seulement un gage de sécurité juridique et de limitation des risques mais elle permet aussi d'accroitre le capital de confiance vis-à-vis des personnes concernées. Elle constitue ainsi un facteur de compétitivité pour les entreprises qui peuvent se prévaloir de pratiques vertueuses et de mise en place d'outils ou d'acteurs tels que les correspondants Informatique et Libertés, les labels ou les règles internes d'entreprise qui témoignent de leur engagement pour un déploiement d'innovations prenant en compte la protection des données.

La CNIL considère que cette réglementation peut être un outil pour bâtir une innovation durable se déployant dans une société préservant les valeurs humanistes dont l'Europe a tout lieu de s'enorgueillir. Il est bien naturel que l'Europe ne se laisse pas dicter un modèle économique et qu'elle entende peser dans le débat mondial pour faire passer le message d'un choix de société dans laquelle nous voulons vivre et qui est en définitive le soubassement de cette législation.

Certes la réglementation n'est pas tout. Comme le disent les Américains, il n'y a pas de « *silver bullet* », pas une seule balle, pas une seule solution. En matière de protection des données, il y a tout à la fois matière à poursuivre l'éducation au numérique des citoyens, des dirigeants, à sensibiliser les professionnels, encourager le déploiement d'innovations technologiques permettant de préserver la vie privée et les libertés, que ce soit des mécanismes d'anonymisation, de limitation de la durée de conservation des données, de gestion par les individus des données qui les concernent… Il y a là aussi un champ infini d'explorations sur le plan technique. ■

2

La technologie au service de la confiance ?

Philippe Tassi (1972), Directeur général adjoint de Médiamétrie, Vice Président d'ENSAE Alumni

« C'est un joli mot : confiance » aurait pu chanter Jean Ferrat. Sur ce mot reposent beaucoup d'équilibres de notre société, dont celui de la société connectée actuelle et à venir.

Les données ou statistiques, détenues ou élaborées par des administrations ou des entreprises, ont en général été construites à partir d'informations individuelles, ce qui pose la question de la protection des sources, c'est-à-dire de la vie privée, compte tenu des progrès constants de la science et des processus de traitement.

Comment établir et maintenir la confiance du grand public, partie prenante numéro un, tout en respectant l'équilibre entre promesse de confidentialité et utilisation des données recueillies ?

L'établissement de la confiance s'appuie sur deux conditions nécessaires et fondamentales : l'une est réglementaire, et montre que beaucoup d'Etats ont pris conscience depuis longtemps de la nécessité d'établir des garde-fous juridiques ; l'autre s'appuie sur la déontologie et l'éthique des acteurs.

Une troisième condition existe, et repose sur les technologies qui peuvent aider à déployer des obstacles techniques empêchant la diffusion de données contre le gré de leur auteur.

Le grand public peut-il avoir confiance dans les procédures technologiques permettant de sécuriser les bases de données ? La réponse est OUI. Néanmoins, le maintien à niveau des méthodes de protection nécessite une attention permanente.

Cryptage et codage : premières approches toujours d'actualité

Comment protéger des informations ? C'est une question à laquelle les réponses sont anciennes. Les mots clés sont : transformer, masquer le contenu des messages envoyés par un émetteur vers un récepteur. Cela permet de limiter la possibilité pour un tiers, un « adversaire », d'accéder à une information confidentielle.

Sont donc apparues voici quelques milliers d'années (Sumer, Babylone, Egypte) des techniques de cryptage ou de chiffrement : une suite de symboles est remplacée au niveau de l'émetteur par une autre suite selon un système connu du récepteur, ce qui permet à ce dernier seul de comprendre le message de départ. Le propos n'est pas ici d'écrire l'histoire des méthodes de cryptage, mais d'en faire comprendre les principes, les enjeux et les conséquences (Voir figure 1).

Le vrai destinataire connaît la technique de cryptage et est donc capable de retrouver le message initial. Un tiers, qui aurait intercepté la lettre, est *a priori* incapable d'en comprendre le sens.

Si le cryptage implique le décryptage, pour que le destinataire puisse comprendre le message original, il a très vite aidé également au développement des tentatives pour casser la clé de cryptage.

A commencé alors un jeu de gendarme et voleur (adversaire, « tierce partie », hacker…), traduisant le couple cryptographie-cryptanalyse, le premier étant l'appellation générique de techniques de protection par chiffrement d'un contenu initial, le second étant celle des travaux et moyens nécessaires à faire céder cette protection pour découvrir ce même contenu chiffré.

Le couple cryptographie-cryptanalyse est un bel exemple d'un algorithme protection-effraction dont on peut se demander s'il sera convergent un jour.

