Données personnelles : rien à cacher, mais beaucoup à perdre

Antoine Boutet

Données personnelles : rien à cacher, mais beaucoup à perdre

Antoine Boutet
Univ. Lyon, INSA-Lyon, Inria, CITI, Lyon, France
antoine.boutet@insa-lyon.fr

Nos données personnelles circulent sur Internet : nom, adresses, coordonnées bancaires ou de sécurité sociale, localisation en temps réel... et les affaires qui y sont liées se font une place pérenne dans le débat public, du scandale Facebook-Cambridge Analytica [1] au vol de données à la Croix-Rouge [2], en passant par les récents blocages d'hôpitaux par des rançongiciels [3] (ou ransomware) et l'interdiction de l'application TikTok pour les fonctionnaires de plusieurs pays [4].

Mais si l'on sait de plus en plus que nos données personnelles sont "précieuses" et offrent des possibilités sans précédent en matière de commercialisation et d'innovation, il est parfois difficile de saisir ou d'expliquer pourquoi il faudrait les protéger.

Quels sont les risques liés à la divulgation de mes données personnelles ?

Le premier risque concerne la perte du contrôle sur nos propres données. C'est ce qui arrive par exemple quand on autorise le traçage par des sites ou des applications : on autorise l'enregistrement de nos activités sur le Web ou sur notre smartphone (pages visitées, géolocalisation) et l'échange de ces données, et, une fois cet accord donné, nous n'avons plus aucun pouvoir sur la circulation de nos données.

Ces informations sont utilisées le plus souvent pour du profilage qui permet d'alimenter l'économie de la publicité personnalisée [5] régie dorénavant par des

[1] https://theconversation.com/cambridge-analytica-et-facebook-le-respect-de-votre-vie-privee-nous-tient-a-coeur-oupas-93755

[2] https://www.clubic.com/antivirus-securite-informatique/actualite-405297-la-croix-rouge-se-fait-voler-les-donnees-de-plus-de-500-000-personnes-hautement-vulnerables.html

[3] https://theconversation.com/cyberattaques-des-hopitaux-que-veulent-les-hackers-192407

[4] https://theconversation.com/tiktok-piratage-de-donnees-ou-piratage-des-cerveaux-200923

[5] https://www.cnil.fr/fr/cookies-et-autres-traceurs/regles/cookie-walls/publicite-ciblee-en-ligne-quels-enjeux-pour-la-protection-des-donnees-personnelles

plates-formes d'enchères valorisant les données relatives aux profils utilisateurs contre des emplacements publicitaires.

Mais, ces informations peuvent également être utilisées à mauvais escient. La connaissance de votre localisation peut aider le passage à l'acte d'un cambrioleur par exemple, et la connaissance de vos centres d'intérêts ou opinion politique peut vous exposer à des opérations d'influence [6].

Le scandale Cambridge Analytica en est un exemple, avec l'exploitation de données personnelles de millions d'utilisateurs Facebook pour des campagnes de désinformation ciblées afin d'influencer des intentions de vote. Plus récemment, les révélations du Monde sur les entreprises de désinformation [7] indiquent que cette pratique n'est pas un cas isolé.

Un autre risque concerne l'hameçonnage [8] : si des informations personnelles sont présentes dans un courriel ou SMS frauduleux, il vous paraîtra plus réaliste et abaissera vos barrières de vigilance. L'hameçonnage sert souvent à infecter la cible avec un rançongiciel [9] (ransomware en anglais) : les cybercriminels utilisent des informations personnalisées pour gagner la confiance des destinataires et les inciter à ouvrir des pièces jointes, ou à cliquer sur des liens ou documents malveillants, ce qui permet dans un second temps de verrouiller les données de la victime et d'en interdire l'accès. Une rançon est ensuite réclamée pour les déverrouiller.

Bien que les attaques par rançongiciel les plus médiatisées concernent des organisations, des hôpitaux par exemple, les particuliers sont également touchés [1].

Dans le cas de l'usurpation d'identité, une personne malveillante utilise des informations personnelles qui permettent de nous identifier ("se logger") sans notre accord : par exemple, en créant un faux profil sur une plate-forme et en rédigeant des commentaires sous l'identité de la victime afin de nuire à sa réputation.

bras sortant d'un écran pour voler une carte bancaire L'usurpation d'identité peut nuire à votre réputation, ainsi qu'à votre porte-monnaie. Brian A. Jackson/Shutterstock À un autre niveau, la surveillance de masse exercée par certains États capture les informations personnelles de leurs citoyens afin d'entraver la liberté d'expression ou de ficher les individus par exemple. Une surveillance accrue peut tendre vers un sentiment d'absence de sphère privée et ainsi brider le comportement des individus.

En Europe, le RGPD (règlement général sur la protection des données) limite la récolte des données personnelles, notamment par les gouvernements, qui doivent justifier d'une raison suffisante pour toute surveillance.

