Techniques d'anonymisation

Benjamin Nguyen

Techniques d'anonymisation

BENJAMIN NGUYEN

Insa[1] Centre Val de Loire et Inria[2] Paris-Rocquencourt

L'opposition entre une donnée qui permet d'identifier une personne et une donnée anonyme n'est pas une opposition absolue. C'est pourquoi il existe plusieurs méthodes d'anonymisation, plus ou moins efficaces. On utilise souvent aujourd'hui la « k-anonymisation », la « l-diversité », ou la « confidentialité différentielle », trois techniques dont les principes sont donnés dans cet article. Les différentes techniques sont à juger à la fois sur la sécurité qu'elles procurent, et sur ce qu'elles laissent subsister comme analyses possibles.

Il existe légalement deux types de données : des données à caractère personnel, et des données anonymes. Les données sont à caractère personnel « *dès lors qu'elles concernent des personnes physiques identifiées directement ou indirectement* » pour citer la CNIL. Au contraire, toute donnée qu'il est impossible d'associer avec une personne physique sera dite « *anonyme.* » Il est intéressant de constater que la loi Française définit une impossibilité forte, puisqu'elle précise que « *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.* » Plus mesuré, le projet de règlement Européen prévoit que « *pour déterminer si une personne est identifiable, il convient de considérer* ***l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre****, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne. Il n'y a pas lieu d'appliquer les principes de protection aux données qui ont été rendues suffisamment anonymes pour que la personne concernée ne soit plus identifiable.* » En d'autre termes, le projet de règlement reconnaît une *quantification* et *gradation* dans la méthode d'anonymisation.

Cette définition sous entend qu'il existe plusieurs méthodes d'anonymisation, plus ou moins *efficaces* (au sens d'une protection plus ou moins « forte »). Pourquoi n'utiliserait-on pas toujours « la meilleure », ou dit autrement, est ce qu'il y a coût à payer pour avoir une anonymisation forte ? On pourrait imaginer qu'une partie du coût serait un coût temporel, i.e. qu'il serait très long de calculer une « bonne » anonymisation. On pourrait aussi penser que la force de l'anonymisation est inversement proportionnelle à la quantité (ou précision) des données publiées i.e. si on publie des informations au niveau d'un département on est à peu près 4 fois moins anonyme que si on les publie au niveau d'une région. Bien que tous ces facteurs entrent en jeu, le facteur déterminant est le *type de modèle d'anonymisation* utilisé.

[1] Institut national des sciences appliquées
[2] Institut national de recherche en informatique et en automatique

Il faut en effet prendre garde, car celui-ci peut restreindre l'exploitation future des données anonymes à certains types de calculs.
Pour que le lecteur puisse facilement se représenter les données manipulées, nous considérons une base de données constituée d'un ensemble d'enregistrements (appelés *n*-uplets) ayant chacun une structure identique, c'est-à-dire les mêmes champs, par exemple : numéro de sécurité sociale, nom, adresse, date de naissance, salaire, etc. (voir Figure 1). On repère dans un *n*-uplet des données dites *sensibles* comme une pathologie médicale, le salaire voire l'adresse.

Numéro de sécurité sociale ***(Identifiant)***	**Age**	**Code postal**	**Sexe**	**Pathologie** ***(Donnée sensible)***
2023475123123	75	75005	F	Cancer
2067875123123	40	75012	F	Grippe
1101175123123	12	78000	M	Grippe

Figure 1. Une base de données personnelles

Dans cet article, nous allons décrire cinq types de modèles d'anonymisation, qui cherchent à cacher ou briser le lien existant entre une personne du monde réel, et ses données sensibles : la pseudonymisation, le *k*-anonymat, la *l*-diversité, la *t*-proximité et la *differential privacy* (confidentialité différentielle, au sens du calcul différentiel). Nous illustrerons leur utilisation possible et leur degré de protection.

