Pour la Science, © Shutterstock.com/VoodooDot, Valua Vitaly, mjaud

1. LES REQUÊTES TAPÉES DANS UN MOTEUR DE RECHERCHES renseignent sur l'identité de celui qui les soumet. Ainsi, supposons que l'on sache que l'individu 6208 s'intéresse à l'arbre généalogique des Dupont et qu'il a cherché un coiffeur et un boulanger rue Férou. En croisant ces requêtes avec des données disponibles sur Internet, telles les pages blanches, on découvre qu'un certain Marc Dupont habite rue Férou. On peut alors avoir accès aux requêtes de cet individu et s'apercevoir, par exemple, qu'il s'intéresse à la dépression (sans doute en souffre-t-il) ou qu'il a consulté des pages où l'on propose des rencontres discrètes. Sa vie privée est menacée.

individu mal intentionné. En outre, les analystes se contentent de moins en moins de statistiques sur les données personnelles et demandent d'accéder directement à celles-ci, afin d'augmenter la précision de leurs études.

La conjonction de ces deux facteurs a exacerbé la nécessité d'une protection robuste. Au cours des dix dernières années, de nombreux chercheurs ont étudié la façon de publier des données tout en préservant la vie privée des individus concernés. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée, afin de réduire la probabilité que l'on puisse réidentifier la personne correspondante ou accéder à des informations sensibles la concernant.

Les faiblesses de la pseudonymisation

Comment rendre anonyme un jeu de données ? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650 000 utilisateurs sur une période de trois mois (officiellement pour les rendre disponibles aux chercheurs de divers domaines). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots clefs des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont identifié la personne portant le pseudonyme 4417749. Elle avait tapé des centaines de requêtes, tels « paysagiste à Lilburn, Géorgie », « individus dont le nom de famille est Arnold » ou « hommes célibataires de 60 ans ». Les journalistes ont croisé ces mots clefs avec des données disponibles sur le Web et ont peu à peu reconstitué l'identité probable de l'individu correspondant, une veuve de 62 ans habitant Lilburn, qu'ils ont fini par retrouver.

De nombreux cas similaires illustrent l'échec de la pseudonymisation à rendre les données anonymes. Remplacer les seuls champs directement identifiants, tels que le nom et le prénom ou le numéro de sécurité sociale, ne prémunit pas contre une réidentification ultérieure *(voir la figure 2)*. Pourtant, la pseudonymisation de données reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

Prévenir les croisements de données

Au début des années 2000, Latanya Sweeney, de l'Université Carnegie Mellon, aux États-Unis, a proposé une méthode, nommée *k*-anonymat, pour prévenir les réidentifications *via* des croisements de données. Elle a d'abord montré qu'il était parfois possible de retrouver le titulaire d'un jeu de données médicales pseudonymisé à partir des champs {Sexe, Date de naissance, Code postal}. En effet, ces champs sont aussi présents dans d'autres jeux de données comprenant le nom du titulaire *(voir la figure 2)*. En outre, la combinaison des renseignements correspondants est

LES AUTEURS

Tristan ALLARD est postdoctorant à l'Inria (Institut national de recherche en informatique et en automatique).

Benjamin NGUYEN est chercheur à l'Inria et maître de conférences à l'Université de Versailles Saint-Quentin-en-Yvelines.

Philippe PUCHERAL est chercheur à l'Inria et professeur à l'Université de Versailles Saint-Quentin-en-Yvelines.

unique pour la plupart des gens, selon plusieurs études récentes – qui portaient sur les États-Unis, mais c'est probablement aussi le cas ailleurs.

Pour empêcher les réidentifications, L. Sweeney a proposé de scinder les champs du jeu de données en deux catégories : les champs quasi identifiants (QID, pour *Quasi identifiers data*), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD, pour *Sensitive data*), qui doivent rester privés (par exemple, un diagnostic médical). Les valeurs des quasi-identifiants d'un individu sont ensuite rendues identiques à celles d'au moins $(k - 1)$ autres individus dans le jeu de données, pour un entier k convenablement choisi. En d'autres termes, le k-anonymat dissimule chaque individu dans une foule d'au moins k personnes impossibles à distinguer par leurs quasi-identifiants. Tout croisement avec une autre source de données aura ainsi une précision inférieure à $1/k$.

