base de données. Cette démarche est courante en matière de gestion des risques (par exemple, méthode AFNOR ou EBIOS[6]), où l'objectif est de cartographier les risques et de tenter de réduire leur probabilité d'occurrence ou leur impact.

Les techniques d'anonymisation modernes peuvent donc être paramétrées, afin de borner le risque de réidentification. Toutefois, en général, plus la réidentification sera difficile, plus les données seront dégradées, et donc plus leur exploitabilité future sera réduite. Il faut donc comprendre l'anonymisation comme un compromis, entre une protection plus ou moins

6. La méthode EBIOS a été adaptée par la CNIL pour proposer sa méthodologie « PIA » qui effectue une analyse d'impact de vie privée (EIVP) demandée par le RGPD.

grande contre la réidentification, et une utilité plus ou moins grande pour des tâches d'analyse subséquentes ; on pensera en particulier aux opérations d'analyse de données et d'IA (classification, *clustering*).

Dans la suite de ce chapitre, nous présenterons diverses techniques d'anonymisation de l'état de l'art, nous discuterons des protections offertes, et nous évoquerons la qualité des analyses possibles au travers d'exemples réels. Nous conclurons le chapitre par une présentation d'outils issus de travaux de recherche permettant l'analyse comparative de techniques d'anonymisation.

III – Les acteurs du processus d'anonymisation

Nous commencerons en considérant des données de type « tabulaire », c'est-à-dire un ensemble de lignes (enregistrements ou *nuplets*) d'une base de données (relationnelles), chaque ligne étant composée de plusieurs champs (attributs ou colonnes). Ces données sont réalisées par des producteurs de données : par exemple, dans le cadre de la collecte de données GPS, les producteurs de données sont les téléphones portables des individus ; dans celui de la collecte de données par un compteur Linky, ce « compteur intelligent » collecte les données, lesquelles sont ensuite envoyées à des éditeurs de données qui les publient en *open data* auprès d'entités tierces. Pour ce dernier transfert, il est important qu'elles soient anonymisées.

Le producteur de données est supposé toujours être de confiance dans le processus d'anonymisation, puisque l'on considère que s'il est un individu, alors il est concerné par la donnée. Par conséquent, si des techniques de protection sont mises en œuvre, le producteur n'a aucun intérêt à ne pas les respecter, et de même, le producteur ne cherche pas à récupérer la donnée, puisqu'il y a déjà accès. Plusieurs hypothèses de confiance sont possibles pour l'éditeur :

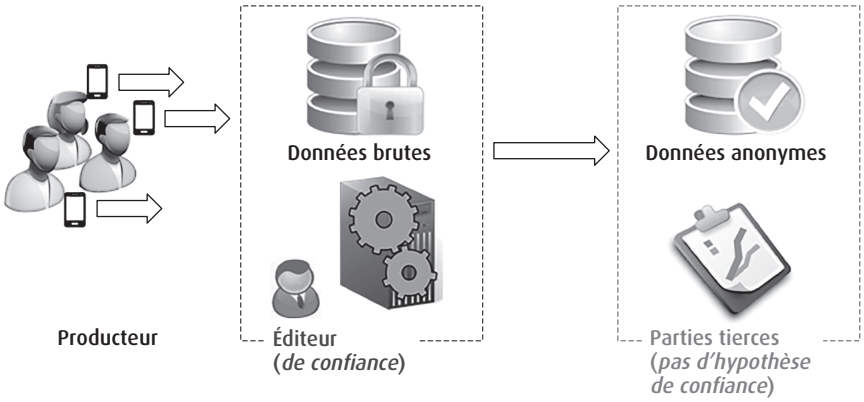
– première hypothèse dite « Conf » : l'éditeur est complètement de confiance, par exemple un hôpital ou un fournisseur d'énergie qui gérerait lui-même le processus d'anonymisation, ou encore le producteur de données lui-même ;
– deuxième hypothèse dite « HmC » : l'éditeur est dit « honnête mais curieux », c'est-à-dire qu'il va respecter le protocole d'anonymisation, mais cherchera à récupérer des données auquel il n'a pas forcément droit. Ce modèle peut s'appliquer à un *cloud* commercial qui proposerait un service d'anonymisation ;

– troisième hypothèse dite « modèle Furtif » : elle correspond à un adversaire qui est prêt à ne pas respecter les règles, tant qu'il n'a aucune chance (ou une très faible) de se faire repérer.

