modélisant les informations obtenues. Avec un nombre de requêtes suffisant, l'attaque est ravageuse, car on reconstruit toute la base de données : ce n'est plus la simple identification d'un individu unique, même s'il est gouverneur, mais par exemple l'ensemble d'un fichier de patients.

Schématiquement, la confidentialité différentielle consiste à cacher la contribution de tout individu possible au résultat d'une requête en ajoutant un bruit aléatoire aux données. En d'autres termes, le résultat d'une analyse sera sensiblement le même que l'on retire les données privées d'un individu particulier de l'ensemble des données ou qu'on les y inclut *(voir la figure ci-contre)*. De la sorte, tout individu devient « inrepérable ». L'amplitude du bruit ajouté dépend d'un paramètre quantifiant le « niveau de vie privée » (noté ε) et d'un paramètre de « sensibilité de la requête » quantifiant l'impact d'un individu sur les résultats de la requête. La valeur du paramètre ε est fixée par l'administrateur. Il s'agit d'un compromis entre, d'une part, la précision des résultats obtenus, en un mot l'utilité des données, et, d'autre part, la protection de la vie privée.

Ainsi, pour une requête sur la somme des salaires d'un nombre d'individus dans une base de données, par exemple l'ensemble du personnel d'une université, la confidentialité différentielle ajoute un nombre d'euros déterminé aléatoirement à partir d'une loi de statistiques (comme une distribution de Laplace) et proportionnel à la contribution d'un individu. De la sorte, impossible de connaître le salaire de tel ou tel membre du campus. Et ce sera d'autant plus le cas que la valeur du paramètre ε, celui qui contrôle le niveau de vie privée, sera élevée.

UN BUDGET DE VIE PRIVÉE

La confidentialité différentielle présente deux avantages cruciaux : elle est composable et résiste à un post-traitement, c'est-à-dire là où le bât blessait avec les modèles précédents. Toutefois, il a été montré au début des années 2000 que, même bruitées, on peut finir par reconstituer une base de données pour peu que l'on dispose de beaucoup d'informations. Pour contrer cet obstacle, et c'est un changement d'habitude à acquérir, la confidentialité différentielle impose une borne sur la quantité d'informations totale révélée par plusieurs calculs effectués sur les mêmes données. La conséquence en est que l'on peut voir dans le paramètre ε une sorte de budget de vie privée. À mesure que les requêtes se suivent, les valeurs de ε de chacune s'additionnent tandis que le « niveau de vie privée garanti » se dégrade. Aussi, un ε total, une « somme à ne pas dépasser », est-il fixé par l'administrateur.

La seule façon de poursuivre au-delà est d'utiliser une autre base de données. C'est possible, par exemple, en subdivisant un groupe initial de 100 000 individus en quatre parties égales. Cependant, cette méthode oblige à réfléchir à la construction des bases de données pour éviter les biais.

Si la confidentialité différentielle originale est aujourd'hui la référence, plus de 200 variantes ont depuis été proposées, chacune prenant en compte différents besoins, modèles d'attaquant ou types de données. En voici quelques-unes.

Une variante populaire est la (ε, δ)-confidentialité différentielle qui consiste à autoriser le mécanisme à moins bien cacher les contributions (proportionnellement à δ). Cet assouplissement de la définition originale autorise à introduire moins de bruit, et donc à faire des économies sur le budget de vie privée. Dans une seconde variante, la confidentialité différentielle personnalisée, chaque individu définit son exigence en matière de vie privée, par exemple par le biais d'un formulaire de consentement, plutôt que d'en imposer une globale. Cependant, l'hétérogénéité qui en résulte complique l'exploitation de la base de données.