[6] https://theconversation.com/comment-lusage-de-vos-donnees-peut-influencer-les-elections-140001

[7] https://www.lemonde.fr/pixels/article/2023/02/16/les-reseaux-sociaux-pierre-angulaire-des-operations-d-influence-et-d-intoxication_6161995_4408996.html

[8] https://theconversation.com/cyberattaques-des-hopitaux-que-veulent-les-hackers-192407

[9] https://theconversation.com/rancongiciels-vos-donnees-en-otage-159975

Chacun d'entre nous a une empreinte numérique unique

Ces problèmes touchent chacun d'entre nous. En effet, dans un monde de plus en plus numérique où nous générons quotidiennement des données à travers notre navigation sur Internet, nos smartphones, ou nos montres connectées, nous avons tous une "empreinte numérique unique".

En clair, il est généralement possible de réidentifier quelqu'un juste à partir des "traces" que nous laissons derrière nous sur nos appareils numériques.

Par exemple, l'observation aléatoire de quatre lieux visités seulement représente une signature unique pour 98% des individus [2, 3]. Cette unicité est généralisable dans un grand nombre de comportements humains.

Cacher l'identité du propriétaire de données personnelles uniquement derrière un pseudonyme n'est pas une protection suffisante face au risque de réidentification, il est nécessaire d'anonymiser les données.

Données synthétiques, apprentissage fédéré : les nouvelles méthodes pour protéger les données personnelles

Tels les membres d'un "black bloc" essayant d'être indistinguables entre eux en s'habillant de manière identique dans une manifestation houleuse, l'anonymisation de données a pour but d'éviter qu'une personne ne se démarque du reste de la population considérée, afin de limiter l'information qu'un cyberattaquant pourrait extraire.

Dans le cas de données de géolocalisation, on pourrait par exemple modifier les données afin que plusieurs utilisateurs partagent les mêmes lieux visités, ou alors introduire du bruit pour ajouter une incertitude sur les lieux réellement visités.

Mais cette anonymisation a un coût car elle "déforme" les données et diminue leur valeur : une trop grande modification des données brutes dénature l'information véhiculée dans les données anonymisées. De plus, pour s'assurer de l'absence d'une empreinte réidentifiante, les modifications nécessaires sont très importantes [4] et souvent incompatibles avec nombre d'applications.

Trouver le bon compromis entre protection et utilité des informations anonymisées reste un challenge. À l'heure où certains voient les données comme le nouveau pétrole du XXIe siècle, l'enjeu est de taille car une donnée anonyme n'est plus considérée comme une donnée personnelle et échappe au RGPD, ce qui veut dire qu'elle peut être partagée sans consentement du propriétaire.

Cette difficulté de trouver un compromis acceptable entre protection et utilité des données au travers de mécanismes d'anonymisation a fait évoluer les pratiques. De nouveaux paradigmes de protection des données personnelles ont vu le jour.

Une première tendance consiste à générer des données synthétiques reproduisant les mêmes propriétés statistiques que les vraies données.

Ces données générées de manière artificielle ne sont par conséquent pas liées à une personne et ne seraient plus encadrées par le RGPD. Un grand nombre d'entreprises voient en cette solution des promesses de partage d'information moins limitées. En pratique, les risques résiduels des modèles de génération synthétique ne sont pas négligeables et sont encore à l'étude [10].

Une autre solution limitant le risque de partage de données personnelles est l'apprentissage fédéré. Dans l'apprentissage machine conventionnel, les données sont centralisées par une entité pour entraîner un modèle.

Dans l'apprentissage fédéré, chaque utilisateur se voit attribuer un modèle qu'il entraîne localement sur ses propres données. Il envoie ensuite le résultat à une entité qui s'occupe d'agréger l'ensemble des modèles locaux. De manière itérative, cet apprentissage décentralisé permet de créer un modèle d'apprentissage sans divulguer de données personnelles.

Ce nouveau paradigme de protection des données personnelles suscite beaucoup d'engouement [11]. Cependant, plusieurs limitations subsistent, notamment sur la robustesse face aux acteurs malveillants qui souhaiteraient influencer le processus d'entraînement. Un participant pourrait par exemple modifier ses propres données pour que le modèle se trompe lors d'une tâche de classification particulière [12].

References

[1] A. Kujawa, W. Zamora, J. Umawing, J. Segura, W. Tsing, M. Rivero, Hasherezade, C. Boyd, P. Arntz, and D. Ruiz, "Cybercrime tactics and techniques: Ransomware retrospective," 2019. [Online]. Available: https://www.malwarebytes.com/wp-content/uploads/sites/2/2023/09/ctnt-2019-ransomware_august_final.pdf

[2] A. Boutet, S. Ben Mokhtar, and V. Primault, "Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets," LIRIS UMR CNRS 5205, Research Report, Oct. 2016. [Online]. Available: https://hal.science/hal-01381986

[3] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013. [Online]. Available: http://dx.doi.org/10.1038/srep01376

[4] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th*

[10] https://files.inria.fr/ipop/

[11] https://www.emergenresearch.com/industry-report/federated-learning-market

[12] https://www.lemonde.fr/pixels/article/2021/09/04/des-personnes-noires-confondues-avec-des-singes-par-un-algorithme-de-facebook_6093366_4408996.html

Annual International Conference on Mobile Computing and Networking, ser. MobiCom '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 145–156. [Online]. Available: https://doi.org/10.1145/2030613.2030630