GROUPE DE TRAVAIL «ARTICLE 29» SUR LA PROTECTION DES DONNÉES

0829/14/FR
WP216

Avis 05/2014 sur les Techniques d'anonymisation

Adopté le 10 avril 2014

LE GROUPE DE TRAVAIL SUR LA PROTECTION DES PERSONNES À L'ÉGARD DU TRAITEMENT DES DONNÉES À CARACTÈRE PERSONNEL

institué par la directive 95/46/CE du Parlement européen et du Conseil du 24 octobre 1995,

vu les articles 29 et 30 de ladite directive,

vu son règlement intérieur,

A ADOPTÉ LE PRÉSENT AVIS:

RÉSUMÉ

Dans le présent avis, le groupe de travail «Article 29» analyse l'efficacité et les limites des techniques d'anonymisation existantes dans le contexte juridique de la protection des données dans l'Union et formule des recommandations pour l'utilisation de ces techniques en tenant compte du risque résiduel d'identification inhérent à chacune d'elles.

Le groupe de travail «Article 29» convient de l'intérêt potentiel de l'anonymisation, notamment comme stratégie permettant aux citoyens et à la société en général de bénéficier des avantages des «données ouvertes», tout en atténuant les risques pour les personnes concernées. Cependant, les études de cas et les recherches publiées ont montré combien il est difficile de créer un ensemble de données vraiment anonymes en conservant suffisamment d'informations sous-jacentes pour les besoins de la tâche concernée.

Au regard de la directive 95/46/CE et d'autres instruments juridiques pertinents de l'Union, l'anonymisation est le résultat du traitement des données personnelles afin d'empêcher, de façon irréversible, toute identification. Ce faisant, les responsables du traitement des données doivent tenir compte de plusieurs éléments, en prenant en considération l'ensemble des moyens «susceptibles d'être raisonnablement mis en œuvre» à des fins d'identification (soit par le responsable du traitement, soit par un tiers).

L'anonymisation constitue un traitement ultérieur des données à caractère personnel; à ce titre, elle doit satisfaire à l'exigence de compatibilité au regard des motifs juridiques et des circonstances du traitement ultérieur. De plus, si les données anonymisées sortent du champ d'application de la législation sur la protection des données, les personnes concernées peuvent néanmoins avoir droit à une protection au titre d'autres dispositions (comme celles qui protègent la confidentialité des communications).

Les principales techniques d'anonymisation, à savoir la randomisation et la généralisation, sont décrites dans le présent avis. Il y est notamment question d'ajout de bruit, de permutation, de confidentialité différentielle, d'agrégation, de k-anonymat, de l-diversité et de t-proximité. Les principes, les points forts et les points faibles de ces techniques sont expliqués, de même que les erreurs courantes et les échecs qui se rapportent à l'utilisation de chaque technique.

L'avis examine la fiabilité de chaque technique sur la base de trois critères:

i) est-il toujours possible d'isoler un individu?

ii) est-il toujours possible de relier entre eux les enregistrements relatifs à un individu? et

iii) peut-on déduire des informations concernant un individu?

La connaissance des principales forces et faiblesses de chaque technique peut être utile pour décider comment concevoir un processus d'anonymisation adéquat dans un contexte donné.

Il est aussi question de la pseudonymisation, afin d'éviter certains écueils et idées fausses: la pseudonymisation n'est pas une méthode d'anonymisation. Elle réduit simplement la corrélation d'un ensemble de données avec l'identité originale d'une personne concernée et constitue par conséquent une mesure de sécurité utile.

La conclusion du présent avis est que les techniques d'anonymisation peuvent apporter des garanties en matière de respect de la vie privée et peuvent servir à créer des procédés d'anonymisation efficaces, mais uniquement si leur application est correctement conçue – ce qui suppose que les conditions préalables (le contexte) et les objectif(s) du processus d'anonymisation soient clairement définis de façon à parvenir à l'anonymisation visée, tout en produisant des données utiles. Le choix de la solution optimale devrait s'opérer au cas par cas, en utilisant éventuellement une combinaison de techniques différentes, sans perdre de vue les recommandations pratiques formulées dans cet avis.

Enfin, les responsables du traitement des données devraient être conscients qu'un ensemble de données anonymisées peut encore présenter des risques résiduels pour les personnes concernées. En effet, d'une part, l'anonymisation et la ré-identification sont des domaines de recherche très actifs où de nouvelles découvertes sont régulièrement publiées et, d'autre part, même des données anonymisées, comme les statistiques, peuvent servir à étoffer des profils existants, créant ainsi de nouveaux problèmes en termes de protection des données. C'est pourquoi l'anonymisation ne doit pas être considérée comme un exercice ponctuel: il appartient aux responsables du traitement des données de réévaluer régulièrement les risques associés.

1 Introduction

Alors que les appareils, les capteurs et les réseaux engendrent des volumes considérables et de nouveaux types de données et que le coût de leur stockage devient négligeable, les perspectives de réutilisation de ces données suscitent dans le public un intérêt et une demande qui ne cessent de croître. Les «données ouvertes» peuvent apporter des avantages évidents à la société, aux citoyens et aux organisations, à condition cependant que soit respecté le droit de chacun à la protection de ses données à caractère personnel et de sa vie privée.

L'anonymisation peut constituer une bonne stratégie afin de préserver ces avantages tout en atténuant les risques. Dès lors qu'un ensemble de données est vraiment anonymisé et que les individus ne sont plus identifiables, la législation européenne sur la protection des données ne s'applique plus. Toutefois, les études de cas et les travaux de recherche publiés font clairement ressortir toute la difficulté de créer un ensemble de données vraiment anonymes à partir d'une profusion de données à caractère personnel, en conservant suffisamment d'informations sous-jacentes pour les besoins de la tâche concernée. Par exemple, un ensemble de données considérées comme anonymes peut être combiné avec un autre ensemble de données de telle façon qu'un ou plusieurs individus deviennent identifiables.

Dans le présent avis, le groupe de travail «Article 29» analyse l'efficacité et les limites des techniques d'anonymisation existantes dans le contexte juridique de la protection des données dans l'Union et formule des recommandations pour une utilisation prudente et responsable de ces techniques afin de mettre en place un processus d'anonymisation.

2 Définitions et analyse juridique

2.1. Définitions dans le contexte législatif de l'Union

La directive 95/46/CE mentionne l'anonymisation au considérant 26 pour exclure les données anonymisées du champ d'application de la législation sur la protection des données:

> *«considérant que les principes de la protection doivent s'appliquer à toute information concernant une personne identifiée ou identifiable; 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; que les principes de la protection ne s'appliquent pas aux données rendues anonymes d'une manière telle que la personne concernée n'est plus identifiable; que les codes de conduite au sens de l'article 27 peuvent être un instrument utile pour fournir des indications sur les moyens par lesquels les données peuvent être rendues anonymes et conservées sous une forme ne permettant plus l'identification de la personne concernée»*[1].

[1] Il est à noter, de surcroît, que c'est aussi l'approche suivie dans le projet de règlement de l'Union sur la protection des données, au considérant 23: «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.»

Une lecture attentive du considérant 26 permet d'en tirer une définition conceptuelle de l'anonymisation. Le considérant 26 signifie que, pour rendre des données anonymes, il faut en retirer suffisamment d'éléments pour que la personne concernée ne puisse plus être identifiée. Plus précisément, les données doivent être traitées de façon à ne plus pouvoir être utilisées pour identifier une personne physique en recourant à «l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre», soit par le responsable du traitement, soit par un tiers. Un facteur important est que le traitement doit être irréversible. La directive ne précise pas comment un tel processus d'anonymisation devrait ou pourrait être exécuté[2]. L'accent est mis sur le résultat: il faut faire en sorte que les données ne permettent pas d'identifier la personne concernée par «l'ensemble» des moyens «susceptibles» d'être «raisonnablement» employés. Il est fait référence aux codes de conduite en tant qu'instrument utile pour envisager des mécanismes possibles d'anonymisation et de conservation des données sous une forme «ne permettant plus» l'identification de la personne concernée. La directive place donc manifestement la barre très haut.

La directive «vie privée et communications électroniques» (directive 2002/58/CE) évoque aussi l'anonymisation et les données anonymes dans une optique très similaire. Le considérant 26 indique:

> *«Il convient également d'effacer ou de rendre anonymes les données relatives au trafic utilisées pour la commercialisation de services de communications ou pour la fourniture de services à valeur ajoutée, lorsque les services en question ont été fournis.»*

En conséquence de quoi, l'article 6, paragraphe 1, dispose:

> *«Les données relatives au trafic concernant les abonnés et les utilisateurs traitées et stockées par le fournisseur d'un réseau public de communications ou d'un service de communications électroniques accessibles au public doivent être effacées ou rendues anonymes lorsqu'elles ne sont plus nécessaires à la transmission d'une communication sans préjudice des paragraphes 2, 3 et 5, du présent article ainsi que de l'article 15, paragraphe 1.»*

De plus, conformément à l'article 9, paragraphe 1:

> *«Lorsque des données de localisation, autres que des données relatives au trafic, concernant des utilisateurs ou abonnés de réseaux publics de communications ou de services de communications électroniques accessibles au public ou des abonnés à ces réseaux ou services, peuvent être traitées, elles ne le seront qu'après avoir été rendues anonymes ou moyennant le consentement des utilisateurs ou des abonnés, dans la mesure et pour la durée nécessaires à la fourniture d'un service à valeur ajoutée.»*

Le raisonnement qui sous-tend ces dispositions est que le résultat de l'anonymisation, en tant que technique appliquée aux données à caractère personnel, devrait être, dans l'état actuel de la technologie, aussi permanent qu'un effacement, c'est-à-dire qu'il devrait rendre impossible tout traitement de données à caractère personnel[3].

---

[2] Cette notion est approfondie en p. 9 du présent avis.

[3] Il convient de rappeler ici que l'anonymisation est aussi définie dans des normes internationales – notamment la norme ISO 29100 – comme étant le «processus par lequel des informations personnellement identifiables (IPI) sont irréversiblement altérées de telle façon que le sujet des IPI ne puisse plus être identifié directement ou indirectement, que ce soit par le responsable du traitement des IPI seul ou en collaboration avec une quelconque autre partie» (ISO 29100:2011). L'irréversibilité de l'altération subie par les données personnelles pour ne plus permettre l'identification directe ou indirecte est donc aussi un élément déterminant pour l'ISO. De ce point de vue, la norme présente une convergence considérable avec les principes et les notions qui sous-tendent la

2.2. Analyse juridique

L'analyse du libellé des dispositions relatives à l'anonymisation dans les principaux instruments de protection des données de l'Union permet de retenir quatre aspects essentiels:

- L'anonymisation peut être le résultat du traitement de données à caractère personnel dans le but d'empêcher irréversiblement l'identification de la personne concernée.

- Plusieurs techniques d'anonymisation peuvent être envisagées; il n'y a pas de normes prescriptives dans la législation de l'Union.

- Les éléments contextuels ont leur importance: il faut prendre en considération «l'ensemble» des moyens «susceptibles» d'être «raisonnablement» utilisés à des fins d'identification par le responsable du traitement ou par des tiers, en prêtant une attention particulière aux moyens que l'état actuel de la technologie a rendu récemment «susceptibles» d'être «raisonnablement» mis en œuvre (compte tenu de l'évolution de la puissance de calcul et des outils disponibles).

- Pour apprécier la validité d'une technique d'anonymisation, il faut tenir compte du facteur de risque qui lui est inhérent – et notamment des utilisations possibles des données «anonymisées» au moyen de cette technique – et évaluer la gravité et la probabilité de ce risque.

L'expression «technique d'anonymisation» est employée dans le présent avis, plutôt que les termes «anonymat» ou «données anonymes», afin d'insister sur le risque résiduel de ré-identification inhérent à toute mesure technique ou organisationnelle visant à rendre des données «anonymes».

2.2.1. Licéité du processus d'anonymisation

Premièrement, l'anonymisation est une technique appliquée aux données à caractère personnel afin d'empêcher irréversiblement leur identification. L'hypothèse de départ est donc que les données à caractère personnel doivent avoir été collectées et traitées dans le respect de la législation applicable en matière de conservation des données sous une forme identifiable.

Dans ce contexte, le processus d'anonymisation, désignant le traitement de telles données à caractère personnel en vue de les rendre anonymes, constitue un «traitement ultérieur». À ce titre, ce traitement doit satisfaire au critère de compatibilité conformément aux lignes directrices proposées par le groupe de travail «Article 29» dans son avis 03/2013 sur la limitation des finalités[4].

Il s'ensuit que la base juridique de l'anonymisation peut, en principe, résider dans l'un des motifs mentionnés à l'article 7 (notamment l'intérêt légitime poursuivi par le responsable du

---

directive 95/46/CE. Cela s'applique aussi aux définitions que l'on peut trouver dans certaines législations nationales (par exemple, en Italie, en Allemagne et en Slovénie), qui insistent sur le caractère non identifiable et qui font référence à l'«effort disproportionné» qu'exigerait une ré-identification (D, SI). La loi française relative à la protection des données prévoit néanmoins que les données conservent un caractère personnel même si la ré-identification de la personne concernée est rendue très difficile et improbable – c'est-à-dire qu'il n'y a pas de disposition renvoyant au critère du «caractère raisonnable».

[4] Avis 03/2013 du groupe de travail «Article 29», disponible à l'adresse: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

traitement des données) pour autant que les exigences de qualité des données visées à l'article 6 de la directive soient également satisfaites, en tenant dûment compte des circonstances spécifiques et de tous les facteurs mentionnés dans l'avis du groupe de travail «Article 29» sur la limitation des finalités[5].

D'un autre côté, il faut mentionner les dispositions énoncées à l'article 6, paragraphe 1, point e), de la directive 95/46/CE (mais aussi à l'article 6, paragraphe 1, et à l'article 9, paragraphe 1, de la directive «vie privée et communications électroniques»), qui soulignent la nécessité de conserver des données à caractère personnel «sous une forme permettant l'identification» pendant une durée n'excédant pas celle nécessaire à la réalisation des finalités pour lesquelles elles sont collectées ou pour lesquelles elles sont traitées ultérieurement.