Selon le type d'hypothèse de confiance, les techniques d'anonymisation mises en œuvre seront différentes, ce qui mène à deux types d'architecture d'anonymisation distincts :

– architecture centralisée (« Cent ») : l'éditeur, en général de type « Conf », récupère les données, et suit un processus d'anonymisation. L'intérêt de l'anonymisation centralisée est que l'entité qui anonymise connaît l'ensemble des données, produisant ainsi des résultats de la meilleure qualité possible ;

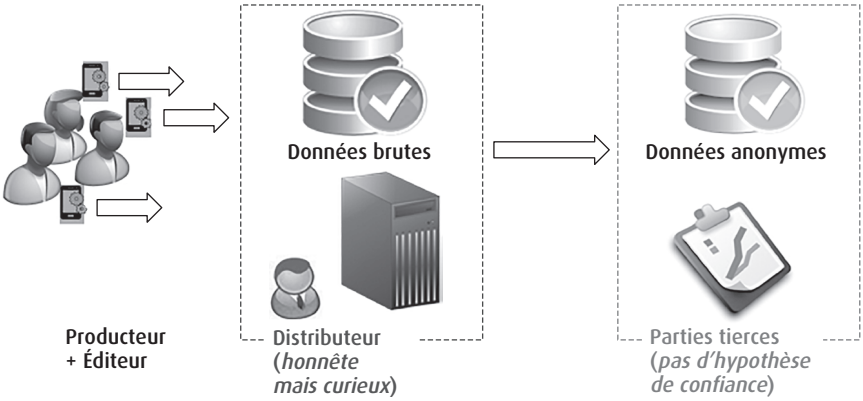
Figure 1 – Éditeur (T), Architecture (C)



Source : D'après Allard, Nguyen, Pucheral.

– architecture locale – ou décentralisée) – (« Loc ») : le producteur et l'éditeur constituent une même entité, et sont nombreux. L'anonymisation est réalisée de manière indépendante des données des autres producteurs. Une nouvelle entité (distributeur), de type « HmC », est utilisée pour collecter et diffuser les données.

Figure 2 – Éditeur (T), Distributeur (HbC), Anonymisation (L)



Source : D'après Allard, Nguyen, Pucheral.

Toutes les techniques d'anonymisation peuvent être vues comme opérant certaines transformations sur une base de données, effectuées par l'éditeur : 1) suppression des *nuplets* ; 2) ajout de *nuplets* ; 3) modification des *nuplets*. Les transformations peuvent être opérées de manière déterministe ou aléatoire.

IV – Les modèles d'anonymisation

A – Le *k-anonymat*

Le *k-anonymat* est un modèle où les données anonymes respectent la contrainte suivante :

Définition 1 [*k-anonymat*] : une base de données est dite « k-anonyme » pour les quasi-identifiants (A_1, …, A_n) si chaque enregistrement dans la base de données possède les mêmes valeurs de ces quasi-identifiants qu'au moins *k-1* autres enregistrements[7].

De nombreux algorithmes permettent d'atteindre le *k-anonymat*, généralement en adoptant une architecture centralisée[8]. Certaines propositions, faisant usage de matériel sécurisé, sont capables de gérer des modèles

7. Le nom de la technique exprime le fait qu'un individu est anonyme puisqu'il est dissimulé dans une foule de *k* individus dont il est indistinguable.

8. On peut citer l'algorithme dit « de Mondrian » présenté par LeFevre K., DeWitt D.J. et Ramakrishnan R., *Mondrian Multidimensional K-Anonymity*, 2006, ICDE, p. 25 Cet algorithme divise l'espace des attributs en rectangles, ce qui lui donne une apparence des tableaux bien connus de Piet Mondrian.

d'éditeur de type « HmC » ou« Furtif »[9], l'enjeu ici étant de pouvoir garantir une meilleure sécurité contre un modèle d'attaquant plus puissant.

Le *k-anonymat* étant le premier modèle historique d'anonymisation, de très nombreuses applications l'ont utilisé pour publier des données. Nous avons cité, par exemple, les données médicales, où le *k* choisi est souvent assez petit, de l'ordre de 5 à 15. L'ordre de grandeur du risque de réidentification est borné par 1/k (soit 7 à 20 % de risque, dans le pire des cas, de retrouver la personne correspondant à une donnée sensible). En pratique, ce risque maximum n'est atteint que pour une toute petite fraction de la base de données (moins de 1 %), et le risque moyen de réidentification est beaucoup plus faible.

