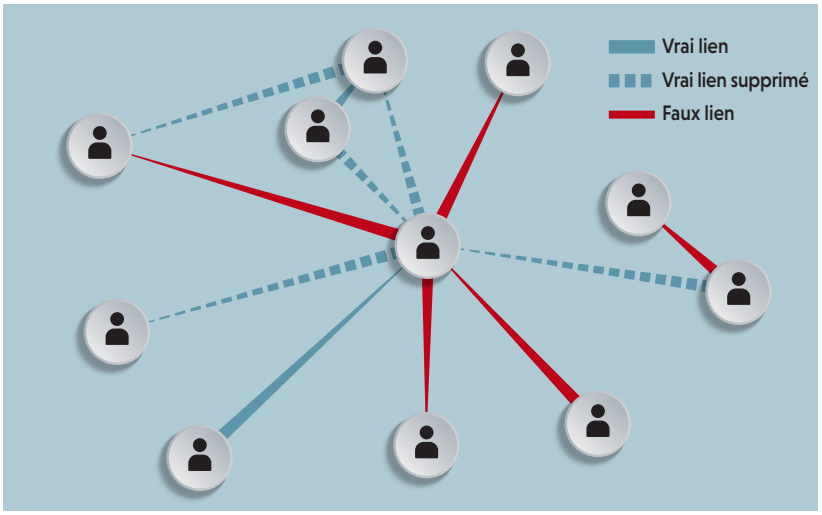
York, et ses collègues, quantifie la connaissance préalable qu'a l'attaquant de sa cible par la probabilité de trouver la valeur convoitée avant l'observation du jeu de données: si l'attaquant recherche le diagnostic médical d'une personne nommée Dupont, qu'il ne connaît de sa cible que son nom et qu'il sait que 10% des Dupont ont un cancer, il a 1 chance sur 10 de ne pas se tromper en affirmant que sa cible a un cancer. Cette connaissance est dite *a priori*. Après la découverte des données, l'attaquant a plus de chances de trouver la donnée qu'il cherche; la connaissance de sa cible qui en résulte, connaissance dite *a posteriori*, est quantifiée par la probabilité de lui associer la bonne valeur sensible.

C'est ici que la théorie bayésienne intervient, car elle permet de calculer des probabilités conditionnelles: les probabilités de déduire telle ou telle information d'un jeu de données sachant qu'on a telle ou telle connaissance préalable. La confidentialité des données publiées est caractérisée par la différence entre connaissance *a posteriori* et connaissance *a priori*, autrement dit par la connaissance supplémentaire sur la cible apportée par les données. Assurer un niveau adéquat de protection revient à limiter cette valeur. Cependant, cette technique est difficile à mettre en œuvre, car elle suppose de connaître exactement ce que l'attaquant sait de sa cible.

La l-diversité, méthode conçue la même année et par les mêmes chercheurs, est davantage applicable. Elle consiste à imposer la présence de l valeurs sensibles distinctes dans chaque groupe (*voir l'encadré page 102*). Un attaquant qui ne connaît de sa cible que le groupe auquel elle appartient ne peut trouver sa valeur sensible avec une probabilité supérieure à $1/l$.

De nombreuses autres techniques ont été élaborées depuis, afin de s'adapter à la diversité des données, des attaques et des attaquants. La «t-proximité» vise à créer des groupes au sein desquels la distribution des données est à peu près la même que dans la population globale.



Les graphes de réseau social représentent les individus par des points et leurs liens par des traits. Pour qu'ils ne dévoilent pas la vie privée des individus, on peut leur appliquer un algorithme dit de confidentialité différentielle, consistant, par exemple, à supprimer de nombreux vrais liens et à les remplacer par autant de faux liens.

La «(c, k)-sûreté» prend en compte la capacité de l'attaquant à formuler k déductions logiques du type: «Si Adrien a la grippe, alors Bettie aussi.» La «3D-confidentialité» considère que l'attaquant connaît *a priori* trois types de données: l valeurs sensibles que sa cible n'a pas, k valeurs sensibles que d'autres individus ont et m déductions logiques entre individus. La «m-invariance» protège contre les comparaisons entre les publications successives d'un même jeu de données, telles certaines informations statistiques sur un hôpital, dont les variations reflètent les arrivées, les départs et les évolutions des patients.