Une course à la complexification a commencé. Pour l'illustrer, en repartant du chiffrement de Jules César, il suffit d'accroître le nombre d'alphabets de transcription (Voir figure 1).

Quelques lignes sur la cryptanalyse : le déchiffrement d'un message crypté a donné lieu à de multiples avancées dès le premier millénaire, mais il y a peu de doute sur le fait que la formalisation des méthodes de cryptanalyse est due à l'école des mathématiciens arabes des neuvième et dixième siècles. Comme l'écrit Simon Singh, « *la cryptanalyse ne pouvait être inventée que par une civilisation ayant atteint un niveau élevé de connaissances dans plusieurs disciplines, dont les mathématiques, la statistique et la linguistique* ». Un nom semble se dégager, celui d'Abu Yusuf Al Kindi, dont on a retrouvé en 1987 le traité intitulé *Manuscrit sur le déchiffrement des messages cryptographiques*, et qui fonde son approche sur l'analyse statistique des fréquences de lettres, de signes et de syllabes.

La cryptographie atteint ses lettres de noblesse reconnue sous Louis XIV, avec le très célèbre « Grand Chiffre » du Cabinet Noir du roi. Louis XIV confie à un mathématicien d'Albi, Antoine Rossignol des Roches, plus tard aidé de son fils Bonaventure puis de son petit-fils Antoine-Bonaventure, la mise au point d'un code d'une extrême complexité, dont la clé disparaitra avec la famille Rossignol. Il faudra attendre deux siècles pour le casser. Le nom de *rossignol* est d'ailleurs

Figure 1 : EXEMPLE DE CRYPTAGE

Un exemple de cryptage bien connu est celui utilisé par Jules César pour communiquer avec ses généraux pendant la la guerre des Gaules.

Le principe est simple : il s'agit d'établir une correspondance entre deux alphabets, un alphabet initial classique et un alphabet d'écriture, décalé.

Ainsi pour transformer le mot *tandem*, le deuxième est décalé de trois lettres et commence par d pour finir par c. Chaque lettre du mot original est remplacée par sa correspondante dans l'alphabet n°2.

Ainsi, le t est donc remplacé par w, le a par d, le n par q, etc…

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
↓ ↓ ↓ ↓ ↓ ↓
D E F G H I J K L M N O P Q R S T U V W X Y Z A B C

Le mot *tandem* ainsi crypté devient *wdqghp*.

resté dans le langage courant pour désigner un crochet capable d'ouvrir de multiples serrures.

Bien sûr, les cryptographes savent qu'existent les cryptanalystes, et un jeu de vrai – faux message va rapidement apparaître. L'auteur d'un message sachant qu'il risque d'être décrypté et donc compris par un tiers sera tenté d'y inclure de fausses informations pour tromper l'ennemi. Ainsi, une lettre de Louis XIV écrite avec le « Grand Chiffre » a été décodée à la fin du XIXème siècle, dans laquelle le roi désigne le général Vivien de Bulonde comme étant le célèbre « homme au masque de fer », puni à la suite de trahisons et fautes militaires. Mais de nombreux historiens considèrent que cette lettre – absolument authentifiée – n'est peut-être qu'un leurre pour cacher une vérité toute autre.

N'oublions pas de mentionner également le passage à la mécanisation du chiffrement, dont l'un des initiateurs fut au début du XIXème siècle l'anglais Charles Babbage, visionnaire de l'ordinateur, et, bien évidemment, le génie d'Alan Turing, à la fois pour ses avancées théoriques et son génie de cryptanalyste pendant la guerre 1939-1945.

Bien plus proches de nous, enfin, il faut mentionner les travaux de Mme Shafi Goldwasser et Silvio Micali, du MIT, sur la notion de sécurité sémantique. Ils ont mis au point le premier système à chiffrement probabiliste, à sécurité absolue.

Après un pré-print en 1982 – dans lequel est présente une question majeure : *What is a secret ?* –, leur publication initiale date de 1984, dans le *Journal of Computer and Systems Sciences*.

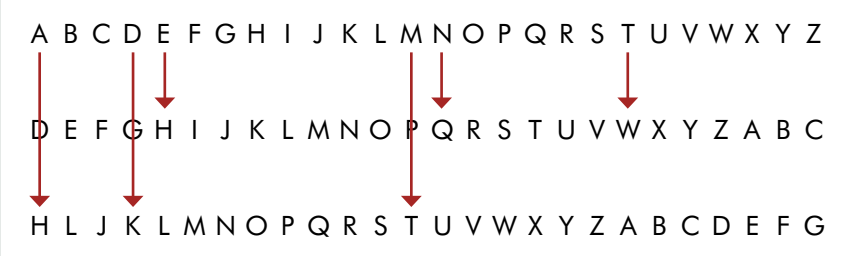
Ils ont reçu en mars 2013 le *Alan Turing Award* (considéré comme le prix Nobel pour l'informatique, de même que la médaille Fields l'est pour les mathématiques) au titre de l'année 2012, prix décerné par l'ACM (*Association for Computing Machinery*).