Dans le domaine de la statistique nationale française, les données sont souvent publiées sous forme de tableaux qui fusionnent dans une même case les données de nombreux individus ou entités. Pour ce type de publication, certaines cases peuvent être masquées lorsqu'il n'y a pas suffisamment d'individus ou si la contribution d'un individu est majoritaire. Ces modèles de protection sont assez proches du concept du *k-anonymat*. Notons que le modèle du *k-anonymat* a été étudié comme un moyen de publier des micro-données, c'est-à-dire de données non agrégées par l'INSEE[10]. En effet, la mise en œuvre en pratique du *k-anonymat* pose plusieurs difficultés : « Plusieurs points méthodologiques cruciaux sont difficiles à résoudre quand on s'intéresse à la protection des données individuelles, comme la quantification de la réduction du risque de réidentification lors de la mise en place de méthodes perturbatrices : à quel moment peut-on considérer que les données sont "suffisamment" perturbées et ne présentent plus de risque de réidentification ? De même, le choix des variables quasi identifiantes utilisées pour quantifier le risque est crucial et non trivial. »[11] Pour résumer, si le principe du *k-anonymat* est simple à comprendre, son utilisation met en jeu des choix complexes comme l'identification des

9. Cette approche est détaillée par Allard T., Nguyen B. et Pucheral P., *"MetaP : Revisiting Privacy-Preserving Data Publishing Using Secure Devices"*, *Distributed and Parallel Databases* (DAPD) 2014, vol. 32, nº 2, p. 191-244.
10. Une micro-donnée étant une donnée « atomique » qui correspond en général à un individu (ou une entreprise), l'intérêt ici est de publier un grand jeu de micro-données, plutôt que des tableaux représentant des données agrégées (regroupement par tranche d'âge, etc.). À ce sujet, un rapport a été produit en 2016 par Bergeat M., *La gestion de la confidentialité pour les données individuelles*, série Documents de travail « Méthodologie Statistique », de la Direction de la méthodologie et de la coordination statistique et internationale. Le document est disponible au lien suivant : *www.insee.fr/fr/statistiques/fichier/2535625/M201607.pdf* Toutefois à notre connaissance, ces techniques n'ont pas encore été utilisées dans des études publiées.
11. *Ibid.*, p. 44.

quasi-identifiants et le réglage de la valeur de *k* susceptible d'offrir une protection suffisante.

B – La confidentialité différentielle (*differential privacy*)

Le modèle de confidentialité différentielle (*differential privacy*, ou DP) a été introduit en 2006 par Cynthia Dwork[12], afin de proposer un modèle d'anonymisation qui ne dépende pas des connaissances de l'attaquant, et permette de protéger la présence d'un individu dans le jeu de données tout en permettant d'apprendre des informations utiles. Par exemple, une analyse d'un jeu de données permettra d'apprendre que les chances de contracter un cancer du poumon sont multipliées par 10 pour les fumeurs, mais pas de savoir si un individu en particulier a participé à l'étude. Toutefois, si on sait par ailleurs qu'une personne fume, on pourra en déduire que cette personne a un risque accru. Le modèle de la DP permet donc de réaliser des études permettant de découvrir des corrélations entre des variables.

Un seul paramètre, traduisant le risque de réidentification maximal autorisé, doit être configuré. D'une manière générale, la DP utilise un concept de base de données adjacentes. Ce concept se décline selon la structure des données étudiées. Dans le cas des bases de données tabulaires, on considère que deux bases de données sont adjacentes si et seulement si les deux bases de données ne diffèrent que par un seul *nuplet*. Ce peut être parce qu'une base a un *nuplet* en plus ou en moins, ou bien parce qu'il est différent.

Tableau 1 – Deux bases de données tabulaires adjacentes

Base de données $\mathbf{D_1}$				Base de données $\mathbf{D_2}$		
ID	**Âge**	**Diagnostic**		**ID**	**Âge**	**Diagnostic**
1	32	COVID		1	32	COVID
2	67	GRIPPE		2	67	GRIPPE
3	25	COVID		3	25	COVID
				4	*78*	*GRIPPE*

La DP peut être vue comme une contrainte probabiliste sur un mécanisme de traitement de données. Par exemple, si on considère le traitement consistant à compter le nombre de personnes qui sont diagnostiquées à la covid, ce processus (qui inclut une part d'aléa) doit *grosso modo* retourner

12. Dwork C., *"Differential Privacy"*, *ICALP* 2006, n° 2, p. 1-12.

le même résultat, qu'on l'applique à la base de données D_1 ou à la base de données D_2. Formellement, la définition de la DP est la suivante :

Définition 2 : Soit ε un réel positif, D_1 et D_2 deux bases de données adjacentes, et Ω le résultat de l'exécution d'un mécanisme aléatoire ALG. On dit que ALG respecte la contrainte de ε-confidentialité différentielle si et seulement si $\Pr[ALG(D_1) = \Omega] \leq \exp(\varepsilon) \times \Pr[ALG(D_2) = \Omega]$.

