Comment protéger la vie privée

Prénom Nom auteur

Presque tous nos comportements laissent des empreintes numériques. De nouvelles techniques sont développées pour que l'analyse de ces nouveaux gisements de données ne compromettent pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information : tous ces comportements et aspects de nos vies laissent des empreintes numériques, sur Internet ou sur divers appareils. Elles 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 pendant plusieurs années. Cette manne d'information représente une opportunité 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 réaliser des études globales de santé publique : l'analyse du nombre de fois où le mot « grippe » est tapé dans un moteur de recherche dans une certaine région (reconstituée approximativement avec l'adresse IP) renseigne ainsi sur l'importance locale de l'épidémie. De même, les données de mobilité pourraient être enregistrées par les GPS et aider à optimiser l'aménagement urbain. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements démographiques, la quantité et la diversité des informations disponibles aujourd'hui en multiplient l'intérêt potentiel.

Cette manne d'information fait cependant 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 tout de même entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre (légalement ou non) une partie des données personnelles de leurs utilisateurs, etc. Certaines données sont même accessibles à tous sur Internet : le moteur de recherche de Facebook (uniquement disponible en anglais pour l'instant) permet par exemple de trouver les utilisateurs qui aiment tel sport, ont tel âge, habitent telle région... Autre exemple, le site Netflix, qui recommande des films en fonction des goûts de chacun, a publié un certain nombre de données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui renvoient les films recommandés.

Un compromis entre utilité et protection

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne concernée par un jeu de données, même quand il ne contient pas ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec des préjudices multiples : discrimination au

© Pour la Science - n° 000 - Mois 2012

crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique, etc. Protéger les données personnelles sans les rendre inutilisables pour les analystes et les utilisateurs autorisés est donc un enjeu majeur à l'ère du tout numérique. Leur diffusion s'avère un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus. C'est cet « art » que nous examinerons ici.

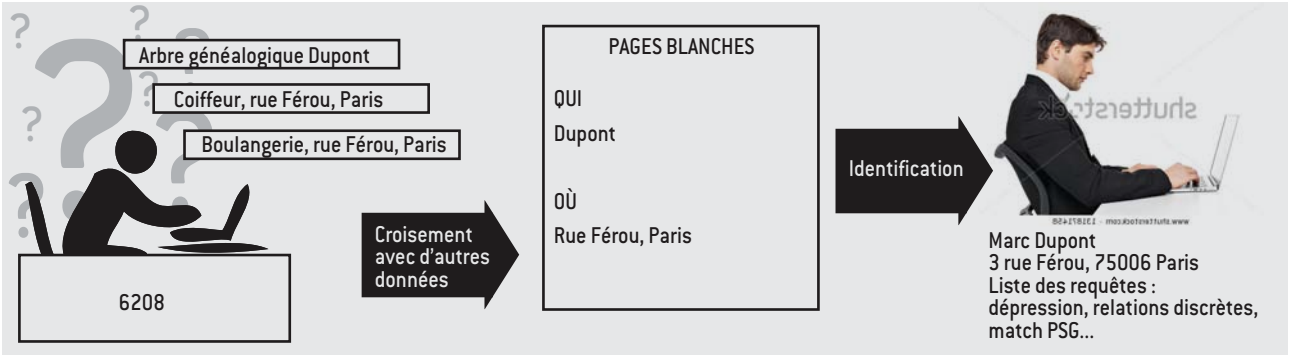
Le souci de la protection des données s'est accru pendant la deuxième moitié du XX^{e} siècle, à mesure que l'informatique se répandait. Il a entrainé la naissance d'un domaine de recherche spécifique dans les années 1970, afin de garantir un anonymat plus ou moins complet aux titulaires des données. Un ensemble de statistiques sur des données personnelles peut révéler beaucoup d'informations sur un individu précis : par exemple, si un recensement liste le revenu moyen par habitant avec un découpage du territoire en carrés de 200 mètres de côté, le propriétaire d'un grand domaine risque d'être seul dans la bande de territoire considérée ; un individu qui le sait peut donc déduire son revenu de l'analyse des statistiques. On voit ici que la quantité de renseignements extractible de données publiées dépend de ce que l'on sait par ailleurs. Les ordinateurs ayant rendu les informations disponibles plus nombreuses et plus accessibles, le danger s'est multiplié.