Le codage de Chappe

Une autre méthode pour masquer le contenu d'un message en clair est le codage. Illustrons par le codage de Chappe. Claude Chappe (1763-1805), ingénieur et inventeur, conçoit un système visuel de télégraphie aérienne, utilisant des pitons ou des édifices visibles de loin, et constitué d'un réseau de stations distantes de 10 kms en moyenne (Galilée a inventé la

Figure 2 : EXEMPLE DE CRYPTAGE

Reprenons l'exemple précédent de chiffrement du mot *tandem*, en ajoutant un deuxième alphabet décalé de 7 lettres, commençant par la lettre h pour s'achever par g.



Le principe sera de remplacer les lettres impaires (t, n, e) par leur correspondante dans l'alphabet 1, et les lettres paires (a, d, m) dans l'alphabet 2. La lettre t est transcrite par w, le a par h, le n par q, le d par k, etc. Le mot *tandem* ainsi crypté sera écrit *whqkht*.

Figure 3 : UNE STATION DE CHAPPE

Figure 2 : PREMIER TÉLÉGRAMME OFFICIEL UTILISANT LE RÉSEAU ET LE CODE DE CHAPPE

Ce message est envoyé à la Convention Nationale – assemblée qui gouverna la France de septembre 1792 à octobre 1795. Il annonce la victoire des forces révolutionnaires françaises le 30 août 1794 sur les troupes autrichiennes à Condé-sur-Escaut (« Condé être restitué à République. Reddition avoir eu lieu ce matin à six heures », avec les codes associés).

lunette astronomique en 1609). Le principe est d'utiliser un tryptique articulé, avec sémaphore, permettant le transfert de station en station d'un message masqué en utilisant un code décrit dans un livre de 92 pages et 92 lignes chacune. Les signaux sont en paires, le signal 1 indique la page et le signal 2 la ligne dans la page, ce qui autorise le récepteur final du message à le retranscrire.

Après le suicide de Chappe, en 1805, Napoléon Ier fait construire en 1806 un système de surveillance des côtes depuis le rivage (conçu par Louis Jacob, amiral et homme politique). Le poste de guet s'appelle un sémaphore, et la communication est faite par signaux optiques, inspirée du télégraphe de Chappe.

Principes et techniques de protection et de confidentialité

Les méthodes de cryptage ou de chiffrement, anciennes, sont toujours au centre des recherches et procédés pour dissimuler une information, et les travaux de Goldwasser et Micali en montrent toute l'actualité et l'efficacité.

Nous allons maintenant aborder d'autres techniques de masquage de contenus tout en présentant la philosophie qui préside maintenant à la recherche sur la sécurité des bases de données, fussent-elles *small* ou *big*.

L'approche de Tore Dalenius

Au cours du dernier quart du XXème siècle, un précurseur de la réflexion sur les bases de données est le grand statisticien et spécialiste de sondages suédois Tore Dalenius.

Dans un article publié dans *Statistik Tidskrift* en 1977, intitulé « *Towards a methodology for statistical disclosure control* », Dalenius pose le principe suivant : « *Accéder à une base de données ne doit pas permettre d'apprendre plus de choses sur un individu que ce qui pourrait être appris sans accéder à cette base de données.* »

Dit autrement, si la publication d'un agrégat statistique T de la variable X permet de connaître la valeur de X pour une personne plus précisément que sans accès à T, alors il y a eu faille de confidentialité.

Ce principe a le mérite d'être simple et compréhensible. Cependant, il peut être mis en défaut par l'existence d'information auxiliaire, élément important de la théorie des sondages.

Par exemple, supposons qu'existe une base de données D contenant des agrégats statistiques qui sont les tailles moyennes des hommes adultes de diverses nationalités. On y ajoute une information auxiliaire : « M. A fait 2 cm de plus que l'Italien moyen ». Le seul accès à l'information auxiliaire donne très peu d'éléments sur M. A, mais en croisant les deux informations, on connaît la taille exacte de M. A.

