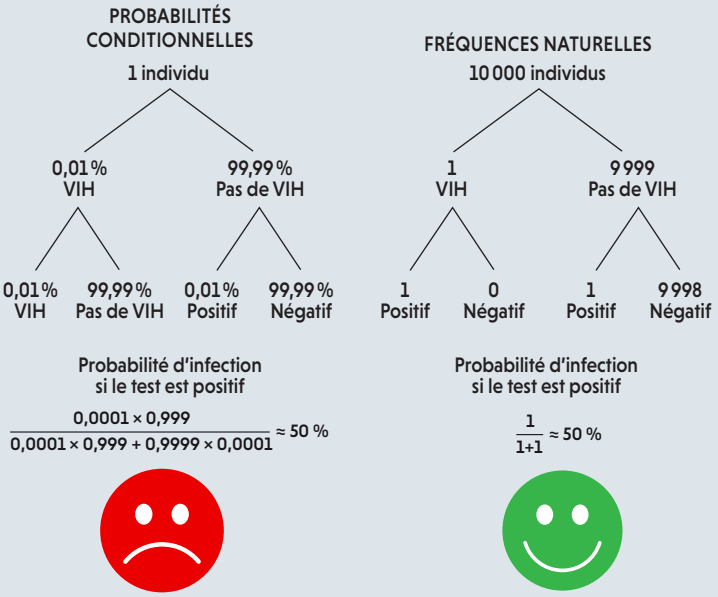
PRÉVOIR UNE INFECTION

Si votre test de dépistage du VIH est positif et que vous êtes un homme à faible risque d'infection (hétérosexuel, non drogué…), quel est le risque que vous soyez effectivement porteur du virus ? Les probabilités conditionnelles *(à gauche)* proposent un calcul compliqué. En revanche, en se fondant sur les fréquences naturelles *(à droite)*, on obtient facilement la réponse : sur 10 000 hommes, on s'attend à ce qu'un seul soit infecté et que son test soit donc positif ; parmi les 9 999 qui ne sont pas infectés, 1 devrait aussi avoir un résultat positif. En conséquence, on a deux tests positifs, mais un seul individu infecté. Donc la probabilité d'infection donnée par un test positif n'est pas de 100 %, mais de 50 %.



Les taux de mortalité sont des indicateurs plus fiables de la valeur des programmes de dépistage que les taux de survie à cinq ans. Si l'on se fie à ces taux, un homme doit-il faire systématiquement un dosage de l'antigène PSA ? Un fumeur doit-il passer systématiquement un scanner des poumons ? Il est vrai que ces deux examens détectent plus de cancers à un stade précoce ; mais aucun des deux ne permet de réduire la mortalité.

LES ÉPIDÉMIES DE DIAGNOSTICS

Les gens considèrent souvent les tests de dépistage comme des garants de leur santé. Toutefois, des examens supplémentaires peuvent conduire à des interventions médicales inutiles dont les effets sont parfois délétères. Et pour les nombreux patients inutilement diagnostiqués, le traitement a forcément des conséquences indésirables. Une épidémie de diagnostics peut être aussi dangereuse pour la santé que la maladie.

Les erreurs d'interprétation des statistiques seraient moins fréquentes si les chercheurs, les médecins et les médias utilisaient des données chiffrées directes au lieu de nombres qui prêtent à confusion : le risque absolu au lieu du risque relatif, les fréquences naturelles à la place des probabilités conditionnelles, et les taux de mortalité plutôt que les taux de survie à cinq ans. En outre, nous devons mieux éduquer les jeunes à la science du risque et de l'incertitude.

Comme le suggérait H. G. Wells, les statistiques devraient être enseignées en même temps que la lecture et l'écriture. De fait, aux États-Unis, l'Association nationale des enseignants de mathématiques insiste pour que l'enseignement des statistiques et des probabilités commence à l'école primaire. Si les enfants apprenaient que le monde n'est pas fait de certitudes et ce d'une façon ludique, les statistiques seraient mieux comprises.

Au lieu d'apprendre aux étudiants comment appliquer des formules de probabilité pour résoudre des problèmes virtuels, les professeurs devraient leur montrer comment utiliser les statistiques pour résoudre des problèmes concrets. Par exemple, ils pourraient leur enseigner l'usage des statistiques et des probabilités quand il s'agit de décider comment se comporter face aux drogues, à la consommation d'alcool, à la conduite automobile, aux biotechnologies et à d'autres questions importantes pour la vie quotidienne.