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 de multiples sources d'information (réseaux sociaux, annuaires, listes de diplômés, etc.). Lors de la publication de données, il est devenu crucial de prendre en compte les connaissances annexes

L'ESSENTIEL

- **De nombreuses données personnelles sont publiées sur Internet ou transmises à divers organismes en indiquant leurs titulaires par un pseudonyme.**
- **Cela ne suffit pas à garantir l'anonymat des données, d'où une menace sur la vie privée.**
- **On développe alors de multiples techniques pour protéger les données sensibles sans les rendre inexploitables.**

Brouillon, à refaire !!!!



LES requêtes tapées dans un moteur de recherches renseignent sur l'identité de celui qui les soumet. Le moteur de recherche AOL en a publié plusieurs millions en 2006, en identifiant leurs émetteurs par des pseudonymes (des nombres). En croisant ces requêtes avec des données disponibles sur Internet, des journalistes ont retrouvé plusieurs personnes. Ils ont alors eu accès à un pan essentiel de leur vie privée, puisqu'ils disposaient de toutes les requêtes associées à leur pseudonyme.

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 de leur propriétaire. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée afin de réduire la probabilité de réidentification de leur propriétaire ou d'accès à des informations sensibles le concernant. .

L'échec de la pseudonymisation

Comment rendre un jeu de données anonymes ? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de son propriétaire par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisé l'entreprise AOL, en août 2006, lorsqu'elle a mis 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 d'utilisateur étaient remplacés par des nombres aléatoires, mais les mots-clés des recherches étaient laissés tels quels. Quelques jours plus tard, des journalistes du New York Times ont identifié plusieurs utilisateurs, tel celui portant le pseudonyme « 4417749 ». Ce dernier 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 clés avec des données disponibles sur le Web et ont peu à peu reconstitué l'identité probable de la propriétaire des données, 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 publication de données pseudonymisées reste en pratique autorisée par la loi, sans doute car ses failles sont mal comprises par les législateurs, qui considèrent de telles données comme anonymes.

Prévenir les croisements de données

Au début des années 2000, Latanya Sweeney, de l'Université de 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 souvent 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 3)*. En outre, la combinaison des renseignements correspondants est unique pour la majorité des gens : selon une étude récente, la proportion est de 80 pour

LES AUTEURS

Prénom NOMVXCV auteur textebvcx de biographiexbvv, odiasi ut que vbvel idemped qubcxxis voloreptbvatis si

PRÉNOM NOMVXCV auteur textebvcx de biographiexbvv, odiasi ut que vbvel idemped qubcxxis voloreptbvatis si

Prénom NOMVXCV auteur textebvcx de biographiexbvv, odiasi ut que vbvel idemped qubcxxis voloreptbvatis si

© Pour la Science - n° 000 - Mois 2012

cent aux États-Unis ; elle est probablement voisine en France.

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, d'une part, et les champs sensibles (SD, pour *Sensitive data*), qui doivent rester privés (par exemple un diagnostic), d'autre part. 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. 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 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'âge [15-20] 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 parcourt 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. L'exemple le plus connu est l'algorithme de Mondrian *(voir l'encadré page xx)*.

