RISKS OF ANONYMIZED AND AGGREGATED DATA

Posted on December 1, 2021

Categories: Insights, Publications

Data drives many business decisions in today's digital economy. How that data is used is facing greater scrutiny, in particular when that data can identify specific individuals. As a result, businesses are seeking alternative ways to use data in a way that, they hope, will allow them to continue to reap the benefits of using such data while also staying on the right side of all applicable privacy requirements. Many businesses, for example, use technology to aggregate data for a number of reasons including making their marketing and product development processes more efficient and effective. Relatedly, companies will often seek to anonymize the data they collect in order to try to avoid the application of privacy requirements. However, simply using anonymized and/or aggregated data does not insulate a business from the risk of privacy violations, it may instead just give a business a false sense of security with respect to that risk. If a business anonymizes and simply aggregates collected data into a group of unidentified data points, how can it be at risk? In this bulletin, we will touch on the risks and considerations that a business should focus on when using such data in its operations.

What Does It Mean To Be Identifiable?

The restrictions on the collection, use, and disclosure of data in privacy laws across the globe are triggered when data can be used to identify a specific person.[1] For example, British Columbia's *Personal Information Protection Act* ("***PIPA***") provides protection for information which falls within the definition of "personal information".[2] Personal information is defined as "information about an identifiable individual" and it is generally thought to include primary identifiers such as one's name, age, address, fingerprints, ethnic origin, and marital status.[3] Canada's federal privacy legislation, the *Personal Information Protection and Electronic Documents Act* ("***PIPEDA***"),[4] applies the same concept regarding personal information.

Risks and Challenges of Anonymization

Anonymization, or de-identification, refers to a process that removes information capable of identifying individuals or their households from collected data.[5] The risk with anonymizing data is that it can often be re-identified – where anonymized data is matched with available information to discover the individual to whom it belongs. However, there are a number of practices that can be used to help reduce the risk of re-identification. For example, statistical "white noise" can be introduced to obscure the connections between

data elements, or obfuscation can render data less accessible.[6] Many organizations struggle with finding the right balance of anonymization largely because, while greater anonymization of data affords better privacy protection, the usefulness of that data is correspondingly reduced. The trick is finding an optimal state between the two extremes.[7]

While anonymizing data is a strong start to avoiding violating an individual's privacy, "personal information" is often defined quite broadly such that certain types of data are not truly capable of being anonymized. For example, sensor data collected from passive smart home devices poses particular challenges to traditional methods of anonymization. While voice or video data can be obscured, and digital profiles containing primary identifiers can be segregated or encrypted, the nature of sensor data makes it challenging to de-identify. Sensor data is a collection of a user's activities where specific personally identifiable elements cannot be easily removed or obscured.[8] As a result, sensor data is more prone to re-identification due to the unique imperfections and irregularities within the sensor.[9] Basically, sensors are susceptible to having slight flaws or differences between them, and those flaws can act like a fingerprint to identify data that comes from a particular device.

The Implications of Aggregate Data

Alongside the risks of anonymization come the risks of such data being used and disclosed in the aggregate. Much of the data collected by smart home devices, wearables, and other Internet of Things ("**IoT**") technologies is not directly identifiable, but still may be deeply personal, and may create an identifiable profile when aggregated. The purpose for the collection of such data is crucially linked to the function of most IoT devices – to better understand the behaviour, habits, and preferences of the user.[10] The combined mass, however, creates a picture of the user that can lead to identifiable personal information for a specific individual.

Aggregated data, which combines various discrete data points specific to a particular individual, can provide substantial and surprising inferences about private behaviours and habits that an individual never intended to share.[11] These unintended consequences are exacerbated by the developments in AI that allow data processors to extract data trends and relationships that were previously inconceivable by data scientists.[12]

One phenomenon, known as "sensor fusion", will likely become more prominent as the market uptake of IoT devices increases and their presence multiplies within the home. Sensor fusion is where data from two sensing devices can reveal greater information, and perhaps unexpected inferences, when that data is combined.[13] This phenomenon may also mean that a sensor within an IoT device is used for purposes beyond its intended and original use, particularly when used alongside other IoT devices.[14] Sensor fusion raises legitimate concerns regarding whether an individual has provided or can provide informed consent, where unintended uses could not be adequately communicated to the user in advance, and creates risks for those selling and

incorporating such technologies in their businesses.

These risks remain even when information is de-identified, largely due to the fact that the distinctive nature of this data makes it relatively easy to identify the individual to whom the data belongs.[15] In fact, the Office of the Privacy Commissioner ("**OPC**") has been critical of an approach which characterizes technologies which anonymize data at particular points in their use as offering anonymity where identification of an individual, while highly improbable, is not impossible.[16] Further, the Supreme Court of Canada has made it clear that when considering a user's reasonable expectation of privacy, it is not enough to only consider each data point in isolation, but consider what the whole may reveal about the personal habits and choices of the individual behind the data.[17]

Can You Freely Use Anonymized And Aggregated Data?

While truly anonymized data, whether in an aggregated form or not, can be freely used and shared, the ability to glean personal information from both anonymized and aggregated data creates risks for using and disclosing such data for commercial purposes because there is always a risk of re-identification. Privacy laws currently rely on the assumption that it is possible to distinguish between what is "personally identifiable information" and anonymized or aggregated data,[18] however this assumption does not entirely absolve a company from risk.

