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

dans des données médicales pseudonymisées par une simple comparaison avec, justement, une liste électorale !

Une autre forme d'attaque est l'inférence. Celle dite « d'attribut » consiste à déduire une information sensible à partir des données anonymisées. Se protéger de cette menace est important même dans les situations où le risque de réidentification est nul. Ainsi, déterminer l'origine ethnique d'une personne peut conduire à une discrimination pour l'obtention d'un prêt bancaire. Quant à l'inférence « d'appartenance », elle a pour objectif de décider si un individu particulier fait partie d'un ensemble de données sensibles, même anonymisées, comme une cohorte de patients atteints d'un cancer.

Récemment, ces attaques par inférence ont été adaptées au contexte des modèles d'apprentissage, du type de ceux utilisés en intelligence artificielle. Ce ne sont pas alors les données elles-mêmes qui sont visées, mais une structure, comme un réseau de neurones, qui a été élaborée à partir des données. Ici, l'objectif est de déterminer si les données d'un individu cible faisaient partie de celles qui ont servi à l'apprentissage, par exemple d'un système de classification des tumeurs en oncologie.

PREMIERS ÉCHECS

Motivés par les scandales de réidentification, les premiers modèles formels de vie privée ont cherché à contrecarrer les attaques dues à l'existence des quasi-identifiants. Les algorithmes d'anonymisation ainsi conçus partitionnent un jeu de données de sorte qu'un ensemble de contraintes soit satisfait dans chaque morceau. Le *k*-anonymat, par exemple, requiert que les quasi-identifiants d'une même partition soient identiques et que chaque partition contienne les données d'au moins *k* individus. Une attaque de réidentification à partir du quasi-identifiant devient donc imprécise, d'autant plus que la valeur de *k* est grande.

Autre modèle formel pionnier, la *l*-diversité nécessite que chaque partition contienne un ensemble de données sensibles suffisamment divers : au moins *l* données sensibles distinctes, ou bien que la valeur de la donnée sensible la plus fréquente ne soit pas beaucoup plus présente que les données sensibles les moins fréquentes. La *l*-diversité espère ainsi lutter contre des attaques d'inférence d'attribut qui associeraient simplement à un quasi-identifiant la donnée sensible la plus fréquente dans sa partition.

Cette famille de modèles est aujourd'hui sur le déclin, victime des failles et des limitations qui ont été découvertes progressivement. D'abord, ces modèles ne sont nativement pas composables. Ce qui signifie que des analyses successives sur un même jeu de données, certes anonymisées, peuvent néanmoins briser la protection. Ensuite, ces modèles ne sont pas assurés contre le post-traitement : les critères de protection ignorent les algorithmes qui anonymisent les jeux de données. Or le comportement de ces programmes est souvent déterministe (il fournit toujours le même résultat à partir de données identiques) et dépendant des données. Certaines attaques tirent parti de ces caractéristiques pour faire de la « rétro-ingénierie » afin de retrouver la base de données initiale à partir de sa version anonymisée.

BEAUCOUP DE BRUIT POUR RIEN ?

La situation de l'anonymisation des données a radicalement changé avec la confidentialité différentielle, inventée en 2006 par Cynthia Dwork, alors chez Microsoft, conjointement avec d'autres chercheurs, originellement pour contrer les attaques d'un autre type, celles dites de « reconstruction de base de données » : l'idée est de générer un grand nombre de requêtes statistiques sur des sous-ensembles variés de la base de données (nombre de patients qui ont la grippe, ceux dont le numéro de sécurité sociale commence par « 1 70 » et qui ont la grippe…), puis de résoudre un système d'équations