En soi, cette exigence insiste sur le fait que les données à caractère personnel devraient, à tout le moins, être anonymisées «par défaut» (sous réserve de dispositions juridiques différentes, comme celles mentionnées dans la directive «vie privée et communications électroniques» à propos des données relatives au trafic). Si le responsable du traitement des données souhaite conserver ces données à caractère personnel après que les finalités du traitement original ou ultérieur ont été réalisées, des techniques d'anonymisation devraient être appliquées de façon à empêcher irréversiblement l'identification.

Par conséquent, le groupe de travail «Article 29» considère que l'anonymisation, en tant que traitement ultérieur de données à caractère personnel, peut être jugée compatible avec les finalités originales du traitement, à condition que le processus d'anonymisation soit de nature à produire des informations anonymisées au sens décrit dans le présent document.

Il faut ajouter que l'anonymisation doit demeurer conforme aux contraintes juridiques rappelées par la Cour de justice dans son arrêt C-553/07 (College van burgemeester en wethouders van Rotterdam/M.E.E. Rijkeboer), concernant la nécessité de conserver les données sous une forme identifiable pour permettre, par exemple, l'exercice des droits d'accès des personnes concernées. La Cour a jugé que *«[l]'article 12, sous a), de la directive [95/46/CE] impose aux États membres de prévoir un droit d'accès à l'information sur les destinataires ou les catégories de destinataires des données ainsi qu'au contenu de l'information communiquée non seulement pour le présent, mais aussi pour le passé. Il appartient aux États membres de fixer un délai de conservation de cette information ainsi qu'un accès corrélatif à celle-ci qui constituent un juste équilibre entre, d'une part, l'intérêt de la personne concernée à protéger sa vie privée, notamment au moyen des voies d'intervention et de recours prévus par la directive et, d'autre part, la charge que l'obligation de conserver cette information représente pour le responsable du traitement.»*

Cela vaut en particulier dans le cas où un responsable du traitement des données s'appuie sur l'article 7, point f), de la directive 95/46/CE, en ce qui concerne l'anonymisation: l'intérêt légitime poursuivi par le responsable du traitement des données doit toujours être mis en balance avec les droits et les libertés fondamentales des personnes concernées.

[5] Cela signifie notamment qu'il faut procéder à une appréciation matérielle à la lumière de toutes les circonstances pertinentes, en prêtant une attention particulière aux facteurs-clés suivants:
a) la relation entre les finalités pour lesquelles les données à caractère personnel ont été collectées et les finalités du traitement ultérieur;
b) le contexte dans lequel les données à caractère personnel ont été collectées et les attentes raisonnables des personnes concernées à propos de leur utilisation ultérieure;
c) la nature des données à caractère personnel et l'impact du traitement ultérieur sur les personnes concernées;
d) les garanties appliquées par le responsable du traitement pour assurer un traitement équitable et éviter tout impact excessif sur les personnes concernées.

Par exemple, une enquête des autorités néerlandaises chargées de la protection des données DPA en 2012-2013 sur le recours à des technologies d'inspection approfondie des paquets par quatre opérateurs de téléphonie mobile a mis en évidence un fondement juridique, au titre de l'article 7, point f) de la directive 95/46/CE, justifiant l'anonymisation des contenus des données relatives au trafic dès que possible après la collecte de ces données. En effet, l'article 6 de la directive «vie privée et communications électroniques» stipule que les données relatives au trafic concernant les abonnés et les utilisateurs traitées et stockées par le fournisseur d'un réseau public de communications ou d'un service de communications électroniques accessibles au public doivent être effacées ou rendues anonymes aussi rapidement que possible. Dans ce cas, dès lors que l'article 6 de la directive «vie privée et communications électroniques» le permet, il existe un fondement juridique correspondant dans l'article 7 de la directive sur la protection des données. L'inverse est vrai également: si un type de traitement de données n'est pas autorisé au titre de l'article 6 de la directive «vie privée et communications électroniques», l'article 7 de la directive sur la protection des données ne peut constituer un fondement juridique.

2.2.2. Caractère potentiellement identifiable des données anonymisées

Le groupe de travail «Article 29» a examiné en détail le concept de données à caractère personnel dans son avis 4/2007 sur les données à caractère personnel, en concentrant son attention sur les éléments constitutifs de la définition figurant à l'article 2, point a), de la directive 95/46/CE, et notamment la mention «identifiée ou identifiable» qui fait partie de cette définition. Dans ce contexte, le groupe de travail «Article 29» a aussi conclu que les «"données anonymisées" sont donc des données anonymes qui concernaient auparavant une personne identifiable, mais ne permettent plus cette identification».

De ce fait, le groupe de travail «Article 29» a déjà fait ressortir que le critère des «moyens susceptibles d'être raisonnablement mis en œuvre» évoqué par la directive doit être appliqué pour apprécier si le procédé d'anonymisation est suffisamment fiable, c'est-à-dire si l'identification est devenue «raisonnablement» impossible. Le contexte et les circonstances propres à chaque cas spécifique ont un impact direct sur le caractère identifiable. Dans l'annexe technique jointe au présent avis, les conséquences du choix de la technique la plus appropriée sont analysées.

Ainsi qu'il a déjà été signalé, les recherches, les outils et la puissance de calcul évoluent. Il n'est par conséquent ni réaliste ni utile de dresser une liste exhaustive des circonstances dans lesquelles l'identification n'est plus possible. Cependant, certains facteurs-clés méritent d'être pris en considération et illustrés.

Premièrement, il peut être avancé que les responsables du traitement des données devraient concentrer leur attention sur les moyens concrets qui seraient nécessaires pour inverser la technique d'anonymisation, notamment en termes de coût et de savoir-faire requis pour mettre en œuvre ces moyens, et sur l'appréciation de leur probabilité et de leur gravité. Par exemple, les responsables du traitement des données devraient mettre en balance leurs efforts d'anonymisation et les coûts résultants (en termes de temps et de ressources) avec la disponibilité croissante, à peu de frais, des moyens techniques permettant d'identifier des individus dans des ensembles de données, l'accessibilité publique croissante d'autres ensembles de données (comme ceux mis à disposition dans le cadre de politique de «données ouvertes»), et les nombreux exemples d'anonymisation incomplète entraînant par la suite des

effets négatifs, parfois irréparables, pour les personnes concernées[6]. Il est à noter que le risque d'identification peut augmenter avec le temps et dépend aussi des progrès des technologies de l'information et des communications. Les règles juridiques doivent donc, le cas échéant, être formulées d'une manière technologiquement neutre et tenir compte, dans l'idéal, des capacités d'évolution des technologies de l'information[7].

Deuxièmement, «les moyens» qu'il convient de considérer «pour déterminer si une personne est identifiable» sont ceux «susceptibles d'être raisonnablement mis en œuvre, soit par le responsable du traitement, soit par une autre personne». Il est donc crucial de comprendre que, dans le cas où un responsable du traitement des données n'efface pas les données originales (identifiables) au niveau des événements individuels et transmet une partie de cet ensemble de données (par exemple après avoir supprimé ou masqué les données identifiables), l'ensemble de données résultant constitue encore des données à caractère personnel. Ce n'est que si les données sont agrégées par le responsable de leur traitement à un niveau où les événements individuels ne sont plus identifiables que l'ensemble de données résultant peut être qualifié d'anonyme. Par exemple: si une organisation collecte des données sur des déplacements individuels, les habitudes de voyage au niveau des événements individuels pourraient encore être considérées comme des données à caractère personnel pour toute partie intéressée, tant que le responsable du traitement des données (ou un tiers) continue à avoir accès aux données brutes originales, même si les identifiants directs ont été supprimés de l'ensemble de données transmis à des tiers. Mais si le responsable du traitement des données efface les données brutes et ne transmet à des tiers que des statistiques agrégées à un niveau supérieur, par exemple «le lundi, sur le trajet X, le nombre de passagers est supérieur de 160 % à celui du mardi», ces données pourraient être qualifiées d'anonymes.

Une solution d'anonymisation efficace doit empêcher toutes les parties d'isoler un individu dans un ensemble de données, de relier entre eux deux enregistrements dans un ensemble de données (ou dans deux ensembles de données séparés) et de déduire des informations de cet ensemble de données. D'une manière générale, il ne suffit donc pas de supprimer directement des éléments qui sont, en eux-mêmes, identifiants pour garantir que toute identification de la personne n'est plus possible. Il sera souvent nécessaire de prendre des mesures supplémentaires pour empêcher l'identification, toujours en fonction du contexte et des finalités du traitement auquel sont destinées les données anonymisées.

EXEMPLE:
Les profils génétiques constituent un exemple de données à caractère personnel qui, en raison du caractère unique de certains profils, peuvent présenter un risque d'identification si la seule technique utilisée est la suppression de l'identité du donneur. Il a déjà été démontré dans la littérature[8] que la combinaison de ressources génétiques publiquement disponibles (par exemple, des registres généalogiques, des notices nécrologiques, les résultats obtenus en interrogeant des moteurs de recherche) et des métadonnées concernant les données d'ADN (date du prélèvement, âge, lieu de résidence) peut révéler l'identité de certains individus même si cet ADN a été donné «anonymement».

[6] Il est intéressant de relever que les amendements du Parlement européen au projet de règlement général sur la protection des données récemment proposé (21 octobre 2013) mentionnent spécifiquement au considérant 23: «Pour établir si des moyens sont raisonnablement susceptibles d'être mis en œuvre afin d'identifier une personne physique, il convient de considérer l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte à la fois des technologies disponibles au moment du traitement et de l'évolution de celles-ci.».

[7] Voir l'avis 4/2007 du groupe de travail «Article 29», p. 16.

[8] Voir John Bohannon, «Genealogy Databases Enable Naming of Anonymous DNA Donors», *Science*, vol. 339, n° 6117 (18 janvier 2013), p. 262.

Les deux grandes familles de techniques d'anonymisation – la randomisation et la généralisation des données[9] – ont leurs lacunes; cependant, chacune d'elles peut être appropriée, selon les circonstances et le contexte, pour atteindre la finalité souhaitée sans compromettre le droit des personnes concernées au respect de leur vie privée. Il faut insister sur le fait que l'«identification» ne désigne pas simplement la possibilité de retrouver le nom et/ou l'adresse d'une personne, mais inclut aussi la possibilité de l'identifier par un procédé d'individualisation, de corrélation ou d'inférence. De surcroît, pour que la législation sur la protection des données s'applique, peu importe que les intentions soient celles du responsable du traitement des données ou de celui à qui elles sont destinées. Du moment que les données sont identifiables, les règles en matière de protection des données s'appliquent.

Quand des tiers traitent un ensemble de données auquel une technique d'anonymisation a été appliquée (données anonymisées et communiquées par le responsable de leur traitement à l'origine), ils ne sont pas tenus d'observer les exigences de protection des données pour autant qu'ils ne puissent pas identifier (directement ou indirectement) les personnes concernées dans l'ensemble de données original. Cependant, les tiers doivent prendre en compte les facteurs contextuels et circonstanciels mentionnés précédemment (y compris les spécificités des techniques d'anonymisation appliquées par le responsable du traitement des données à l'origine) pour décider comment ils comptent exploiter et, en particulier, combiner ces données anonymisées pour leur propre usage – car les conséquences résultantes peuvent entraîner différents types de responsabilité de leur part. Dans le cas où ces facteurs et ces caractéristiques sont de nature à comporter un risque inacceptable d'identification des personnes concernées, le traitement entre de nouveau dans le champ d'application de la législation en matière de protection des données.

La liste présentée plus haut ne se veut en aucune façon exhaustive, mais vise plutôt à donner une orientation générale pour l'appréciation du caractère potentiellement identifiable d'un ensemble de données selon les différentes techniques d'anonymisation disponibles qui lui sont appliquées. Tous les aspects mentionnés ci-dessus peuvent être considérés comme autant de facteurs de risque que doivent peser aussi bien les responsables du traitement qui anonymisent les ensembles de données que les tiers qui exploitent ces ensembles de données «anonymisés» pour leur propre usage.

2.2.3. Risques de l'utilisation de données anonymisées

Quand ils envisagent de recourir à des techniques d'anonymisation, les responsables du traitement des données doivent tenir compte des risques suivants.

- Un piège à éviter en particulier est de considérer les données pseudonymisées comme équivalentes à des données anonymisées. La section consacrée à l'analyse technique expliquera que les données pseudonymisées ne peuvent être assimilées à des informations anonymisées puisqu'elles continuent à permettre l'individualisation d'une personne concernée et la corrélation entre différents ensembles de données. Le pseudonymat n'est pas de nature à empêcher qu'une personne concernée soit identifiable et reste donc dans le champ d'application du régime juridique de la protection des données. Cela vaut en particulier dans le contexte des recherches scientifiques, statistiques ou historiques[10].

[9] Les principales caractéristiques de ces deux techniques d'anonymisation et leurs différences sont décrites à la section 3 ci-après («Analyse technique»).
[10] Voir aussi l'avis 4/2007 du groupe de travail «Article 29», p. 19 à 21.

EXEMPLE:
L'«affaire AOL (America On Line)» illustre de manière typique les idées fausses qui entourent la pseudonymisation. En 2006, une base de données contenant vingt millions de mots-clés figurant dans les recherches effectuées par plus de 650 000 utilisateurs au cours d'une période de 3 mois a été diffusée publiquement, sans autre mesure destinée à préserver la vie privée que le remplacement de l'identifiant d'utilisateur AOL par un attribut numérique. À la suite de quoi, l'identité et la localisation de certains utilisateurs ont été rendues publiques. Les requêtes transmises à un moteur de recherches, surtout si elles peuvent être couplées avec d'autres attributs, comme les adresses IP ou d'autres paramètres de configuration, ont un potentiel d'identification très élevé.

- Une deuxième erreur serait de considérer que les individus n'ont plus aucune garantie dès lors que les données ont été correctement anonymisées (ayant satisfait à l'ensemble des conditions et critères mentionnés ci-dessus et sortant, par définition, du champ d'application de la directive sur la protection des données) – d'abord et avant tout parce que d'autres actes législatifs peuvent s'appliquer à l'utilisation de ces données. Par exemple, l'article 5, paragraphe 3, de la directive «vie privée et communications électroniques» interdit le stockage d'«informations» de quelque type que ce soit (y compris les informations non personnelles) et l'accès à ces informations sur l'équipement terminal d'un abonné ou d'un utilisateur sans son accord, dans le cadre du principe plus large de la confidentialité des communications.