La pseudonymisation

La pseudonymisation consiste à supprimer les champs *directement* identifiants des enregistrements, et à rajouter à chaque enregistrement un nouveau champ, appelé *pseudonyme*, dont la caractéristique est qu'il doit rendre impossible tout lien entre cette nouvelle valeur et la personne réelle. Pour créer ce pseudonyme, on utilise souvent une *fonction de hachage* que l'on va appliquer à l'un des champs identifiants (par exemple le numéro de sécurité sociale), qui est un type de fonction particulier qui rend impossible (ou tout du moins extrêmement difficile) le fait de déduire la valeur initiale. On voit ainsi que deux entités possédant des informations sur une même personne, identifiée par son numéro de sécurité sociale, pourraient partager ces données de manière anonyme en *hachant* cet identifiant. Il est également possible d'utiliser tout simplement une fonction aléatoire pour générer un identifiant unique pour chaque personne, mais nous verrons plus bas que cela ne résout pas tous les problèmes.
Le gros avantage de la pseudonymisation est qu'il n'y a aucune limite sur le traitement subséquent des données. Tant que l'on traite des champs qui ne sont pas directement identifiants, on pourra exécuter exactement les mêmes calculs qu'avec une base de données non-anonyme. Ainsi, on montre dans la Figure 2 un exemple de calcul de la moyenne d'âge pour une pathologie donnée. L'utilisation de données pseudonymisées ne nuit pas à ce calcul.

Pseudonyme

Donnée sensible

ID	Age	CP	Sexe	Pathologie
1	75	75005	F	Cancer
2	40	75012	F	Grippe
3	12	78000	M	Grippe

Pathologie	MOY(age)
Cancer	75
Grippe	26

Figure 2. Pseudonymisation et exemple de calcul

Toutefois, la pseudonymisation n'est pas reconnue comme un moyen d'anonymisation, car elle ne donne pas un niveau de protection suffisamment élevé : la combinaison d'autres champs peut permettre de retrouver l'individu concerné. Sweeney l'a mis en évidence aux Etats-Unis en 2001 en croisant deux bases de données, une base de données médicale pseudonymisée et une liste électorale avec des données nominatives. Le croisement a été effectué non pas sur des champs directement identifiants, mais sur un triplet de valeurs : code postal, date de naissance et sexe, qui est unique pour environ 80% de la population des Etats-Unis[3] ! Elle a ainsi pu relier des données médicales à des individus (en l'occurrence le gouverneur de l'Etat).

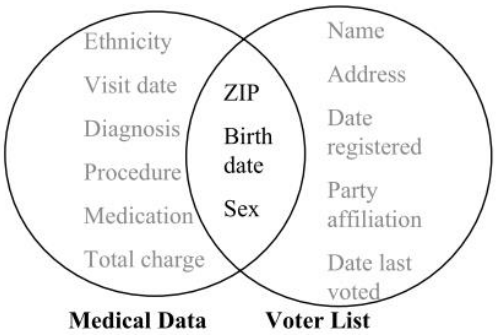

Figure 3. Un exemple de recoupement d'une base anonyme (source Sweeney 2002)

Le *k*-anonymat

[3] Autrement dit : une personne qui appartient à ce groupe de 80% de la population est seule à posséder son triplet code postal - date de naissance – sexe. Dans le complément de 20%, les personnes partagent leurs triplets avec une ou plusieurs autres personnes.

Afin de se protéger contre ce type d'attaque, appelée *record linkage*[4], Sweeney a proposé la technique de *k*-anonymat. Celle-ci va flouter la possibilité de lier un *n*-uplet anonyme à un *n*-uplet non anonyme de la manière suivante : 1) déterminer les ensembles d'attributs (appelés *quasi*-identifiants) qui peuvent être utilisés pour croiser les données anonymes avec des données identifiantes ; puis 2) réduire le niveau de détail des données de telle sorte qu'il y a au moins *k n*-uplets différents qui ont la même valeur de *quasi*-identifiant, une fois celui-ci généralisé (on dit alors que les individus font partie de la même *classe d'équivalence*). « Généraliser » signifie en fait « enlever un degré de précision » à certains champs. Ainsi, il est impossible d'être sûr à plus d'une chance sur *k* qu'on a bien lié un individu donné avec son *n*-uplet anonyme. L'avantage du *k*-anonymat est que l'analyse des données continue de fournir des résultats exacts, à ceci près qu'on ne peut pas dissocier les individus d'un groupe. Dans la Figure 4, nous montrons un exemple de généralisation des champs activité et âge d'une base de données médicales sur des étudiants et enseignants d'une université. Les étudiants sont identifiés par leur niveau d'étude (L3, M1, etc.), qui se généralise en « étudiant », et les enseignants par leur position académique (doctorant, maître de conférences, etc.), qui se généralise en « enseignant ». Nous traçons dans cette Figure l'origine de chaque *n*-uplet flouté.

