LES RISQUES DE L'ANONYMISATION ET DE L'AGRÉGATION DE DONNÉES

Publié le 6 décembre, 2021

Catégories: Perspectives, Publications

Dans notre économie numérique, bon nombre de décisions d'affaires se fondent sur des données. Or, l'utilisation des données est étroitement encadrée, surtout lorsqu'elles permettent d'identifier une personne. Les entreprises cherchent donc de nouvelles façons de tirer profit des données tout en respectant les exigences de protection des renseignements personnels. Ainsi, elles ont recours à l'agrégation des données pour diverses raisons, notamment pour améliorer leurs processus de développement de produits et de marketing. Elles peuvent aussi anonymiser les données recueillies pour être dispensées des exigences en matière de protection des renseignements personnels. Mais attention : ces deux méthodes n'éliminent pas le risque de violation de la vie privée; elles procurent plutôt un faux sentiment de sécurité. En effet, on n'imagine pas que des données anonymisées et agrégées formant un groupe de points de données non identifiées puissent poser un risque. Nous verrons dans le présent bulletin les facteurs à prendre en considération afin d'éviter les pièges liés à l'utilisation de telles données.

Qu'est-ce qu'un renseignement personnel?

Partout dans le monde, des lois régissent la collecte, l'utilisation et la communication des données permettant d'identifier une personne[1]. Par exemple, la *Personal Information Protection Act* (PIPA) de la Colombie-Britannique protège les données qui entrent dans la catégorie des renseignements personnels[2], soit ceux concernant un individu identifiable, dont les principaux sont le nom, l'âge, l'adresse, les empreintes digitales, l'origine ethnique et l'état matrimonial[3]. La *Loi sur la protection des renseignements personnels et les documents électroniques* (LPRPDE) du gouvernement fédéral[4] applique la même définition des renseignements personnels.

Les risques et les limites de l'anonymisation des données

L'anonymisation, ou dépersonnalisation désigne le processus qui consiste à supprimer les renseignements permettant d'identifier une personne ou son domicile des données recueillies[5]. Néanmoins, il est souvent possible de réidentifier les données anonymisées, en y associant des informations qui permettent d'identifier la personne concernée. Plusieurs méthodes contribuent à réduire ce risque, par exemple l'introduction d'un

« bruit blanc » statistique en vue d'obscurcir les liens entre les éléments de données, ou encore le brouillage des données pour en restreindre l'accès[6].

Mais il n'est pas facile de trouver le juste équilibre en matière d'anonymisation. En effet, plus les données sont anonymisées, plus elles sont protégées, mais leur utilité est réduite en conséquence. Il s'agit donc de déterminer le degré optimal d'anonymisation entre les deux extrêmes[7].

Si l'anonymisation s'avère efficace pour protéger la vie privée, elle est toutefois impraticable pour certains types de données, vu l'étendue de la notion de « renseignement personnel ». Citons, entre autres, les données recueillies par les dispositifs de maison intelligente, difficiles à anonymiser par les méthodes traditionnelles. On peut obscurcir les données audio et vidéo, séparer ou chiffrer les profils numériques contenant des identificateurs principaux, mais la nature même des données récoltées par les capteurs complique leur dépersonnalisation.

Les données provenant de capteurs constituent un ensemble d'informations sur les activités de l'utilisateur dont on ne peut facilement supprimer ou obscurcir les renseignements personnels[8]. Par conséquent, les imperfections et irrégularités des capteurs rendent ces données plus susceptibles de réidentification[9]. En effet, les capteurs ont de légères différences entre eux, et chacun a ses propres petits défauts qui, à l'instar d'empreintes digitales, permettent de reconnaître les données qui en sont issues.

Les incidences de l'agrégation des données

Tout comme l'anonymisation, l'utilisation et la communication de données agrégées comportent des risques. La plupart des données que récoltent les dispositifs de maison intelligente, les technologies prêt-à-porter et celles de l'Internet des objets (IDO) ne permettent pas directement l'identification, mais n'en demeurent pas moins très personnelles et peuvent constituer un profil identificatoire sous forme agrégée. Le but de la collecte de telles données est intimement lié à la fonction de la plupart des dispositifs IDO : mieux connaître les comportements, les habitudes et les préférences de l'utilisateur[10]. Cette masse de données combinées trace cependant un portrait de l'utilisateur qui peut fournir des renseignements personnels permettant de l'identifier.

À partir de données agrégées regroupant nombre de points de données discrètes sur une personne en particulier, on peut tirer des conclusions substantielles et étonnantes sur ses comportements et habitudes, alors qu'elle n'avait aucune intention de communiquer de tels renseignements[11]. Ces conséquences indésirables sont exacerbées par les progrès de l'intelligence artificielle (IA), grâce auxquels les responsables du traitement des données dégagent des tendances et établissent des liens autrefois inconcevables pour les experts en science des données[12].