- Une troisième négligence résulterait aussi de ne pas envisager l'impact que des données correctement anonymisées peuvent avoir sur les individus dans certaines circonstances, en particulier dans le cas du profilage. La vie privée des personnes est protégée par l'article 8 de la CEDH et par l'article 7 de la Charte des droits fondamentaux de l'Union européenne; de ce fait, même si la législation sur la protection des données ne s'applique plus à ce type de données, l'usage qui est fait des ensembles de données anonymisées et mises à la disposition de tiers peut entraîner une atteinte à la vie privé. Une prudence particulière s'impose dans la manipulation d'informations anonymisées, surtout lorsque ces informations servent (souvent en combinaison avec d'autres données) à prendre des décisions qui produisent des effets (même indirectement) sur les individus. Ainsi qu'il a déjà été signalé dans le présent avis et comme l'a clairement précisé le groupe de travail «Article 29» notamment dans son avis sur la notion de «limitation des finalités» (avis 03/2013)[11], les attentes légitimes des personnes concernées quant au traitement ultérieur de leurs données doivent être appréciées à la lumière des facteurs contextuels pertinents – comme la nature de la relation entre les personnes concernées et les responsables du traitement des données, les obligations juridiques applicables, la transparence des opérations de traitement.

3 Analyse technique, fiabilité des technologies et erreurs typiques

Il existe différentes pratiques et techniques d'anonymisation, avec des degrés de fiabilité variables. Cette section portera sur les principaux points que les responsables du traitement des données doivent prendre en considération quand ils choisissent d'appliquer une technique donnée, au regard notamment des garanties offertes par cette technique, en tenant compte de l'état actuel de la technologie et en envisageant trois risques essentiels en matière d'anonymisation:

[11] Disponible à l'adresse http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

- *l'individualisation*, qui correspond à la possibilité d'isoler une partie ou la totalité des enregistrements identifiant un individu dans l'ensemble de données;
- *la corrélation,* qui consiste dans la capacité de relier entre elles, au moins, deux enregistrements se rapportant à la même personne concernée ou à un groupe de personnes concernées (soit dans la même base de données, soit dans deux bases de données différentes). Si une attaque permet d'établir (par exemple, au moyen d'une analyse de corrélation) que deux enregistrements correspondent à un même groupe d'individus, mais ne permet pas d'isoler des individus au sein de ce groupe, la technique résiste à l'«individualisation», mais non à la corrélation;
- *l'inférence*, qui est la possibilité de déduire, avec un degré de probabilité élevé, la valeur d'un attribut à partir des valeurs d'un ensemble d'autres attributs.

Une solution résistant à ces trois risques offrirait par conséquent une protection fiable contre les tentatives de ré-identification utilisant les moyens les plus susceptibles d'être raisonnablement mis en œuvre par le responsable du traitement des données ou par des tiers. Le groupe de travail «Article 29» souligne, à cet égard, que les recherches en matière de techniques d'anonymisation se poursuivent et font apparaître invariablement qu'aucune technique n'est, en soi, infaillible. En termes généraux, on distingue deux grandes approches de l'anonymisation: la première repose sur la ***randomisation*** tandis que la seconde se fonde sur la ***généralisation***. D'autres notions sont aussi abordées dans le présent avis, comme la *pseudonymisation*, la *confidentialité différentielle*, la *l-diversité* et la *t-proximité*.

Le vocabulaire suivant est employé dans cette section: un ensemble de données se compose des différents enregistrements relatifs à des individus (les personnes concernées). Chaque enregistrement se rapporte à une personne concernée et comporte une série de valeurs (ou «entrées», par exemple: 2013) pour chaque attribut (par exemple, l'année). Un ensemble de données est donc un groupe d'enregistrements qui peuvent être présentés tantôt sous la forme d'un tableau (ou de plusieurs tableaux), tantôt sous la forme d'un graphique annoté/pondéré, ce qui est de plus en plus souvent le cas aujourd'hui. Les exemples donnés dans le présent avis porteront sur des tableaux, mais ils s'appliquent aussi à d'autres représentations graphiques des enregistrements. Les combinaisons d'attributs qui se rapportent à une personne concernée ou à un groupe de personnes concernées peuvent être désignées par le terme de «quasi-identifiants». Dans certains cas, un ensemble de données peut comporter des enregistrements multiples pour un même individu. Un «attaquant» est un tiers (c'est-à-dire ni le responsable du traitement des données ni un sous-traitant) qui accède aux enregistrements originaux par accident ou de manière intentionnelle.

3.1. Randomisation

La randomisation est une famille de techniques qui altèrent la véracité des données afin d'affaiblir le lien entre les données et l'individu. Si les données sont suffisamment incertaines, elles ne peuvent plus être rattachées à un individu en particulier. En elle-même, la randomisation ne réduira pas la singularité de chaque enregistrement, qui sera toujours dérivé d'une seule personne concernée, mais elle peut apporter une protection contre les attaques/risques relevant de l'inférence et peut être combinée avec des techniques de généralisation pour offrir de meilleures garanties de respect de la vie privée. Des techniques supplémentaires peuvent se révéler nécessaires pour empêcher qu'un enregistrement permette d'identifier un individu.

3.1.1. Ajout de bruit

La technique d'ajout de bruit est particulièrement utile quand des attributs peuvent avoir un effet négatif important sur des individus et consiste à modifier des attributs dans l'ensemble de données pour les rendre moins précis, tout en conservant la distribution générale. Pour traiter un ensemble de données, un observateur supposera que les valeurs sont exactes, mais ce ne sera vrai qu'à un certain degré. Par exemple, si la taille d'un individu a été mesurée à l'origine au centimètre près, l'ensemble de données anonymisées peut présenter une précision de ± 10 cm seulement. Si cette technique est appliquée efficacement, un tiers ne sera pas en mesure d'identifier un individu ni ne pourra restaurer les données ou discerner de quelque autre façon comment les données ont été modifiées.

L'ajout de bruit devra ordinairement être combiné avec d'autres techniques d'anonymisation comme la suppression des attributs évidents et des quasi-identifiants. Le niveau de bruit devrait dépendre du niveau d'information requis et de l'impact que la divulgation des attributs protégés aurait sur le respect de la vie privée des individus.

3.1.1.1. Garanties

- *Individualisation:* Il reste possible d'isoler les enregistrements correspondant à un individu (peut-être de manière non identifiable), même si les enregistrements sont moins fiables.
- *Corrélation:* Il reste possible relier entre eux les enregistrements correspondant au même individu, mais ces enregistrements sont moins fiables et il peut donc arriver qu'un enregistrement réel soit relié à un enregistrement ajouté artificiellement (c'est-à-dire à un «bruit»). Dans certains cas, une attribution erronée pourrait exposer une personne concernée à un niveau de risque considérable, voire plus élevé que celui résultant d'une attribution correcte.
- *Inférence:* Une attaque par inférence est peut-être possible, mais le taux de succès sera moins élevé et certains faux positifs (et faux négatifs) sont plausibles.

3.1.1.2. Erreurs courantes

- Ajout de bruit incohérent: Si le bruit n'est pas sémantiquement viable (c'est-à-dire s'il est disproportionné et ne respecte pas la logique entre les attributs d'un ensemble), un attaquant ayant accès à la base de données sera en mesure de le filtrer et, dans certains cas, de recréer les entrées manquantes. De plus, si l'ensemble de données est trop clairsemé[12], il peut arriver qu'il reste possible de relier les entrées de données bruitées avec une source extérieure.
- Supposer que l'ajout de bruit est suffisant: l'ajout de bruit est une mesure complémentaire qui rend plus difficile la récupération des données à caractère personnel par un attaquant. À moins que le bruit ne soit plus élevé que le niveau d'information contenu dans l'ensemble de données, il ne faut pas supposer que l'ajout de bruit représente une solution d'anonymisation qui se suffit à elle-même.

3.1.1.3. Échecs de l'ajout de bruit

Une expérience de ré-identification très connue est celle réalisée sur la base de données des clients du fournisseur de contenu vidéo Netflix. Des chercheurs ont analysé les

[12] Cette notion est examinée plus en détail dans l'annexe, en p. 33.

propriétés géométriques de cette base de données composée de plus de 100 millions d'évaluations, sur une échelle de 1 à 5, attribuées à plus de 18 000 films par près de 500 000 utilisateurs, qui avait été rendue publique par la société, après avoir été «anonymisée» conformément à la politique interne de l'entreprise en matière de confidentialité, en supprimant toutes les informations d'identification des utilisateurs hormis les évaluations et les dates. Un bruit avait été ajouté dans la mesure où les évaluations avaient été légèrement augmentées ou diminuées.

Malgré ces précautions, il est apparu que 99 % des enregistrements des utilisateurs pouvaient être identifiés de manière unique dans l'ensemble de données en prenant comme critères de sélection 8 évaluations et des dates comportant une marge d'erreur de 14 jours, tandis qu'un abaissement des critères de sélection (2 évaluations, avec une marge d'erreur de 3 jours dans les dates) permettait encore d'identifier 68 % des utilisateurs[13].

3.1.2. Permutation

Cette technique, qui consiste à mélanger les valeurs des attributs dans un tableau de telle sorte que certaines d'entre elles sont artificiellement liées à des personnes concernées différentes, est utile quand il est important de conserver la distribution exacte de chaque attribut dans l'ensemble de données.

La permutation peut être considérée comme une forme spéciale d'ajout de bruit. Dans une technique de bruit classique, les attributs sont modifiés au moyen de valeurs aléatoires. La production d'un bruit cohérent peut se révéler une tâche difficile et le simple fait de modifier légèrement les valeurs des attributs risque de ne pas garantir la confidentialité adéquate. Au lieu de quoi, les techniques de permutation altèrent les valeurs au sein de l'ensemble de données en les échangeant simplement d'un enregistrement à un autre. Cet échange garantira que la fourchette et la distribution des valeurs resteront les mêmes, mais non les corrélations entre les valeurs et les individus. Si deux ou plusieurs attributs sont liés par une relation logique ou une corrélation statistique et sont permutés indépendamment l'un de l'autre, ce lien sera détruit. Il peut donc être important de permuter un ensemble d'attributs de façon à ne pas briser la relation logique, faute de quoi un attaquant pourrait identifier les attributs permutés et inverser la permutation.

Par exemple, si l'on examine un sous-ensemble d'attributs dans un ensemble de données médicales, comme les «motifs d'hospitalisation/symptômes/service concerné», les valeurs seront dans la plupart des cas liées par une forte relation logique et la permutation d'une seule des valeurs serait par conséquent détectée et pourrait même être inversée.

À l'instar de l'ajout de bruit, la permutation risque de ne pas garantir en soi l'anonymisation et devrait toujours être combinée à la suppression des attributs évidents/quasi-identifiants.

3.1.2.1. Garanties

- *Individualisation:* Comme dans le cas de l'ajout de bruit, il reste possible d'isoler les enregistrements correspondant à un individu, mais ces enregistrements sont moins fiables.

[13] Narayanan, A., et Shmatikov, V. (mai 2008), «Robust de-anonymization of large sparse datasets», in *Security and Privacy, 2008, SP 2008, IEEE Symposium on* (p. 111 à 125), IEEE.

- *Corrélation:* Si la permutation affecte des attributs et des quasi-identifiants, elle peut empêcher de relier «correctement» des attributs entre eux tant à l'intérieur qu'à l'extérieur d'un ensemble de données, mais elle autorise toujours une corrélation «incorrecte», puisqu'une entrée réelle peut se trouver associée à une personne concernée différente.
- *Inférence:* Des déductions peuvent encore être tirées de l'ensemble de données, en particulier si les attributs sont corrélés entre eux ou ont des relations logiques fortes; toutefois, sans savoir quels attributs ont été permutés, l'attaquant doit envisager la possibilité que son inférence se fonde sur une hypothèse et seule une inférence probabiliste demeure possible.

3.1.2.2. Erreurs courantes

- Sélection du mauvais attribut: La permutation des attributs non sensibles ou ne comportant pas de risques n'apporterait pas de gain significatif en termes de protection des données à caractère personnel. En effet, si les attributs sensibles/à risque restent associés à la valeur originale, un attaquant aura toujours la possibilité d'extraire des informations sensibles à propos des individus.
- Permutation aléatoire des attributs: Si deux attributs sont fortement corrélés, le fait de permuter les attributs au hasard n'offrira pas de garanties solides. Cette erreur courante est illustrée dans le tableau 1.
- Supposer que la permutation est suffisante: À l'instar de l'ajout de bruit, la permutation ne garantit pas en elle-même l'anonymat et devrait être combinée avec d'autres techniques comme la suppression des attributs évidents.

3.1.2.3. Échecs de la permutation

L'exemple qui suit montre que la permutation aléatoire des attributs offre de piètres garanties de confidentialité quand il existe des liens logiques entre différents attributs. Après la tentative d'anonymisation, il est facile de déduire le revenu de chaque individu selon sa situation professionnelle (et l'année de sa naissance). Par exemple, un examen direct des données permet de soutenir que le PDG est très probablement né en 1957 et perçoit la rémunération la plus élevée, tandis que le chômeur est né en 1964 et a le revenu le moins élevé.

Année	Sexe	Situation professionnelle	Revenu (permuté)
1957	M	Ingénieur	70 000
1957	M	PDG	5 000
1957	M	Sans emploi	43 000
1964	M	Ingénieur	100 000
1964	M	Directeur	45 000

Tableau 1. Un exemple d'anonymisation inefficace par permutation d'attributs corrélés

3.1.3. Confidentialité différentielle

La confidentialité différentielle[14] fait partie de la famille des techniques de randomisation, avec une approche différente: si, dans les faits, l'insertion de bruit intervient à l'avance, quand

[14] Dwork, C. (2006), «Differential privacy», in *Automata, languages and programming* (p. 1 à 12), Springer Berlin Heidelberg.

l'ensemble de données est censé être publié, la confidentialité différentielle peut être utilisée quand le responsable du traitement des données produit des aperçus anonymisés d'un ensemble de données tout en conservant une copie des données originales. Ces aperçus anonymisés sont ordinairement produits au moyen d'un sous-ensemble de requêtes à l'intention d'un tiers en particulier. Le sous-ensemble comprend un bruit aléatoire délibérément ajouté a posteriori. La confidentialité différentielle indique au responsable du traitement des données quel niveau de bruit il doit ajouter, et sous quelle forme, pour obtenir les garanties nécessaires[15]. Dans ce contexte, il sera particulièrement important d'assurer un contrôle permanent (au moins pour chaque nouvelle requête), afin de repérer toute possibilité d'identifier un individu dans l'ensemble des résultats de la requête. Il faut cependant préciser que les techniques de confidentialité différentielle ne modifient pas les données originales et que, par conséquent, tant que les données originales sont conservées, le responsable du traitement des données reste en mesure d'identifier des individus dans les résultats des requêtes de confidentialité différentielle, compte tenu de l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre. Ces résultats doivent donc aussi être considérés comme des données à caractère personnel.

