INTRODUCTION À L'ANONYMISATION ET À LA PSEUDONYMISATION

FICHE INFO DU PPDT

INTRODUCTION

Avant de partager ou de publier des données, le responsable de traitement a souvent intérêt à les minimiser, les pseudonymiser, et/ou les anonymiser. Même durant la collecte des données (manuellement, via une application mobile, une page web etc.), il a intérêt à appliquer ces techniques. Cette approche offre en effet les avantages suivants:

- *Les éventuels accès par des hackers seraient moins graves.*
- *Les accès légitimes par le personnel interne seraient proportionnés, moins intrusifs, et moins risqués.*
- *Les obligations légales seraient mieux respectées.*
- *En cas d'une "bonne" anonymisation, le traitement des données ne serait plus soumis aux lois sur la protection des données (et ainsi soumis à moins de contraintes).*

Dans cette fiche informative, nous traitons de la différence entre pseudonymisation, anonymisation, et de comment les appliquer. Nous expliquons aussi divers aspects qui y sont liés. Cette fiche est une introduction de base. Elle vise les lecteurs curieux, les développeurs qui veulent débuter et savoir par où commencer, et les non-techniciens qui aimeraient connaître les aspects techniques de la protection des données. A la fin de la fiche, nous fournissons des références plus approfondies sur le sujet.

ANONYMISATION ET PSEUDONYMISATION DE BASE

Considérons qu'un responsable de traitement a une table de données personnelles:

Nom	Age	Sexe
Anaïs Panchaud	24	F
Stéphanie Gomez	35	F
Claude Dupont	83	H

Table 1: données d'origine

Nous y voyons des "identifiants" (prénoms, noms), et des "attributs" (âge, sexe). L'approche d'anonymisation "de base" est de remplacer les identifiants (prénoms, noms) par des identifiants "difficiles à lier aux noms originaires".

Nom	Age	Sexe
Minnie Mouse	24	F
Dora L'Exploratrice	35	F
Mickey Mouse	83	H

Table 2: données "cachées"

Pour garder un moyen de passage des données "anonymisées" de la Table 2 à celles de la Table 1, le responsable de traitement pourrait garder une table de correspondance (Table 3).

Nom	Pseudonyme
Anaïs Panchaud	Minnie Mouse
Stéphanie Gomez	Dora L'Exploratrice
Claude Dupont	Mickey Mouse

Table 3: Table de correspondance

A noter que cette table doit rester hors d'accès pour les utilisateurs de la Table 2 (sinon, l'anonymisation serait inutile!) Tant que cette table de correspondance existe (hors d'accès des utilisateurs de la Table 2, bien sûr), nous parlons de données **pseudonymisées** [RGPD, GR19], puisqu'il existe encore un moyen de passer de ces données "cachées" aux données d'origine.

Si cette table de correspondance est détruite, ou perdue, sans aucun moyen de la récupérer, les données de la Table 2 sont alors **anonymisées**, puisqu'il n'y a plus de moyen de faire le lien avec les données d'origine.

COMMENT AUTOMATISER ÇA?

Quand nous sommes face à des listes de millions de noms, ou à des flux de données de plusieurs milliers d'entrées par seconde, construire des tables de correspondance comme celle de la Table 3 n'est plus pratique. Nous pourrions alors imaginer des fonctions qui permutent et/ou décalent les lettres afin de passer des noms d'origine aux pseudonymes (ex. "Anaïs Panchaud" -> "AsjfdghJassdhgT"). De telles fonctions existent déjà, parmi lesquelles les "fonctions de hachage", telle que la fonction SHA256. Si on passe "Anaïs Panchaud" à cette fonction, elle nous génère:

SHA256("Anaïs Panchaud") = 13368F4[...]38AE0286400

Pratiquement il n'y a pas moyen d'inverser "13368F4[...]38AE0286400" et retrouver "Anaïs Panchaud". C'est pour cela que ces fonctions s'appellent "fonctions de hachage cryptographiques *irréversibles*".

