(subventions des bibliothèques). De plus, la pression financière exercée sur les bibliothèques menace la diffusion des résultats.

Le passage au numérique a permis de s'affranchir des contraintes de la distribution papier. L'appropriation forte et rapide de ce nouveau média par certaines communautés de chercheurs a favorisé l'émergence de réalisations qui ont bouleversé l'accès à la littérature scientifique.

## De *ArXiv* au libre accès

Les physiciens ayant besoin de partager vite les résultats de leurs recherches, l'un d'eux, Paul Ginsparg, a créé en 1991, au Laboratoire américain Los Alamos de l'Université de Californie, aux États-Unis, le premier serveur d'archives ouvertes, *hep-th (High Energy Physics - Theory)*, aujourd'hui connu sous le nom d'*ArXiv* (prononcé *arkaïv*). Ce serveur, sur lequel les chercheurs déposent leurs articles (publiés ou non) sous la forme de fichiers, a servi de modèle à d'autres projets répondant à la définition de ce que, un peu plus tard, on a nommé l'*Open Access*, le libre accès aux publications scientifiques.

Ce nouveau concept a été formalisé en 2002 par une déclaration, la *Budapest Open Access Initiative* (BOAI), qui définit deux stratégies pour mettre en accès libre les publications des chercheurs : la voie verte, c'est-à-dire la mise en ligne des écrits sur le Web par leurs auteurs, et la voie dorée – la publication dans des revues électroniques en accès ouvert. Dès 2003, les grandes institutions de la recherche et de l'enseignement supérieur s'engagent à soutenir les politiques d'accès ouvert, notamment en incitant leurs chercheurs à déposer leurs publications sur des serveurs d'archives ouvertes au sein des bibliothèques universitaires et des institutions. En juillet 2012, la Commission européenne a demandé aux États membres de définir une politique intitulée « Pour un meilleur accès aux informations scientifiques : dynamiser les avantages des investissements publics dans le domaine de la recherche ».

En France, la voie dorée est représentée notamment par le Centre pour l'édition électronique ouverte (CLEO), qui gère Openedition, un portail de revues en sciences humaines et sociales. Pour la voie verte, le CNRS a créé en 2000 le Centre pour la communication scientifique directe (CCSD), dont la principale mission est la gestion de l'archive ouverte HAL (Hyper articles en ligne) interconnectée aux grands serveurs internationaux tel *Pubmed Central*, de l'Institut américain de la santé, proposant de nombreuses archives en sciences de la vie.

Contrairement à *ArXiv* dont elle s'est pourtant inspirée, HAL rassemble des publications issues de tous les domaines scientifiques, déposées par les auteurs eux-mêmes dans plus de 60 pour cent des cas ou par des professionnels de la documentation. Ces dépôts sont faits dans le respect du droit de chaque acteur de la publication. Cette plate-forme, dont l'illustration page ci-contre est l'emblème, est utilisée par l'ensemble des établissements français, universitaires ou de recherche. Faisant partie de la mise en œuvre de la politique demandée par la Commission européenne, elle est aussi soutenue par le ministère de l'Enseignement supérieur et de la Recherche.

Ce projet permet de prendre en compte des aspects techniques lourds, telle la possibilité de garantir l'accessibilité aux textes par des liens URL solides et d'apporter des solutions au problème de la pérennité à long terme des données par des mises à jour régulières des fichiers déposés.

Les éditeurs considèrent le libre accès comme une menace pour leur modèle économique. Ils tentent de différer le dépôt dans une archive ouverte en imposant aux auteurs un embargo de six mois à deux ans après la date de publication, mais leur proposent aussi de payer, avant publication, pour que leur article soit en libre accès. Ce modèle est dit hybride, car si les revues sont disponibles sur abonnement, le téléchargement de certains articles est gratuit. Les revues en libre accès bousculent les modèles économiques. Certaines reportent vers l'auteur les coûts de production de la revue. Le financement n'intervient donc plus en aval de la publication sous forme d'abonnements, mais en amont grâce au paiement par l'auteur d'une somme forfaitaire pour chaque article accepté. En général, cette somme est subventionnée et prise en charge par son institution ou par l'organisme de financement de la recherche.

Aujourd'hui, les enjeux du libre accès et de la pérennité se généralisent aux données produites au cours de la recherche, tels les résultats d'expériences ou les données statistiques. Cela concourt à la validation des résultats en facilitant leur reproduction par d'autres équipes, ainsi qu'au libre accès de l'ensemble des recherches financées par des fonds publics. ■

### HAL en chiffres

**2 800** nouveaux fichiers y sont déposés par mois.

**236 000** fichiers y étaient hébergés en septembre 2013.

**1,5 téraoctet** de données y est stocké.

**286 000** auteurs y ont déposé au moins un article.

**2 500** collections (institutionnelles, thématiques…) y ont été créées.

### ■ À ÉCOUTER

Le jeudi 31 octobre de 14h à 15h, Christine Berthaud reviendra sur les enjeux du libre accès dans la partie Actualités de l'émission **La marche des sciences**, sur *France Culture*.
**www.franceculture.fr**

### ■ LES AUTEURS

**Christine BERTHAUD est directrice du Centre pour la communication scientifique directe (CCSD, CNRS-UPS2275), à Villeurbanne.**

