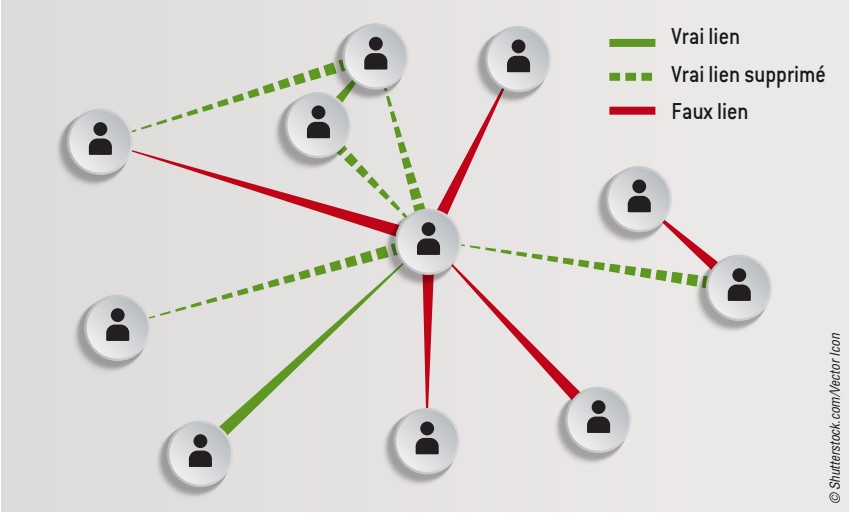

3. LES GRAPHES DE RÉSEAU SOCIAL représentent les individus par des points et leurs liens par des traits. Pour qu'ils ne dévoilent pas la vie privée des individus, on peut leur appliquer un algorithme dit de confidentialité différentielle, consistant, par exemple, à supprimer de nombreux vrais liens et à les remplacer par autant de faux liens. 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.

les épidémiologistes peuvent 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 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, ce qui les protége 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 réponses sincères, 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 si un sondé a dit la vérité ou non, mais ne connaîtrait pas sa réponse. En pratique, l'entité centrale stocke tout de même les réponses, mais sous une forme chiffrée.

Aucune entité centrale ne doit tout savoir

Des algorithmes regroupés sous le nom de *Secure Multi-Party Computation* (littéralement, calcul multipartite sécurisé), 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). 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 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

■ BIBLIOGRAPHIE

T. Allard *et al.*, **MET$_A$P: 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.

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 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.

Avec notre équipe, nous nous sommes attaqués à ce problème. Nous avons développé une architecture de gestion de données personnelles fondée sur des dispositifs individuels sécurisés (telles des clefs USB dotées de processeurs sécurisés, insérables dans un smartphone, une tablette ou un ordinateur). 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, d'une façon respectueuse de 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, la science de l'assainissement reste la recherche du meilleur compromis entre utilité et confidentialité des données, dont la protection absolue est illusoire. Seule la destruction des données peut garantir qu'un attaquant n'en tirera aucune information sensible (cette destruction est 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. Il faudra y apporter des réponses multidisciplinaires, impliquant informaticiens, juristes et analystes. ■