Cependant, si un utilisateur sait déjà que les noms d'origine (Table 1) sont ceux de ses collègues de travail (par exemple), il pourrait facilement passer les noms à la fonction de hachage, l'un après l'autre, et retrouver que 13368F4[...]38AE0286400 correspond à "Anaïs Panchaud". C'est-ce que nous appelons inversion par "force brute".

Pour pallier ce problème, il suffit d'ajouter une "clé", secrète, aux entrées de la fonction de hachage:

SHA256("Anaïs Panchaud, CC5688CDB26") = 84AC65D9D[...]2ADDC060B

Ainsi, le responsable de traitement évite l'inversion des résultats par "brute force", **tant que la clé est gardée secrète**, à l'écart des utilisateurs des données d'origine.

Tant que cette clé secrète est gardée quelque part, il est possible d'inverser les données cachées et retrouver les données d'origine. Comme dans le cas précédent de "table de correspondance" que le responsable de traitement garde quelque part, nous avons alors des données **pseudonymisées**. Si cette clé secrète est perdue, ou détruite, impossible à récupérer, les données seraient alors **anonymisées** puisqu'il n'y a plus moyen de retrouver les données d'origine correspondantes.

Note:

Considérons qu'une Entreprise A possède des données d'origine, qu'elle "anonymise" avant de les passer à une Entreprise B, tout en gardant la clé secrète (ou une table de correspondance) hors d'accès de l'Entreprise B. Techniquement, c'est de la pseudonymisation. Cependant, légalement, si cette clé secrète est rendue "impossible à récupérer" pour l'Entreprise B, on pourrait parler d'anonymisation. Ça reste une question ouverte, à traiter au cas par cas.

DE-ANONYMISATION

Les actions sur les identifiants personnels (PI, les noms par ex.), ne sont pas toujours suffisantes pour l'anonymisation. Reprenons la Table 2 avec une colonne supplémentaire, celles des villes/villages.

Nom	Age	Sexe	Ville
Minnie Mouse	24	F	Genève
Dora L'Exploratrice	35	F	Lausanne
Mickey Mouse	83	H	Collex-Bossy

Table 4: données "cachées", avec ville

Nous pouvons voir que les attributs combinés (âge, sexe, ville) pourraient facilement permettre d'identifier certaines personnes dans la table, peu importe l'"anonymisation" appliquée sur les noms. Ces données supplémentaires s'appellent PII (Personal Identifying Information) qui peuvent aussi servir à ré-identifier certaines personnes dans la table (Mickey Mouse = Claude Dupont).

Cette ré-identification n'est pas limitée seulement aux petits villages (Collex-Bossy) ou aux âges exceptionnels (83). Prenons par exemple les données suivantes.

Nom	Age	Sexe	Ville	Profession	Origine
Minnie Mouse	24	F	Genève	Infirmière	CH
Dora L'Exploratrice	35	F	Lausanne	Dentiste	ES
Mickey Mouse	83	H	Collex-Bossy	Retraité	CH

Table 5: données "cachées", avec ville, profession, et origine

Lausanne compte des centaines de milliers d'habitants. L'Espagne en compte des dizaines de millions. Il y a des milliers de dentistes en Suisse. Des dizaines de milliers d'habitants ont 35 ans. Quelque 4 millions d'habitants sont des femmes. Mais la combinaison (35, F, Lausanne, Dentiste, ES) pourrait être unique, attribuable à "Stéphanie Gomez" moyennant quelques recherches supplémentaires.

Ce qui nous amène à la constatation que tant que quelqu'un peut combiner plus de données supplémentaires, il peut mieux affiner l'identification. C'est ce qui rend la tâche difficile pour celui/celle qui partage des données "anonymisées", sans savoir quelles données supplémentaires pourraient y être combinées, menant ainsi à la ré-identification des personnes.

A noter que les PII ne se limitent pas aux données des "registres officiels". Elles peuvent inclure des données de géolocalisation, de comportement, etc. Par exemple, si on a les données de géolocalisation d'un pendulaire qui réside à Vevey, et qui travaille à Bern-Brunnen, ces données de géolocalisation pourraient servir à l'identifier malgré les tailles individuelles de Vevey et de Bern-Brunnen.