**Agnès MAGRON est gestionnaire de communauté au sein du CCSD.**

### ■ SUR LE WEB

**ArXiv, archive ouverte en physique et en mathématiques :**
arxiv.org

**HAL, archive ouverte couvrant tous les domaines scientifiques :**
hal.archives-ouvertes.fr

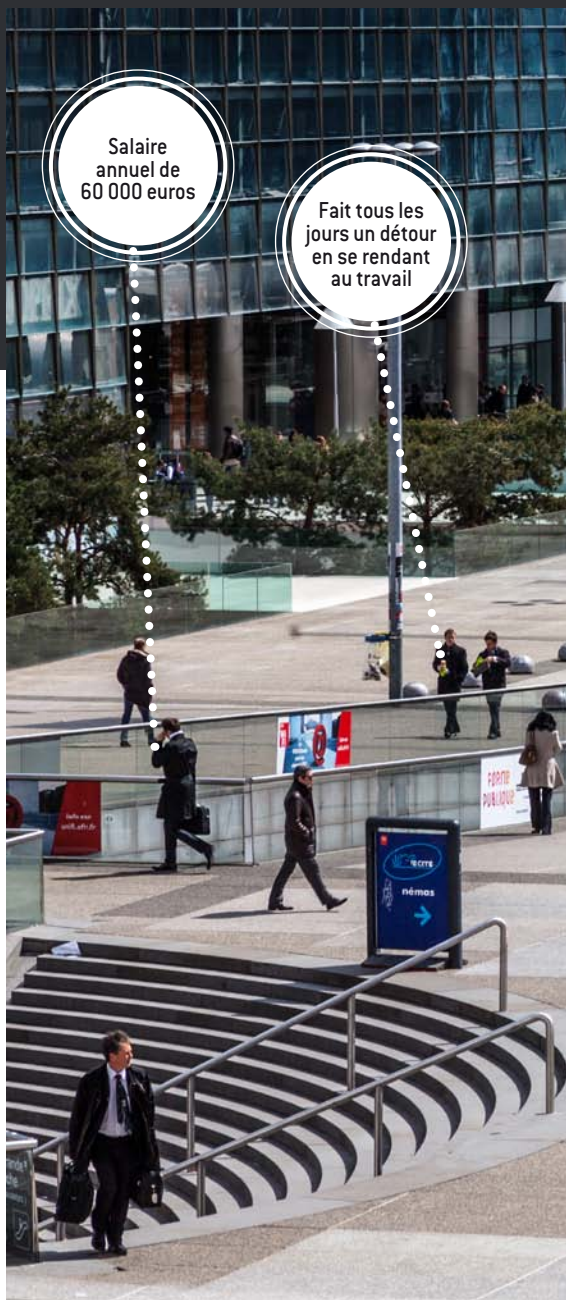
**Le CCSD, concepteur de HAL et chargé de sa gestion :**
ccsd.cnrs.fr

**Le Centre pour l'édition électronique ouverte :**
cleo.openedition.org

# Orienter la société grâce aux **données massives**

**Les traces numériques que chacun de nous laisse derrière lui chaque jour pourraient devenir un cauchemar pour notre vie privée… ou servir à construire un monde plus sain et plus prospère.**

**Alex Pentland**




**1. LES TRACES NUMÉRIQUES** que nous laissons derrière nous permettent de retracer et de prédire nos comportements individuels.

Dès le milieu du XIXe siècle, la croissance urbaine rapide, encouragée par la révolution industrielle, avait engendré des problèmes sociaux et environnementaux majeurs. Les villes y ont répondu en construisant des systèmes centralisés pour fournir de l'eau, de la nourriture et de l'énergie, favoriser le commerce, faciliter les transports, maintenir l'ordre et donner un accès aux soins. Aujourd'hui, ces solutions plus que centenaires sont de moins en moins adaptées. Nos grandes villes sont surpeuplées et encombrées par les bouchons. Et nous sommes confrontés à de nouveaux défis, en particulier pour nourrir et loger une population mondiale qui devrait croître de deux milliards d'habitants, tout en limitant au mieux l'impact du réchauffement climatique.

Ces problèmes typiques du XXIe siècle appellent une réflexion nouvelle. Mais de nombreux économistes et sociologues abordent encore les questions sociales avec des concepts d'un autre âge, tels que les marchés et les classes, dans des modèles simplifiés qui réduisent les interactions sociales à des règles générales et ignorent les comportements individuels. Il faut examiner la situation plus en profondeur et prendre en compte les détails les plus fins des interactions sociales.

Les technologies numériques nous le permettent aujourd'hui. Elles offrent une vue unique sur les milliards d'interactions individuelles par lesquelles les gens échangent des idées, de l'argent, des biens ou des informations. Le Laboratoire de dynamique humaine que je dirige au MIT (l'Institut de technologie du Massachusetts) explore cette masse de données pour tenter de faire émerger les lois qui régissent ces interactions. Par cette approche, nous commençons à pouvoir prédire certains phénomènes tels que les chutes des cours de la Bourse, les résultats d'élections politiques ou les épidémies de grippe. L'analyse des données sociales peut nous aider à bâtir des systèmes financiers stables, des institutions politiques qui fonctionnent, des systèmes de soins efficaces et économiques, et bien d'autres choses. Mais nous devons d'abord prendre la pleine mesure du poten-

tiel des données massives, le « *Big Data* », et construire un cadre juridique pour les exploiter à bon escient. La capacité à suivre, prédire et même influer sur le comportement des individus et des groupes n'est pas sans rappeler le feu prométhéen : elle peut servir au meilleur comme au pire.

## Le pouvoir de prédiction des miettes numériques

Pendant que nous vaquons à nos occupations quotidiennes, tels des Petits Poucets, nous semons dernière nous des miettes virtuelles. Les personnes que nous appelons au téléphone, les endroits où nous allons, les aliments que nous mangeons ou les produits que nous achetons laissent des traces numériques. Ces petits détails en disent bien plus long sur notre vie que ce que nous choisissons d'en révéler nous-mêmes. Nos publications sur les réseaux sociaux ne livrent que des informations que nous avons choisi de divulguer à d'autres personnes, sélectionnées sur la base des critères du moment. Les traces numériques, en revanche, enregistrent la réalité de notre comportement.