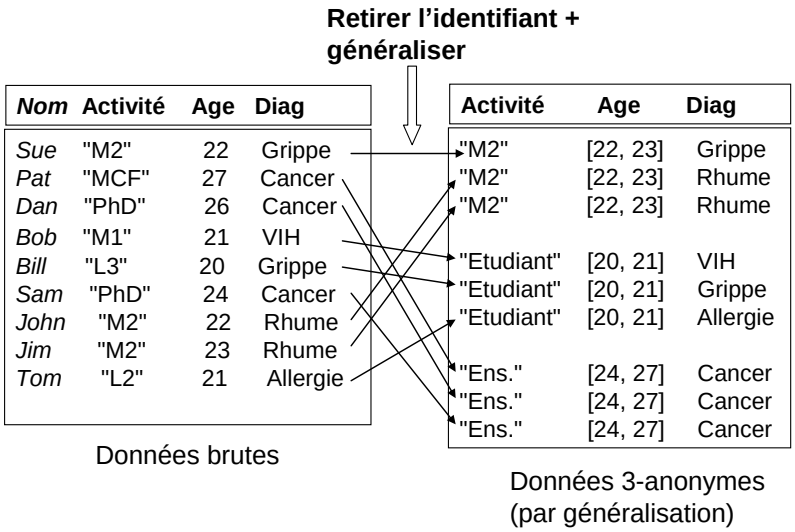

Figure 4. Anonymisation d'une table sur des données universitaires

Toutefois, une certaine quantité d'information sera déjà dévoilée, en particulier de l'information négative : si on connaît le *quasi-identifiant* d'une personne, on pourra exclure tout un ensemble de valeurs, ou bien savoir qu'elle a de plus grandes chances d'avoir une certaine valeur sensible. Certains cas peuvent aussi apparaître : si tous les individus d'une classe d'équivalence possèdent les mêmes valeurs sur un champ intéressant l'attaquant, alors celui-ci sera capable d'identifier cette valeur. Par exemple, en considérant les données de la Figure 4, on peut déduire qu'un enseignant ayant un âge entre 24 et 27 ans a forcément le cancer. Si on sait que Sam est un doctorant de 24 ans, alors on peut en déduire qu'il a le cancer.

Enfin, un problème technique important subsiste pour réaliser le *k*-anonymat : être capable de déterminer les généralisations à effectuer pour produire les *quasi*-identifiants, ce qui peut être fait soit par un expert humain qui connaît le domaine, ou bien par un calcul informatique, souvent très coûteux pour une base de données réelle.

[4] Liaison entre enregistrements

La *l*-diversité

Comme on l'a vu à la Figure 4, il est possible de déduire des informations dans certains cas pathologiques, sans faire le moindre croisement, par exemple si tous les individus d'une classe possèdent la même valeur sensible. Le modèle de la *l*-diversité répond à ce problème, en rajoutant une contrainte supplémentaire sur les classes d'équivalence : non seulement au moins *k n*-uplets doivent apparaître dans une classe d'équivalence, mais en plus le champ sensible associé à la classe d'équivalence doit prendre au moins *l* valeurs distinctes[5]. Dans l'exemple de la Figure 5, on voit que pour constituer de telles classes on doit parfois regrouper ensemble des étudiants et des enseignants. Leur activité est alors désignée de façon encore plus générale (« université »). Notons qu'on peut également lister les valeurs possibles, par exemple avoir une modalité « Étudiant ou Doctorant » (Etu/PhD).