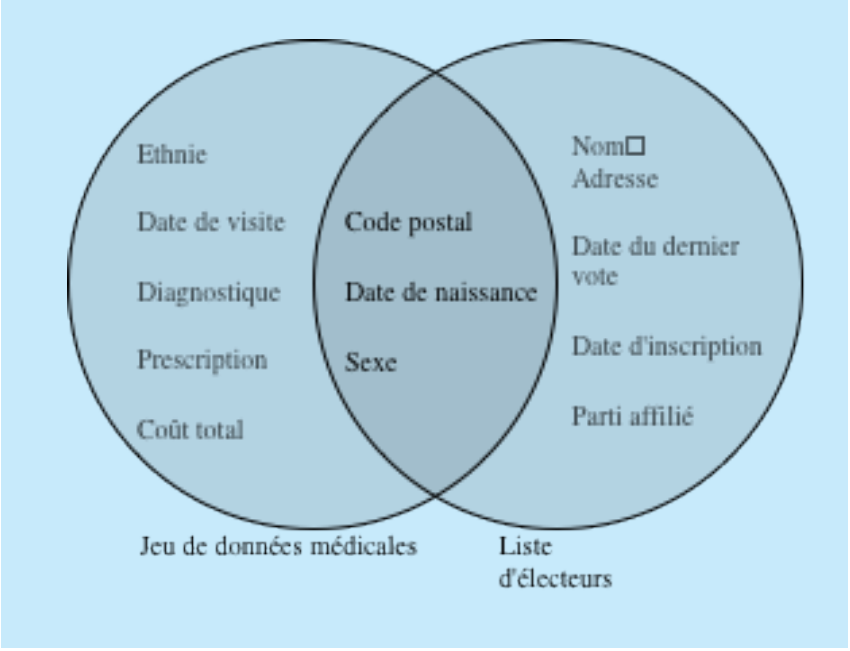
Cependant, le k-Anonymat est loin de résoudre tous les problèmes. Considérons la situation suivante : un attaquant dispose d'un jeu de données k-Anonyme et recherche la valeur sensible (tel le diagnostic médical) d'un individu cible dont il connaît le quasi-identifiant. Il peut retrouver le groupe auquel 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 a 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 ayant 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 qui ?], 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 : s'il 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 ; de même, on quantifie la connaissance résultante de sa cible, dite *a posteriori*, 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 : la probabilité de déduire telle information d'un jeu de données sachant qu'on a 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 imposer une valeur maximale à 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 – ceux-ci pouvant même avoir des connaissances a priori différentes. La l-Diversité, conçue la même année [par qui ?], est plus applicable. Elle consiste à imposer *l* valeurs sensibles distinctes dans chaque groupe (voir l'encadré page xx). Sans autre connaissance de sa cible, un attaquant ne peut trouver sa valeur sen-



LE titulaire d'un dossier médical (pseudonymisé et publié sur Internet) peut souvent être retrouvé en croisant plusieurs jeux de données. En effet, la plupart du temps, la combinaison {Code Postal, Date de naissance, Sexe} est unique. Il suffit alors de trouver un autre jeu de données qui la contient et qui renferme aussi les coordonnées précises de la personne (telle une fiche de liste électorale, disponible sur simple demande à la mairie) pour identifier cette dernière.

sible avec une probabilité supérieure à 1/*l*.

De multiples autres techniques ont été élaborées au cours des années suivantes, 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. 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 hopital, 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, qui construit 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 xx). Tous ces modèles considèrent qu'un jeu de données est d'autant plus confidentiel qu'il renseigne peu l'attaquant sur sa cible. Volontairement ou non, ce sont des instances du paradigme de non-information formulé en 1977 par le statisticien suédois Tore Dalenius dans le contexte d'assainissement de statistiques [pourriez-vous énoncer ce paradigme et préciser son application dans ce contexte ? cf mail].