L'application de ces techniques commence souvent par celle de l'algorithme de Mondrian. Il sert à construire des groupes k-anonymes, au sein desquels on vérifie la conformité de la distribution des données sensibles vis-à-vis du modèle choisi (*voir l'encadré page 102*).

Volontairement ou non, tous ces modèles sont des exemples du paradigme de non-information, formulé en 1977 par le statisticien suédois Tore Dalenius: selon ce paradigme, un jeu de données est d'autant plus confidentiel qu'il renseigne peu l'attaquant sur sa cible. En 2006, Cynthia Dwork, alors chercheuse à Microsoft, a proposé une nouvelle façon de définir la confidentialité, qualifiée de confidentialité différentielle: un jeu de données serait confidentiel s'il n'est presque pas modifié par l'ajout des données d'un individu, quelles qu'elles soient.

NOYER LES DONNÉES

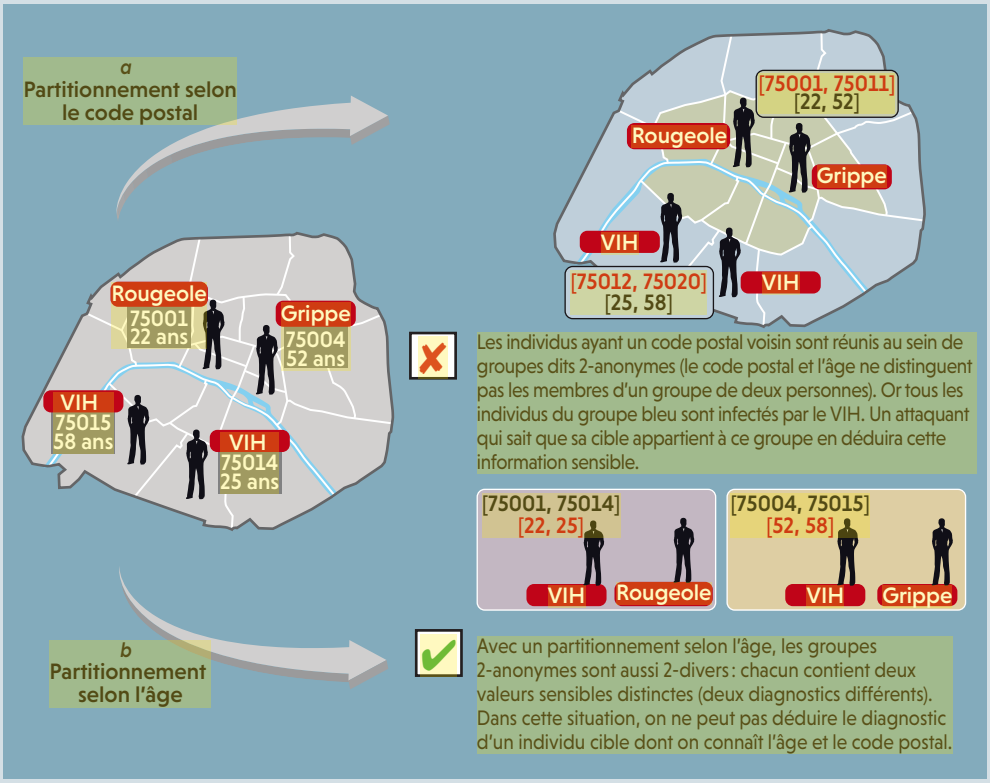
L'assainissement consiste alors à perturber les données de sorte que celles de chaque individu se trouvent noyées dans la perturbation (et non plus dans la foule, comme dans le k-anonymat). On introduit, par exemple, un grand nombre de fausses données (typiquement 100 fois plus que de vraies), afin que le jeu de données qui contient la contribution d'un individu particulier ressemble beaucoup à celui qui ne la contient pas. Certains algorithmes suppriment aussi de vraies données. La protection est assurée par l'impossibilité de distinguer les fausses données des vraies dans le jeu de données assaini, dont on ne peut même pas garantir qu'il contient la contribution d'un individu précis.

L'idée de la confidentialité différentielle est en plein essor. Elle est désormais applicable à des types variés de données, tels les graphes de réseaux sociaux. Ces graphes représentent les individus par des nœuds, connectés par des liens. La perturbation peut alors consister à supprimer de vrais liens et à en introduire de faux. Des informations tel le nombre de liens sont toujours extractibles du graphe, sans que l'on puisse déterminer les connexions précises d'un individu.