Ceci peut être présenté de façon plus mathématique :

Soit S un mécanisme de protection d'une base de données D, et une faille F dans cette protection. Pour toute base de données D, il existe une information auxiliaire Z telle que :

a. Z seule n'a pas d'utilité pour un « adversaire ».

b. Z combinée avec les données protégées de D permet de trouver dans S une faille F avec une probabilité qui tend vers 1.

Quoi qu'il en soit, le principe de Dalenius est très intéressant dans sa formulation et sa finalité, la protection des données individuelles. Et l'information auxiliaire créant une faille de confidentialité, si elle a une existence théorique, doit encore être trouvée et accessible. Que se passe-t-

il si elle-même est protégée, par exemple cryptée de façon très sécurisée ?

Autres techniques de transformation

- La destruction

Comme son nom l'indique, il s'agit ici de supprimer toutes les données individuelles après une période conventionnelle d'exploitation. Méthode définitive, bien sûr, mais qui n'empêche pas de protéger la base de données pendant sa durée de vie active pour éviter qu'un tiers, pirate, accède à la base pendant cette période et continue à l'exploiter ensuite.

Par ailleurs, de nombreuses données ont une longue durée d'intérêt, pouvant dépasser le cadre conventionnel, ne serait-ce que pour les historiens ou les chercheurs.

- L'anonymisation

C'est une méthode *a priori* évidente : il s'agit d'enlever de la base de données les variables permettant d'identifier une personne, comme le nom, l'adresse, le code d'immatriculation, l'adresse IP, les numéros de téléphone, le code PIN, les photos, les éléments biométriques…

Plus généralement, il convient d'ôter aussi les variables permettant, par recoupement ou croisement, de retrouver – même imparfaitement – un élément identifié dans un ensemble, par inférence individuelle.

Néanmoins, on observe que le nombre de failles augmente avec le progrès des méthodes et des algorithmes scientifiques et la probabilité de réidentifier un individu augmente, même après anonymisation.

- Agrégation

Une idée facile à mettre en oeuvre consiste à agréger les données individuelles et à ne rendre publics ou utilisables que des résultats. C'est le principe de la loi de 1951 sur le secret statistique. Base de la statistique, l'agrégation consiste à regrouper les données individuelles en résultats synthétiques.

Pour l'agrégation comme l'anonymisation, les performances des modèles mathématiques remettent en cause ces deux techniques, en autorisant une inférence individuelle.

Deux exemples récents en illustrent leurs limites potentielles.

En 2008, N. Homer, S. Szelinger, M. Redman et alii publient « *Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microararys* » dans PLoS Genetics (Public Library of Science).

Extrait du résumé :

« *We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.* »

Le deuxième exemple est contenu dans l'article "*Identifying participants in the Personal Genome Project by name*", de L. Sweeney, A. Abu et J. Winn. Travaillant sur le Personal Genome Project (PGP) américain, les auteurs montrent qu'ils identifient correctement 84 % des profils ; ce taux monte à 97 % si on relâche la notion d'identification stricte, par exemple en acceptant les diminutifs comme Jim pour James ou Andy pour Andrew.

Obscurcissement

Obscurcissement, assombrissement, obfuscation, autant de synonymes pour désigner des méthodes d'altération des données. On distingue les techniques non perturbatrices et les techniques perturbatrices ou interventionnistes.

Dans la première famille, on peut masquer la valeur de certaines cellules dans un tableau de résultats, enlever des variables pour certains individus, ou partager seulement un échantillon de

données extrait de la base de données générale, ou enfin combiner certaines catégories pour des variables à modalités.

Dans la deuxième famille, on agira directement sur les données, afin d'engendrer du bruit, au sens large. Par exemple, il est possible de modifier certaines variables numériques en les arrondissant, ou en les bloquant par troncature à des seuils maximum ou minimum.

On pourra transformer les variables en leur appliquant un homomorphisme $(X_1, \ldots, X_K) \dashrightarrow (j_1(X_1), \ldots, j_K(X_K))$; ou bien permuter les valeurs d'une même variable entre deux répondants, ou encore perturber les données en y ajoutant un bruit aléatoire.

Certaines transformations (permutation, rotation) laisseront invariantes les statistiques, d'autres non.

Une piste d'investigation particulièrement intéressante porte le nom de données synthétiques (« *synthetic data* »), née des travaux sur les données manquantes de John Little et Donald Rubin.

La confidentialité différentielle

Dernière approche pour établir la confiance, née au début du XXIème siècle, la notion de confidentialité différentielle semble particulièrement porteuse d'espoirs et de richesse.