Une vision alternative de la confidentialité

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 *(voir la figure 4)*. L'assainissement consiste à perturber les données de telle sorte que la contribution de chaque individu se trouve noyée dans la perturbation introduite (et non plus dans la foule, comme dans le k-Anonymat) : on introduit par exemple un grand nombre de fausses données (typiquement cent 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

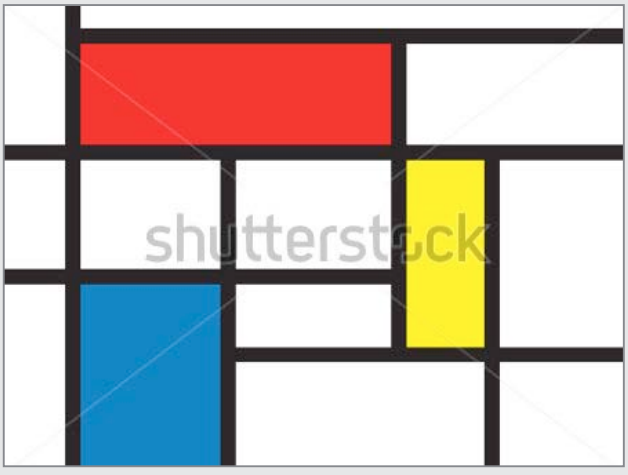
COMMENT RENDRE ANONYME UN JEU DE DONNÉES ?

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. Cette technique n'empêche pas les réidentifications ultérieures via des croisements de données. D'autres techniques ont alors été développées. Le k-anonymat est la plus employée aujourd'hui. Il consiste à brouiller certains champs, dits quasi-identifiants car leur combinaison est parfois unique : l'âge, le sexe, le code postal… On remplace leurs valeurs précises par des intervalles ou des catégories, de façon à ce que 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 est une référence à Piet Mondrian (1872-1944), un peintre hollandais dont les œuvres partitionnent l'espace (ci-contre, une peinture inspirée de 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 des quasi-identifiants, selon lequel il divise les individus en deux groupes : par exemple, ceux dont le code postal est supérieur ou égal à 75003, et les autres. Les deux groupes résultants sont de nouveau divisés en deux (en utilisant le code postal ou un autre quasi identifiant, tel l'âge) 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 chaque individu par l'ensemble de valeurs couvert par son groupe. Chaque individu est alors dit k-anonyme, c'est-à-dire qu'on ne peut le distinguer des k personnes de son groupe.

Le problème est que les données personnelles 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 (les valeurs sensibles étant les données à protéger). 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 l'attaquant est fort, plus l sera élevé, plus il faudra inclure d'individus dans chaque groupe, et moins les statistiques calculées à partir du jeu de données assaini seront fines. Le compromis entre utilité et protection est ici réalisé à travers le choix de l.

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.

© Pour la Science - n° 000 - Mois 2012

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.