Un livre scolaire américain du secondaire raconte l'histoire vraie d'une mère célibataire de 26 ans. À la suite d'un test positif de dépistage du VIH, elle perd son travail, déménage dans un foyer hébergeant d'autres personnes séropositives, a des relations sexuelles non protégées avec l'un d'entre eux, et attrape une bronchite. Son médecin lui prescrit alors un nouveau test de dépistage. Le résultat est négatif, tout comme celui de son échantillon sanguin précédent, qui a été réanalysé. Cette femme a vécu un cauchemar parce que ses médecins n'ont pas compris qu'un résultat positif à ce test n'était pas un verdict définitif, mais qu'il signifiait que cette femme avait une probabilité d'être infectée de 50 %, étant donné son appartenance à un groupe à faible risque.

L'éducation statistique peut changer des vies, aider les gens à prendre de meilleures décisions personnelles, à reconnaître les messages trompeurs et à développer une attitude plus sereine envers leur santé. Comme le recommandait le philosophe Emmanuel Kant : « Osez savoir ! » ■

BIBLIOGRAPHIE

WOLOSHIN *ET AL.*, *Know your chances : Understanding health statistics*, University of California Press, 2008.

A. PLEASANT, Communiquer sur les statistiques et le risque, *SciDev Net*, 15 décembre 2008.

A. FAGERLIN *ET AL.*, Making numbers matter : present and future research in risk communication, *Am. J. of Health and Behav.*, vol. 31, pp. S47-S51, 2007.

G. GIGERENZER, *Calculated risk : how to know when numbers deceive you*, Simon & Schuster, 2003.

B. FALISSARD, *Comprendre et utiliser les statistiques dans les sciences de la vie*, Masson, Abrégés, 3e éd., 2005.

L'ESSENTIEL

- Cachés derrière un pseudonyme, nous laissons de nombreuses données personnelles sur Internet ou à divers organismes.
- C'est insuffisant pour garantir l'anonymat de ces données !
- Diverses techniques sont développées pour protéger les données sensibles sans empêcher leur exploitation.

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.

L'art de préserver L'ANONYMAT

Presque tous nos comportements laissent des empreintes numériques. Les informaticiens conçoivent et développent des techniques pour que l'analyse de ces gisements de données ne compromette pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information: toutes ces facettes de nos vies laissent des empreintes numériques qui sont stockées et classées dans des bases de données gigantesques, telles celles des moteurs de recherche, qui gardent l'historique des requêtes associées à une adresse IP (la carte d'identité d'un appareil connecté) pendant des années. Ces masses d'informations constituent une manne sans précédent pour les sociétés humaines. Dans le domaine de la santé, par exemple, on peut analyser les caractéristiques individuelles de chaque patient, afin de mieux le prendre en charge, ou de mener des études de santé publique. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements, la quantité et la diversité des informations aujourd'hui disponibles en multiplient l'intérêt potentiel… et les dangers.

En effet, ces myriades d'informations font planer une menace, elle aussi sans précédent, sur la vie privée des individus. La plupart du temps, les données ne sont pas publiques, mais elles circulent néanmoins entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre une partie des données personnelles de leurs utilisateurs… Certaines données sont même accessibles à tous sur Internet : par exemple, le site Netflix a publié des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand celui-ci ne contient ni son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec des préjudices multiples : discrimination au crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique… Protéger les données personnelles sans empêcher leur exploitation est devenu un enjeu majeur à l'ère du tout-numérique. La diffusion de ces données est un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus.

UN BESOIN D'ÉQUILIBRE

Le souci de la protection des données s'est accru pendant la seconde moitié du XX^e^ siècle, à mesure que l'informatique se généralisait. Il a fait naître un domaine de recherche spécifique dans les années 1970, visant à garantir un anonymat plus ou moins complet aux titulaires des données. Le phénomène s'est encore accéléré depuis le début des années 2000, avec l'essor d'Internet, qui permet de consulter et de croiser instantanément ou presque de multiples sources d'information (réseaux sociaux, annuaires…). Lors de la publication de données, il est essentiel de prendre en compte les connaissances annexes accessibles à un 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. >