Sa philosophie n'est pas éloignée de celle de Tore Dalenius ; on peut l'énoncer ainsi :

« La probabilité d'une conséquence négative quelconque pour une personne (par exemple, le fait qu'elle se voie refuser un crédit, ou une assurance) ne peut augmenter significativement en raison de sa participation à une base de données. »

Le principe est donc de promettre à une personne susceptible de contribuer à une base de données qu'une tierce partie (« l'adversaire ») ne pourra rien apprendre de plus sur elle, qu'elle accepte d'y participer ou non.

Les deux mots importants dans l'énoncé de la philosophie de la confidentialité différentielle sont *probabilité* et *significativement*, qui nous renvoient au domaine de la statistique et de l'aléatoire.

Faisons une analogie : dans les années cinquante, les travaux sur la qualité de William E. Deming évoquaient la notion de « zéro défaut ». Deux suites ont été données. Une première voie a consisté à ne rechercher que le zéro défaut, et à organiser les processus de production et de contrôle en ce sens. Une alternative a admis l'existence de défauts, dont l'existence est impossible à empêcher, ne serait-ce que parce que l'intervention humaine sera toujours un point faible, mais elle a structuré la production de façon à ce que la probabilité qu'il y en ait zéro soit la plus grande possible.

L'esprit de la confidentialité différentielle est proche de la seconde voie.

On admet la possibilité de failles de confidentialité, soit en raison du maillon faible constant qu'est l'homme (le « traître »), soit en raison des progrès exponentiels des modèles mathématiques qu'il est impossible de prévoir. Un cadre d'interdits sera ainsi rapidement dépassé et caduc en raison du progrès scientifique. Cependant, si des failles sont possibles, il faut tenter d'en diminuer la probabilité de survenance.

L'idée est donc de quantifier le risque (probabilité) de faille dans la promesse de confidentialité, tout en mesurant l'effet des techniques de protection des données et de la vie privée sur la qualité des résultats et analyses statistiques.

Ce concept de confidentialité différentielle fait partie du rapport de J. P. Holdren – E. S. Lander, intitulé « *big data and Privacy : A Technological*

Perspective », remis en mai 2014 au Président B. Obama, par le Conseil sur la Science et la Technologie auprès du Président (PCAST).

En résumé, les approches technologiques existent pour garantir et favoriser la protection des données et aider à l'établissement de la confiance du public contribuant à une base de données. Elles constituent un champ important de recherche. Certes, elles ne sont certainement pas très connues du grand public, mais doivent-elles l'être ?

L'avenir repose sur leur adoption, leur déploiement, leur maintenance et leurs évolutions, chez les gestionnaires de systèmes d'information, en dépit de leur coût de mise en place et de possession. ■

Ethique et déontologie : l'exemple d'Esomar

3

Elisabeth Cosnefroy, Représentante France d'ESOMAR
(the World Association for Social, Opinion and Market Research)

Esomar : une initiative européenne, un périmètre mondial

Esomar représente une profession discrète, qui ne parle pas assez de ce qu'elle a mis en place depuis quasiment ses débuts, mais qui prouve qu'en appliquant une déontologie et une éthique, on peut établir une confiance, et ça fonctionne.

Esomar est, historiquement, une association européenne à l'origine, d'où le E. Elle est depuis devenue mondiale. Esomar est donc actuellement l'association mondiale de tous les professionnels des études de marché, sociétales et d'opinion. L'expression « tous les professionnels » recouvre les deux parties prenantes que sont d'une part les acteurs qui conçoivent et réalisent ces études, donc les instituts et beaucoup d'autres organismes, ainsi que, d'autre part, tous les utilisateurs et commanditaires, les acheteurs, c'est-à-dire les entreprises. Cela recouvre un grand nombre d'interlocuteurs.

Cette association a été créée quasiment dès la naissance de cette profession des études, en 1948, au milieu du XXe siècle. Ce métier des études, apparu un peu avant la guerre – l'American Institute for Public Opinion est créé par Gallup, Crossley et Roper en 1936 –, est resté en sommeil pendant la deuxième guerre mondiale, pour des raisons facilement compréhensibles. Il renaît juste après, avec une volonté ferme et très efficace de créer tout de suite une association, au départ européenne, qui établisse un certain nombre de règles et de bonnes pratiques pour être sûr de pouvoir développer correctement le secteur des études marketing, dans le respect de ce qui anime le plus ses membres, c'est-à-dire l'humanisme.

Aujourd'hui, Esomar est reconnu dans 130 pays, avec près de 5 000 membres. Depuis deux ans, nous nous sommes ouverts à une forme d'homologation beaucoup plus *corporate*, qui n'implique pas seulement des membres individuels au sein d'entreprises, comme les cadres responsables des