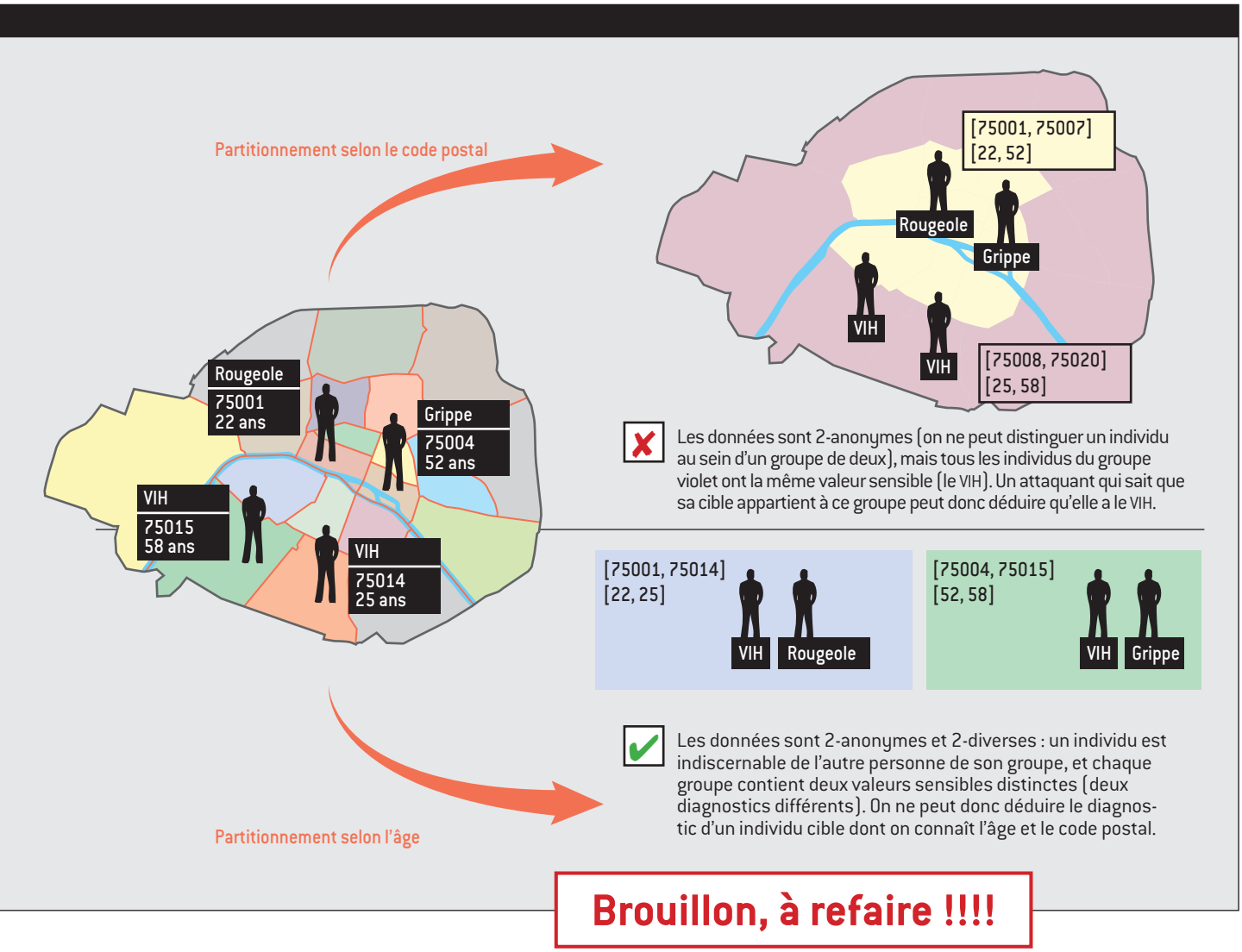
Les fausses données sont élaborées de façon à ce qu'on puisse toujours extraire certaines informations. Prenons l'exemple d'un jeu de cent dossiers médicaux, à partir duquel on souhaite savoir combien de personnes ont la grippe. On perturbe ce jeu de données en y introduisant mille faux dossiers, chacun se voyant attribué de façon équiprobable une maladie parmi quatre (dont la grippe). On sait alors qu'on a environ 250 valeurs fausses pour chacune de ces maladies. En comptant le nombre de grippes dans le jeu de données assaini et en retirant 250 à ce nombre, on obtient le nombre réel de grippe avec une bonne précision. Plus les vraies données sont nombreuses par rapport aux fausses, plus les informations extraites sont précises, mais moins les données sont protégées. Il existe de multiples façons de créer des fausses données, et les algorithmes qui s'en chargent font l'objet de multiples études.

La confidentialité différentielle est en plein essor. Elle est désormais applicable à des types de données variées, telles 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 telles que le nombre de liens sont toujours extractibles du graphe, sans que l'on puisse déterminer les connections précises d'un individu.

