PAR PHILIPPE WOLF (78)

ingénieur général de l'armement, ANSSI

Garantir la **disponibilité**, l'**intégrité** et la **confidentialité** des données

Les trois fonctions principales de la sécurité des systèmes d'information sont la disponibilité, l'intégrité et la confidentialité. Une analyse des remèdes, autour des trois aspects complémentaires que sont la sécurité des infrastructures, la protection des données et la protection des informations produites, ouvre de nouvelles pistes de recherche ou de développement.

Sécurité des infrastructures

La sécurité des infrastructures mettant en œuvre les *big data*, potentiellement sensibles au regard de ce qu'ils manipulent, fait appel principalement aux fonctions de disponibilité et d'intégrité. La résilience doit être une propriété globale de la chaîne (réseaux, baies, procédures, humains) et ne peut s'appréhender, dans son ensemble, qu'avec une analyse holistique et une gestion permanente des risques. Du très classique, même si les questions d'interdépendance et d'éparpillement prennent une importance cruciale due à la complexification des architectures de protection. De plus, le modèle du pair-à-pair se substitue au modèle client-serveur qui facilitait la supervision de sécurité. L'introduction de mécanismes de sécurité sur des couches logicielles qui se standardisent (*openstack*, *hadoop*, etc.) doit pouvoir apporter une résistance nouvelle. Les puissances de calcul requises par les *big data* nécessitent, sauf pour quelques très grosses entreprises, d'externaliser ou, au minimum, de mutualiser stockages et traitements dans l'infonuagique *(cloud computing)*. Le recours au *cloud computing* réclame des précautions même dans le cas de transparence absolue.

Rédiger les contrats sous la protection du droit national pour gérer le risque juridique

Protection des données

Dans le cas des *big data* non ouverts (privés), la *confidentialité* des données stockées ne pose pas de problème particulier si l'entreprise ou l'organisme garde la capacité de gérer ses propres clés de chiffrement ou de signature, de préférence dans un coffre-fort numérique labellisé, ou en confie la gestion à des tiers réellement de confiance. Pour rendre confidentiels les algorithmes de calcul, il manque aujourd'hui un ingrédient essentiel qui serait une implémentation pratique du chiffrement dit homomorphique, c'est-à-dire d'un chiffrement qui donnerait un moyen de réaliser diverses opérations sur le chiffré sans recourir à l'opération de déchiffrement complète. Une avancée dans ce domaine comme sur le calculateur quantique ou à ADN nécessitera, de toutes les manières, de reconcevoir une algorithmique adaptée.

Quatre règles à respecter

Les quatre trayons du « *cloud* maîtrisé ou souverain » sont connus mais pas toujours activés : faire appel à un ou des prestataires de confiance ; être capable d'auditer réellement la solution dans un temps court ; avoir la garantie testée de réversibilité pour changer de prestataire sans pertes, si nécessaire ; rédiger les contrats sous la protection du droit national pour gérer le risque juridique.

REPÈRES

La CNIL propose que « l'appellation *coffre-fort numérique*, ou *coffre-fort électronique*, soit réservée à une forme spécifique d'espace de stockage numérique, dont l'accès est limité à son seul utilisateur et aux personnes physiques spécialement mandatées par ce dernier. Les services de coffre-fort numérique doivent garantir l'intégrité, la disponibilité et la confidentialité des données stockées et impliquer la mise en œuvre des mesures de sécurité décrites dans la recommandation. »

Tolérance au flou

L'intégrité stricte des données n'est plus nécessaire quand il s'agit de manipuler des données non structurées, parfois faussées ou incomplètes, ou de travailler principalement par échantillonnage. Une tolérance au flou, aux calculs approchés et aux mutations rompant le clonage binaire parfait, est un ingrédient porteur d'une meilleure adéquation des *big data* au monde réel qu'ils sont censés nous aider à comprendre.

L'*intégrité* classique qui repose sur la signature numérique doit être, à son tour, révisée. Il existe déjà des dérives potentielles liées aux calculs largement répartis ou en grilles. Le respect des règles internationales de non-prolifération impose un contrôle, préalable de préférence, à un usage dévoyé des puissances calculatoires disponibles. La seule signature des ressources partagées, distribuées, hétérogènes, délocalisées et autonomes ne suffit plus. Des techniques d'obscurcissement (« obfuscation de code ») compliquent le contrôle.