L'un des avantages d'une approche reposant sur la confidentialité différentielle tient au fait que des ensembles de données sont communiqués à des tiers autorisés en réponse à une demande spécifique, plutôt que d'être publiés sous la forme d'un unique ensemble de données. Pour faciliter le contrôle, le responsable du traitement des données peut conserver une liste de toutes les demandes et requêtes afin de vérifier que les tiers n'ont pas accès à des données pour lesquelles ils ne disposent pas d'autorisation. Une requête peut aussi être soumise à des techniques d'anonymisation, incluant l'ajout de bruit ou la substitution, pour mieux garantir la confidentialité. Les recherches se poursuivent en vue de trouver un bon mécanisme interactif de question-réponse, qui soit capable tout à la fois de répondre assez précisément à n'importe quelle question (c'est-à-dire en ajoutant le moins de bruit possible) et de préserver la confidentialité.

Pour limiter les attaques par inférence et par corrélation, il est nécessaire de garder une trace des requêtes soumises par une entité et de surveiller les informations obtenues à propos des personnes concernées; par conséquent, les bases de données à «confidentialité différentielle» ne devraient pas être déployées sur des moteurs de recherche ouverts qui n'offrent aucune traçabilité des entités requérantes.

3.1.3.1 Garanties

- *Individualisation:* Si les résultats se limitent à la production de statistiques et si les règles appliquées à l'ensemble de données sont bien choisies, il ne devrait pas être possible d'utiliser les réponses pour isoler un individu.
- *Corrélation:* En recourant à des requêtes multiples, il pourrait être possible de relier entre elles les entrées relatives à un individu spécifique d'une réponse à l'autre.
- *Inférence:* Il est possible de déduire des informations concernant des individus ou des groupes au moyen de requêtes multiples.

[15] Cf. Ed Felten (2012), «Protecting privacy by adding noise». Internet: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

3.1.3.2. Erreurs courantes

- Ne pas injecter suffisamment de bruit: Afin d'empêcher que des liens puissent être établis avec des connaissances tirées du contexte, la difficulté consiste à fournir le moins d'éléments possibles indiquant si une personne concernée ou un groupe de personnes concernées en particulier a contribué ou non à l'ensemble de données. Le plus difficile, du point de vue de la protection des données, est de parvenir à générer le niveau de bruit approprié à ajouter aux réponses réelles, de façon à protéger la vie privée des individus sans nuire à l'utilité des réponses fournies.

3.1.3.3 Échecs de la confidentialité différentielle

Traitement indépendant de chaque requête: Une combinaison de résultats de requêtes risque de permettre la divulgation d'informations censées rester confidentielles. Si un historique des requêtes n'est pas conservé, un attaquant peut concevoir des questions multiples destinées à interroger une base de données à «confidentialité différentielle» qui réduisent progressivement l'amplitude de l'échantillon résultant jusqu'à ce qu'un caractère spécifique à une seule personne concernée ou à un groupe de personnes concernées finisse par émerger, de façon certaine ou avec un taux de probabilité très élevé. Il faut, en outre, veiller à ne pas commettre l'erreur de penser que les données sont anonymes pour les tiers, alors que le responsable du traitement des données reste en mesure d'identifier la personne concernée dans la base de données originale, compte tenu de l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre.

3.2. Généralisation

La généralisation constitue la seconde famille de techniques d'anonymisation. Cette approche consiste à généraliser, ou diluer, les attributs des personnes concernées en modifiant leur échelle ou leur ordre de grandeur respectif (par exemple, une région plutôt qu'une ville, un mois plutôt qu'une semaine). Si la généralisation peut être efficace pour empêcher l'individualisation, elle ne garantit pas une anonymisation effective dans tous les cas; en particulier, elle requiert des approches quantitatives spécifiques et sophistiquées afin de prévenir la corrélation et l'inférence.

3.2.1. Agrégation et k-anonymat

Les techniques de l'agrégation et du k-anonymat visent à empêcher qu'une personne concernée puisse être isolée en la regroupant avec, au moins, *k* autres individus. Pour ce faire, les valeurs des attributs sont généralisées dans une mesure telle que tous les individus partagent la même valeur. Par exemple, en abaissant la granularité géographique d'une ville à un pays, on inclut un nombre plus élevé de personnes concernées. Les dates de naissance individuelles peuvent être généralisées en une fourchette de dates, ou regroupées par mois ou par année. D'autres attributs numériques (par exemple, les salaires, le poids, la taille ou la dose d'un médicament administrée) peuvent être généralisés au moyen de valeurs d'intervalle (par exemple, salaire de 20 000 à 30 000 EUR). Ces méthodes peuvent être utilisées quand la corrélation de valeurs d'attributs ponctuelles risque de créer des quasi-identifiants.

3.2.1.1. Garanties

- *Individualisation:* Dès lors que les mêmes attributs sont désormais partagés par *k* utilisateurs, il ne devrait plus être possible d'isoler un individu au sein d'un groupe de *k* utilisateurs.

- *Corrélation:* Si la corrélation est limitée, il reste possible de relier les enregistrements par groupe de *k* utilisateurs. Ensuite, au sein de ce groupe, la probabilité que deux enregistrements correspondent aux mêmes pseudo-identifiants est de 1/*k* (ce qui pourrait être nettement plus que la probabilité que ces entrées ne puissent pas être reliées entre elles).
- *Inférence:* Le principal défaut du modèle du k-anonymat est qu'il n'empêche pas un quelconque type d'attaque par inférence. En effet, si tous les *k* individus font partie du même groupe, pour peu que l'on sache à quel groupe appartient un individu, il est facile d'obtenir la valeur de cette propriété.

3.2.1.2. Erreurs courantes

- Négliger certains quasi-identifiants: Le seuil de *k* constitue un paramètre critique dans la technique du k-anonymat. Plus la valeur de *k* est élevée, plus les garanties de confidentialité sont fortes. Une erreur courante consiste à augmenter artificiellement la valeur de *k* en réduisant l'ensemble des quasi-identifiants pris en considération. Un nombre réduit de quasi-identifiants facilite la constitution de groupes de *k* utilisateurs du fait de la capacité d'identification inhérente associée aux autres attributs (surtout si certains d'entre eux sont sensibles ou ont une entropie très élevée, comme dans le cas d'attributs très rares). Le fait de ne pas prendre en considération tous les quasi-identifiants lors de la sélection de l'attribut à généraliser est une erreur critique; si certains attributs peuvent servir à isoler un individu dans un groupe de *k,* la généralisation ne permet pas de protéger certains individus (voir l'exemple du tableau 2).
- Faible valeur de *k*: La recherche d'une faible valeur de *k* se révèle, elle aussi, problématique. Si *k* est trop petit, le coefficient de pondération d'un individu au sein d'un groupe est trop important et les attaques par inférence ont de meilleures chances de succès. Par exemple, si $k=2$ la probabilité que les deux individus partagent la même propriété est plus grande que dans le cas où $k>10$.
- Ne pas regrouper des individus dont le coefficient de pondération est similaire: Le fait de constituer un ensemble d'individus présentant une distribution inégale d'attributs peut aussi créer des problèmes. L'impact des enregistrements correspondant à un individu sur un ensemble de données variera: certains représenteront une fraction considérable des entrées tandis les contributions d'autres resteront assez insignifiantes. Il est donc important de veiller à ce que *k* soit suffisamment élevé pour qu'aucun individu ne représente une fraction trop grande des entrées dans un groupe.

3.1.3.3. Échecs du k-anonymat

Le principal problème lié au k-anonymat est qu'il n'empêche pas les attaques par inférence. Dans l'exemple qui suit, si l'attaquant sait qu'un individu figure dans l'ensemble de données et est né en 1964, il sait aussi que cet individu a fait une crise cardiaque. De plus, si l'on sait que cet ensemble de données a été obtenu auprès d'une organisation française, on peut en déduire que chacun des individus réside à Paris puisque les trois premiers chiffres des codes postaux sont 750*).

Année	Sexe	Code postal	Diagnostic
1957	M	750*	Crise cardiaque
1957	M	750*	Cholestérol
1957	M	750*	Cholestérol
1964	M	750*	Crise cardiaque
1964	M	750*	Crise cardiaque

Tableau 2. Un exemple de k-anonymisation mal conçue

3.2.2. l-diversité/t-proximité

La l-diversité étend le k-anonymat pour faire en sorte qu'il ne soit plus possible d'obtenir des résultats certains au moyen d'attaques par inférence en veillant à ce que, dans chaque classe d'équivalence, chaque attribut ait au moins *l* valeurs différentes.

Un objectif fondamental à atteindre est d'éviter autant que possible les classes d'équivalence caractérisées par une faible variabilité des attributs, de telle sorte qu'un attaquant reste toujours confronté à un degré d'incertitude considérable, malgré les connaissances tirées du contexte dont il pourrait disposer à propos d'une personne concernée.

La l-diversité est utile pour protéger les données contre les attaques par inférence, quand les valeurs des attributs sont bien distribuées. Il faut cependant souligner que cette technique n'empêche pas les fuites d'informations si les attributs au sein d'un segment sont distribués de manière inégale ou ne présentent qu'un faible écart de valeurs ou de contenus sémantiques. En définitive, la l-diversité se prête à des attaques par inférence probabilistes.

La t-proximité constitue un affinement de la l-diversité, en ce sens qu'elle vise à créer des classes d'équivalence qui ressemblent à la distribution initiale des attributs dans le tableau. Cette technique est utile quand il est important de conserver des données aussi proches que possible des données originales; à cet effet, une contrainte supplémentaire est ajoutée à la classe d'équivalence, à savoir que non seulement chaque classe d'équivalence doit comporter au moins *l* valeurs différentes, mais aussi que chaque valeur est représentée autant de fois que nécessaire pour refléter la distribution initiale de chaque attribut.

3.2.2.1. Garanties

- *Individualisation:* À l'instar du k-anonymat, la l-diversité et la t-proximité permettent d'empêcher que les enregistrements relatifs à un individu soient isolés dans la base de données.
- *Corrélation:* La l-diversité et la t-proximité n'apportent pas d'amélioration par rapport au k-anonymat pour ce qui est d'empêcher la corrélation. Le problème reste le même pour n'importe quel regroupement: la probabilité que les mêmes entrées se rapportent à la même personne concernée est plus élevée que $1/N$ (où N est le nombre de personnes concernées dans la base de données).
- *Inférence:* la principale amélioration de la l-diversité et de la t-proximité par rapport au k-anonymat est qu'il n'est plus possible de lancer des attaques par inférence contre une base de données à «l-diversité» ou «t-proximité» avec un degré de certitude de 100 %.

3.2.2.2. Erreurs courantes

- Protéger les valeurs des attributs sensibles en les mélangeant avec d'autres attributs sensibles: Le fait d'avoir deux valeurs pour un attribut dans un groupe ne suffit pas à apporter des garanties de confidentialité. En fait, la distribution des valeurs sensibles dans chaque groupe devrait être semblable à la distribution de ces valeurs dans la population totale ou, du moins, elle devrait être uniforme dans l'ensemble du groupe.

3.2.2.3. Échecs de la l-diversité

Dans le tableau présenté ci-dessous, la l-diversité est assurée pour l'attribut «Diagnostic»; cependant, pour peu que l'on connaisse un individu né en 1964 qui figure dans ce tableau, il reste possible de supposer avec une probabilité très élevée qu'il a fait une crise cardiaque.

Année	Sexe	Code postal	Diagnostic
1957	M	750*	**Crise cardiaque**
1957	M	750*	**Cholestérol**
1957	M	750*	**Cholestérol**
1957	M	750*	**Cholestérol**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Cholestérol**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**
1964	M	750*	**Crise cardiaque**

Tableau 3. Un exemple de l-diversité où les valeurs de «Diagnostic» ne sont pas uniformément distribuées

Nom	Date de naissance	Sexe
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tableau 4. En sachant que ces individus figurent dans le tableau 3, un attaquant pourrait en déduire qu'ils ont fait une crise cardiaque

4. Pseudonymisation

La pseudonymisation consiste à remplacer un attribut (généralement un attribut unique) par un autre dans un enregistrement. La personne physique est donc toujours susceptible d'être identifiée indirectement; par conséquent, la pseudonymisation ne permet pas, à elle seule, de produire un ensemble de données anonymes. Elle est néanmoins examinée dans le présent avis en raison de nombreuses idées fausses et erreurs qui entourent son utilisation.

La pseudonymisation réduit le risque de mise en corrélation d'un ensemble de données avec l'identité originale d'une personne concernée; à ce titre, c'est une mesure de sécurité utile, mais non une méthode d'anonymisation.

Le résultat de la pseudonymisation peut être indépendant de la valeur initiale (comme dans le cas d'un numéro aléatoire généré par le responsable du traitement ou d'un nom choisi par la personne concernée) ou il peut être dérivé des valeurs originales d'un attribut ou d'un ensemble d'attributs, par exemple au moyen d'une fonction de hachage ou d'un système de chiffrement.