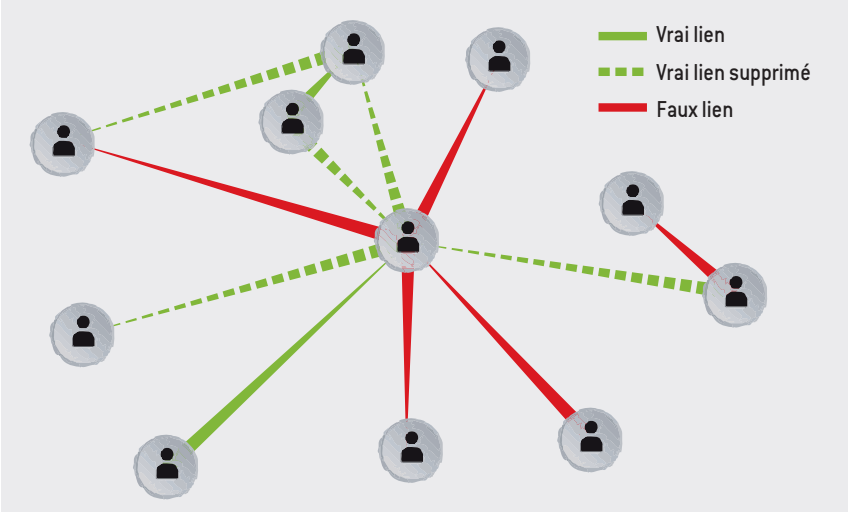
Le succès de la confidentialité différentielle s'explique en partie par sa simplicité : l'attaquant n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, n'est pas nécessaire. Cependant, des travaux récents [auteurs, labos et date] ont montré que le niveau de protection maximal atteignable par cette technique est plus faible, du fait de la non prise en compte des connaissances annexes de l'attaquant et des relations potentielles entre les individus d'un jeu de données.

Un modèle d'assainissement universel 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



Brouillon, à refaire !!!!



UN graphe de réseau social représente 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. Il est alors impossible de savoir avec précision à qui est connecté un individu, mais on peut toujours calculer des valeurs telles que la connectivité moyenne. On voit ici que les méthodes de perturbation des données doivent être choisies en fonction des informations que l'on souhaite pouvoir extraire.

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. Les dossiers médicaux électroniques français et anglais, notamment, fonctionnent ainsi. Un tel processus facilite la production de multiples versions assainies d'un même jeu de données, chacune étant adaptée aux besoins du destinataire et limitée aux usages prévus par la législation. Ainsi, les épidémiologistes peuvent recevoir des jeux de données plus précis que l'industrie pharmaceutique.

Toutefois, ce principe d'assainissement centralisé nécessite une fiabilité totale du gestionnaire de l'entrepôt de données. Or on a constaté de nombreuses fuites d'information, dues à des négligences ou à des attaques internes ou externes (L'*Open Security Foundation*, organisation non gouvernementale américaine, recense les cas les plus significatifs sur son site Internet DataLossDB.org). 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.

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, les protégeant par conséquent avant tout enregistrement. Illustrons ce principe dans le cas d'un sondage politique. On demande aux sondés de répondre par oui ou par non à une enquête d'opinion, en disant la vérité ou un mensonge avec une probabilité connue (par exemple en suivant le résultat d'un tirage à pile ou face). Comme on connaît le pourcentage de bonnes réponses, on peut faire des calculs statistiques sur ces données, mais on ne peut reconstituer l'opinion d'un individu précis.

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 : à protection équivalente, les données sont moins précises, et à précision égale, les données sont moins protégées. 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 – une piste étudiée par plusieurs laboratoires. En d'autres termes, la perturbation à appliquer n'est pas décidée localement (comme dans notre exemple précédent, où chaque sondé décide de la véracité de sa réponse en jouant à pile ou face), mais par une entité centrale. Celle-ci possède des informations sur toutes les réponses, sans connaître les réponses elles-mêmes. Dans le cas du sondage d'opinion, une telle entité saurait qui a dit la vérité, mais ne connaîtrait pas les réponses des sondés. En pratique, l'entité centrale stocke tout de même les réponses, mais sous une forme cryptée.