GENERALISATION

Pour pallier le problème d'identification quand on combine plusieurs attributs (ex. âge, sexe, ville), on peut *généraliser* certaines données avant de les partager. Dans le cas de (83, H, Collex-Bossy) qui identifiait "Claude Dupont", on peut "généraliser" l'âge pour donner une fourchette d'âges [75-90], et le canton (GE) plutôt que la ville. Ainsi ("Mickey Mouse", [75-90], H, GE) pourrait correspondre à des milliers de personnes, cachant mieux l'identité de "Claude Dupont". C'est ce qu'on appelle le *k-anonymat*, où k est le nombre de personnes dans le groupe de confusion ([75-90], H, GE). Les données utiles qui en résultent seront ainsi *agrégées* (nombre de personnes, moyenne d'âge etc.)

A noter que la généralisation doit être soigneusement adaptée au type de données (ville, géolocalisation, âge, etc.) ainsi qu'au but de l'analyse des données. Si un développeur conçoit une application pour les prévisions météo, la généralisation des données GPS au niveau ville serait suffisante à récolter par l'application. Par contre, pour une application de navigation sur les routes, une précision de localisation maximale est appropriée.

L-DIVERSITE ET T-PROXIMITE

Supposons maintenant qu'une employée reçoive une table des données de ses collègues, où les attributs qui pourraient identifier les personnes sont généralisés (2-anonymat dans ce cas). En sus, il y a un attribut "Maladie":

Age	Sexe	Ville	Maladie
[20-30]	F	VD	Hypertension
[20-30]	F	VD	Diabète
[40-50]	H	GE	Cancer
[40-50]	H	GE	Cancer
[40-50]	H	GE	Cancer
[40-50]	F	VS	Cholestérol
[40-50]	F	VS	Diabète
[40-50]	F	VS	Diabète
[40-50]	F	VS	Aucune

Table 6: données 2-anonymisées

Nous voyons que l'employée ne pourrait pas préciser exactement quelle ligne de la table correspond à quel collègue. Elle aura toujours la confusion entre deux collègues, au moins. Cependant, la table montre avec certitude que son collègue de bureau, qui est genevois, est atteint d'un cancer. Malgré le manque d'identifiants, et malgré la généralisation pour "flouter" les attributs susceptibles d'identifier les personnes, l'attribut "Maladie" n'est pas assez diversifié pour empêcher d'apprendre (d' "inférer", le terme technique utilisé plus bas) des informations sensibles à propos de certains groupes.

Pour pallier ce problème, la l-diversité complémente la k-anonymisation en veillant à ce que dans chaque groupe il y ait au moins l valeurs différentes (par exemple en changeant les généralisations des divers attributs). Ainsi, l'employée qui a ces données n'aura pas de certitude concernant la maladie de son collègue de bureau.

Même l'application de la l-diversité (en plus de la k-anonymité) pourrait ne pas être suffisante en terme de protection des données. Supposons qu'après l'application de 2-diversité, le groupe des genevois ressemble à:

[40-50]	H	GE	Cancer
[40-50]	H	GE	Cancer
[40-50]	H	GE	Diabète

Table 7: Groupe des genevois de la Table 6, 2-diversifié

L'employée qui a cette table pourrait toujours "inférer" que les Genevois dans l'entreprise sont plus atteints par le cancer que les autres groupes. C'est là que la *t-proximité* vient affiner la généralisation, en créant des groupes où la distribution des maladies ressemble à la distribution sur la table entière (par exemple: la proportion de Genevois atteints par le cancer est égale à celle de l'entreprise entière).

Ces exemples permettent aussi de voir que, tant que nous poussons vers une *protection des données* plus forte, nous réduisons ainsi *l'utilité* des données. C'est au responsable du traitement de juger *au cas par cas*, à quel point il faut pousser la protection des données tout en gardant un niveau d'utilité acceptable.

RANDOMISATION

Prenons l'exemple d'une grande entreprise et un conseiller qui aimerait savoir (entre autre) l'âge moyen des employés. Afin de bien protéger les données / la sphère privée des personnes on peut imaginer:

- Lors de partage de données, les ressources humaines transmettent au conseiller une liste où les âges sont altérés en ajoutant un nombre aléatoire entre -10 et +10
- Lors de la collecte, le conseiller questionne individuellement les employés, en leur demandant de "mentir" à propos de leur âge en ajoutant un nombre aléatoire entre -10 et +10.