Le paramètre ε quantifie combien de fois il est plus probable que l'observation d'une certaine valeur provienne d'une base de données plutôt qu'une autre. On choisit souvent $\varepsilon < 1$ soit un rapport maximal d'environ 3 dans le pire des cas, néanmoins le choix de ce paramètre est loin d'être « naturel » puisqu'il quantifie une probabilité relative[13]. En effet, il est difficile d'appréhender un risque associé à une possibilité trois fois supérieure de présence d'un individu dans la base par rapport à un autre, la probabilité absolue de présence pouvant rester faible.

L'intérêt majeur de la DP est que la protection est en théorie « totale », c'est-à-dire que la probabilité est intrinsèque au mécanisme de publication de données, et donc ne dépend pas des connaissances de l'attaquant. En réalité, la connaissance de l'attaquant (ou sa capacité de discernement) est intégrée dans la définition de ce que sont des bases de données adjacentes : seules les bases de données qui diffèrent d'un *nuplet* réaliste doivent être prises en compte. La quantification de la probabilité se fait ainsi sur l'ensemble des bases de données adjacentes, c'est-à-dire qu'il faut que les données ainsi générées soient réalistes, sinon on pourra s'en rendre compte et les ignorer.

L'un des exemples d'applications les plus emblématiques de la DP est le recensement américain de 2020[14] réalisé avec un processus respectant les garanties de DP, en remplacement des mécanismes précédents, qui utilisaient en particulier la suppression et l'échange de données. Un des avantages notés par le bureau du recensement américain est que leur approche antérieure utilisait une méthodologie qui devait rester secrète, ce qui n'est jamais une caractéristique acceptable d'un mécanisme de protection formel, tandis que la nouvelle procédure est publique et documentée.

13. Un exemple très simplifié serait le suivant : on souhaite conserver la valeur de l'attribut « âge » des individus originaux de la base, mais pour protéger la micro-donnée, notre mécanisme va procéder à une modification *aléatoire* de la valeur de l'attribut « diagnostic », dont on considère qu'elle peut prendre trois valeurs sensibles : Grippe, Covid et Rhume. Si on choisit $\varepsilon = 1$ alors il faudra modifier la base de données avec un rapport de probabilité inférieur à 3 entre chacune des probabilités. Par exemple, la probabilité de transformer Covid en Covid serait de 0,5, la probabilité de transformer Covid en grippe serait de 0,25 et la probabilité de transformer Covid en Rhume serait de 0,25. Ici, le rapport maximal entre les probabilités est de 2, ce qui vérifie bien que chaque probabilité est inférieure à $\exp(\varepsilon)$ x n'importe quelle autre. Il faudrait également définir les probabilités de transformer Grippe et Rhume dans toutes les autres valeurs en respectant aussi ce rapport de probabilités.

14. Leur méthodologie et les avantages liés à l'utilisation de cette approche sont présentés dans US Census Bureau, *Disclosure Avoidance for the 2020 Census : An Introduction*, US Government Publishing Office, Washington, DC, nov. 2021 (*www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf*).

Un autre exemple d'utilisation de la DP est la publication de données de mobilité, issues de données des opérateurs téléphoniques[15], permettant de calculer les déplacements d'individus tout en protégeant les trajectoires individuelles. Typiquement, une telle approche peut être utilisée pour compter combien de personnes ont quitté la région parisienne pour se rendre dans d'autres régions au moment du premier confinement de 2020, tout en garantissant l'impossibilité de retrouver les données d'un individu en particulier.