Le phénomène connu sous le nom de « fusion de capteurs » prendra de l'importance avec l'adoption croissante des technologies IDO et leur multiplication dans les maisons. La fusion de capteurs consiste à combiner les données de deux capteurs pour obtenir plus d'informations et, peut-être, en tirer des conclusions inattendues[13]. Il peut également s'agir d'utiliser le capteur d'un dispositif d'IDO à des fins autres que celles prévues à l'origine, particulièrement en association avec d'autres dispositifs[14]. Ce phénomène soulève des inquiétudes légitimes quant à la possibilité pour la personne concernée de donner son consentement éclairé à l'utilisation non prévue des données lorsqu'elle n'a pu en être informée à l'avance en bonne et due forme, en plus de poser des risques pour les entreprises qui vendent ou intègrent ces technologies.

Et ces risques demeurent malgré la dépersonnalisation des données, principalement en raison de leur nature distinctive, qui facilite l'identification de la personne concernée[15]. Le Commissariat à la protection de la vie privée (CPVP) a d'ailleurs exprimé ses critiques à l'endroit des technologies qui anonymisent les données à des étapes précises de leur utilisation en prétendant garantir l'anonymat, alors qu'il demeure possible, quoique peu probable, d'identifier une personne[16]. En outre, la Cour suprême du Canada a rendu un avis formel : pour répondre aux attentes raisonnables d'un utilisateur quant à la protection de ses renseignements personnels, il ne suffit pas de considérer isolément chaque point de données; il faut tenir compte du risque que l'ensemble des données puisse révéler les habitudes et les choix d'une personne[17].

Peut-on librement utiliser des données anonymisées et agrégées?

On peut librement utiliser et communiquer les données anonymisées selon les règles, qu'elles soient agrégées ou non, mais la possibilité de recueillir des renseignements personnels à partir des données anonymisées et agrégées pose un risque pour leur utilisation à des fins commerciales, car il reste possible de les réidentifier. Les lois sur la protection de la vie privée en vigueur partent du principe qu'on peut faire la distinction entre des « renseignements permettant d'identifier une personne » et des données anonymisées ou agrégées[18], mais les entreprises n'en sont pas pour autant à l'abri des risques connexes.

On estime que 99,98 % des données anonymisées sont vulnérables à la réidentification, et plus encore lorsqu'elles sont agrégées, comme on l'a vu plus haut[19]. Il est difficile de déterminer si les lois canadiennes applicables tiennent compte de ces risques et, le cas échéant, de quelle façon. On observe une tendance mondiale à l'intégration des données réidentifiables aux lois et règlements protégeant la vie privée. En Europe, le *Règlement général sur la protection des données* (RGPD) considère que les données « pseudonymes », soit les données ne contenant pas d'identifiants, mais susceptibles d'être réidentifiées, relèvent du champ d'application de la loi[20].

De récentes modifications à la *Freedom of Information and Protection of Privacy Act* de la Colombie-Britannique traduisent une volonté d'accorder aux entreprises une plus grande latitude et un meilleur

avantage concurrentiel[21]. De son côté, le gouvernement du Canada semble plutôt vouloir s'aligner sur le RGPD. Ainsi, le fédéral a proposé d'ajouter une disposition interdisant la réidentification des données dans la *Loi sur la protection de la vie privée des consommateurs* (LPVPC), sans toutefois préciser clairement si les données dépersonnalisées sont assujetties à cette loi[22]. Le déclenchement d'élections en septembre 2021 a freiné l'adoption du projet de loi. Depuis, le gouvernement n'a toujours pas réitéré sa proposition. Impossible pour l'instant de savoir si la réidentification des données fera l'objet d'une interdiction, mais le CPVP a laissé entendre que les données pseudonymisées pourraient être visées par l'ensemble des dispositions de la LPRPDE[23].

Nombre d'entreprises tentent d'atténuer ce risque en limitant l'utilisation, la vente et la communication à un petit sous-ensemble de données, sous prétexte qu'à partir d'un ensemble partiel de données, on ne peut avoir la certitude que la personne réidentifiée est bien celle visée[24]. Or, ce risque existe également lorsque l'ensemble de données est très incomplet[25]. Les entreprises doivent donc choisir soigneusement la méthode d'anonymisation des données afin de réduire le risque de réidentification et, par conséquent, de se prémunir contre d'éventuelles poursuites liées à la collecte, à l'utilisation et à la communication de données.

Les lois et obligations relatives à la protection des renseignements personnels et les techniques d'anonymisation des données ne cessent d'évoluer. Difficile alors pour une entreprise de déterminer avec certitude si la méthode utilisée restera acceptable à l'avenir. Il sera donc très important pour les entreprises de prendre toutes les mesures raisonnables pour assurer leur conformité aux lois sur la protection des renseignements personnels en analysant rigoureusement chaque possibilité ou méthode suggérée à la lumière des exigences en vigueur et en menant un examen approfondi des façons dont les données anonymisées ou agrégées pourraient faire l'objet de réidentification.