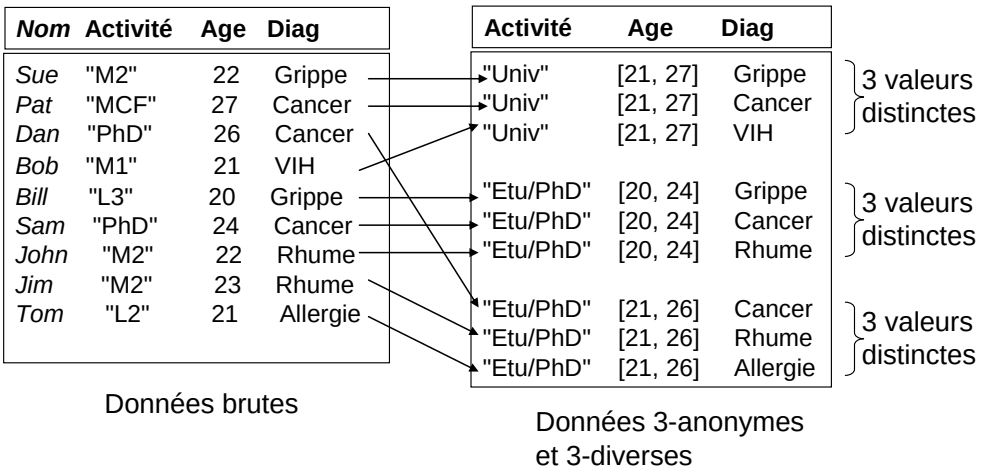

Figure 5. Données *l*-diverses

Cependant, en menant une attaque par croisement du même type que celle de Sweeney, il reste possible de déduire des informations. On voit par exemple dans la Figure 5 qu'on peut déduire qu'un étudiant de 20 ans aura une probabilité 0.33 (soit 1/*k*) d'avoir la grippe, 0.33 d'avoir le cancer et 0.33 d'avoir un rhume... et surtout aucune chance d'avoir une autre pathologie. Si on sait que Bill est la seule personne de la base dans ce cas de figure, alors on peut déduire des informations sensibles à son sujet.

La *t*-proximité

Pour essayer de réduire encore l'information qui peut être observée directement, on introduit le modèle de la *t*-proximité, toujours à partir d'un regroupement de données en classes d'équivalences selon le processus du *k*-anonymat. Ce nouveau modèle est basé sur une connaissance globale de la distribution des données sensibles, c'est-à-dire en ce cas les pathologies, pour essayer de faire coller au mieux les valeurs sensibles d'une classe d'équivalence à cette distribution, et ainsi éviter le problème de déduction d'informations soulevé par la *l*-diversité. Le facteur *t* que nous ne détaillons pas ici, indique dans quelle mesure on se démarque de la distribution globale.

[5] On peut généraliser à plusieurs champs sensibles.

Age	Sexe	Département	Pathologie	Nombre d'individus
<45	M	75	Grippe	400
<45	M	75	Rhume	800
>45	M	75	Grippe	500
>45	M	75	Rhume	1000
<35	F	75	Grippe	300
<35	F	75	Rhume	600
>35	F	75	Grippe	600
>35	F	75	Rhume	1200
…				

Figure 6. *t*-proximité

La *t*-proximité souffre de plusieurs problèmes, le plus important étant sans doute son utilité ! En effet, il parait évident d'exploiter des données *k*-anonymes ou même *l*-diverses pour découvrir des corrélations entre des données appartenant au quasi-identifiant et des données sensibles. Toutefois, le but même de la *t*-proximité est de réduire au maximum ces corrélations, puisque toutes les données sensibles de chaque classe d'équivalence vont se ressembler ! Ainsi, comme on le voit dans la Figure 6, la *t*-proximité permet surtout de répondre à la question suivante : *comment partitionner mes données de telle sorte que toutes les partitions se ressemblent en termes de distribution* ? Par exemple, si on imagine une base de données nationale sur des pathologies, comment regrouper les départements, classes d'âge et sexes, de telle sorte qu'on ait la même distribution des pathologies dans chaque sous-groupe. On peut s'interroger du jeu de données qui résulte de cette opération lorsqu'on souhaite précisément réaliser une analyse qui fait ressortir les facteurs qui différencient les individus.