Enfin, évoquons l'existence d'une technique liée à la DP, appelée « confidentialité différentielle locale » (LDP) qui permet de gérer le cas de figure où l'éditeur est de type « Furtif », alors que la DP classique fait l'hypothèse d'un éditeur « Conf ». Cette technique, déjà utilisée dans les années 1960[16] est également utilisée par Apple et Google, qui propose une implémentation nommée « RAPPOR ». L'approche, très simple, se base sur une modification aléatoire des données collectées avant l'envoi au distributeur. Il s'agit donc d'une méthode compatible avec une approche d'anonymisation locale. L'une des applications de cette technique concerne la construction des courbes de fréquentation de certains endroits sur Google Maps (sites touristiques, magasins, gares, etc.). L'information globale (importance de la fréquentation par tranche horaire) est préservée, mais la donnée personnelle (qui était présent et à quelle date ?) est cachée *via* une perturbation aléatoire.

V – Vers des outils permettant l'évaluation de techniques d'anonymisation en considérant l'utilité

Comme on a l'a vu, il existe de nombreuses techniques d'anonymisation, souvent utilisées dans des contextes applicatifs différents. Toutefois, chaque application nécessite une analyse de sécurité spécifique et des décisions de configuration, ce qui n'est pas nécessairement simple pour une entité souhaitant publier des données. Le premier outil offrant un

15. Voir les travaux réalisés par une équipe de recherche d'Orange Labs : Brunet S., Canard S., Gambs S. et Olivier B., *"Edge-calibrated Noise for Differentially Private Mechanisms on Graphs"*, Annual Conference on Privacy, Security and Trust (PST), 2016, p. 42-49.

16. Lors de travaux en sociologie, afin d'éviter les biais, Warner a proposé en 1965 une méthode appelée « réponse aléatoire », où les participants à une enquête voient leurs réponses modifiées de manière aléatoire, afin que l'enquêteur ne puisse pas connaître leur réponse réelle. Warner S. L., *"Randomised Response : A Survey Technique for Eliminating Evasive Answer Bias"*, *Journal of the American Statistical Association*, *Taylor & Francis* mars 1965, vol. 60, n° 309, p. 63-69.

service d'anonymisation « clé en main » est ARX[17], ciblé initialement sur l'anonymisation de données médicales. L'utilisateur peut anonymiser des données tabulaires, avec divers algorithmes, et optimiser leur qualité par rapport à des métriques d'utilité générale, comme la perte d'information ou l'entropie. Si le logiciel est pratique à utiliser, les métriques générales d'utilité sont une limite importante, car il est difficile de lier la qualité des résultats d'analyses de données comme la classification, à des métriques de perte d'information[18].

Francis a proposé un *framework* nommé « GDA Score »[19] qui permet de soumettre des jeux de données anonymisés et de les confronter à des attaques déjà mises en œuvre, afin de tester leur robustesse. L'évaluation de l'utilité dans GDA Score est plus adaptée, puisqu'elle est directement issue de l'exécution de certaines requêtes sur les données originales et anonymisées, et repose sur une estimation de l'ampleur de leur différence.

Dans une optique similaire, Antoine Boutet et a.[20] ont développé une plateforme permettant d'organiser des compétitions d'anonymisation et de réidentification sur des jeux de données proposés par des entités tierces. Une des particularités de cette approche est qu'elle prend en compte l'utilité des données réelles, par l'évaluation des résultats d'analyse sur les données réelles et anonymes. Ainsi, il est possible de tester à la fois la robustesse de l'anonymisation et la qualité des données produites.

VI – Conclusion

Quelle que soit la technique utilisée, l'anonymisation est donc un compromis entre un certain degré de protection et une certaine conservation d'utilité. Toute la problématique, qui dépend largement de l'application, est donc de savoir s'il est possible d'atteindre un degré de protection suffisant tout en conservant une utilité intéressante. En effet, si cet équilibre n'est pas atteignable, il est préférable de ne pas publier les données qui ne seront pas assez protégées, ou pas assez utiles.

17. Pour une présentation d'ARX (en anglais) : *https://arx.deidentifier.org*
18. Comme montré par Mauger C., Le Mahec G., Dequen G., *"Multi-criteria Optimization Using l-diversity and t-closeness for k-anonymization" DPM/CBT@ESORICS* 2020, p. 73-88.
19. Pour une présentation de GDA Score (en anglais) : *www.gda-score.org*
20. L'approche est présentée dans Boutet A., Cunche M., Gambs S., Nguyen B. et Laurent A., *"DARC : Data Anonymization and Re-identification Challenge", Rendez-vous de la recherche et de l'enseignement de la sécurité des systèmes d'information* (*RESSI*) déc. 2020.