Les techniques de pseudonymisation les plus utilisées sont les suivantes:

- Système cryptographique à clé secrète: dans ce cas, le détenteur de la clé peut aisément ré-identifier chaque personne concernée en décryptant l'ensemble de données, puisque les données à caractère personnel y figurent toujours, quoique sous une forme cryptée. En supposant qu'un système cryptographique conforme à l'état de la technique a été appliqué, le décryptage ne serait possible qu'à condition de connaître la clé.
- Fonction de hachage: il s'agit d'une fonction qui renvoie un résultat de taille fixe, quelle que soit la taille de l'entrée encodée (l'entrée peut être un attribut unique ou un ensemble d'attributs) et qui ne peut être inversée; c'est-à-dire que le risque de récupération des données observé dans le cas du chiffrement n'existe plus. Cependant, si la fourchette dans laquelle se situent les valeurs d'entrée de la fonction de hachage est connue, il est possible de réintroduire ces valeurs dans la fonction de hachage afin d'obtenir la valeur correcte correspondant à un enregistrement en particulier. Par exemple, si un ensemble de données a été pseudonymisé en procédant au hachage du numéro d'identification national, il peut être reconstitué simplement en appliquant la fonction de hachage à toutes les valeurs possibles et en comparant les résultats avec les valeurs figurant dans l'ensemble de données. Les fonctions de hachage sont ordinairement conçues pour être calculées relativement vite et se prêtent donc à des

attaques par force brute[16]. Des tables pré-calculées peuvent aussi être créées pour permettre la reconstitution en masse d'un ensemble volumineux de valeurs de hachage.

L'utilisation d'une fonction de hachage avec salage (où une valeur aléatoire, appelée «sel», est ajoutée à l'attribut qui fait l'objet du hachage) permet de réduire la probabilité de reconstituer la valeur d'entrée. Il reste néanmoins possible, avec des moyens raisonnables, de calculer la valeur originale de l'attribut qui se cache derrière le résultat d'une fonction de hachage avec salage[17].

- Fonction de hachage par clé avec clé enregistrée: il s'agit d'une fonction de hachage particulière qui utilise une clé secrète comme entrée supplémentaire (à la différence d'une fonction de hachage avec salage, où le «sel» n'est généralement pas secret). Un responsable du traitement des données peut ré-exécuter la fonction sur l'attribut en se servant de la clé secrète, mais il est beaucoup plus difficile pour un attaquant de ré-exécuter la fonction sans connaître la clé car le nombre de possibilités à tester est suffisamment grand pour rendre la tâche impraticable.
- Chiffrement déterministe ou fonction de hachage par clé avec suppression de la clé: cette technique équivaut à sélectionner un nombre aléatoire comme pseudonyme pour chaque attribut de la base de données et à supprimer ensuite la table de correspondance. Cette solution permet[18] de réduire le risque de corrélation entre les données à caractère personnel figurant dans l'ensemble de données et celles qui se rapportent au même individu dans un autre ensemble de données, où un pseudonyme différent est utilisé. En supposant qu'un algorithme conforme à l'état de la technique soit appliqué, il sera difficile pour un attaquant, en termes de puissance de calcul requise, de décrypter ou de ré-exécuter la fonction, car cela supposerait d'essayer chaque clé possible, puisque la clé n'est pas disponible.
- Tokenization: cette technique est généralement appliquée dans le secteur financier (même si elle n'y est pas confinée) pour remplacer les numéros d'identification de cartes par des valeurs sans grande utilité pour un attaquant. Elle dérive des techniques précédentes, dans la mesure où elle repose normalement sur l'application de mécanismes de chiffrement à sens unique ou sur l'assignation, au moyen d'une fonction d'index, d'un numéro séquentiel ou d'un nombre produit de manière aléatoire qui n'est pas mathématiquement dérivé des données originales.

4.1. Garanties

- *Individualisation:* Il reste possible d'isoler les enregistrements d'un individu, puisque celui-ci est toujours identifié par un attribut unique qui est le résultat de la fonction de pseudonymisation (= l'attribut pseudonymisé).
- *Corrélation:* La corrélation restera facile entre les enregistrements qui utilisent le même attribut pseudonymisé en référence au même individu. Même si des attributs pseudonymisés différents sont utilisés pour la même personne concernée, la corrélation est encore possible au moyen d'autres attributs. Ce n'est que dans le cas où

[16] Ces attaques consistent à essayer toutes les entrées plausibles afin de constituer des tableaux de correspondance.
[17] Surtout si le type d'attribut est connu (nom, numéro de sécurité sociale, date de naissance, etc.). Pour augmenter la puissance de calcul requise, on pourrait recourir à une fonction de hachage à dérivation de clé, où la valeur calculée est hachée plusieurs fois avec une courte chaîne de «sel».
[18] Tout dépend des autres attributs figurant dans l'ensemble de données et de la suppression des données originales.

aucun autre attribut dans l'ensemble de données ne peut servir à identifier la personne concernée et où tout lien entre l'attribut original et l'attribut pseudonymisé a été éliminé (notamment par la suppression des données originales) qu'aucun recoupement évident ne pourra être fait entre deux ensembles de données qui utilisent des attributs pseudonymisés différents.

- *Inférence:* Les attaques par inférence sur l'identité réelle d'une personne concernée sont possibles au sein de l'ensemble de données ou entre différentes bases de données qui utilisent le même attribut pseudonymisé pour un individu, ou encore dans le cas où les pseudonymes sont transparents et ne masquent pas correctement l'identité originale de la personne concernée.

4.2. Erreurs courantes

- Croire qu'un ensemble de données pseudonymisé est anonymisé: Les responsables du traitement des données supposent souvent qu'il suffit de supprimer ou de remplacer un ou plusieurs attributs pour rendre l'ensemble de données anonyme. De nombreux exemples ont montré que ce n'est pas le cas; le simple fait de modifier l'identité n'empêche pas quelqu'un d'identifier une personne concernée s'il subsiste des quasi-identifiants dans l'ensemble de données, ou si les valeurs d'autres attributs permettent encore d'identifier un individu. Dans bien des cas, il peut se révéler aussi facile d'identifier un individu dans un ensemble de données pseudonymisé qu'à partir des données originales. Des mesures supplémentaires devraient être prises pour pouvoir considérer l'ensemble de données comme anonymisé, notamment la suppression et la généralisation d'attributs, l'effacement des données originales ou du moins leur conservation à un niveau hautement agrégé.
- Erreurs courantes lors de l'utilisation de la pseudonymisation comme technique destinée à réduire la corrélation:

    - Utiliser la même clé dans des bases de données différentes: l'élimination du risque de corrélation entre différents ensembles de données dépend beaucoup de l'utilisation d'un algorithme à clé et du fait qu'un même individu correspondra à différents attributs pseudonymisés dans des contextes différents. Il est donc important d'éviter d'utiliser la même clé dans des bases de données différentes pour pouvoir réduire la corrélation.
    - Utiliser des clés différentes («clés alternées») pour des utilisateurs différents: il pourrait être tentant d'employer des clés différentes pour différents ensembles d'utilisateurs et de changer la clé en fonction de son utilisation (par exemple, se servir de la même clé pour 10 entrées d'enregistrement relatives au même utilisateur). Cependant, si elle n'est pas correctement conçue, cette opération pourrait faire apparaître des motifs, réduisant partiellement les avantages escomptés. Par exemple, l'utilisation alternée d'une clé selon des règles spécifiques pour des individus spécifiques faciliterait la mise en corrélation des entrées correspondant à des individus donnés. De plus, la disparition de données pseudonymisées récurrentes dans la base de données au moment où de nouvelles données apparaissent peut indiquer que les enregistrements se rapportent à la même personne physique.
    - Conserver la clé: si la clé secrète est conservée avec les données pseudonymisées, et si les données sont compromises, l'attaquant peut être en mesure de relier facilement les données pseudonymisées avec leur attribut

original. Il en va de même si la clé est conservée séparément, mais de façon peu sûre.

4.3. Lacunes de la pseudonymisation

- *Soins de santé*

1. Nom, adresse, date de naissance	2. Période de perception d'une prestation d'assistance spéciale	3. Indice de masse corporelle	6. N° de référence dans la cohorte de recherche
[redacted]	< 2 ans	15	QA5FRD4
	> 5 ans	14	2B48HFG
	< 2 ans	16	RC3URPQ
	> 5 ans	18	SD289K9
	< 2 ans	20	5E1FL7Q

Tableau 5. Un exemple de pseudonymisation par hachage (nom, adresse, date de naissance) qui peut être aisément inversée

Un ensemble de données a été créé pour examiner la relation entre le poids d'une personne et la perception d'une prestation d'assistance spéciale. L'ensemble de données original comprenait le nom, l'adresse et la date de naissance des personnes concernées, qui ont été effacés. Le numéro de référence dans la cohorte de recherche a été généré à partir des données supprimées en utilisant une fonction de hachage. Bien que le nom, l'adresse et la date de naissance aient été supprimés du tableau, si l'on connaît le nom, l'adresse et la date de naissance d'une personne concernée, en plus de la fonction de hachage utilisée, il est facile de calculer les numéros de référence dans la cohorte de recherche.

- *Réseaux sociaux*

Il a été démontré[19] que des informations sensibles à propos d'individus spécifiques peuvent être extraites des graphes de réseaux sociaux, malgré les techniques de «pseudonymisation» appliquées à ces données. L'exploitant d'un réseau social a supposé à tort que la pseudonymisation suffisait à empêcher l'identification après la vente des données à d'autres sociétés à des fins de marketing et de publicité. À la place des noms réels, l'exploitant utilisait des pseudonymes, mais ce n'est manifestement pas assez pour anonymiser les profils d'utilisateurs, étant donné que les relations entre les différents individus sont uniques et peuvent servir d'identifiants.

- *Localisation*

Les chercheurs du MIT[20] ont récemment analysé un ensemble de données pseudonymisé couvrant 15 mois de coordonnées mobiles spatiales et temporelles de 1,5 million de personnes sur un territoire d'un rayon de 100 km. Ils ont démontré que quatre points de localisation permettaient d'isoler 95 % de cette population et que deux points seulement suffisaient pour isoler plus de 50 % des personnes concernées (un seul point étant supposé être très probablement le domicile ou le lieu de travail), ce qui laissait très peu de place à la

[19] A. Narayanan et V. Shmatikov, «De-anonymizing social networks», in *30th IEEE Symposium on Security and Privacy*, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen et V. Blondel, «Unique in the Crowd: The privacy bounds of human mobility», *Nature*, n° 1376, 2013.

protection de la vie privée, même si les identités des individus avaient été pseudonymisées en remplaçant leurs attributs réels [….] par d'autres étiquettes.

5. Conclusions et recommandations

5.1. Conclusions

Les techniques d'anonymisation font l'objet de recherches intensives, et le présent document a invariablement montré que chaque technique a ses avantages et ses inconvénients. Le plus souvent, il n'est pas possible de formuler des recommandations minimales quant aux paramètres à utiliser, étant donné que chaque ensemble de données doit être envisagé au cas par cas.

Dans beaucoup de situations, un ensemble de données anonymisées peut encore présenter un risque résiduel pour les personnes concernées. En effet, même quand il n'est plus possible de reconstituer précisément l'enregistrement d'un individu, il reste parfois possible de glaner des informations à propos de cet individu à l'aide d'autres sources d'informations disponibles (publiquement ou non). Il faut souligner qu'au-delà de l'impact direct produit par les conséquences d'un processus d'anonymisation inefficace sur les personnes concernées (désagrément, temps perdu et sentiment de perte de contrôle du fait de l'inclusion dans un groupe sans notification ni accord préalable), d'autres effets secondaires indirects peuvent survenir quand une personne concernée est erronément prise pour cible par un quelconque attaquant, à la suite du traitement de données anonymisées – surtout si les intentions de l'attaquant sont malveillantes. C'est pourquoi le groupe de travail «Article 29» insiste sur le fait que les techniques d'anonymisation peuvent apporter des garanties en matière de respect de la vie privée, mais uniquement si leur application est correctement conçue – ce qui suppose que les conditions préalables (le contexte) et les objectif(s) du processus d'anonymisation soient clairement définis de façon à assurer le niveau d'anonymisation visé.

5.2. Recommandations

- Il existe des limitations inhérentes à certaines techniques d'anonymisation. Ces limitations doivent être envisagées avec attention par les responsables du traitement avant de recourir à une technique donnée pour élaborer un processus d'anonymisation. Il faut prendre en considération les finalités que l'anonymisation vise à atteindre – comme de protéger la vie privée des personnes lors de la publication d'un ensemble de données, ou de permettre l'obtention de certaines informations à partir d'un ensemble de données.
- Aucune des techniques décrites dans le présent document ne satisfait de façon certaine aux critères d'une anonymisation efficace (à savoir, empêcher l'individualisation d'une personne concernée, la corrélation entre les enregistrements se rapportant à un individu et l'obtention par inférence de données concernant un individu). Cependant, dès lors que certains de ces risques peuvent être évités complètement ou partiellement au moyen d'une technique donnée, il est nécessaire de concevoir avec soin l'application d'une technique individuelle à la situation concernée et d'opter pour une combinaison de ces techniques en vue de renforcer la fiabilité du résultat.

Le tableau présenté ci-après donne un aperçu des forces et des faiblesses des techniques considérées au regard des trois exigences fondamentales:

	Reste-t-il un risque d'individualisation?	Reste-t-il un risque de corrélation?	Reste-t-il un risque d'inférence?
Pseudonymisation	Oui	Oui	Oui
Ajout de bruit	Oui	Peut-être pas	Peut-être pas
Substitution	Oui	Oui	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
Hachage/Tokenization	Oui	Oui	Peut-être pas

Tableau 6. Forces et faiblesses des techniques considérées

- La solution optimale devrait être choisie au cas par cas. Une solution (c'est-à-dire un processus d'anonymisation complet) répondant aux trois critères résisterait aux tentatives d'identification utilisant les moyens les plus susceptibles d'être raisonnablement mis en œuvre par le responsable du traitement des données ou par des tiers.
- Lorsqu'un des critères n'est pas rempli par une proposition, il convient de procéder à une évaluation approfondie des risques d'identification. Cette évaluation devrait être soumise à l'autorité compétente si le droit national requiert l'examen ou l'autorisation du processus d'anonymisation par ladite autorité.