La confidentialité différentielle (Différential Privacy)

Nous concluons ce survol des techniques d'anonymisation par la *confidentialité différentielle*, une méthode très en vogue dans les milieux de la recherche en informatique depuis quelques années, car contrairement aux méthodes précédentes, elle est la seule à donner des garanties formelles, c'est-à-dire des preuves mathématiques, sur la possibilité de borner les informations qu'on peut apprendre sur les individus. Cette méthode introduit un échantillonnage des données vraies (avec une probabilité α), et une génération de données fictives avec une probabilité $\beta << \alpha$ (mais ces données doivent naturellement rester réalistes...). Les garanties formelles sont cruciales, et permettent de quantifier le risque de ré-identification des *n*-uplets, d'où l'engouement pour cette méthode. En effet, en observant le jeu de données anonymes, l'information qu'on peut obtenir sur le fait qu'un *n*-uplet soit vrai ou faux est doublement bornée : on n'est jamais sûr qu' un *n*-uplet soit vrai avec une probabilité supérieure à α, ni qu'il soit faux avec une probabilité inférieure à β.

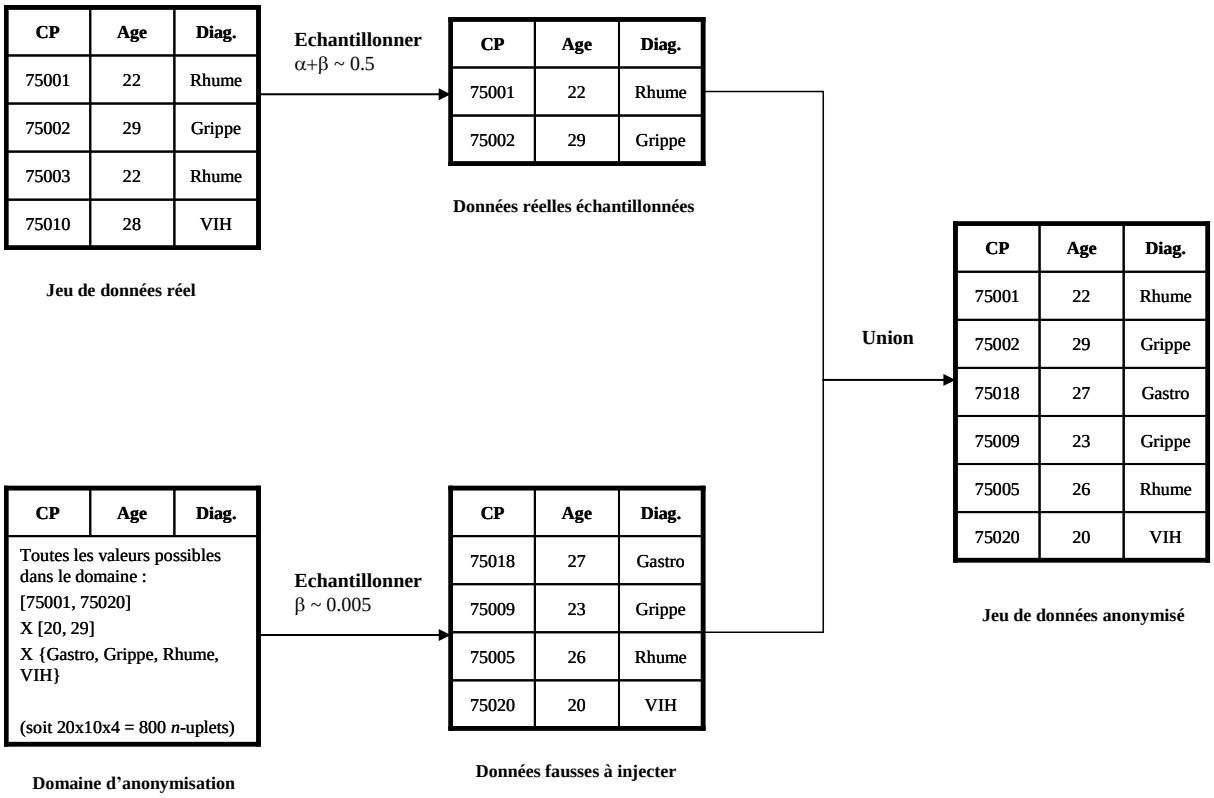

Figure 7. Confidentialité Différentielle

La confidentialité différentielle oblige à calculer un *estimateur* d'un agrégat que l'on souhaite connaître. Prenons l'exemple du calcul du nombre moyen de malades de la grippe par département, et supposons pour simplifier que les données fictives sont générées de manière équiprobable[6]. On peut estimer le nombre total de malades de la grippe par la fonction suivante, dont l'objectif est de soustraire le bruit (connu) introduit :