Des algorithmes regroupés sous le nom de *Secure Multi-Party Computation* (littéralement calcul multipartite sécurisé), qui font intervenir des techniques de cryptographie, visent à assurer la confidentialité de l'assainissement distribué (où les calculs sont répartis entre plusieurs acteurs). Par exemple, si quatre personnes possèdent chacune un nombre qui doit rester privé, mais dont on souhaite que la somme soit calculable, on peut appliquer l'algorithme suivant : la première personne tire un chiffre aléatoire, auquel elle ajoute son nombre, avant de transmettre le résultat à la deuxième personne ; celle-ci y ajoute son nombre, transmet le résultat, et ainsi de suite jusqu'à ce que le

■ BIBLIOGRAPHIE

T. Allard, B. Nguyen, et P. Pucheral, « MET P: revisiting Privacy-Preserving Data Publishing using secure devices », Distributed and Parallel Databases, p. 1-54, 2013. (To appear.)

B.-C. Chen, D. Kifer, K. LeFevre, et A. Machanavajjhala, « Privacy-Preserving Data Publishing », Found. Trends databases, vol. 2, no 1-2, p. 1–167, janv. 2009.

C. Dwork, « Differential privacy », in Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II, Berlin, Heidelberg, 2006, p. 1–12.

Adam Greenfield, Everyware : La révolution de l'ubimédia

L. Sweeney, « k-anonymity: a model for protecting privacy », Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 10, no 5, p. 557–570, oct. 2002.

© Pour la Science - n° 000 - Mois 2012

résultat final revienne à la première personne ; cette dernière retranche le chiffre aléatoire qu'elle avait tiré et dispose alors de la somme des nombres, sans qu'aucun des participants n'ait découvert le nombre des autres. En effet, chacun a transmis sa donnée sous une forme perturbée, mais avec un paramètre commun (le nombre aléatoire tiré au début).

Les mécanismes distribués constituent un pas majeur vers la sécurisation du processus d'assainissement, mais leur mise en œuvre pose encore aujourd'hui des problèmes de passage à l'échelle, car ils nécessitent des calculs importants. Ils sont pour l'instant relégués au traitement de jeux de données de faible taille.

Avec notre équipe, nous nous sommes attaqués à ce problème. Nous avons développé des architectures de gestion des données fondées sur un ensemble de dispositifs personnels (smartphone, tablette, etc.) fonctionnant en réseau. Ces dispositifs, de plus en plus répandus, ont une grande capacité de stockage (plusieurs gigaoctets) et peuvent être dotés de processeurs sécurisés. En pratique, les données seraient stockées localement sur les appareils de l'utilisateur et ne seraient jamais exportées sous une forme non perturbée. Une telle architecture pourrait par exemple assurer la gestion des dossiers médicaux ou des factures. Aucun serveur ne les regrouperait tous, mais les échanges entre une entité centrale et les appareils des utilisateurs permettraient tout de même de collecter certaines informations, d'une façon respectueuse de la vie privée. Nous avons montré que cette architecture assurerait un assainissement de même qualité qu'en environnement centralisé.

Des techniques trop peu appliquées

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, disons neuf fois sur dix, 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 se partagent les miettes. Ces techniques doivent donc continuer à se répandre et se perfectionner.

Pour autant, la science de l'assainissement reste la recherche du meilleur compromis entre utilité et confidentialité des données, dont la protection absolue est une chimère : on ne peut garantir qu'un attaquant ne tirera aucune information sensible des données que par leur destruction définitive (d'ailleurs compliquée, en raison de leur présence sur de nombreux serveurs). 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. Nous devrons y apporter des réponses multidisciplinaires, impliquant informaticiens, juristes et analystes.

Figure 1 (Ouverture) : Shutter 133296158 ou 111979016 (en enlevant l'œil sur la main), ou Corbis 42-24299778

Figure 2 :

Figure 3 :
© L. Sweeney, Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 2002

© Pour la Science - n° 000 - Mois 2012