Afin de réduire les risques d'identification, les bonnes pratiques suivantes devraient être prises en considération:

Bonne pratique d'anonymisation

En général:

- Ne pas se contenter de «publier et oublier». Compte tenu du risque résiduel d'identification, les responsables du traitement des données devraient:
    - 1. identifier les nouveaux risques et réévaluer régulièrement le(s) risque(s) résiduel(s);
    - 2. examiner si les contrôles des risques identifiés sont suffisants et les ajuster en conséquence; ET
    - 3. surveiller et contrôler les risques.
- Parmi ces risques résiduels, la possibilité d'identifier la partie non anonymisée d'un ensemble de données devrait (le cas échéant) être prise en considération, surtout en combinaison avec la partie anonymisée, ainsi que les corrélations possibles entre les attributs (par exemple entre les données relatives à la localisation géographique et celles concernant le niveau de prospérité).

Éléments contextuels:

- Les finalités visées par l'anonymisation d'un ensemble de données devraient être clairement définies, dans la mesure où elles jouent un rôle-clé dans la détermination du risque d'identification.
- Cela va de pair avec la prise en considération de tous les éléments contextuels pertinents – par exemple, la nature des données originales, les mécanismes de contrôle en place (y compris les mesures de sécurité restreignant l'accès aux ensembles de données), la taille de l'échantillon (aspects quantitatifs), la disponibilité de ressources d'informations

publiques (sur lesquelles peuvent s'appuyer les destinataires), la communication envisagée de données à des tiers (limitée ou illimitée, par exemple sur l'internet, etc.).

- Il convient de prendre en considération les attaquants possibles, compte tenu de l'attrait des données pour des attaques ciblées (là encore, le caractère sensible des informations et la nature des données seront des facteurs-clés à cet égard).

Éléments techniques:

- Les responsables du traitement des données devraient divulguer la technique d'anonymisation / la combinaison de techniques appliquée, surtout s'ils prévoient de diffuser l'ensemble de données anonymisées.
- Les attributs évidents (par exemple, rares) / quasi-identifiants devraient être supprimés de l'ensemble de données.
- Si des techniques d'ajout de bruit sont utilisées (dans le cadre de la randomisation), le niveau de bruit ajouté aux enregistrements devrait être déterminé en fonction de la valeur d'un attribut (c'est-à-dire qu'il ne faut pas injecter un bruit démesuré), de l'impact des attributs à protéger pour les personnes concernées et/ou du caractère clairsemé de l'ensemble de données.
- En cas de recours à la confidentialité différentielle (dans le cadre de la randomisation), il convient de tenir compte de la nécessité de conserver une trace des requêtes de façon à détecter celles qui présentent un risque d'intrusion dans la vie privée, car le caractère intrusif des requêtes est cumulatif.
- Si des techniques de généralisation sont appliquées, il est crucial que le responsable du traitement des données ne se limite pas à un critère de généralisation qui reste inchangé pour le même attribut; c'est-à-dire qu'il convient de sélectionner des granularités géographiques ou des intervalles de temps différents. Le choix du critère à appliquer doit être dicté par la distribution des valeurs des attributs dans la population concernée. Toutes les distributions ne se prêtent pas à une généralisation. Autrement dit, il n'existe pas d'approche universelle à suivre en matière de généralisation. Il importe de veiller à la variabilité au sein des classes d'équivalence; par exemple, un seuil spécifique devrait être sélectionné en fonction des «éléments contextuels» mentionnés ci-dessus (taille de l'échantillon, etc.) et, si ce seuil n'est pas atteint, il conviendrait de rejeter l'échantillon concerné (ou de définir un critère de généralisation différent).

ANNEXE
Un aperçu des techniques d’anonymisation

A.1. Introduction

Il existe dans l'Union différentes interprétations de l'anonymat: dans certains pays, la notion correspond à un anonymat informatique (c'est-à-dire qu'il doit être difficile, en termes de puissance de calcul, d'identifier directement ou indirectement une des personnes concernées, même pour le responsable du traitement des données, avec la collaboration de quelque autre partie) et dans d'autres pays, il s'agit d'un anonymat parfait (c'est-à-dire qu'il doit être impossible d'identifier directement ou indirectement une des personnes concernées, même pour le responsable du traitement des données, avec la collaboration de quelque autre partie). Néanmoins, l'«anonymisation» désigne dans les deux cas le processus au moyen duquel des données sont rendues anonymes. La différence réside dans ce qui est considéré comme un niveau acceptable de risque de ré-identification.

Divers usages peuvent être envisagés pour les données anonymisées: enquêtes sociales, analyses statistiques, développement de nouveaux services/produits. Aussi générales que soient les finalités poursuivies, il arrive parfois que ces activités puissent avoir un impact sur certaines personnes concernées, annulant le caractère prétendument anonyme des données traitées. On peut en citer de nombreux exemples, depuis le lancement d'actions de marketing ciblées jusqu'à la mise en œuvre de mesures publiques fondées sur les profils, les comportements ou les schémas de mobilité des utilisateurs[21].

Malheureusement, hormis les déclarations générales, il n'existe pas de système de mesure suffisamment évolué qui permette d'apprécier à l'avance le temps ou les efforts nécessaires pour parvenir à une ré-identification après le traitement, ou encore de sélectionner la procédure la plus appropriée à mettre en place si l'on veut réduire la probabilité qu'une base de données diffusée renvoie à un ensemble identifié de personnes concernées.

L'«art de l'anonymisation», ainsi qu'on désigne parfois ces pratiques dans la littérature scientifique[22], est une nouvelle branche scientifique encore balbutiante et il existe beaucoup de pratiques visant à réduire la capacité d'identification des ensembles de données; il doit être bien clair, cependant, que la majorité de ces pratiques n'empêchent pas la mise en relation des données traitées avec les personnes concernées. Dans certaines circonstances, les tentatives d'identification d'ensembles de données censés être anonymes ont été couronnées de succès; dans d'autres situations, des faux positifs ont été constatés.

Il existe, en gros, deux approches différentes: l'une se fonde sur la généralisation de l'attribut, l'autre sur la randomisation. Un examen des détails et des subtilités de ces pratiques nous aidera à mieux comprendre le potentiel d'identification des données et jettera un nouvel éclairage sur la notion même de données à caractère personnel.

A.2. L'«anonymisation» par randomisation

En matière d'anonymisation, une option consiste à modifier les valeurs réelles pour empêcher que les données anonymisées puissent être mises en relation avec les valeurs originales. Cet objectif peut être atteint au moyen de nombreuses méthodes, qui vont de l'injection de bruit à

[21] Par exemple, dans le cas de TomTom aux Pays-Bas (voir l'exemple expliqué à la section 2.2.3).

[22] Jun Gu,Yuexian Chen, Junning Fu, HuanchunPeng, Xiaojun Ye, *Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes in Computer Science*, Springer, volume 6261, 2010, p. 385 à 399.

la substitution de données (permutation). Il faut aussi souligner que la suppression d'un attribut équivaut à une forme extrême de randomisation dudit attribut (lequel est alors entièrement couvert par le bruit).

Dans certaines circonstances, l'objectif du traitement n'est pas tant de publier un ensemble de données randomisées, mais plutôt de permettre l'accès aux données au moyen de requêtes. Le risque pour les personnes concernées vient dans ce cas de la probabilité qu'un attaquant soit en mesure d'obtenir des informations en transmettant une série de requêtes différentes, sans que le responsable du traitement des données en ait connaissance. Pour garantir l'anonymat des individus auxquels se rapporte l'ensemble de données, il faudrait qu'il ne soit pas possible de conclure qu'une personne concernée a contribué à l'ensemble de données, de façon à rompre le lien avec d'éventuelles informations tirées du contexte qu'un attaquant pourrait avoir en sa possession.

En ajoutant du bruit, selon les modalités appropriées, à la réponse à une requête, il est possible de réduire encore le risque de ré-identification. Cette approche, désignée dans la littérature par les termes de «confidentialité différentielle»[23], s'écarte des méthodes décrites précédemment en ce qu'elle donne à ceux qui publient des données un plus grand contrôle sur l'accès aux données par rapport à une diffusion publique. L'ajout de bruit vise deux objectifs principaux: premièrement, protéger la vie privée des personnes concernées reprises dans l'ensemble de données et, deuxièmement, préserver l'utilité des informations communiquées. En particulier, l'ampleur du bruit doit être proportionnelle au niveau des requêtes (des réponses trop précises à de trop nombreuses requêtes concernant des individus augmentent le risque d'identification). Pour être efficace, l'application de la randomisation doit aujourd'hui être envisagée au cas par cas, car aucune technique ne constitue une méthodologie à toute épreuve. Il existe des exemples de fuites d'informations à propos des attributs d'une personne concernée (figurant ou non dans l'ensemble de données), alors même que le responsable du traitement considérait l'ensemble de données comme randomisé.

Il peut être utile d'examiner des exemples précis pour mettre en lumière les failles possibles de la randomisation, en tant que procédé d'anonymisation. Ainsi, dans le cas d'un accès interactif, des requêtes jugées anodines en termes de confidentialité pourraient représenter un risque pour la vie privée des personnes concernées. En fait, si l'attaquant sait qu'un sous-groupe S d'individus figure dans l'ensemble de données qui contient des informations relatives à l'incidence d'un attribut A dans une population P, en posant simplement les deux questions «Combien d'individus dans la population P possèdent l'attribut A?» et «Combien d'individus dans la population P, hormis ceux repris dans le sous-groupe S, possèdent l'attribut A?», il peut être possible de déterminer (par soustraction) le nombre d'individus du sous-groupe S qui possèdent effectivement l'attribut A – soit de façon certaine, soit par inférence probabiliste. En tout cas, le respect de la vie privée des individus du sous-groupe S pourrait être gravement compromis, selon la nature de l'attribut A.

On peut aussi considérer que la publication d'un ensemble de données peut présenter un risque pour la vie privée d'une personne concernée qui n'y est pas reprise, s'il existe un lien connu entre cette personne et des données figurant dans cet ensemble. Par exemple, si l'on sait que «la valeur de l'attribut A de la cible diffère dans une quantité X de la valeur moyenne de la population», en demandant simplement à l'administrateur de la base de données d'exécuter une opération – anodine en termes de confidentialité – d'extraction de la valeur

[23] Cynthia Dwork, *Differential Privacy, International Colloquium on Automata, Languages and Programming*, ICALP, 2006, p. 1 à 12

moyenne de l'attribut A, l'attaquant peut déduire avec exactitude des données à caractère personnel relatives à une personne concernée en particulier.

L'opération qui consiste à injecter certaines inexactitudes relatives dans les valeurs réelles d'une base de données doit être conçue correctement. Il faut ajouter suffisamment de bruit pour protéger la confidentialité, mais pas trop non plus, afin de préserver l'utilité des données. Par exemple, si le nombre de personnes concernées présentant un attribut particulier est très réduit ou si l'attribut a un caractère hautement sensible, il peut être préférable de s'en tenir à une fourchette, ou à une phrase générale, du genre «un petit nombre de cas, peut-être même zéro», plutôt que d'indiquer le chiffre exact. De cette façon, même si le mécanisme d'ajout de bruit est connu à l'avance, la vie privée des personnes concernées est respectée, puisqu'il subsiste un degré d'incertitude. Du point de vue de l'utilité, si l'inexactitude est correctement conçue, les résultats restent utiles à des fins d'analyse statistique ou de prise de décision.

La randomisation d'une base de données et l'accès en confidentialité différentielle requièrent une réflexion plus poussée. Premièrement, la juste dose de distorsion peut varier considérablement selon le contexte (le type de requête, la taille de la population de la base de données, la nature de l'attribut et le risque d'identification inhérent) et aucune solution universelle ne peut être envisagée. De plus, le contexte peut changer avec le temps et le mécanisme interactif devrait être modifié en conséquence. Le calibrage du bruit nécessite une surveillance des risques cumulatifs qu'un mécanisme interactif représente pour le respect de la vie privée des personnes concernées. Le mécanisme d'accès aux données devrait donc disposer d'alertes qui se déclenchent lorsqu'un budget «coût pour la confidentialité» est épuisé et que les personnes concernées pourraient être exposées à des risques spécifiques si une nouvelle requête est transmise, afin d'aider le responsable du traitement des données à déterminer le niveau approprié de distorsion qu'il y a lieu d'injecter à chaque fois dans les données à caractère personnel.

D'un autre côté, il faut aussi envisager le cas où des valeurs d'attributs sont supprimées (ou modifiées). Une solution souvent employée pour traiter certaines valeurs d'attributs atypiques est la suppression de l'ensemble de données relatives aux individus atypiques ou des valeurs atypiques. Dans ce dernier cas, il est important de veiller à ce que l'absence de la valeur ne devienne pas en elle-même un élément permettant l'identification d'une personne concernée.

Passons à présent à la randomisation par substitution d'attribut. Une des principales idées fausses qui circulent à propos de l'anonymisation consiste à l'assimiler au chiffrement ou au codage à clé. Cette erreur repose sur deux suppositions, à savoir a) que lorsque certains attributs d'un enregistrement dans une base de données (par exemple, le nom, l'adresse, la date de naissance) font l'objet d'un chiffrement ou qu'une chaîne apparemment aléatoire leur est substituée à la suite d'une opération de codage à clé, comme une fonction de hachage, cet enregistrement est «anonymisé», et b) que l'anonymisation est plus efficace si la longueur de la clé est appropriée et si l'algorithme de chiffrement est conforme à l'état de la technique. Cette idée fausse est largement répandue parmi les responsables du traitement des données et appelle des éclaircissements, car elle se rapporte aussi à la pseudonymisation et à ses risques prétendument moindres.

Premièrement, les objectifs de ces techniques sont radicalement différents: le chiffrement, en tant que mesure de sécurité, vise à garantir la confidentialité d'un canal de communication entre des parties identifiées (êtres humains, appareils ou éléments logiciels/matériels) afin d'éviter une interception ou une divulgation involontaire. Le codage à clé correspond à une traduction sémantique des données qui dépend d'une clé secrète. L'objectif de

l'anonymisation, en revanche, est d'éviter l'identification d'individus en empêchant que des attributs puissent être mis en relation avec une personne concernée à son insu.