Nous sommes des animaux sociaux, et notre comportement n'est jamais aussi unique que nous aimerions le croire. Les gens que vous appelez, à qui vous envoyez des textos et avec qui vous passez du temps (même les habitants de votre quartier, que vous reconnaissez sans les avoir directement rencontrés) vous ressemblent probablement à de nombreux égards. Mon équipe de recherche a montré qu'il est possible de prédire votre risque de développer un diabète simplement en examinant la liste des restaurants que vous fréquentez et des gens avec qui vous sortez. Les mêmes données révèlent le genre de vêtements que vous avez tendance à acheter ou votre capacité à rembourser un emprunt. Et parce que


© Shutterstock.com/Radu Razvan

### L'ESSENTIEL

- L'organisation de la société s'appuie sur des principes datant du début de la révolution industrielle. Pour répondre aux défis du XXIe siècle, une organisation sociale plus efficace est nécessaire.
- L'analyse de la masse de données numériques que nous laissons derrière nous au quotidien fait émerger de nouvelles idées pour s'attaquer aux problèmes sociaux.
- Mais nous devons d'abord donner aux individus plus de contrôle sur leurs informations avant d'exploiter celles-ci.

notre comportement change quand nous sentons que nous tombons malade – nous n'allons pas aux mêmes endroits, nous n'achetons pas les mêmes choses, nous n'appelons pas les mêmes personnes et nous faisons des requêtes différentes sur Internet –, il est possible grâce à l'analyse de données de dresser une carte qui prédit heure par heure l'endroit où vous risquez le plus d'attraper la grippe, par exemple.

Parmi les motifs qui se dégagent des données massives, ceux qui éclairent le plus le fonctionnement de la société sont ceux qui concernent le flux d'idées et d'informations entre les personnes. Nous pouvons mesurer et visualiser ce flux en étudiant les interactions sociales (conversations en face à face, appels téléphoniques, messages sur les réseaux sociaux...), les comportements d'achat des individus (*via* les données des cartes bancaires) ou les déplacements (*via* les données de géolocalisation par GPS). Le flux d'idées est central dans la compréhension de la société, non seulement parce qu'être informé en temps utile est crucial pour construire des systèmes efficaces, mais aussi parce que le partage et l'association des idées sont au fondement de l'innovation. En général, les communautés coupées du reste de la société stagnent.

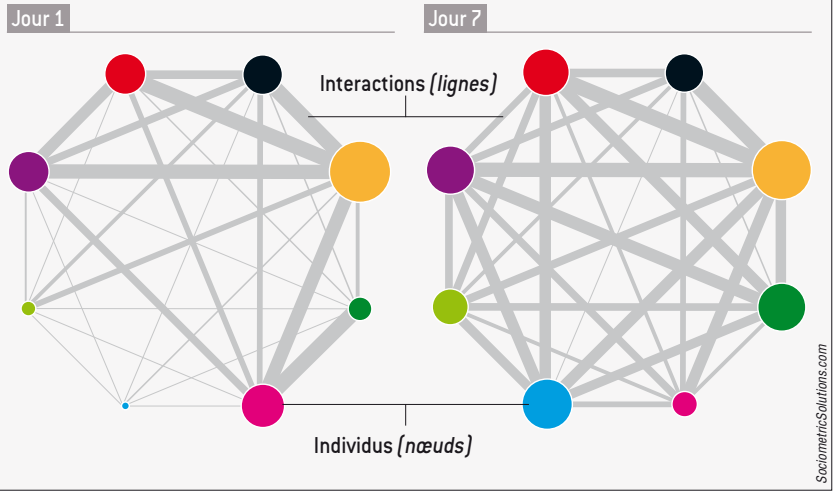
L'une des découvertes les plus surprenantes que nous avons faites est que les motifs formés par le flux d'idées sont directement reliés à l'augmentation de la productivité et à l'innovation. Les groupes et les organisations dont les membres interagissent et tissent des liens en dehors du groupe ont globalement une productivité plus élevée, une plus grande créativité, et leurs membres vivent même plus longtemps et en meilleure santé. Cette règle s'observe d'ailleurs chez toutes les espèces sociales. L'échange d'information semble essentiel à la bonne santé de toutes les sociétés.

On peut se représenter les entreprises ou les organisations comme des machines à idées. Ces machines agrègent et diffusent des idées, principalement par des interactions entre individus. Deux grandeurs permettent de mesurer la valeur d'un flux d'idées. La première est l'engagement, c'est-à-dire la proportion d'échanges d'individu à individu au sein du groupe. La relation entre l'engagement et la productivité est simple : un niveau d'engagement élevé prédit une productivité élevée du groupe, quel que soit (ou presque) le sujet sur lequel travaille le groupe ou le type de personnalité de ses membres. Le second facteur est l'exploration, c'est-à-dire la capacité des membres du groupe à établir de nouvelles connexions avec des personnes extérieures et à rapporter au sein du groupe de nouvelles idées provenant de l'extérieur. L'exploration est un bon indicateur à la fois de l'innovation et de la production créative.

Dans des expériences de terrain menées dans des entreprises, mes étudiants et moi avons mesuré le niveau d'engagement et d'exploration en équipant les employés de badges sociométriques, des dispositifs qui enregistrent les interactions en face à face. Nous avons découvert qu'en augmentant l'engagement au sein d'un groupe, on peut améliorer la productivité de façon spectaculaire tout en réduisant le stress. Par exemple, après avoir appris qu'un centre d'appels organisait les pauses de ses employés de telle façon qu'une seule personne prenne sa pause à la fois, j'ai convaincu le directeur de prévoir des pauses café avec plusieurs personnes en même temps, afin d'augmenter l'engagement entre les employés. Ce simple changement s'est traduit par un gain de productivité de 11 millions d'euros par an.

## Dynamique de groupe

En suivant les interactions sociales à l'aide de badges d'identification équipés de capteurs ou grâce aux données de téléphonie mobile, on peut mettre en évidence les dynamiques de groupe, ce qui permet de visualiser comment les membres d'une équipe collaborent. Les diagrammes ci-dessous illustrent la dynamique d'un groupe de réflexion de huit personnes étudié par les auteurs. La taille des cercles représente l'ampleur de la communication d'un membre avec le reste du groupe. À la fin de chaque journée, le diagramme de leurs interactions était présenté aux membres du groupe. En utilisant ces diagrammes, le groupe a pu identifier les points faibles et les corriger. L'équipe est ainsi devenue plus interactive et donc plus productive.