Risques d'identification

Un pan croissant des *big data* touche aux données personnelles quand ils n'en sont pas le carburant premier. Les progrès des moteurs de recherche intelligents permettent d'identifier facilement une personne à partir d'un nombre très réduit de caractères, cela d'autant plus que l'intimité est littéralement mise à nu sur les réseaux sociaux. On retrouve, à une échelle nouvelle, de vieux problèmes d'inférences par déduction, induction, abduction ou adduction dans les bases de données classiques. Les croisements de données permettent des attaques par canaux auxiliaires sémantiques – attaques qui ne visent pas directement les protections théoriques mais leur implémentation pratique – structure redoutée en SSI. On arrivait à négliger ou à juguler les canaux cachés numériques : ce n'est plus le cas avec les canaux sémantiques.

Quatre critères sécuritaires

Les critères communs pour l'évaluation de la sécurité des technologies de l'information introduisent dès 1999, sous l'impulsion du Dr Pfitzmann, des fonctions de sécurité pour la protection des données personnelles. Elles sont au nombre de quatre. L'*anonymat* garantit qu'un sujet peut utiliser une ressource ou un service sans révéler son identité d'utilisateur. La *possibilité d'agir sous un pseudonyme* garantit qu'un utilisateur peut utiliser une ressource ou un service sans révéler son identité, mais peut quand même avoir à répondre de cette utilisation. L'*impossibilité d'établir un lien* garantit qu'un utilisateur peut utiliser plusieurs fois des ressources ou des services sans que d'autres soient capables d'établir un lien entre ces utilisations. La *non-observabilité* garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres, en particulier des tierces parties, soient capables d'observer que la ressource ou le service est en cours d'utilisation. Ces fonctions font l'objet de travaux algorithmiques novateurs, principalement en Europe, mais tardent à s'implanter dans les traitements numériques de masses qui vont passer rapidement aux traitements d'informations en masses.

Protection des informations

On ne peut éliminer le rôle du sujet dans la production de l'information, ou parfois de la connaissance, par les *big data*. « La signification d'une information est toujours relative » (Jean Zin). Il s'agit de mesurer l'intelligibilité, la vérifiabilité et la traçabilité, d'estimer la responsabilité contractuelle, de gérer les conflits d'influences, de distinguer les fausses nouvelles, bref, de résister au mirage des *big data* simplistes. Des amendes records touchent aujourd'hui des institutions financières. Elles sanctionnent des infractions à répétition qui n'auraient pas été possibles sans l'*obscurcissement numérique*, technique consistant à cacher des informations en les noyant dans

Des amendes records touchent aujourd'hui des institutions financières

Anonymat et santé

La sphère santé-social accumule les difficultés malgré les promesses des *big data* (études épidémiologiques, dossier médical personnel, optimisation des systèmes sociaux). Le constat de départ est qu'il n'y a pas de confidence (médicale) sans confiance (singulière). Il faut alors distinguer la confidentialité-discrétion partageable par du chiffrement réversible de la « confidentialité-séclusion » qui exige des fonctions à sens unique. Mais, dans ce dernier cas, la pseudo-anonymisation réversible serait parfois préférable à une véritable anonymisation irréversible, dans le cas, par exemple, de détection d'une maladie orpheline ou d'une grave épidémie où il faudrait retrouver l'individu porteur. Il manque clairement un modèle de sécurité partagé.