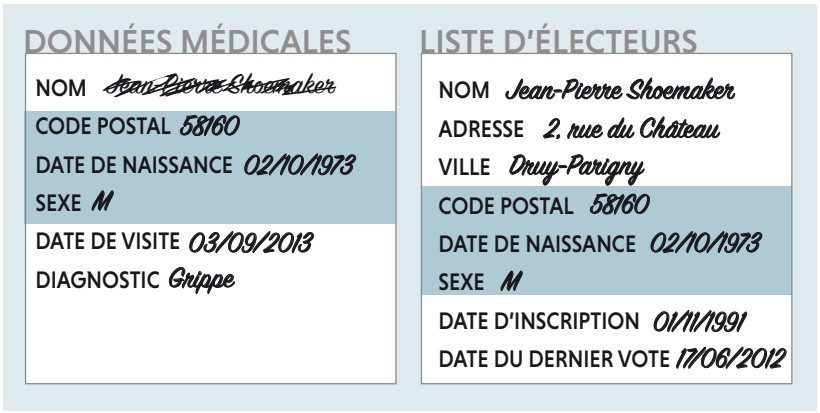
> Comment rendre anonyme un jeu de données? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650000 utilisateurs sur une période de trois mois (pour les rendre disponibles aux chercheurs). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots-clés des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont réussi à identifier la personne portant le pseudonyme 4417749 en croisant les mots-clés de ses requêtes avec des données disponibles sur le Web.

ASSAINIR LES DONNÉES

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 ci-dessus*). Pourtant, la pseudonymisation reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

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. En outre, la combinaison des renseignements correspondants est unique pour la plupart des gens, selon plusieurs études récentes.

Pour empêcher les réidentifications, Latanya Sweeney a proposé de scinder les champs du jeu de données en deux catégories: les champs quasi identifiants (notés QID), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD), 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 $1/k$.



Le titulaire d'un dossier médical dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données. En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(en rouge)* est unique. Il suffit alors de trouver un autre jeu de données, par exemple une liste électorale *(à droite)*, qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière.

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. Il serait simple d'élaborer un algorithme qui parcoure 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 102*).

Cependant, le *k*-anonymat ne résout pas tous les problèmes. Considérons la situation suivante: un attaquant dispose d'un jeu de données *k*-anonymes 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 un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer: la valeur sensible n'est pas protégée.

BROUILLER LES PISTES

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.

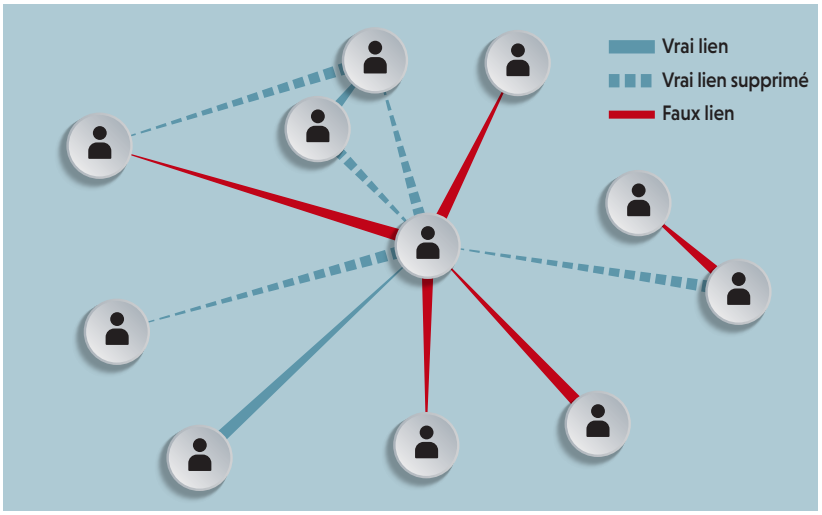
Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'université Cornell, à New

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 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