## L'AUTEUR

Alex PENTLAND est directeur du Laboratoire de dynamique humaine du MIT, l'Institut de technologie du Massachusetts, aux États-Unis, et codirige les projets du Forum économique mondial en matière de *Big Data* et données personnelles.

Les canaux de communication « riches», en particulier les interactions en face à face, ont beaucoup plus de poids que les canaux de communication électroniques. En d'autres termes, les courriels ne pourront jamais complètement remplacer les réunions et les conversations.

Nous avons également découvert que des phases successives d'exploration et d'engagement (les membres du groupe échangent entre eux, puis vont à la recherche d'informations nouvelles, les transmettent au groupe, et ainsi de suite) sont systématiquement associées à une meilleure production créative. Mes collègues ont observé ce schéma oscillant dans les interactions en face à face au sein d'organismes de recherche, et ont utilisé ces mesures pour identifier avec précision les journées les plus créatives des chercheurs. La même approche fonctionne avec les équipes virtuelles, dont les membres sont éparpillés dans divers laboratoires.

Des motifs similaires dans le flux d'information permettent de prédire la productivité globale d'une ville ou d'une région entière. L'engagement dans la communauté et l'exploration hors de celle-ci sont même reliés à des grandeurs sociales telles que l'espérance de vie, le taux de criminalité ou la mortalité infantile. Les quartiers qui sont des ghettos de l'information s'en sortent aussi mal que les ghettos physiques, alors que les quartiers dont les habitants sont engagés les uns avec les autres et connectés aux communautés environnantes sont plus prospères, et leurs habitants en meilleure santé.

## Optimiser le flux d'idées

L'utilisation des données massives pour diagnostiquer les problèmes et prédire les succès est une chose. Mais ce qui est passionnant, c'est que l'on peut aussi utiliser ces données pour concevoir des organisations (entreprises, institutions politiques, collectivités) qui fonctionnent mieux que celles existant.

Ce potentiel s'illustre facilement pour les entreprises. En mesurant le flux d'idées, il est en général possible d'identifier des changements simples qui améliorent la productivité et la créativité. Le service de publicité d'une banque allemande avait par exemple des difficultés à lancer des campagnes publicitaires efficaces pour les nouveaux produits, et la banque souhaitait comprendre pourquoi. Lorsque nous avons étudié la situation avec des badges d'identification sociométriques, nous nous sommes aperçus que les différents services échangeaient beaucoup de courriers électroniques, mais que presque personne ne parlait aux employés du service commercial. La raison en était simple : il était situé à un autre étage. Inévitablement, le service de la publicité concevait des campagnes que le service commercial n'était pas en mesure de mettre en œuvre. Quand la direction a vu notre diagramme représentant cette rupture du flux d'information, elle a pris conscience qu'il fallait déménager le service commercial au même étage que les autres départements, ce qui régla le problème.

Wen Dong et Alex Pentland

**2. EN EXPLOITANT LES DONNÉES DES TÉLÉPHONES PORTABLES,** on peut dresser la carte des zones où les individus sont le plus susceptibles d'attraper la grippe *(ici un exemple sur le campus de l'Institut de technologie du Massachusetts, probabilité croissante du rouge au violet)*.

Augmenter l'engagement au sein d'un groupe n'est cependant pas une solution miracle. En fait, le faire sans développer l'exploration peut même être une source de problèmes. Par exemple, quand j'ai mesuré avec le postdoctorant Yaniv Altschuler le flux d'information au sein du réseau social d'investisseurs *eToro*, nous avons constaté qu'à partir d'un certain point, les gens deviennent tellement interconnectés que le flux d'idées est dominé par les boucles de rétroaction. Certes, tout le monde échange des idées, mais ce sont les mêmes qui reviennent sans cesse. Les traders sont dans une sorte de chambre résonnante. Ce travail en circuit fermé expliquerait l'apparition de bulles financières, comme la bulle Internet du début des années 2000, provoquée par des investissements déconnectés de la réalité économique.

Néanmoins, nous avons montré qu'il est possible d'influer sur le flux d'idées au sein d'une communauté grâce à de petites incitations. Des « coups de pouce » peuvent inciter des personnes isolées à s'engager davantage avec les autres ; ou encourager des personnes prisonnières de la façon de penser du groupe à s'informer hors de leurs cercles habituels. Dans une expérience impliquant 2,7 millions de petits investisseurs du réseau *eToro*, nous avons « régulé » la communauté en donnant à certains investisseurs des coupons de réduction qui les encourageaient à explorer les idées d'un panel d'investisseurs plus diversifié. Résultat : le comportement de l'ensemble du groupe est resté dans les limites raisonnables de la « sagesse du plus grand nombre ». Le plus remarquable est que nous avons réussi à augmenter la rentabilité globale de plus de six pour cent alors que nous n'avions octroyé les incitations qu'à un petit nombre d'investisseurs.

Le développement d'un flux d'idées peut également contribuer à résoudre la « tragédie des biens communs », bien connue en économie, c'est-à-dire une situation où l'intérêt pour chacun de profiter le plus possible d'une ressource commune dépasse le prix que chacun doit payer pour utiliser cette ressource. Le secteur des assurances santé en est un bon exemple. Les personnes qui négligent de faire de l'exercice, de surveiller leur alimentation, de prendre

leurs médicaments ou qui ont des comportements à risque engendrent des frais médicaux élevés, qui sont répercutés par les assurances sur le prix des cotisations payées par chacun. Ce surcoût individuel est cependant trop faible pour que quiconque agisse afin de résoudre le problème.