Ni le chiffrement ni le codage à clé ne se prêtent, en eux-mêmes, à l'objectif de rendre une personne concernée non identifiable, puisque les données originales, entre les mains du responsable du traitement au moins, peuvent encore être consultées ou reconstituées par déduction. La seule traduction sémantique de données à caractère personnel, telle qu'elle est appliquée dans le cas du codage à clé, n'exclut pas la possibilité de rétablir la structure originale des données, en exécutant l'algorithme en sens inverse ou au moyen d'attaques par force brute, selon la nature des mécanismes employés, ou à la suite d'une violation de données. Un chiffrement conforme à l'état de la technique peut garantir une meilleure protection des données, c'est-à-dire empêcher leur consultation par des entités qui ignorent la clé de décryptage, mais il ne se traduit pas nécessairement par une anonymisation. Tant que la clé ou les données originales sont accessibles (même dans le cas de leur conservation par un tiers de confiance, tenu par contrat d'assurer un service de séquestre), la possibilité d'identifier une personne concernée n'a pas été éliminée.

Le fait de se fonder exclusivement sur la robustesse du mécanisme de chiffrement comme mesure du degré d'«anonymisation» d'un ensemble de données est trompeur, car de nombreux autres facteurs techniques et organisationnels affectent la sécurité générale d'un mécanisme de chiffrement ou d'une fonction de hachage. La littérature fait état de beaucoup d'attaques couronnées de succès, qui contournent totalement l'algorithme, en exploitant les faiblesses de la conservation des clés (par exemple, l'existence d'un mode par défaut moins sécurisé) ou d'autres facteurs humains (par exemple, des mots de passe faibles permettant de récupérer la clé). Enfin, un système de chiffrement sélectionné, avec une clé d'une taille donnée, est conçu pour garantir la confidentialité pendant une certaine période (la taille de la plupart des clés actuelles devra être revue vers 2020), tandis qu'un processus d'anonymisation ne devrait pas être limité dans le temps.

Il peut être intéressant d'examiner en détail les limites de la randomisation (ou de la substitution et de la suppression) de l'attribut, à la lumière de divers exemples d'anonymisation par randomisation inefficaces recensés ces dernières années et des raisons qui expliquent ces échecs.

Un cas bien connu de diffusion d'un ensemble de données mal anonymisé est celui du prix Netflix[24]. En examinant un enregistrement générique dans une base de données dont plusieurs attributs relatifs à une personne concernée ont été randomisés, chaque enregistrement peut encore être scindé en deux sous-enregistrements comme suit: {attributs randomisés, attributs en clair}, où les attributs en clair peuvent constituer n'importe quelle combinaison de données qui ne sont pas censées avoir un caractère personnel. Dans le cas de l'ensemble de données du prix Netflix, il est à noter que chaque enregistrement peut être représenté par un point dans un espace multidimensionnel, où chaque attribut en clair est une coordonnée. En appliquant cette technique, tout ensemble de données peut être considéré comme une constellation de points dans cet espace multidimensionnel, qui présente parfois un caractère très clairsemé, c'est-à-dire que les points sont distants les uns des autres. En fait, ils peuvent être si éloignés qu'après avoir divisé l'espace en vastes régions, chaque région ne contient qu'un seul enregistrement. Même l'injection de bruit ne parvient pas à rapprocher suffisamment les enregistrements pour qu'ils partagent la même région multidimensionnelle. Dans l'expérience de Netflix, par exemple, 8 évaluations de films attribuées au cours d'une période s'étendant sur 14 jours

[24] Arvind Narayanan, Vitaly Shmatikov, «Robust De-anonymization of Large Sparse Datasets», in *IEEE Symposium on Security and Privacy*, 2008, p. 111 à 125

suffisaient à rendre les enregistrements uniques. Après l'ajout de bruit aux évaluations et aux dates, aucune superposition de régions ne pouvait être constatée. Autrement dit, cette même sélection de 8 films évalués constituait une empreinte digitale des évaluations attribuées, qui n'était pas partagée par deux personnes concernées dans la base de données. Sur la base de cette observation géométrique, les chercheurs ont comparé l'ensemble de données prétendument anonyme de Netflix avec une autre base de données publique contenant des évaluations de films (IMDB) et découvert ainsi des utilisateurs qui avaient attribué des évaluations pour les mêmes films au cours de la même période. Comme la majorité des utilisateurs présentaient une correspondance biunivoque, les informations auxiliaires récupérées dans la base de données IMDB ont pu être importées dans l'ensemble de données publiées par Netflix, de façon à identifier tous les enregistrements censés être anonymes.

Il est important de souligner qu'il s'agit d'une propriété générale: la part résiduelle de toute base de données «randomisée» conserve un potentiel d'identification très élevé, selon la rareté de la combinaison des attributs résiduels. C'est une mise en garde que les responsables du traitement des données devraient toujours avoir à l'esprit en choisissant la randomisation comme moyen de parvenir à l'anonymisation recherchée.

De nombreuses expériences de ré-identification de ce type ont été menées selon une approche similaire de projection de deux bases de données sur le même sous-espace. C'est une méthode de ré-identification très puissante, qui a récemment été appliquée dans de nombreux domaines différents. Par exemple, une expérience d'identification réalisée à l'encontre d'un réseau social[25] a exploité le graphe social d'utilisateurs pseudonymisés au moyen d'étiquettes. Dans ce cas, les attributs utilisés à des fins d'identification étaient les listes de contacts des différents utilisateurs, puisqu'il avait été démontré que la probabilité que deux individus aient une liste de contacts identique est très faible. Sur la base de cette hypothèse intuitive, il a été constaté qu'un sous-graphe de liens internes comportant un nombre très limité de nœuds constitue une empreinte topologique exploitable, cachée au sein du réseau, et qu'une large portion de l'ensemble du réseau social peut être identifiée dès lors que ce sous-réseau a été délimité. Pour ne donner que quelques chiffres sur les performances d'une attaque similaire, il a été démontré qu'en utilisant moins de 10 nœuds (qui peuvent déboucher sur des millions de configurations de sous-réseaux différentes, chacun constituant potentiellement une empreinte topologique) un réseau social de plus de 4 millions de nœuds pseudonymisés et de 70 millions de liens peut être vulnérable à des attaques de ré-identification susceptibles de compromettre un grand nombre de relations. Il faut ajouter que cette approche de ré-identification n'est pas limitée au contexte spécifique des réseaux sociaux, mais est suffisamment générale pour pouvoir être adaptée à d'autres bases de données où les relations entre les utilisateurs sont enregistrées (par exemple, un répertoire téléphonique, une messagerie électronique, des sites de rencontre, etc.).

Un autre moyen d'identifier un enregistrement supposé anonyme repose sur l'analyse du style de rédaction (stylométrie)[26]. Plusieurs algorithmes ont déjà été mis au point pour extraire des mesures de texte analysé, qui couvrent la fréquence d'utilisation d'un mot particulier, l'occurrence de constructions grammaticales spécifiques et le type de ponctuation. Toutes ces propriétés peuvent être utilisées pour associer un texte censé être anonyme au style de rédaction d'un auteur identifié. Des chercheurs ont extrait le style de rédaction de plus de 100 000 blogs et sont aujourd'hui capables d'identifier automatiquement l'auteur d'un

[25] L. Backstrom, C. Dwork, et J. M. Kleinberg, «Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography», compte rendu de la 16e Conférence internationale sur le World Wide Web WWW'07, p. 181 à 190 (2007).

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/

message avec une précision approchant déjà de 80 %; l'efficacité de cette technique devrait aussi se renforcer avec l'exploitation d'autres indications, comme la localisation ou des métadonnées contenues dans le texte.

Le potentiel d'identification au moyen de la sémantique d'un enregistrement (à savoir la part résiduelle non randomisée d'un enregistrement) est un problème qui mérite davantage de considération de la part des milieux de la recherche et de l'industrie. L'exemple récent d'une tentative couronnée de succès visant à rétablir les identités de donneurs d'ADN (2013)[27] montre que peu de progrès ont été accomplis depuis la célèbre affaire AOL (2006), où une base de données contenant vingt millions de mots-clés figurant dans les recherches effectuées par plus de 650 000 utilisateurs au cours d'une période de 3 mois avait été diffusée publiquement. À la suite de quoi, l'identité et la localisation de certains utilisateurs avaient été rendues publiques.

Les données de localisation constituent une autre famille de données dont l'anonymat est rarement garanti par le simple fait de supprimer les identités des personnes concernées ou par le chiffrement partiel de certains attributs. Les schémas de mobilité sont peut-être suffisamment uniques pour que la partie sémantique des données de localisation (le lieu où la personne concernée se trouvait à un certain moment), même en l'absence d'autres attributs, permette de révéler bon nombre de caractéristiques d'une personne concernée[28]. Cela a été démontré bien des fois dans des travaux universitaires représentatifs[29].

À cet égard, il faut se garder de considérer les pseudonymes comme un moyen d'assurer une protection adéquate des personnes concernées contre les fuites d'identité ou d'attribut. Si la pseudonymisation se fonde sur le remplacement d'une identité par un autre code unique, il serait naïf de supposer qu'un tel procédé constitue une solution d'anonymisation fiable, sans tenir compte de la complexité des méthodes d'identification et des multiples contextes dans lesquels elles pourraient être appliquées.

A.3. L'«anonymisation» par généralisation

Un exemple simple peut contribuer à clarifier l'approche fondée sur la généralisation de l'attribut.

Prenons le cas d'un responsable du traitement des données qui décide de publier un simple tableau contenant trois éléments d'information, ou attributs: un numéro d'identification, unique pour chaque enregistrement, une identification de localisation, qui relie la personne concernée au lieu où elle vit, et une identification de propriété, qui spécifie la propriété de la personne concernée. Supposons en outre que cette propriété correspond à une valeur parmi deux valeurs distinctes, indiquée de façon générique par {P1, P2}:

[27] Les données génétiques constituent un exemple particulièrement important de données sensibles, qui peuvent être exposées à un risque de ré-identification si le seul mécanisme censé les «anonymiser» est la suppression des identités des donneurs. Voir l'exemple cité à la section 2.2.2 ci-dessus. Voir aussi John Bohannon, «Genealogy Databases Enable Naming of Anonymous DNA Donors», *Science*, vol. 339, n° 6117 (18 janvier 2013), p. 262.

[28] Ce problème a été pris en compte dans certaines législations nationales. Par exemple, en France, les statistiques de localisation publiées sont anonymisées par des techniques de généralisation et de permutation. Ainsi, l'INSEE publie des statistiques qui sont généralisées en agrégeant toutes les données au niveau d'une superficie de 40 000 mètres carrés. La granularité de l'ensemble de données est suffisante pour préserver l'utilité des données et des permutations empêchent des attaques de ré-identification dans les zones clairsemées. D'une manière plus générale, l'agrégation de cette famille de données et la permutation apportent de solides garanties contre les attaques par inférence et les tentatives de ré-identification (http://www.insee.fr/fr/).

[29] De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. et Blondel, V.D., «Unique in the Crowd: The privacy bounds of human mobility», *Nature*, 3, 1376 (2013).

Identification par n° d'ordre	Localisation	Propriété
N° 1	Rome	P1
N° 2	Madrid	P1
N° 3	Londres	P2
N° 4	Paris	P1
N° 5	Barcelone	P1
N° 6	Milan	P2
N° 7	New York	P2
N° 8	Berlin	P1

Tableau A1. Échantillon de personnes concernées, avec leur localisation et leurs propriétés P1 ou P2

Si quelqu'un, qu'on appellera «l'attaquant», sait à l'avance qu'une personne concernée spécifique (la cible), qui vit à Milan, figure dans le tableau, il lui suffit d'examiner le tableau pour apprendre aussi que sa propriété est P2, le n° 6 étant la seule personne concernée identifiée par cette localisation.

Cet exemple très sommaire montre les principaux éléments de toute procédure d'identification appliquée à un ensemble de données qui a fait l'objet d'un processus d'anonymisation supposé. C'est-à-dire qu'un attaquant se trouve (accidentellement ou délibérément) en possession de connaissances tirées du contexte à propos de certaines ou de toutes les personnes concernées dans un ensemble de données. L'attaquant s'efforce de relier ces connaissances tirées du contexte avec les données figurant dans l'ensemble de données publié pour avoir une idée plus claire des caractéristiques de ces personnes concernées.

Afin de rendre moins efficace ou moins immédiate la mise en relation des données avec une forme quelconque de connaissances tirées du contexte, le responsable du traitement des données pourrait intervenir sur l'identification de localisation, en remplaçant la ville où vivent les personnes concernées par une zone plus large, comme le pays. De cette façon, le tableau se présenterait comme suit:

Identification par n° d'ordre	Localisation	Propriété
N° 1	Italie	P1
N° 2	Espagne	P1
N° 3	Royaume-Uni	P2
N° 4	France	P1
N° 5	Espagne	P1
N° 6	Italie	P2
N° 7	États-Unis	P2
N° 8	Allemagne	P1

Tableau A2. Généralisation du tableau A1 par nationalité

Avec cette nouvelle agrégation de données, les connaissances tirées du contexte dont dispose l'attaquant à propos d'une personne concernée identifiée (disons: «la cible vit à Rome et figure dans le tableau») ne permettent pas de parvenir à une conclusion claire concernant sa propriété, puisque les deux Italiens mentionnés dans le tableau ont des propriétés distinctes, respectivement P1 et P2. L'attaquant se trouve confronté à une incertitude de 50 % quant à la propriété de l'entité cible. Ce simple exemple montre l'effet de la généralisation sur la

pratique d'anonymisation. En fait, si ce procédé de généralisation peut être efficace pour réduire de moitié la probabilité d'identifier une cible italienne, il est sans effet dans le cas de cibles vivant à d'autres endroits (aux États-Unis, par exemple).

De plus, un attaquant peut encore obtenir des informations sur une cible espagnole. Si les connaissances tirées du contexte sont du type «la cible vit à Madrid et figure dans le tableau» ou «la cible vit à Barcelone et figure dans le tableau», l'attaquant peut en déduire avec 100 % de certitude que la cible a la propriété P1. Par conséquent, la généralisation ne garantit pas le même niveau de confidentialité ou de résistance aux attaques par inférence à toute la population de l'ensemble de données.