$$Nb_{Rhume_{estimé}} = \frac{(Nb_{Rhume_{anonyme}} - \beta \times Nb_{Rhume_{domaine}})}{\alpha} = (2 - 200 \times 0.005)/0.5 = 2$$

Le taux d'erreur peut également être estimé. Cependant, seules certaines fonctions d'agrégation peuvent être calculées avec une erreur bornée : moyenne, nombre total, etc. En revanche, on voit bien que calculer la valeur maximale d'une donnée numérique ne fait pas sens.

Outre cette restriction, le problème principal de la mise en œuvre de la confidentialité différentielle réside dans la vraisemblance des données fictives. Ainsi, cette technique s'applique surtout lorsqu'on cherche à protéger des données de géolocalisation, où il est facile de générer des données fausses « plausibles », et où les fonctions qu'on peut calculer avec cette technique d'anonymisation restent utiles (en particulier la densité et la distance). En revanche, comme on le voit sur l'exemple, il parait plus difficile d'exploiter cette méthode d'anonymat sur des données médicales.

Conclusion

[6] Il faudrait en réalité baser cette génération sur la distribution réelle des maladies.

Le problème de l'anonymisation des données, en vue d'assurer leur innocuité tout en permettant une analyse poussée, reste un problème ouvert. Même s'il existe des solutions pour garantir une certaine protection des données d'un individu (confidentialité différentielle), celles-ci sont difficiles à mettre en œuvre dans tous les domaines. Aussi, et même si elles ne permettent pas de réduire totalement le risque, on pourra de manière pratique recourir à des techniques de *k*-anonymat et *l*-diversité. En revanche, contrairement à ce que peut laisser penser son nom, la pseudonymisation n'est pas une technique d'anonymisation à proprement parler, et ne doit pas être utilisée en tant que telle.
Enfin, notons que de nombreux chercheurs, dont les équipes d'Alex Pentland aux Etats-Unis, ou Philippe Pucheral en France, préconisent d'utiliser la décentralisation du traitement des données, pour lutter entre autres contre le problème du croisement de données à l'insu des personnes concernées. Décentraliser les données signifie que chaque personne concernée par les données gère elle même leur stockage dans un serveur personnel de données (par exemple dans un espace personnel sur le *cloud*, ou dans du matériel sécurisé à son domicile) et autorise ou refuse les traitements.
En effet, puisque l'anonymisation parfaite n'existe pas, il est fondamental que l'utilisateur soit bien informé sur le modèle utilisé, ses performances et ses risques, et qu'il puisse adhérer, en connaissance de cause, aux traitements effectués.

Références

L. Sweeney : " k-anonymity: a model for protecting privacy" *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

A. Machanavajjhala , D. Kifer , J. Gehrke , et M. Venkatasubramaniam : "L-diversity : Privacy beyond k-anonymity" *ACM Transactions on Knowledge Discovery from Data*, 1(1):2007.

N. Li, T. Li, S. Venkatasubramanian : " t-closeness: Privacy beyond k-anonymity and l-diversity", *International Conference on Data Engineering*, 2007.

C. Dwork : "Differential Privacy", *International Colloquium on Automata, Languages and Programming*, 2006.

A. Pentland : Social Physics: how good ideas spread : the lessons from a new science - *Penguin Press*, 2014.

T. Allard, N. Anciaux, L. Bouganim, Y. Guo, L. Le Folgoc, B. Nguyen, P. Pucheral, Ij. Ray, Ik. Ray, S. Yin : Secure Personal Data Servers: a Vision Paper, dans *Very Large Data Bases*, 3(1): 25-35, 2010.