La solution habituelle est d'identifier les individus qui abusent du bien commun et de leur infliger des pénalités (ou leur accorder des incitations) pour les pousser à mieux se comporter. Cette approche est onéreuse et porte rarement ses fruits. En revanche, le doctorant Ankur Mani et moi avons montré qu'en encourageant un engagement accru entre les individus, on parvient à réduire le problème. La clé est de proposer de petites incitations financières à ceux qui interagissent le plus avec les contrevenants, en les récompensant eux plutôt que le contrevenant pour leur comportement plus vertueux. Dans les situations réelles (par exemple, pour pousser les gens à faire des économies d'énergie), nous avons montré que l'approche fondée sur la pression sociale est jusqu'à quatre fois plus efficace que les méthodes habituelles.

## Une nouvelle donne pour les données

Cette même approche peut être utilisée pour la mobilisation sociale (dans les situations d'urgence ou quand un effort coordonné est nécessaire pour atteindre un but commun). En 2009, à l'occasion des 40 ans d'Internet, l'Agence américaine de recherche pour la défense (DARPA) a mené une expérience pour voir comment les réseaux sociaux et Internet pouvaient servir à une mobilisation d'urgence à l'échelle des États-Unis. L'agence offrait une récompense de 40 000 dollars à l'équipe qui trouverait le plus rapidement dix gros ballons rouges disséminés à travers le pays. Quelque 4 000 équipes se sont inscrites. Elles ont presque toutes adopté l'approche la plus simple, à savoir offrir une récompense à quiconque leur signalerait la présence d'un ballon. Mon équipe a choisi une stratégie différente. Nous avons décidé de partager le prix entre les personnes qui verraient un ballon et ceux qui auraient utilisé leurs réseaux sociaux pour recruter ces personnes. Cette stratégie, similaire à celle de la pression sociale évoquée précédemment, encourageait les gens à utiliser leurs réseaux sociaux autant que possible. Nous avons gagné la compétition en localisant tous les ballons en seulement neuf heures.

Pour mettre en pratique ces idées et améliorer le fonctionnement des systèmes sociaux en exploitant les données numériques personnelles, il est d'abord nécessaire d'établir ce que j'appelle une « nouvelle donne pour les données » : garantir que les données nécessaires soient facilement disponibles tout en protégeant la vie privée des citoyens. Selon moi, la clé est de traiter les données personnelles comme un bien : les individus auraient un droit de propriété sur les données qui les concernent. Que signifie « posséder » ses propres données ? En 2007, j'ai suggéré une analogie avec les principes de la propriété, de l'utilisation et de la cession dans le droit anglais :

*Vous avez le droit de posséder les données qui portent sur vous*. Indépendamment de l'entité qui recueille les données, celles-ci vous appartiennent, et vous pouvez y accéder à tout moment.

*Vous avez le droit au contrôle total de l'utilisation de vos données*. Cette utilisation ne peut se faire que par accord actif, et les termes doivent en être clairement expliqués. Si vous n'êtes pas satisfait de la façon dont une entreprise utilise vos données, vous pouvez lui en retirer l'accès.

*Vous avez le droit de vous débarrasser de vos données ou de les distribuer*. Vous pouvez faire détruire vos données ou les faire redéployer ailleurs.

## Des laboratoires vivants

Ces cinq dernières années, au Forum économique mondial, j'ai contribué à animer le débat sur ces principes entre responsables politiques, dirigeants d'entreprises multinationales et groupes de défense des citoyens. Les réglementations sont en train d'évoluer. Aux États-Unis, la nouvelle charte de respect de la vie privée des consommateurs, et dans l'Union européenne, le projet de directive sur la protection des données personnelles, accordent ainsi aux individus un meilleur contrôle de leurs données, tout en encourageant une plus grande ouverture.

Pour la première fois dans l'histoire, nous pouvons observer la société avec une précision suffisante pour envisager de construire des systèmes sociaux plus performants que ceux ayant existé jusqu'à présent. Les données massives promettent

### ■ BIBLIOGRAPHIE


A. Pentland, **Society nervous system : building effective governements, energy and public health,** *Computer*, vol. 45(1), pp. 31-38, 2012.

A. Madan *et al.*, **Sensing the health state of a community,** *IEEE Pervasive Computing*, vol. 11(4), pp. 36-45, 2012.

S. Moturu *et al.*, **Using social sensing to understand the links between sleep, mood, and sociability,** *Proc. of IEEE SocialCom*, pp. 208-214, 2011.

A. Madan *et al.*, **Pervasive sensing to model political opinions in face-to-face networks,** *Proc. of the 9th international conference on Pervasive computing*, pp. 214-231, 2011.

A. Madan *et al.*, **Social sensing : Obesity, unhealthy eating and exercise in face-to-face networks,** *Proc. of ACM Wireless Health*, pp. 104-110, 2010.

A. Madan *et al.*, **A social sensing to model epidemiological behavior change,** *Proc. of ACM Ubicomp*, pp. 291-300, 2010.

A. Pentland et *al.*, **Life in the network : the coming age of computational social science,** *Science*, vol. 323, pp. 721-723, 2009.

**Laboratoire de dynamique humaine au MIT :** http://hd.media.mit.edu/

de transformer nos vies autant que l'invention de l'écriture ou d'Internet.

Bien sûr, cette transition vers une société orientée par les données est un défi. Dans un monde de données illimitées, la méthode scientifique à laquelle nous sommes habitués ne fonctionne plus : il y a tellement de connexions et de corrélations potentielles entre les données que les outils statistiques classiques fournissent souvent des résultats absurdes. L'approche scientifique classique donne de bons résultats quand l'hypothèse est claire et que les données sont conçues pour répondre à la question. Mais dans la complexité inextricable des systèmes sociaux à grande échelle, il existe souvent des milliers d'hypothèses raisonnables, et il est impossible d'ajuster les données à toutes ces hypothèses en même temps.

Ainsi, à l'avenir, nous devrons analyser les données du monde réel pour faire émerger les lois sous-jacentes plus systématiquement et plus en amont que jusqu'ici. Nous devons construire des « laboratoires vivants » où nous pourrons tester nos idées pour construire des sociétés orientées par les données.