Approximately 99.98% of anonymized data may be capable of re-identification and, as explored above, the risks of re-identification are heightened when data is aggregated. [19] It is currently uncertain whether, and how, Canadian privacy legislation may consider these risks. There is a global trend towards incorporating re-identifiable data under privacy protections. The GDPR, for example, considers "pseudonymous data", which is data that does not contain direct identifiers but is capable of re-identification, as being within the scope of the law.[20]

In British Columbia, however, recent amendments to the *Freedom of Information and Protection of Privacy Act* indicate a willingness for legislation that gives business flexibility and a greater competitive edge.[21] Federally, on the other hand, it appears there may be some willingness to follow the GDPR's lead. The federal government had proposed to introduce a prohibition against re-identifying data in the *Consumer Privacy Protection Act* ("***CPPA***"), but was not clear whether de-identified data would be subject to the *CPPA*.[22] Due to the calling of the September 2021 election, the *CPPA* was not passed into law. As the federal government has not yet reintroduced similar legislation following the election, we cannot say with certainty at this time whether there will again be a prohibition against re-identifying data, but the OPC has suggested that pseudonymous data could fall within the current provisions of *PIPEDA*.[23]

Many companies are attempting to mitigate this risk by using, selling, or otherwise sharing only a small subset

of data, arguing that by providing incomplete data sets, those seeking to re-identify an individual related to such data set cannot be sure the right person was identified.[24] However, these risks can arise even where the data set is largely incomplete.[25] Thus, companies who collect and use anonymized data should consider the means by which they are anonymizing their data to reduce the risk of re-identification and, in turn, potential liability for its collection, use, and disclosure.

The legislation and requirements around the protection of personal information and the techniques available to anonymize such personal information are constantly evolving. As a result, it is difficult for businesses to currently know for certain whether a particular approach will be acceptable going forward. Moving forward, it is important that businesses carefully analyze each opportunity or suggested approach in light of the current requirements and with a full review and assessment of the potential ways in which any such anonymized or aggregated data may be re-identified to ensure that it has taken all reasonable steps to remain in compliance with all privacy requirements.

If you have any questions or concerns regarding your business' use of anonymized and/or aggregated data, we recommend reaching out to our Privacy and Data Protection team.

[1] 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 at 104, 107.
[2] *Personal Information Protection Act*, SBC 2003, c 63 [PIPA], ss 6 – 9.
[3] *Ibid*, s 1; Ministry of Citizens' Services, "Guide to the *Personal Information Protection Act*", available online: *Office of the Chief Information Officer of British Columbia*.
[4] *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5, s 2(1).
[5] Gilad Rosner, "De-Identification as Public Policy" (2020) 3:3 Journal of Data Protection & Privacy 1 at 3 – 4.
[6] Tschider, *supra* note 1 at 105.
[7] The optimal state between anonymization and open use is often referred to as the "Goldilocks principle"; see Rosner, *supra* note 5 at 7.
[8] Tschider, *supra* note 1 at 107.
[9] Scott R Peppet, "Regulating the Internet of Things: First Steps Towards Managing Discrimination, Privacy, Security, and Consent" (2014) 93 Texas Law Rev 85 at 93 – 94.
[10] *Ibid* at 16.
[11] Peppet, *supra* note 9 at 121 – 122.
[12] Tschider, *supra* note 1 at 96.
[13] Peppet, *supra* note 9 at 93.
[14] *Ibid* at 121.
[15] Peppet, *supra* note 9 at 128 – 129.

[16] *Privacy review of the COVID Alert exposure notification application*, Office of the Privacy Commissioner of Canada, July 31, 2020.
[17] *R v Spencer*, 2016 SCC 43.
[18] Peppet, *supra* note 9 at 94.
[19] Luc Rocher, Julien M. Hundrickx & Yves-Alexandre de Montjoye, "Estimating the success of re-identification in incomplete data sets using generative models" (2019) 10 *Nature Communications* 3069; see also Arvind Narayanan & Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets" (2008), IEEE Symposium on Security and Privacy, pp 111–125 and Arvind Narayanan & Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets: a Decade Later" (2019) unpublished research paper, available online: *PDF*.
[20] See *General Data Protection Regulation* (EU), 2016/679, recital 75; Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques" (2014), Technical Report, Article 29, online: *European Commission* at 10.
[21] Bill 22, *Freedom of Information and Protection of Privacy Amendment Act, 2021*, 2nd Sess, 42nd Parl, British Columbia, 2021 (first reading) [Bill 22].
[22] See Bill C-11, *An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts*, 2nd Sess, 43rd Parl, Canada, 2020 (first reading) s 75.
[23] *A discussion paper exploring potential enhancements to consent under the Personal Information Protection and Electronic Documents Act*, Policy and Research Group of the Office of the Privacy Commissioner of Canada, May 2016
[24] See Gregory J Matthews & Ofer Harel, "Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy" (2011) 5 Stat Surv 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), online: *SSRN Electronic Journal*.
[25] Rocher, Hundrickx & de Montjoye, *supra* note 18 at 2.

by Robert Piasentin and Kristen Shaw (Articled Student)

A Cautionary Note

The foregoing provides only an overview and does not constitute legal advice. Readers are cautioned against making any decisions based on this material alone. Rather, specific legal advice should be obtained.