Comment protéger les données concernant un individu tout en obtenant des informations utiles sur une population à laquelle il appartient ? C'est le paradoxe que résout la confidentialité différentielle.

Anonymat garanti

Tristan Allard, Louis Béziaud et Sébastien Gambs


7298470568305267901232 10

En bref

> Pour préserver l'anonymat des données personnelles, des méthodes longtemps utilisées, non fondées formellement, ont montré leurs limites.

> Aujourd'hui, la « confidentialité différentielle » et ses variantes pallient ces défauts et sont largement plébiscitées.

> Elles consistent à ajouter un bruit aléatoire aux données de façon à cacher la contribution de n'importe quel individu.

> Ce modèle de vie privée est désormais utilisée par Apple, Google et d'autres géants du numérique.

Le 22 mai 2021, Guillaume Rozier, informaticien, a été décoré de l'Ordre national du mérite pour avoir créé deux sites, l'un pour suivre la progression de la pandémie de Covid-19, l'autre afin de trouver facilement un créneau de vaccination disponible. Cela n'aurait pas été possible sans la mise à disposition en *open data* de données publiques, notamment de la part du ministère de la Santé. Au-delà de cet exemple, l'intérêt de rendre accessibles des informations, parfois personnelles, est aujourd'hui largement reconnu.

Cependant, il importe d'offrir aux individus concernés des garanties fortes de protection de leur vie privée. En Europe, et plus particulièrement en France, elles sont offertes par un arsenal juridique constitué dans le sillage du règlement général sur la protection des données (le RGPD) pour favoriser la diffusion et la réutilisation de telles données. L'anonymisation est la clé de voûte de cet édifice juridique. Elle rend possible non seulement le partage des données, mais aussi leur valorisation sans avoir besoin de revenir vers les individus pour obtenir leur consentement.

Le récent rapport de la mission du député Éric Bothorel sur la politique publique de la donnée, des algorithmes et des codes sources, fait état de cette difficile tension entre protection et exploitation. Néanmoins, les progrès de ces dernières années dans le développement de techniques d'anonymisation robustes suggèrent que le défi n'est pas insurmontable.

Historiquement, les approches non formelles, c'est-à-dire empiriques et non fondées sur des modèles mathématiques de vie privée rigoureux, à la problématique de l'anonymisation ont prévalu. Parmi elles, la pseudonymisation a longtemps été jugée digne de confiance : elle consiste à remplacer les informations directement identifiantes, comme un numéro de sécurité sociale, par une chaîne d'octets aléatoire, une sorte de pseudonyme. Cependant, il manque à ces approches un composant essentiel : un modèle de vie privée solide, c'est-à-dire la garantie exigée des algorithmes d'anonymisation, exprimée par une équation. Elle formalise le type de garanties offertes et le niveau de protection atteint (et parfois aussi le type d'attaquant auquel le protocole de protection résiste). Bien entendu, un modèle formel de vie privée ne fonctionne pas à 100 %, mais il explicite les garanties de protection. Mais contre quoi faut-il se prémunir ?

L'EMPIRE CONTRE ATTAQUE

Une grande partie des attaques se focalisent sur la réidentification, c'est-à-dire identifier l'identité associée à une donnée censée être anonymisée, dans un but, par exemple de profilage ou d'usurpation d'identité. La réidentification est souvent rendue possible par la présence dans les données de quasi-identifiants, en l'occurrence un ensemble d'attributs dont la combinaison des valeurs pour un individu est en général unique. Ainsi, l'association (code postal, genre et date de naissance) dans une base de données pseudonymisée peut être un quasi-identifiant conduisant, avec l'aide d'une autre base de données, comme une liste électorale, à l'identité d'un individu. Plusieurs affaires ont fait beaucoup de bruit dans les années 2000 et mis en lumière le peu de protection qu'assure la pseudonymisation : le gouverneur du Massachusetts a été identifié