Un exemple d'un tel laboratoire vivant est la Cité des données ouvertes que nous venons d'inaugurer à Trente, en Italie, avec la coopération des élus, de *Telecom Italia*, de *Telefonica*, de la fondation Bruno Kessler et de l'Institut pour la conception dirigée par les données (IDDD). Le but de ce projet est de démultiplier le flux d'idées au sein de la ville de Trente. Des logiciels tel que notre système openPDS (*Personal Data Store*), qui implémente les principes de la nouvelle donne pour les données, permettent aux individus de partager en toute sécurité des données personnelles (telles des informations relatives à leur santé ou à leur famille) en contrôlant où va l'information et quelle utilisation en est faite. Par exemple, une des applications d'openPDS encourage le partage entre les familles des bonnes pratiques concernant l'éducation des jeunes enfants. Comment les familles dépensent-elles leur argent ? Quels sont la fréquence de leurs sorties et leur niveau de socialisation ? Une fois que l'individu a donné son accord, ses données peuvent être rassemblées, rendues anonymes et partagées automatiquement en toute sécurité avec d'autres familles.

**POUR LA PREMIÈRE FOIS DANS L'HISTOIRE**
**nous pouvons observer la société avec une précision suffisante pour envisager de construire des systèmes sociaux plus performants.**

Nous pensons que de telles expériences montreront que les avantages potentiels d'une société orientée par les données valent la peine de faire ces efforts et de prendre ces risques. Imaginez : nous pourrions anticiper les krachs financiers, détecter et prévenir les épidémies, utiliser au mieux les ressources naturelles et encourager la créativité pour prospérer. Ce rêve pourrait devenir bientôt une réalité... si nous évitons les écueils avec prudence. ■

# Comment préserver l'anonymat

Tristan Allard, Benjamin Nguyen et Philippe Pucheral

**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 compromettent pas la vie privée.**

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information : toutes ces facettes 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 (qui identifie un appareil connecté à Internet) pendant plusieurs années. Ces masses d'informations représentent une manne sans précédent pour les sociétés humaines *(voir l'article d'Alex Pentland dans ce numéro)*. 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 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 les adresses IP) renseigne ainsi sur l'importance locale de l'épidémie. De même, les données de mobilité pourraient être enregistrées par les systèmes de géolocalisation (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'informations 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 des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

## Un compromis entre utilité et protection

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand ce jeu ne contient pas son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec


Pour la Science, © Shutterstock.com/Krichanut

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, etc. Protéger les données personnelles sans les rendre inexploitables est donc un enjeu majeur à l'ère du tout numérique. La diffusion de ces données se révèle être 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 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. 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 extractibles de données publiées dépend de ce que l'on sait par ailleurs. Le développement des 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 ou presque 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 accessibles à un

## 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.**
- **De nombreuses techniques sont développées pour protéger les données sensibles sans les rendre inexploitables.**


Pour la Science, © Shutterstock.com/VoodooDot, Valua Vitaly, mjaud

**1. LES REQUÊTES TAPÉES DANS UN MOTEUR DE RECHERCHES** renseignent sur l'identité de celui qui les soumet. Ainsi, supposons que l'on sache que l'individu 6208 s'intéresse à l'arbre généalogique des Dupont et qu'il a cherché un coiffeur et un boulanger rue Férou. En croisant ces requêtes avec des données disponibles sur Internet, telles les pages blanches, on découvre qu'un certain Marc Dupont habite rue Férou. On peut alors avoir accès aux requêtes de cet individu et s'apercevoir, par exemple, qu'il s'intéresse à la dépression (sans doute en souffre-t-il) ou qu'il a consulté des pages où l'on propose des rencontres discrètes. Sa vie privée est menacée.

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.

## Les faiblesses de la pseudonymisation

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 650 000 utilisateurs sur une période de trois mois (officiellement pour les rendre disponibles aux chercheurs de divers domaines). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots clefs des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont identifié la personne portant le pseudonyme 4417749. Elle 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 clefs avec des données disponibles sur le Web et ont peu à peu reconstitué l'identité probable de l'individu correspondant, 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 pseudonymisation de données reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

## Prévenir les croisements de données

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 *(voir la figure 2)*. En outre, la combinaison des renseignements correspondants est


### 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.

unique pour la plupart des gens, selon plusieurs études récentes – qui portaient sur les États-Unis, mais c'est probablement aussi le cas ailleurs.

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, et les champs sensibles (SD, pour *Sensitive data*), 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 $1/k$.

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'âges [15, 20] (c'est-à-dire de 15 à 20 ans) 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 parcourre 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 66)*.

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 souffre de ce cancer : la valeur sensible n'est pas protégée.

## Brouiller les données confidentielles

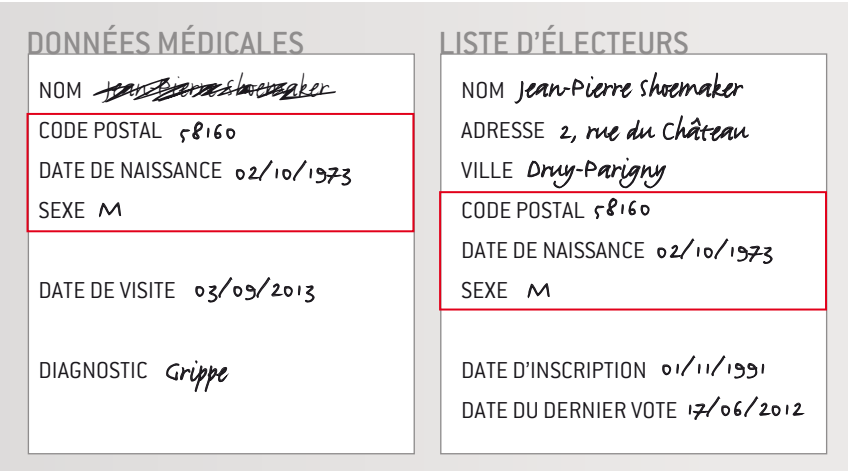
De multiples techniques ont succédé au $k$-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre 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 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 ; 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. C'est d'autant plus difficile qu'on ignore souvent qui sera le ou les attaquants – ces derniers pouvant même avoir des connaissances *a priori* différentes.