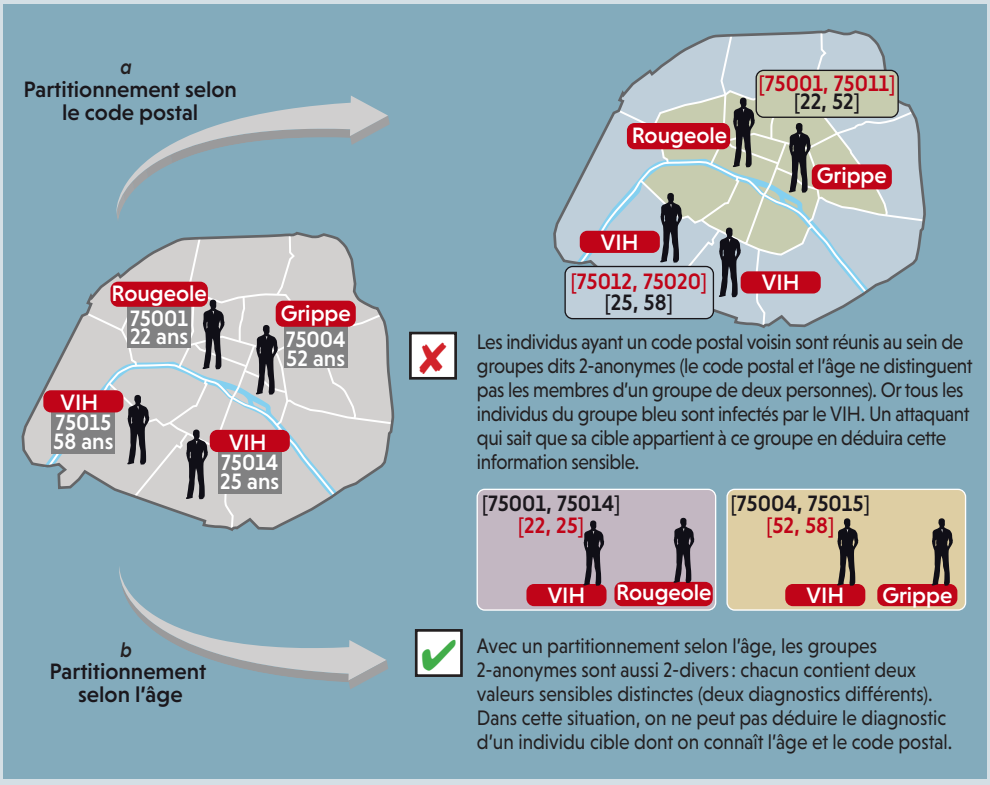
L'assainissement consiste alors à perturber les données de sorte que celles de chaque individu se trouvent noyées dans la perturbation (et non plus dans la foule, comme dans le k-anonymat). On introduit, par exemple, un grand nombre de fausses données (typiquement 100 fois plus que de vraies), afin que 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

NOZHA BOUJEMAA

« Les algorithmes sont-ils transparents et éthiques ? Pour nous en assurer, nous avons besoin d'outils adaptés. »

Un algorithme peut-il être juste ?

Nozha Boujemaa : Pour répondre, un référentiel est indispensable. C'est précisément une des difficultés que l'on rencontre lorsqu'on s'intéresse à la qualité des algorithmes. La question à se poser est de savoir : « Pour qui ? ». La notion d'algorithme juste, ou loyal (d'une façon générale, on parlera de transparence), est différente selon que l'on se place du point de vue du consommateur du service adossé à l'algorithme ou du producteur de ce service.

Le consommateur individuel est un citoyen, et de fait les questions de transparence algorithmique ont souvent été soulevées par des affaires de libertés individuelles et de droits civiques. Quand l'utilisateur est une entreprise ou bien même les services de l'État, les problèmes possibles sont ceux, d'une part, de la concurrence déloyale, des abus de position dominante et, d'autre part, de souveraineté nationale.

BIO EXPRESS

17 JUILLET 1963
Naissance

2010-2015
directrice du centre de recherche Inria Saclay Ile-de-France.

2017
Directrice du projet *TransAlgo*, de l'Inria et de l'institut Dataia.

La question de la transparence est essentiellement celle de l'asymétrie informationnelle, entre celui qui propose un service logiciel et celui qui l'utilise, le second en sachant beaucoup moins que le premier. La réduction de cette asymétrie est cruciale : on doit savoir comment nos données sont calculées et par où elles transitent. Notons que la discrimination ou les biais potentiels ne sont pas toujours intentionnels.

Ces problématiques sont le cœur du projet *TransAlgo* que vous coordonnez. Comment est-il né ?

Nozha Boujemaa : Cette initiative fait suite à la loi pour une République numérique (promulguée le 7 octobre 2016) proposée par l'ancienne secrétaire d'État au numérique Axelle Lemaire, ainsi qu'au rapport commandé au Conseil général de l'économie (CGE) sur la gouvernance des algorithmes. La loi prévoit