Le succès de la confidentialité différentielle s'explique en partie par sa simplicité: l'attaquant >

MONDRIAN ET L'ANONYMISATION

Pour compliquer voire empêcher l'identification d'un individu à partir de ses données, une des meilleures solutions est l'algorithme dit de Mondrian. Pourquoi une telle référence au peintre néerlandais ?



Pour rendre le titulaire d'un jeu de données impossible à identifier, on se contente souvent de remplacer les champs tels que le nom ou le numéro de sécurité sociale par un pseudonyme. C'est insuffisant. D'autres techniques ont alors été développées, dont celle du *k*-anonymat, la plus employée aujourd'hui. Elle consiste à brouiller certains champs, dits quasi-identifiants parce que leur combinaison est parfois unique : l'âge, le sexe, le code postal… On remplace leurs valeurs précises par des intervalles de valeurs ou des catégories, de façon à ce que la combinaison résultante soit partagée par au moins *k* personnes.

Pour appliquer cette technique, on utilise souvent l'algorithme de Mondrian, élaboré en 2006. Son nom renvoie au partitionnement de l'espace cher au peintre hollandais Piet Mondrian. De même, l'algorithme partitionne l'espace des quasi-identifiants, où ces derniers sont indiqués sur des axes et où les individus sont repérés par des points. L'algorithme commence par choisir un quasi-identifiant, selon lequel il divise les individus en deux groupes, par exemple ceux dont le code postal est inférieur ou égal à 75011, et les autres *(a)*. Si *k* est supérieur à 2, les groupes résultants sont de nouveau divisés en deux pour former quatre groupes. Ce découpage continue jusqu'à ce qu'aucun groupe ne puisse plus être divisé sans englober moins de *k* individus. La dernière étape consiste à remplacer les quasi-identifiants de chacun par l'ensemble de valeurs couvert par son groupe.

Le problème est que les données personnelles sensibles à protéger, tel le diagnostic médical, peuvent être identiques au sein d'un groupe. Un attaquant qui sait que sa cible appartient à ce groupe a alors accès à sa valeur sensible. Plusieurs techniques visent à y remédier. La *l*-diversité, par exemple, consiste à construire des groupes ayant au moins *l* valeurs sensibles. On choisit la valeur de *l* en fonction de ce que l'attaquant sait de sa cible, c'est-à-dire du nombre maximal de valeurs sensibles qu'on l'estime capable d'éliminer. Plus il sait de choses, plus *l* sera élevé, plus il faut inclure d'individus dans chaque groupe afin d'avoir une diversité suffisante, et moins les statistiques calculées à partir du jeu de données assaini sont fines. Le choix de *l* traduit un compromis entre utilité et protection.

En pratique, on construit souvent les groupes grâce à l'algorithme de Mondrian. À chaque division d'un groupe en deux, on vérifie que les nouveaux groupes contiennent au moins *l* valeurs sensibles distinctes. Dans le cas contraire, on revient en arrière et on partitionne le groupe selon un autre quasi-identifiant *(b)*.

n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, n'est pas nécessaire. Mais cette simplicité est à double tranchant: des travaux publiés en 2011 ont montré que la non-prise en compte des connaissances annexes de l'attaquant et des relations potentielles entre les individus d'un jeu de données ouvre des failles dans la protection.

Un modèle d'assainissement universel des données reste donc chimérique. Chaque modèle a ses défenseurs et ses détracteurs, et est plus ou moins efficace selon la situation.

Outre le modèle d'assainissement, l'architecture de gestion des données a une importance. Les données nominatives sont souvent extraites du système informatique assurant leur usage quotidien (tel le serveur d'un centre de soin), puis copiées sous une forme pseudonymisée dans un entrepôt de données. C'est à partir de cet entrepôt que seront produits à la demande des jeux de données assainis pour différents destinataires. En France et en Angleterre, où le système de dossier médical personnel national fonctionne ainsi, les épidémiologistes peuvent par exemple recevoir des jeux de données plus précis que les industriels pharmaceutiques.

Toutefois, ce principe d'assainissement centralisé nécessite une fiabilité totale du gestionnaire de l'entrepôt de données. Or on constate régulièrement des fuites d'information sur de nombreux serveurs, dues à des négligences ou à des attaques. Même si les données stockées dans les entrepôts sont pseudonymisées, nous avons vu que cela n'apporte pas une protection suffisante. Il est donc légitime de s'interroger sur les risques de la centralisation.