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 66)*. 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 $1/l$.




**2. LE TITULAIRE D'UN DOSSIER MÉDICAL** dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données (de tels dossiers circulent parfois entre les hôpitaux et divers instituts de recherche ou de statistique). En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(encadrée en rouge)* est unique. Il suffit alors de trouver un autre jeu de données qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière. C'est le cas, par exemple, d'une liste électorale *(à droite)*, disponible sur simple demande à la mairie.

De nombreuses 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 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 ci-dessous)*.

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.

L'assainissement consiste alors à perturber les données de telle 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.

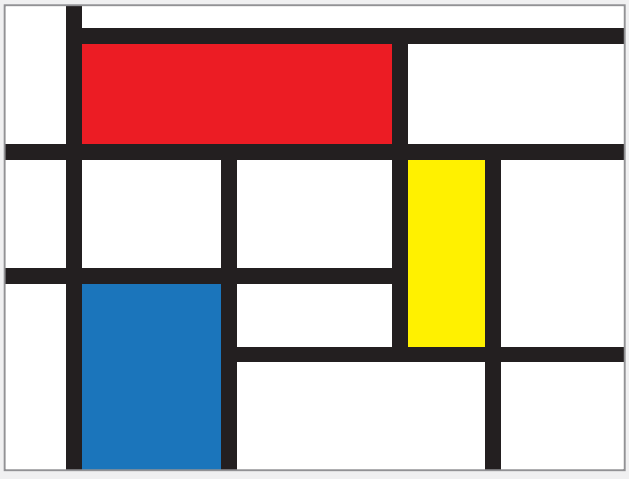
## Une vision alternative de la confidentialité

Les fausses données sont élaborées de façon à ce qu'on puisse toujours extraire certaines informations. Prenons l'exemple d'un jeu de 100 dossiers médicaux, à partir duquel on souhaite savoir combien de personnes ont la grippe. On perturbe ce jeu de données en y introduisant 1 000 faux dossiers, chacun se voyant attribué de façon équiprobable une maladie parmi quatre (dont la grippe).

### 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. Cela n'empêche pas toujours les réidentifications ultérieures *via* des croisements de données. D'autres techniques ont alors été développées. Celle du *k*-anonymat est 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 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), peintre hollandais dont les œuvres partitionnent l'espace *(ci-contre, un dessin inspiré 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.



© Shutterstock/Reinhold Leitner

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 *(ci-contre en haut à droite)*. Si *k* est supérieur à 2, les 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 chacun par l'ensemble de valeurs couvert par 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 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 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 *(en bas à droite)*.

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.

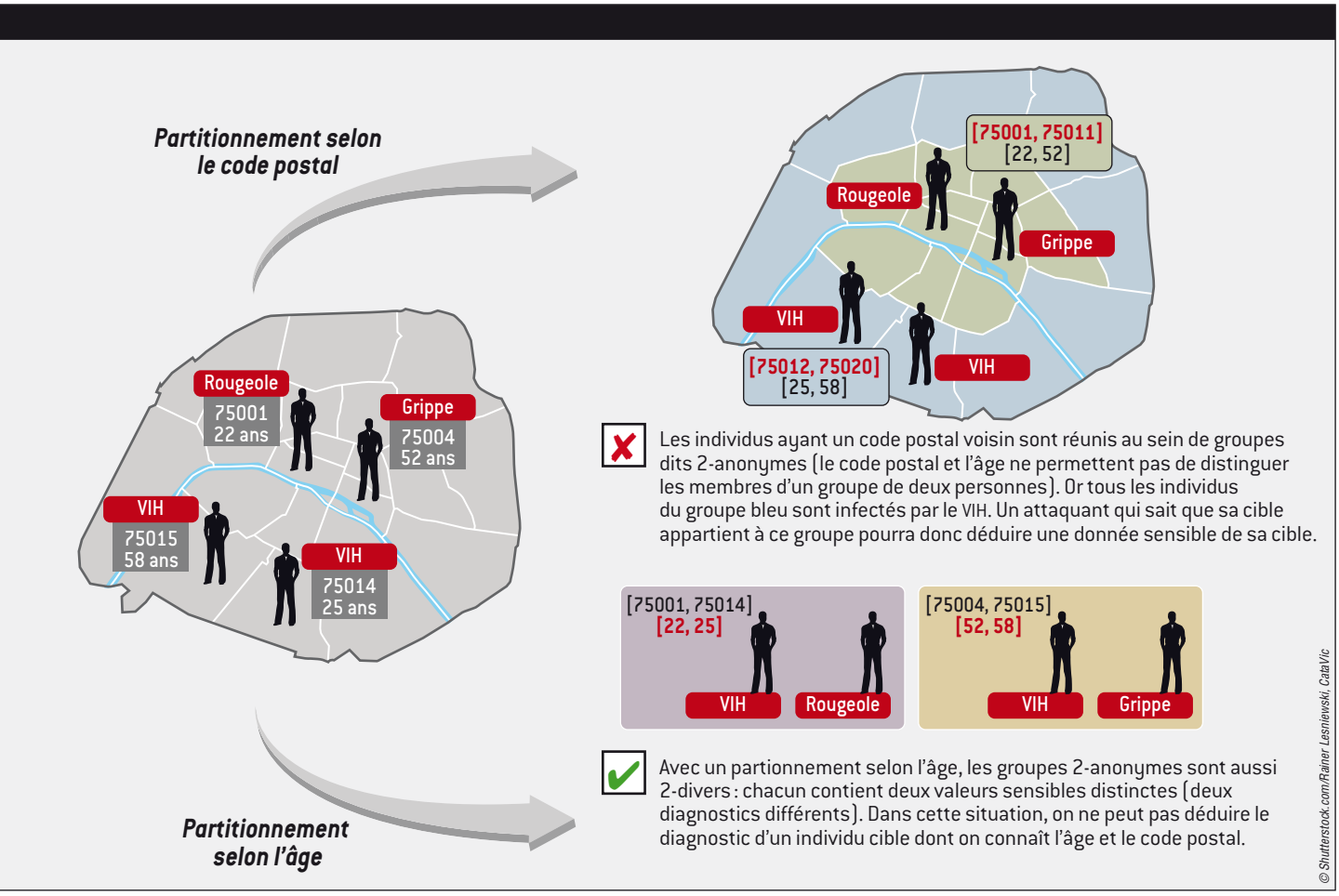
Le succès de la confidentialité différentielle s'explique en partie par sa simplicité : l'attaquant n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, 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.