Sciences du danger et *big data*

Il est intéressant de noter que les cindyniques, ou sciences du danger, commencent à investiguer le champ de l'information. Elles proposent un regard à cinq dimensions, examinant à la fois la dimension des données (axe statistique), la dimension des modèles (axe épistémique), les finalités de l'acteur (axe téléologique), l'axe des règles, normes, codes auxquels est soumis (ou que s'impose) l'acteur et les valeurs (éthiques, morales) de l'acteur (axe axiologique).

une masse de données. L'obésité, sans diète, nourrit et amplifie cette obscurité. De plus, les biais cognitifs des *big data*, voulus ou non, aveuglent une saine compréhension des enjeux de sécurité. La capacité d'absorption humaine étant limitée, un différentiel de plus en plus grand se créera avec les capacités attendues des robots-programmes. Tant que les résultats espérés ne seront pas là, la tendance sera de complexifier les traitements par une massification encore plus grande des données et par l'ajout de paramètres aux automates. Alors qu'il faudrait, au contraire, modéliser, analyser, expliquer et mieux cibler et cribler les données utiles et rationaliser cette intelligence artificielle. Cette tendance à l'entropie porte en elle le germe des « accidents de la connaissance » signalés par l'essayiste Paul Virilio. À brasser trop large et trop gros, on oublie les fonctions essentielles et on bride l'engagement.

Pour faire des* big data *un outil de progrès, il faut en maîtriser les dérives

Une nouvelle approche de la SSI

Les *big data* ouvrent aussi des perspectives nouvelles en SSI, qui passent d'abord par la mutualisation des compétences devant une menace multiforme qui s'adapte très vite aux mutations technologiques. Dans cette lutte aujourd'hui inégale entre défenseurs et attaquants, l'analyse des signaux faibles est largement prônée. Les *big data* semblent adaptés à cette détection d'anomalies sur l'échelle dite des sources ouvertes. Ils préparent l'analyse des significations (la sémantique) des affrontements cyber. Ils fournissent un faisceau d'indices permettant aux analystes d'évaluer l'origine des attaques. Ils doivent aussi servir à anticiper les usages malveillants des technologies microrobotiques constitutives de l'Internet des objets. Enfin, ils doivent offrir des simulations dynamiques d'attaques, les plus proches du réel, pour en déduire les mécanismes de contre-réaction les plus pertinents. Plusieurs écueils constitutifs des *big data* sont à éviter ici. Il ne s'agit ni de remplacer la précision des données par leur masse, ni de remplacer la recherche de causes par celle de coïncidences ou de corrélations. Il faut se méfier du retour de certaines illusions bien connues des informaticiens expérimentés, comme l'apprentissage, les réseaux de neurones, voire certains aspects de l'intelligence artificielle dans lesquels les hypothèses implicites (structure du réseau de neurones, biais de la collecte servant à l'apprentissage) ne peuvent être ignorées. Appliqué, par exemple, à l'identification de suspects ou de cibles en sécurité civile, cela semble être porteur de très graves dangers pour les sociétés. Mais la SSI ne se réduit pas, malheureusement, aux architectures de systèmes. L'assemblage de composants sécurisés ne garantit pas la solidité du tout ; au contraire, la complexité facilite le travail de l'attaquant dans la recherche d'un chemin d'attaque. *A contrario*, la monoculture technologique favorise le contrôle centralisé mais cette facilité fragilise également.

Protéger la cyberdiversité

Une analogie s'impose. La diversité des espèces est le plus grand rempart immunitaire contre la perte d'un écosystème. De même, la cyberdiversité, si malmenée par quelques écosystèmes numériques fermés dont aucun n'est européen, reste le constituant principal d'une véritable défense en profondeur.

Éthique des *big data* ?

Un rapport gouvernemental récent affirme qu'il est impératif « d'assurer la sécurité des données ». Pour faire des *big data* un outil de progrès sociétal, par exemple pour les villes intelligentes ou *smart cities* (eau, transports, énergie, commerce électronique), il faut en maîtriser les dérives. On pourrait paraphraser le célèbre *Code is Law (Le Code fait loi)* de Lawrence Lessig par « *Microcode is law in cyberspace* ». La France ou l'Europe voudront-elles revenir dans le jeu technologique ? Une opportunité se présente avec le probable remplacement du silicium par le carbone (graphène). Quoi qu'il en soit, des règles d'éthiques sont à poser. La France et la vieille Europe sont héritières des vertus de « dignité, de réserve et de droiture » (Épictète).
Puissent-elles engager la maîtrise et la domestication des robots logiciels des *big data* sur une régulation s'inspirant de ces principes en gardant l'homme au centre des enjeux. ■