C'est la *randomisation*, souvent applicable pour des résultats statistiques. Cette approche convient au calcul de l'âge moyen (la finalité), tout en:

- réduisant l'intrusion dans la vie privée des individus. A noter qu'il ne s'agit pas d'anonymisation ou de pseudonymisation. La randomisation change les valeurs des "attributs". Les personnes *pourraient* rester identifiables, mais leurs informations ne sont pas individuellement précises, ce qui est une autre technique de protéger la sphère privée.
- réduisant les risques d'identification dus aux combinaisons comme (âge, sexe, ville) décrits précédemment: En randomisant (83, H, Collex-Bossy) à (75, H, Genève), les possibilités seraient plus larges que l'unique "Claude Dupont" de Collex-Bossy. C'est un effet similaire à celui de la généralisation décrite précédemment ([75-90], H, GE).

CONFIDENTIALITE DIFFERENTIELLE

Considérons l'exemple précédent du conseiller qui questionne individuellement les employés sur leurs âges, en leur demandant de "mentir" en ajoutant un nombre aléatoire entre -10 et 10. Si le conseiller répète la question à chaque employé plusieurs fois, il finirait par avoir une idée assez précise des âges (quand le nombre de questions augmente, la moyenne des réponses de chaque employé se rapprocherait de la valeur d'âge exacte).

Pour se protéger contre cette "inférence", chaque employé devrait:

- se souvenir des réponses qu'il a données dans le passé
- indiquer au conseiller qu'il ne peut plus répondre, car sinon son âge serait facile à estimer.

La confidentialité différentielle est une technique de randomisation qui indique au responsable du traitement:

- le niveau de randomisation (ou de "bruit") à appliquer aux données avant de répondre à des requêtes
- les garanties de protection que ça va donner
- ainsi que les limitations à respecter ("privacy-budget" est le terme technique).

La confidentialité différentielle offre divers avantages (comme la garantie formelle/mathématique d'anonymat), mais sous certaines contraintes pratiques (par exemple pour garder les traces de qui a passé quelle requête sur quelles données). Par ailleurs, comme pour les techniques précédemment décrites, il faut aussi trouver le juste équilibre entre le bruit ajouté aux données tout en gardant un certain niveau d'utilité.

ANALYSE DE LA FIABILITE DES TECHNIQUES D'ANONYMISATION, SELON LE GROUPE ART. 29

Le Groupe de travail "Article 29" est le groupe de travail européen indépendant qui traitait les questions relatives à la protection de la vie privée et aux données à caractère personnel jusqu'à l'entrée en vigueur du RGPD. Le groupe avait adopté trois points à prendre en considération pour évaluer le niveau d'anonymisation fourni par une technique donnée [GR19]:

- l'**individualisation** ("Singling-out"), qui est la possibilité d'isoler des enregistrements identifiant une personne dans l'ensemble de données
- la **corrélation** ("Linkability"), qui est la possibilité de relier des enregistrements (au moins deux) se rapportant à la même personne ou à un groupe de personnes, dans une ou plusieurs bases de données.
- l'**inférence** ("Inference"), qui est la possibilité de déduire la valeur d'un attribut correspondant à une personne, à partir des valeurs d'autres attributs.

Pour chaque méthode d'anonymisation (citée dans la section précédente, ou autres), nous pouvons évaluer sa performance en analysant ces trois points: individualisation, corrélation, et inférence. Une solution qui résiste à ces trois risques offrirait alors une protection fiable contre les tentatives de ré-identification. La table ci-dessous résume ces 3 points, pour les méthodes d'anonymisation décrites précédemment [GR19].