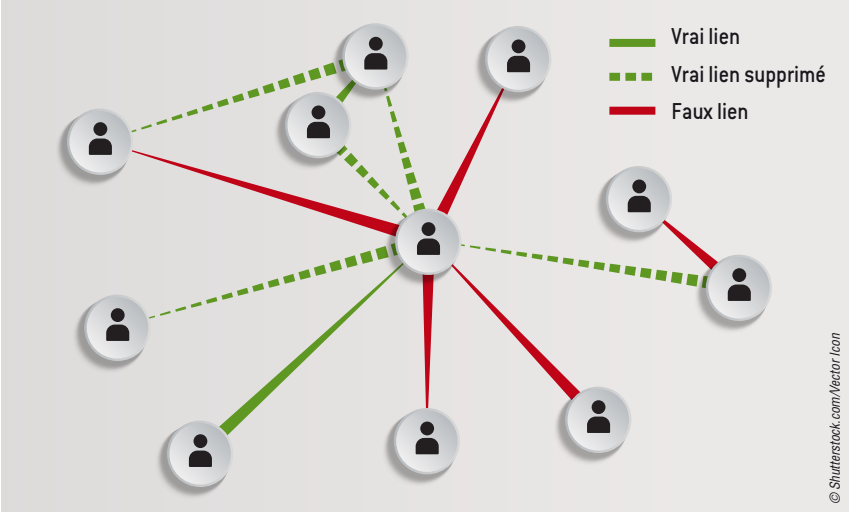
## UN MODÈLE D'ASSAINISSEMENT

**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.

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. ■

# La protection des informations

Rémy Chrétien et Stéphanie Delaune

**Un bon cryptage des données sensibles ne suffit pas à les protéger. Il faut aussi que les protocoles de communication utilisés pour coder ces données soient dépourvus de failles, dans leur logique comme dans leur mise en œuvre.**

Le 6 septembre 2013, le monde apprenait avec stupéfaction l'existence de *Bullrun*. Ce programme de l'Agence de sécurité américaine, la NSA *(National Security Agency)*, vise à décoder un grand nombre d'informations cryptées circulant sur Internet. À partir des documents secrets que leur a livrés le lanceur d'alerte Edward Snowden, le *New York Times* et le *Guardian* révélaient qu'une percée technique avait multiplié la puissance de *Bullrun* en 2010, de sorte que les transactions et communications en ligne protégées des banques et de la plupart des grands fournisseurs de service sur Internet étaient désormais accessibles « en grands volumes » aux agences de renseignement *(voir la figure 1)* ; on apprenait aussi que ces agences disposent de superordinateurs pour décrypter par la « force brute » les données qui résistent…

# sensibles

© Domaine public

Aujourd'hui, sans vraiment le savoir, nous réalisons tous quotidiennement des opérations de cryptage. Des chiffrements permettent, par exemple, de protéger notre numéro de carte bancaire lors d'un achat en ligne ; on les retrouve dans les protocoles (les procédures de communication) utilisés en téléphonie mobile ou pour lire les passeports électroniques. Pour autant, toutes ces données sensibles, nos données, sont-elles vraiment protégées ? Où sont les points faibles ? Nous allons analyser dans cet article ce qui conditionne la qualité des systèmes de sécurisation des données utilisés de façon routinière.

Sur un réseau, tout se passe comme si les informations protégées ne circulaient qu'à l'intérieur de coffres-forts réalisés dans un matériau solide : le chiffre (le cryptage). Toutefois, quelle que soit la solidité mathématique du chiffrement opéré, encore faut-il que l'assemblage des pièces qui constituent les coffres-forts en circulation ne laisse pas apparaître de fente exploitable par une personne mal intentionnée… Finalement, l'expédition de coffres-forts d'un côté et de leurs clefs de l'autre n'est pas une pratique sans risques, et il arrive qu'on y commette des erreurs, voire qu'on laisse traîner les clefs…

## Chiffre et compagnie

Le chiffrement, nous l'avons dit, est l'analogue du matériau constituant les parois du coffre-fort. Il en existe de nombreux types, chacun caractérisé par des contraintes d'utilisation et des niveaux de sécurité différents. Quoi qu'il en soit, la qualité d'une primitive cryptographique, c'est-à-dire d'un algorithme de cryptage particulier, se mesure à sa robustesse et à sa facilité de mise en œuvre. Tout comme les parois d'un coffre-fort peuvent être plus ou moins épaisses, la robustesse d'une primitive cryptographique peut être plus ou moins importante. Une primitive sera considérée comme faible quand il est facile de découvrir comment inverser son chiffrement, inversion qui permet de révéler les données protégées *(voir l'encadré page suivante)*.

**1. CES DÔMES RECOUVRENT LES ANTENNES** utilisées (ici au Japon) par le système mondial d'interception des communications privées et publiques des agences de renseignement des États-Unis, du Royaume-Uni, du Canada, de l'Australie et de la Nouvelle-Zélande.