L'ASSAINISSEMENT DISTRIBUÉ

Les statisticiens ont proposé un mécanisme d'assainissement décentralisé dès les années 1960, pour parer aux réticences à leur confier des données personnelles sensibles. Le principe est de perturber les données de chacun au moment de leur collecte, ce qui les protège avant tout enregistrement.

Cependant, on sait aujourd'hui qu'une perturbation indépendante de chaque donnée ne permet pas d'atteindre le niveau de qualité d'un assainissement réalisé sur le jeu de données dans son ensemble. La centralisation des données personnelles est-elle alors nécessaire à un assainissement de qualité? Non, car il est aussi possible d'élaborer des mécanismes décentralisés où les perturbations ne sont pas indépendantes. En d'autres termes, la perturbation à appliquer n'est pas décidée localement, mais par une entité centrale. Celle-ci possède des informations sur toutes les réponses, sans connaître les réponses elles-mêmes.

Des algorithmes regroupés sous le nom de Secure Multi-Party Computation, qui font souvent intervenir des techniques de cryptographie, visent à assurer la confidentialité de l'assainissement distribué (où les calculs sont répartis entre plusieurs acteurs).

Les mécanismes distribués constituent un pas majeur vers la sécurisation du processus d'assainissement, mais leur mise en œuvre pose des problèmes de passage à l'échelle, car ils nécessitent des calculs importants et une forte connectivité des participants. Ils sont pour l'instant relégués au traitement de jeux de données peu volumineux.

Face à ce problème, notre équipe a développé une architecture de gestion de données personnelles fondée sur des dispositifs individuels, telles des clés USB, sécurisés. Nous avons montré que les calculs nécessaires aux algorithmes d'assainissement peuvent être répartis entre ces dispositifs et une entité centrale sans que celle-ci ne dispose des données personnelles non chiffrées. La puissance de calcul de l'entité centrale assure l'applicabilité à grande échelle et, grâce au fait qu'elle coordonne les dispositifs personnels, ceux-ci n'ont pas besoin d'être interconnectés en permanence. Cette architecture est capable d'appliquer des algorithmes d'assainissement à des jeux de données massifs, correspondant à plusieurs millions d'individus.

Une telle architecture décentralisée pourrait assurer la gestion des dossiers médicaux, des factures ou de tout autre type de dossier personnel. En pratique, les données seraient stockées localement sur les appareils de l'utilisateur et ne seraient jamais exportées sous une forme non perturbée ou non chiffrée. Aucun serveur ne les regrouperait toutes, mais les échanges entre l'entité centrale et les appareils des utilisateurs permettraient tout de même de collecter certaines informations, en respectant la vie privée.

Pendant la dernière décennie, une grande diversité de modèles et d'algorithmes d'assainissement de données ont été élaborés, afin de mieux prendre en compte les capacités de l'attaquant. Aujourd'hui, la pseudonymisation reste massivement utilisée, alors qu'elle ne garantit pas une protection suffisante; dans le reste des cas, la *k*-anonymisation est le plus souvent appliquée, et les techniques plus complexes sont encore exotiques. Ces techniques doivent donc continuer à se répandre et à se perfectionner.

Pour autant, l'assainissement reste le meilleur compromis entre utilité et confidentialité des données, dont la protection absolue est illusoire. L'analyse des nouveaux gisements de données personnelles pouvant apporter des bénéfices notables, l'assainissement de données est une question sociétale avant d'être un défi scientifique. ■

BIBLIOGRAPHIE

T. ALLARD *ET AL.*, METAP: Revisiting privacy-preserving data publishing using secure devices, *Distributed and Parallel Databases*, pp. 1-54, 2013 (à paraître).

B. C. CHEN *ET AL.*, Privacy-preserving data publishing, *Found. and Trends in Databases*, vol. 2, n° 1-2, pp. 1-167, 2009.

A. GREENFIELD, *Everyware. La révolution de l'ubimédia*, Pearson, 2007.

C. DWORK, Differential privacy, Proceedings of the 33rd international conference on Automata, *Languages and Programming*, vol. 2, pp. 1-12, 2006.

L. SWEENEY, *k*-Anonymity: A model for protecting privacy, *Int. J. Uncertain, Fuzziness Knowl.-Based Syst.*, vol. 10(5), pp.557-570, 2002.

L'Open Security Foundation, organisation non gouvernementale américaine, recense les cas les plus significatifs sur son site Internet DataLossDB.org