On sait alors qu'on a environ 250 valeurs fausses pour chacune de ces maladies. En comptant le nombre de grippés dans le jeu de données assaini et en lui soustrayant 250, on obtient le nombre réel de cas 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 nombreuses façons de créer des fausses données, et les algorithmes qui s'en chargent font l'objet de multiples études.

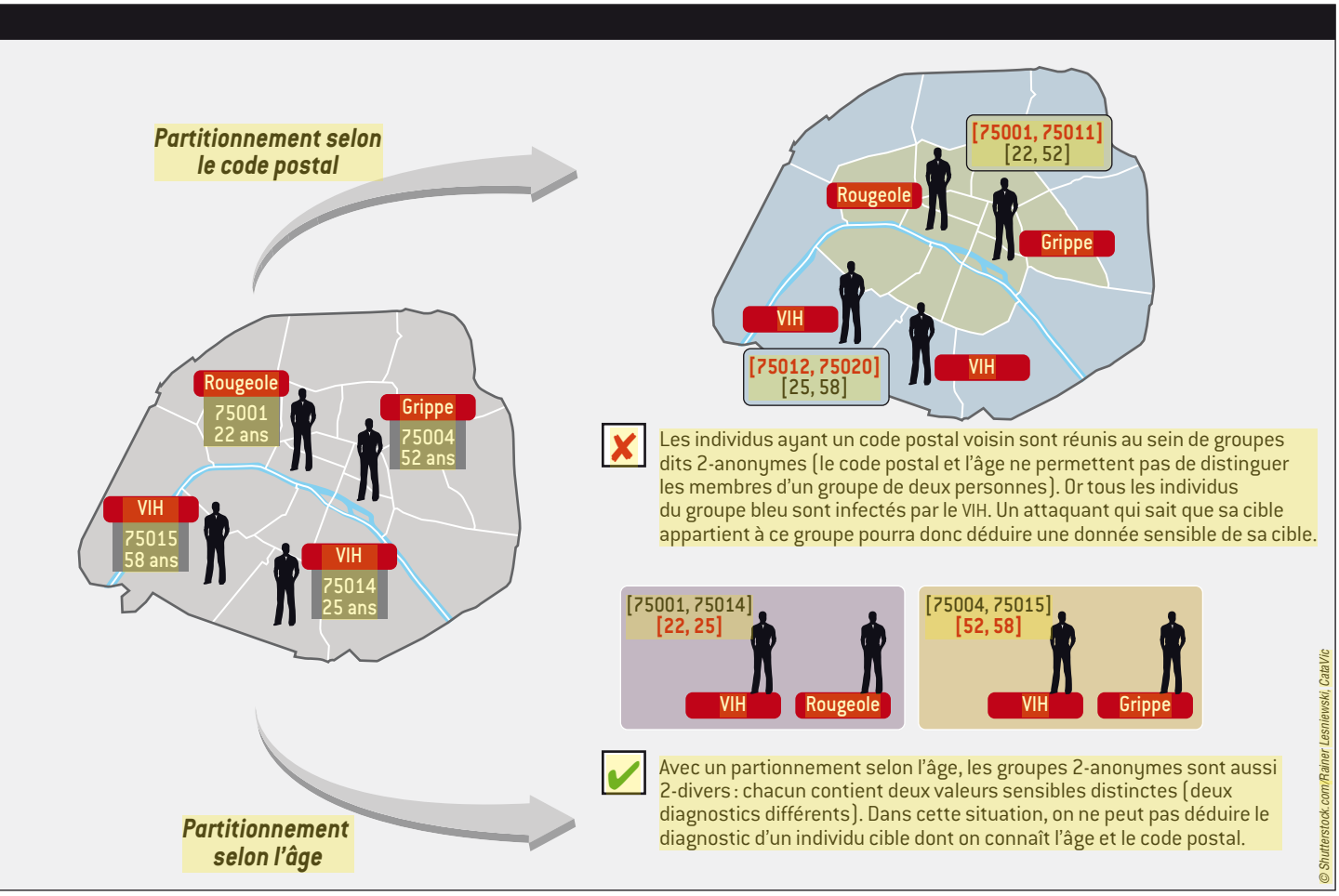
L'idée de la confidentialité différentielle est en plein essor. Elle est désormais applicable à des types variés de données, tels les graphes de réseaux sociaux *(voir la figure 3)*. Ces graphes représentent les individus par des nœuds, connectés par des liens. La perturbation peut alors consister à 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 connexions précises d'un individu.

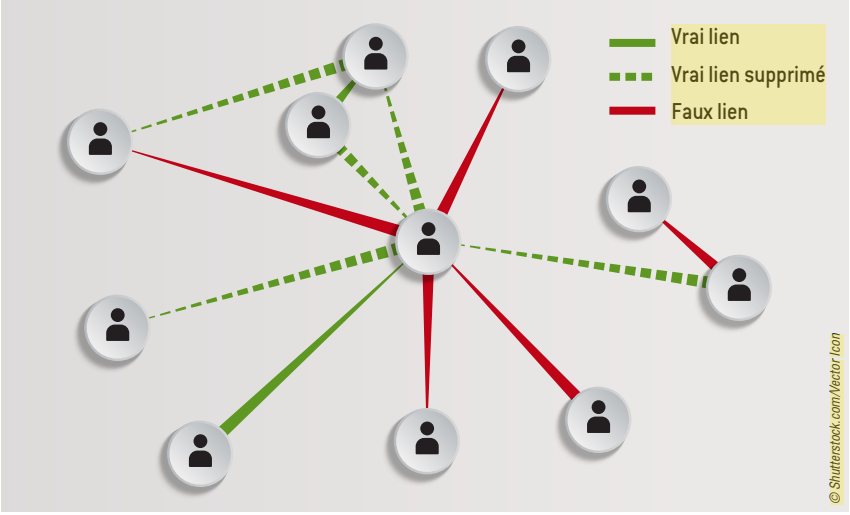
Le succès de la confidentialité différentielle s'explique en partie par sa simplicité : l'attaquant n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, n'est pas nécessaire. Mais cette simplicité est à double tranchant : dans des travaux publiés en 2011, A. Machanavajjhala et ses collègues 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 étant plus ou moins efficace selon la situation.

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, le système de dossier médical personnel national fonctionne 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,





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.