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.