	Reste-il un risque...		
	d'individualisation?	**de corrélation?**	**d'inférence?**
Pseudonymisation	Oui	Oui	Oui
Randomisation	Oui	Peut-être pas	Peut-être pas
Agrégation ou k-anonymat	Non	Oui	Oui
l-diversité	Non	Oui	Peut-être pas
Confidentialité différentielle	Peut-être pas	Peut-être pas	Peut-être pas

Table 8: Évaluation de diverses techniques d'anonymisation

En plus de l'analyse des techniques d'anonymisation utilisant ces trois points, le même document [GR19] liste les erreurs courantes pour chacune de ces techniques. Nous recommandons à tout responsable de traitement de données personnelles cette lecture approfondie.

DONNEES SYNTHETIQUES

Les données synthétiques, comme leur nom l'indique, sont des données "artificielles", produites à partir de données personnelles en essayant autant que possible d'avoir les mêmes propriétés.

Comment ça marche?

Le responsable de traitement ne partage pas les données personnelles claires avec des tiers. Par contre, il les utilise pour entraîner une machine (du "machine learning") à générer un autre ensemble de données, *synthétiques*, qui ont les mêmes propriétés que les données d'origine. Ces données synthétiques ne sont en principe pas des données personnelles. Ainsi, elles peuvent être partagées avec des tiers, sans trop se soucier de la protection de la vie privée.

On trouve sur le marché plusieurs produits générateurs de données synthétiques, produisant des données en principe "anonymisées", sur la base des données originelles que le responsable du traitement détient. Cependant, le niveau d'anonymat en résultant est encore contestable [SOT]. Plus précisément, il est difficile de générer des données synthétiques *pour des requêtes génériques*, tout en assurant un niveau de protection adéquat. Par exemple, il est possible de générer des données synthétiques sur les déplacements des personnes, à la base de données réelles, où certaines propriétés (ex. distances moyennes des déplacement, durées) sont maintenues. Par contre, si le responsable de traitement souhaite générer des données qui ressemblent aux originelles selon *toutes* les propriétés (ex. distances moyennes, durées, nombre de déplacements entre toutes les communes suisses, les données démographiques combinées etc.) il risque d'avoir des cas synthétiques qui révèlent exactement les cas originels, alors personnels, pas assez anonymisés. Là aussi, nous voyons que le responsable de traitement a un compromis à faire entre le nombre de propriétés à garder (qui correspondent aux requêtes prévues) entre les données originelles et synthétiques, et le niveau d'anonymat souhaité. C'est un sujet de recherche encore actif.

CONCLUSIONS

Il existe un grand nombre de techniques d'anonymisation (certaines simples, d'autres assez compliquées) qui apportent aux responsables de traitement un certain degré de protection efficace et/ou anonymat, si leur usage est correctement conçu. Ceci suppose:

- une bonne connaissance des données en question, du contexte d'utilisation, des objectifs, des finalités;
- une combinaison de diverses techniques;
- la recherche du bon équilibre entre utilité et protection;
- la nécessité d'analyser la protection/anonymisation au cas par cas. Les solutions génériques sont risquées;
- la nécessité de ne pas négliger les risques résiduels après l'anonymisation;
- une mise à jour régulière: de nouvelles techniques ainsi que de nouvelles failles sont découvertes de temps à autre;
- la nécessité de ne pas considérer l'anonymisation comme absolue et stable avec le temps. Ce qui est considéré comme anonyme aujourd'hui pourrait être facilement re-identifiable dans le futur. Il faut réévaluer régulièrement les risques associés [GR19].

QUELQUES RESSOURCES SUPPLEMENTAIRES

[GR19] Groupe de travail "article 29" sur la protection des données, "Avis 05/2014 sur les Techniques d'anonymisation", avril 2014.

[PPDM] C. Aggrawal et P. Yu, "Privacy-Preserving Data Mining, Models and Algorithms", Springer, 2008.

[PPDP] B. Fung K. Wang, A. Fu, et P. Yu, "Introduction to Privacy-Preserving Data Publishing, Concepts and Techniques", CRC Press, 2011.

[RGPD] Règlement (UE) 2016/679 du parlement européen et du conseil, 2016.

[SOT] T. Stadler, B. Oprisanu, et C. Troncoso, "Synthetic Data -- Anonymisation Groundhog Day", 2020.

AUTEUR

Dr. Imad Aad, Centre pour la Confiance Numérique (https://www.c4dt.org/), EPFL.

PPDT – 13.09.2021