The Ignorance of Anonymisation to Protect Privacy

Privacy Foundation New Zealand

Privacy in Internet Economy Working Group

Marcin Betkier, Natasha Mazey

Opportunities

Organisations collect large amounts of data from individuals in the course of business. They often collect that data using a range of methods from surreptitiously or indirectly monitoring online activities (e.g. via access to video, audio or photographic content) to direct observations and overt requests for information (e.g. in order to proceed through customs and border control, survey participation). Data sourced by direct means are usually collected and used in a fair and transparent way and generally serves a legitimate primary purpose. This is a common legal requirement across the world, including the new New Zealand Privacy Act 2020 which expressly states organisations should not collect identifying information when it is not required for their lawful purpose.[1]

Personal data may often have a 'second life' being reused for other purposes. The incentive for this comes with the increasing demand for big data analytics to identify new trends, opportunities and inefficiencies. This is coupled with the evolution of technologies which support the convenience and accessibility of collecting individuals' personal information and availability of rich data insights such as artificial intelligence, machine learning and predictive behaviour analytics. Previously collected data could be a goldmine of opportunity for organisations looking to innovate new products and services through an informed, data driven research and development methodology or monitoring trends in relevant social issues. Examples may include data about Covid-19 or flu spread, child poverty, or access and use of sexual health facilities. There is an increasing trend for global laws to place responsibilities on organisations to use personal information in ways that is not collected for its primary purpose for collection. Often this includes ensuring individuals are "not identified" (Principle 10 of the New Zealand Privacy Act 1993, unchanged in the new Act of 2020).[2]

Risks

Ensuring individuals are not identifiable can be difficult to achieve guarantee using a standardised approach. Legal interpretation of these requirements can also be ambiguous. However, to take advantage of these opportunities and advance

[1] See s 22 of Privacy Act 2020, Information privacy principle 1(2).

[2] See s 22 of Privacy Act 2020, Information privacy principle 10 (1)(b).

technologies, products and services while balancing individuals' privacy rights and interests, private and public organisations need to be mindful of:

- Breach of purpose limitation principles, which is a foundation of New Zealand Information Privacy Principles 1 and 10 and many other global privacy laws, by using personal data for secondary purposes for which individuals have not been informed or have not given their consent in an informed or meaningful way.
- Potential privacy risks and harms to individuals and organisations arising from the use, and possible misuse, of their personal information. These include: embarrassment, loss of opportunities, discrimination, identity fraud, surveillance risks or even physical harm for individuals; operational business, reputational, regulatory and financial risks for organisations, and alignment with corporate responsibility and ethic commitments.
- Ensuring appropriate and effective de-identifying techniques and controls which consider the specific nature of data collected and re-identification risk for as long as the data is retained.

Considering these factors, organisations need to carefully evaluate and minimise privacy risks and harms in consideration of existing and proposed de-identification or anonymisation techniques. Appropriate use of these techniques can balance social, business and individual privacy interests.

However, close observation of the use of personal information in the COVID-19 pandemic, contact tracing procedures and the spotlight on health data in recent months[3] has highlighted a gap in our understanding and assumptions for what level of protection is afforded when personal data is "de-identified" or "anonymised", and what these terms actually mean.

De-identification vs anonymisation

The terms 'de-identification' and 'anonymisation' are both becoming increasingly popular as privacy regulations continue to mature, and privacy expectations increase. However, we don't seem to appreciate the difference of the two concepts.

De-identification involves removing or disguising data of the elements that can identify individuals on their own or in combination with other data elements. Anonymisation involves transforming the data into a state in which it is unable to identify individuals or be reverse engineered to enable identification. In other words, it should not be possible to reverse the process of anonymisation itself, or to identify individuals by combining and matching the anonymised data with other datasets.

The difficulty of achieving "anonymised" data is often underestimated and increases as we advance our technologies and accumulated datasets. This leads to two key issues appearing when organisations incorrectly classify data as anonymous. Firstly,

[3] See, for example, Henry Cooke and Thomas Manch "National MP Hamish Walker admits passing on leaked Covid-19 patient info from former party president Michelle Boag" (7 July 2020) Stuff.co.nz <www.stuff.co.nz>; also, Privacy Commissioner *Inquiry into Ministry of Health disclosure of Covid-19 Patient Information* (2020).

they assume privacy risks are largely, if not completely, mitigated. Consequently, appropriate safeguards are not put in place to continue to protect individuals' privacy interests.[4] Secondly, organisations naively provide false assurances to individuals on which individuals then rely to inform their choices and actions.

Believing that personal data is perfectly anonymised and providing assurances about this is likely to aggravate individuals' distress upon being notified of a privacy breach. They may be surprised to learn that information that indirectly relates to them was exposed (e.g. credit card number, phone number) or that they are a victim of identity theft, fraud, financial theft, physical harm or inappropriate surveillance. It may happen that individuals are not notified of this exposure, because organisations believe breached privacy poses no direct risk because it was "anonymised". In such cases, they are left failing to comprehend the extent of their information exposure, how the breach happened and without sufficient knowledge to further safeguard themselves to mitigate further risk or harm.