Les algorithmes du k-anonymat se fondent sur le principe de généralisation, qui consiste à remplacer les valeurs précises des quasi-identifiants par des ensembles de valeurs. Ceux-ci peuvent être des intervalles numériques ou des catégories. Par exemple, l'intervalle d'âges [15, 20] (c'est-à-dire de 15 à 20 ans) englobe plus d'individus qu'un âge précis. De même, la catégorie « France » englobe plus d'individus que la catégorie « Bourgogne ».

Il serait simple d'élaborer un algorithme qui parcourre les quasi-identifiants et forme un ensemble les incluant tous, tel {âge de 0 à 150 ans, habitant sur la Terre}. Mais les données seraient trop dégradées (c'est-à-dire trop imprécises) pour être utiles aux analystes. On a alors développé des algorithmes dits par partitionnement : ils forment des ensembles d'au moins k individus en regroupant les quasi-identifiants voisins, qu'ils remplacent ensuite par un intervalle ou une catégorie les incluant tous. Le plus utilisé est l'algorithme dit de Mondrian *(voir l'encadré page 66)*.

Cependant, le k-anonymat est loin de résoudre tous les problèmes. Considérons la situation suivante : un attaquant dispose d'un jeu de données k-anonyme et recherche la valeur sensible (tel le diagnostic médical) d'un individu cible dont il connaît le quasi-identifiant. Il peut retrouver le groupe auquel appartient sa cible, et donc l'ensemble des valeurs sensibles de ce groupe. La protection apportée par le k-anonymat est d'autant plus faible que le nombre de ces valeurs est petit. Dans le cas extrême, supposons que les k individus du groupe partagent la même valeur sensible, par exemple qu'ils aient tous un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer : la valeur sensible n'est pas protégée.

Brouiller les données confidentielles

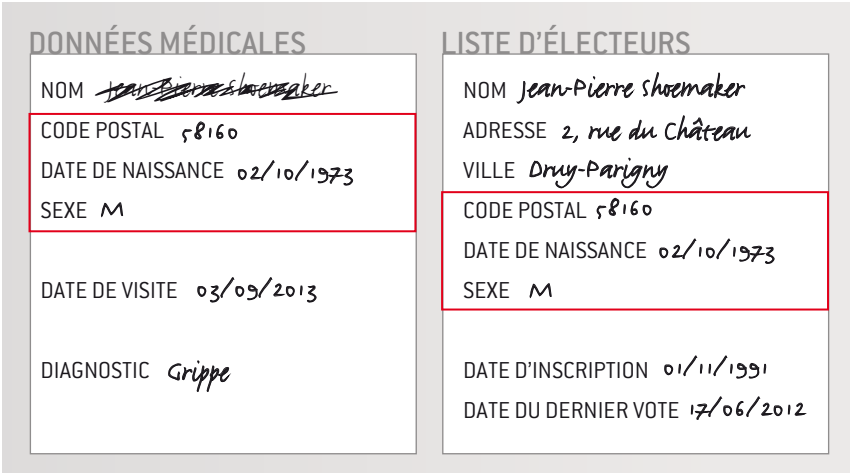
De multiples techniques ont succédé au k-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre en compte des attaques potentielles réalistes.

Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'Université Cornell à New 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 dix pour cent des Dupont ont un cancer, il a une chance sur dix 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. C'est d'autant plus difficile qu'on ignore souvent qui sera le ou les attaquants – ces derniers pouvant même avoir des connaissances *a priori* différentes.

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 66)*. 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 $1/l$.



2. LE TITULAIRE D'UN DOSSIER MÉDICAL dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données (de tels dossiers circulent parfois entre les hôpitaux et divers instituts de recherche ou de statistique). En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(encadrée en rouge)* est unique. Il suffit alors de trouver un autre jeu de données qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière. C'est le cas, par exemple, d'une liste électorale *(à droite)*, disponible sur simple demande à la mairie.