Si vous avez des questions ou des préoccupations concernant votre utilisation de données anonymisées ou agrégées, nous vous invitons à consulter notre équipe spécialisée dans la protection de la vie privée et des données.

[1][ps2id id='1' target='/]Charlotte A. Tschider, « Regulating the Internet of Things: Discrimination, Privacy, and Cybersecurity in the Artificial Intelligence Age » (2018), 96:1 *Denv U Law Rev* 87, p. 104, 107.
[2][ps2id id='2' target='/] *Personal Information Protection Act*, SBC 2003, ch. 63 [PIPA], art. 6 à 9.
[3][ps2id id='3' target='/] *Ibid.*, art. 1; Ministry of Citizens' Services, « Guide to the *Personal Information Protection Act* », en ligne : *Office of the Chief Information Officer of British Columbia*.
[4][ps2id id='4' target='/] *Loi sur la protection des renseignements personnels et les documents électroniques*, L.C. 2000, ch. 5, par. 2(1).
[5][ps2id id='5' target='/] Gilad Rosner, « De-Identification as Public Policy » (2020), 3:3 *Journal of Data Protection & Privacy* 1, p. 3, 4.

[6][ps2id id='6' target='/']Tschider, *supra*, note 1, p. 105.
[7][ps2id id='7' target='/'] Le juste milieu entre l'anonymisation et l'utilisation libre relève du « principe de Boucles d'or »; voir Rosner, *supra*, note 5, p. 7.
[8][ps2id id='8' target='/'] Tschider, *supra*, note 1, p. 107.
[9][ps2id id='9' target='/'] Scott R. Peppet, « Regulating the Internet of Things: First Steps Towards Managing Discrimination, Privacy, Security, and Consent », (2014), 93 *Texas Law Rev* 85, p. 93, 94.
[10][ps2id id='10' target='/'] *Ibid.*, p. 16.
[11][ps2id id='11' target='/'] Peppet, *supra*, note 9, p. 121 et 122.
[12][ps2id id='12' target='/'] Tschider, *supra*, note 1, p. 96.
[13][ps2id id='13' target='/'] Peppet, *supra*, note 9, p. 93.
[14][ps2id id='14' target='/'] *Ibid.*, p. 121.
[15][ps2id id='15' target='/'] Peppet, *supra*, note 9, p. 128, 129.
[16][ps2id id='16' target='/'] *Examen des répercussions sur la vie privée de l'application Alerte COVID*, Commissariat à la protection de la vie privée du Canada, 31 juillet 2020.
[17][ps2id id='17' target='/'] *R.* c. *Spencer*, 2014 CSC 43.
[18][ps2id id='18' target='/']Peppet, *supra*, note 9, p. 94.
[19][ps2id id='19' target='/'] Luc Rocher, Julien M. Hundrickx et Yves-Alexandre de Montjoye, « Estimating the success of re-identification in incomplete data sets using generative models » (2019), 10 *Nature Communications* 3069; voir aussi Arvind Narayanan et Vitaly Shmatikov, « Robust De-anonymization of Large Sparse Datasets » (2008), *IEEE Symposium on Security and Privacy*, p. 111 à 125 et Arvind Narayanan et Vitaly Shmatikov, « Robust De-anonymization of Large Sparse Datasets: a Decade Later » (2019), rapport de recherche non publié, en ligne : *PDF*.
[20][ps2id id='20' target='/'] Voir le *Règlement général sur la protection des données* (UE), 2016/679, considérant 75; Groupe de travail « article 29 » sur la protection des données, « Avis 05/2014 sur les Techniques d'anonymisation » (2014), en ligne: *Commission européenne*, p. 10.
[21][ps2id id='21' target='/'] Projet de loi 22, *Freedom of Information and Protection of Privacy Amendment Act, 2021*, 2e session, 42e législature, Colombie-Britannique, 2021 (première lecture) [projet de loi 22].
[22][ps2id id='22' target='/'] Voir le projet de loi C-11, *Loi édictant la Loi sur la protection de la vie privée des consommateurs et la Loi sur le Tribunal de la protection des renseignements personnels et des données et apportant des modifications corrélatives et connexes à d'autres lois*, 2e session, 43e législature, Canada, 2020 (première lecture), art. 75.
[23][ps2id id='23' target='/'] *Document de discussion sur les améliorations possibles au consentement sous le régime de la Loi sur la protection des renseignements personnels et les documents électroniques*, Groupe des politiques et de la recherche du Commissariat à la protection de la vie privée du Canada, mai 2016.

[24][ps2id id='24' target='/] Voir Gregory J. Matthews et Ofer Harel, « Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy » (2011), *Stat Surv* 5, 1–29 (2011); Daniel Barth-Jones, « The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now » (2012), en ligne : *SSRN Electronic Journal*.
[25][ps2id id='25' target='/] Rocher, Hundrickx et de Montjoye, *supra*, note 18, p. 2.

Par Robert C. Piasentin et Kristen Shaw (stagiaire en droit)

Mise en garde