En suivant ce raisonnement, on pourrait être tenté de conclure qu'une généralisation accrue serait utile pour empêcher toute mise en relation – par exemple une généralisation par continent. De cette façon, le tableau se présenterait comme suit:

Identification par n° d'ordre	**Localisation**	**Propriété**
N° 1	Europe	P1
N° 2	Europe	P1
N° 3	Europe	P2
N° 4	Europe	P1
N° 5	Europe	P1
N° 6	Europe	P2
N° 7	Amérique du Nord	P2
N° 8	Europe	P1

Tableau A3. Généralisation du tableau A1 par continent

Avec ce genre d'agrégation, toutes les personnes concernées dans le tableau, hormis celle qui vit aux États-Unis, seraient protégées contre les attaques par corrélation et les tentatives d'identification, et toute information tirée du contexte du type «la cible vit à Madrid et figure dans le tableau» ou «la cible vit à Milan et figure dans le tableau» aboutirait à un certain niveau de probabilité quant à la propriété qui s'applique à la personne concernée (P1 avec une probabilité de 71,4 % et P2 avec une probabilité de 28,6%), plutôt qu'à une mise en relation directe. Mais cette généralisation supplémentaire s'opère au prix d'une perte évidente et radicale d'informations: le tableau ne permet pas de découvrir des corrélations potentielles entre les propriétés et la localisation, c'est-à-dire d'apprécier si un lieu spécifique serait plus susceptible de déclencher l'une des deux propriétés, puisqu'il ne fait apparaître que les distributions dites «marginales», à savoir la probabilité absolue de l'occurrence des propriétés P1 et P2 dans l'ensemble de la population (respectivement 62,5 % et 37,5 % dans notre exemple) et dans chaque continent (respectivement, comme il a été indiqué, 71,4 % et 28,6% en Europe et 100 % et 0 % en Amérique du Nord).

L'exemple montre aussi que le recours à la généralisation affecte l'utilité pratique des données. Il existe aujourd'hui certains outils qui permettent de déterminer au préalable (c'est-à-dire avant qu'un ensemble de données soit rendu public) quel est le niveau de généralisation de l'attribut le plus approprié, de façon à réduire les risques d'identification des personnes concernées dans un tableau sans affecter exagérément l'utilité des données publiées.

k-anonymat

Le k-anonymat est une technique fondée sur la généralisation des attributs qui vise à prévenir les attaques par corrélation. Cette pratique est issue d'une expérience de ré-identification menée à la fin des années 1990, aux États-Unis, où une entreprise privée, active dans le secteur de la santé, a rendu public un ensemble de données censé être anonymisé. Cette anonymisation consistait à effacer les noms des personnes concernées, mais l'ensemble de données contenait encore des informations d'ordre médical et d'autres attributs comme le code postal (identification de localisation indiquant où vivaient les personnes concernées), le sexe et la date de naissance complète. Le même triplet {code postal, sexe, date de naissance complète} figurait aussi dans d'autres registres accessibles au public (par exemple, la liste électorale) et a donc pu être utilisé par un chercheur pour mettre en relation l'identité de certaines personnes concernées avec les attributs de l'ensemble de données publié. Les connaissances tirées du contexte dont disposait l'attaquant (le chercheur) pouvaient être énoncées comme suit: «Je sais que la personne concernée figurant dans la liste électorale avec un triplet {code postal, sexe, date de naissance complète} spécifique est unique. Il existe un enregistrement correspondant à ce triplet dans l'ensemble de données publié.» Il a été constaté empiriquement[30] que la grande majorité (plus de 80 %) des personnes concernées dans le registre public utilisé pour cette expérience de recherche étaient associées de façon univoque à un triplet spécifique, ce qui rendait l'identification possible. Par conséquent, les données n'étaient pas correctement anonymisées dans ce cas.

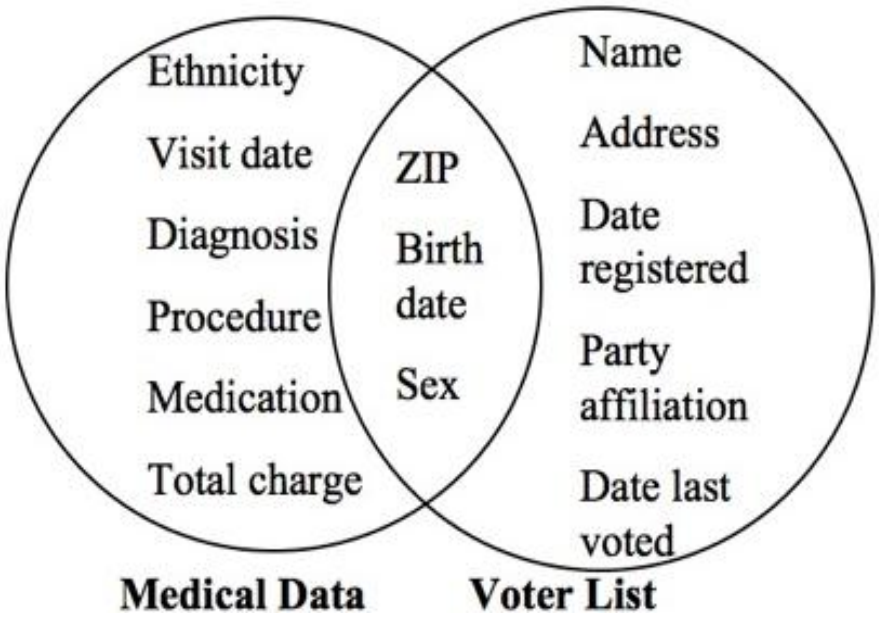

Figure A1. Ré-identification par corrélation entre les données

Afin de réduire l'efficacité d'attaques par corrélation similaires, il a été avancé que les responsables du traitement devraient d'abord examiner l'ensemble de données et regrouper les attributs qui pourraient raisonnablement être exploités par un attaquant pour mettre le tableau publié en relation avec une autre source auxiliaire; chaque groupe devrait inclure au moins *k* combinaisons identiques d'attributs généralisés (c'est-à-dire qu'il devrait représenter une classe d'équivalence d'attributs). Les ensembles de données ne seraient ensuite publiés qu'après avoir été répartis dans de tels groupes homogènes. Les attributs retenus en vue de

[30] L. Sweeney, «Weaving Technology and Policy Together to Maintain Confidentiality», *Journal of Law, Medicine & Ethics*, 25, n° 2 et 3 (1997), p. 98 à 110

leur généralisation sont appelés «quasi-identifiants», puisque leur connaissance, en clair, entraînerait l'identification immédiate des personnes concernées.

De nombreuses expériences d'identification ont démontré la faiblesse de tableaux k-anonymisés mal conçus. Cela peut être dû, par exemple, au fait que les autres attributs d'une classe d'équivalence sont identiques (comme dans le cas de la classe d'équivalence des personnes concernées espagnoles dans l'exemple du tableau A2) ou que leur distribution est très déséquilibrée, avec une forte prévalence d'un attribut spécifique, ou encore au fait que le nombre d'enregistrements dans une classe d'équivalence est très réduit, ce qui permet dans les deux cas une inférence probabiliste, ou qu'il n'existe pas de différence «sémantique» significative entre les attributs en clair des classes d'équivalence (ainsi, la mesure quantitative de ces attributs pourrait être effectivement différente, mais très proche en termes numériques, ou ils pourraient relever d'une gamme d'attributs sémantiquement similaires, par exemple, une même catégorie de risque de crédit ou une même famille de pathologies), de telle sorte que l'ensemble de données peut encore laisser filtrer une grande quantité d'informations sur les personnes concernées exploitables par des attaques par corrélation[31]. Un point important sur lequel il faut insister ici est que, dans tous les cas où les données sont clairsemées (si, par exemple, il y a peu d'occurrences d'une propriété spécifique dans une zone géographique) et où une première agrégation ne permet pas de regrouper les données avec un nombre suffisant d'occurrences de propriétés différentes (si, par exemple, il reste un petit nombre d'occurrences de quelques propriétés seulement qui peuvent être localisées dans une zone géographique), il est nécessaire de procéder à une agrégation d'attributs supplémentaire pour atteindre l'anonymisation recherchée.

l-diversité

À partir de ces observations, des variantes du k-anonymat ont été proposées au fil des années, et certains critères techniques de renforcement de la pratique d'anonymisation par généralisation ont été mis au point, en vue de réduire les risques des attaques par corrélation. Ils se fondent sur des propriétés probabilistes des ensembles de données. En particulier, une contrainte supplémentaire est ajoutée, à savoir que chaque attribut d'une classe d'équivalence apparaît au moins à *l* reprises, de telle sorte qu'un attaquant reste toujours confronté à un degré d'incertitude considérable concernant les attributs, malgré les connaissances tirées du contexte dont il pourrait disposer à propos d'une personne concernée. Cela revient à dire qu'un ensemble de données (ou un segment) doit posséder un nombre minimal d'occurrences d'une propriété sélectionnée: ce procédé permet d'atténuer le risque de ré-identification. Tel est l'objectif de la pratique d'anonymisation à l-diversité. Un exemple de cette pratique est donné dans les tableaux A4 (les données originales) et A5 (le résultat du traitement). Comme on le voit, en traitant correctement l'identification de localisation et les âges des individus du tableau A4, le processus de généralisation d'attributs se traduit par une augmentation considérable de l'incertitude quant aux attributs réels de chaque personne concernée dans l'enquête. Par exemple, même si l'attaquant sait qu'une personne concernée figure dans la première classe d'équivalence, il ne peut déterminer avec plus de certitude si une personne a la propriété X, Y ou Z, puisqu'il existe au moins un enregistrement dans cette classe (et dans n'importe quelle classe d'équivalence) présentant ces propriétés.

[31] Il faut souligner que des corrélations peuvent aussi être établies une fois que les enregistrements ont été regroupés par attributs. Quand le responsable du traitement des données sait quels sont les types de corrélations qu'il souhaite vérifier, il peut sélectionner les attributs les plus pertinents. Par exemple, les résultats des enquêtes du centre PEW ne sont pas vulnérables à des attaques par inférence à granularité fine et restent très utiles pour la recherche de corrélations entre les données démographiques et les interêts (http://www.pewinternet.org/Reports/2013/Anonymity-online.aspx).

Numéro d'ordre	Identification de localisation	Âge	Propriété
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tableau A4. Un tableau où les individus sont regroupés par localisation, âge et trois propriétés X, Y et Z

Numéro d'ordre	Identification de localisation	Âge	Propriété
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tableau A5. Un exemple de version à 1-diversité du tableau A4

t-proximité:

L'approche désignée par le terme «t-proximité» prend en considération le cas particulier des attributs qui sont distribués de manière inégale au sein d'un segment ou qui ne présentent qu'un faible écart de valeurs ou de contenus sémantiques. C'est une amélioration supplémentaire de l'anonymisation par généralisation consistant à organiser les données de façon à créer des classes d'équivalence qui reflètent autant que possible la distribution initiale des attributs dans l'ensemble de données original. À cet effet, une procédure en deux étapes est appliquée comme suit. Le tableau A6 constitue la base de données originale comprenant les enregistrements en clair des personnes concernées, groupées par localisation, âge, salaire et deux familles de propriétés sémantiquement similaires, respectivement {X1,X2, X3} et {Y1, Y2, Y3} (par exemple, des classes de risque de crédit similaires, des maladies similaires). Premièrement, le tableau est *l-diversifié* avec une valeur *l*=1 (tableau A7), en regroupant des enregistrements en classes d'équivalence sémantiquement similaires qui présentent un faible niveau d'anonymisation ciblée; puis il est traité en vue d'obtenir une t-proximité (tableau A8) et une plus grande variabilité au sein de chaque segment. En fait, après cette deuxième étape, chaque classe d'équivalence comprend des enregistrements des deux

familles de propriétés. Il est à noter que l'identification de localisation et l'âge ont des granularités différentes dans les diverses étapes du processus: il s'ensuit que chaque attribut peut nécessiter des critères de généralisation différents pour parvenir à l'anonymisation recherchée, ce qui à son tour requiert, de la part des responsables du traitement des données, un ajustement spécifique et un calcul informatique approprié.

Numéro d'ordre	Identification de localisation	Âge	Salaire	Propriété
1	1127	29	30 000	X1
2	1112	22	32 000	X2
3	1128	27	35 000	X3
4	1215	43	50 000	X2
5	1219	52	120 000	Y1
6	1216	47	60 000	Y2
7	1115	30	55 000	Y2
8	1123	36	100 000	Y3
9	1117	32	110 000	X3

Tableau A6. Un tableau où les individus sont regroupés par localisation, âge, salaire et deux familles de propriétés

Numéro d'ordre	Identification de localisation	Âge	Salaire	Propriété
1	11**	2*	30 000	X1
2	11**	2*	32 000	X2
3	11**	2*	35 000	X3
4	121*	>40	50 000	X2
5	121*	>40	120 000	Y1
6	121*	>40	60 000	Y2
7	11**	3*	55 000	Y2
8	11**	3*	100 000	Y3
9	11**	3*	110 000	X3

Tableau A7. Une version à 1-diversité du tableau A6

Numéro d'ordre	Identification de localisation	Âge	Salaire	Propriété
1	112*	<40	30 000	X1
3	112*	<40	35 000	X3
8	112*	<40	100 000	Y3
4	121*	>40	50 000	X2
5	121*	>40	120 000	Y1
6	121*	>40	60 000	Y2
2	111*	<40	32 000	X2
7	111*	<40	55 000	Y2
9	111*	<40	110 000	X3

Tableau A8. Une version à t-proximité du tableau A6

Il faut préciser que l'objectif de la généralisation des attributs des personnes concernées par des procédés aussi élaborés ne peut parfois être atteint que pour un petit nombre d'enregistrements et non pour l'ensemble d'entre eux. Les bonnes pratiques devraient veiller à ce que chaque classe d'équivalence contienne plusieurs individus et qu'aucune attaque par

inférence ne reste possible. En tout cas, cette approche requiert un examen approfondi des données disponibles de la part des responsables du traitement, ainsi qu'une analyse combinatoire de diverses alternatives (par exemple, des fourchettes d'amplitudes différentes, une granularité différente en termes de localisation ou d'âge, etc.). Autrement dit, l'anonymisation par généralisation ne peut être le résultat d'une tentative rudimentaire qui consisterait, pour les responsables du traitement des données, à remplacer des valeurs d'attributs analytiques dans un enregistrement par des fourchettes. Des approches quantitatives plus spécifiques sont nécessaires, de façon par exemple à évaluer l'entropie des attributs au sein de chaque segment ou à mesurer la distance entre les distributions originales des attributs et la distribution dans chaque classe d'équivalence.