To illustrate the ongoing misunderstanding of anonymisation, on 25th June 2020 the New Zealand Ministry of Health assured the public that information about COVID-19 case management and quarantine activities was "anonymised" with the use of NHI numbers to identify and track New Zealanders and other individuals entering New Zealand. They stated the NHI "allows us to anonymise health information linked to patients and then allows patients or their tests or results to be traced and tracked by health services."[5] However, this information is not anonymised; it is de-identified. Health services can 're-identify' data and match this to a single individual, along with their other sensitive health records, address and contact information.

Use of the NHI does not protect individuals' privacy in the event of exposure or breach of COVID-19 records, it merely adds an additional step to individually identify an individual and access their associated records. The knowledge and use of NHI information in relation to COVID-19 management and quarantine activities may still present risks to individuals without directly identifying them, such as phishing and fraud depending on the users' motives.

The failure of perfect anonymisation

The main problem we face is a common misconception that anonymisation only requires 'de-identification' with the removal of certain limited data elements such as first and last name, or in some cases public identifiers such as driver's licence number, health numbers (e.g. NHI) or IRD numbers.

This thinking, unfortunately, is flawed. There have been well known failures of anonymisation when datasets were released to the public, such as de-identification

[4] See also the 'release-and-forget model in Paul Ohm "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (2010) 57 UCLA Law Review 1701 at 1707 ff.
[5] Ministry of Health media release, 25 June 2020; "3 new cases of COVID-19"; https://www.health.govt.nz/news-media/media-releases/3-new-cases-covid-19-4

of Australian Medicare data released in 2016.[6] Perfect anonymisation is very difficult to attain.

Anonymisation requires an evaluation of an entire dataset and the different combinations of data which are reasonably identifiable.[7] This assessment also needs to be applied at an organisational level, considering other datasets in their control which could be combined and matched for re-identification. Organisations need to consider the possibility of the data being released or exposed to other unintended users and evaluate their likely and reasonable access to other data or knowledge which may be available to identify the dataset. In a recent example, a COVID-19 patient was identified based only on the name of the town of residence and their association to a cluster of COVID cases relating to the Hereford cattle conference. This individual was the only Hereford breeder in their town. This was reasonably public knowledge and local community members of the town were aware of this fact.[8] Alternatively, in a business context, a remuneration report for a company which reports salaries by role would not be anonymous for the only person with the role "Diversity and Inclusion Advisor." This could be easily re-identified using the company name and role in a LinkedIn search.

True anonymisation requires an irreversible process which prevents re-identification. It requires a current state assessment, and future assessments to test whether anonymisation still exists over time as more data is collected by an organisation, or technologies advance which may make re-identification processes more accessible and likely. For example, "anonymous" data about New York taxi trips were re-identified within an hour because the standard one-way hash function (MD5) was easily reversed because the taxi numbers and licence numbers conformed to a specific, well-known pattern.[9] Anonymisation should be assumed to be a moving target over time. As a result, organisations cannot simply 'anonymise' data and forget about protecting it. Privacy and data protection laws will still apply to the information (or data) as long as it relates to an identifiable individual.

Can we rely on de-identification?

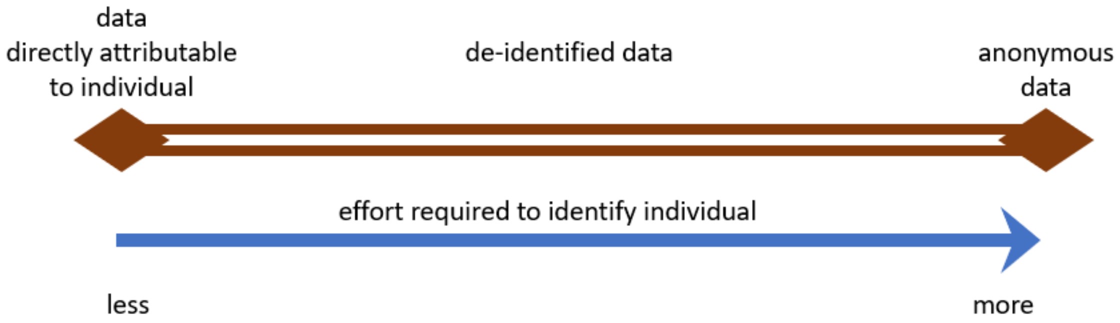
There is a continuum between data that is directly identifiable (attributable) to the individual, and data that is anonymous. The middle space is generally "de-identified" data. It varies in the level of protection is may offer.

[6] "Not so anonymous: Medicare data can be used to identify individual patients, researchers say - ABC News" (18 December 2017) <www.abc.net.au>; see also examples from 2000s like AOL, or Netflix Ohm, above n 4, at 1717 ff.; also, a broader discussion in Ira S Rubinstein and Woodrow Hartzog "Anonymization and Risk" (2016) 91 Washington Law Review 703.

[7] See e.g. Latanya Sweeney "Simple demographics often identify people uniquely" (2000) 671 Health (San Francisco) 1.

[8] Privacy Commissioner, above n 3, at 7.

[9] Alex Hern "New York taxi details can be extracted from anonymised data, researchers say" *The Guardian* (27 June 2014) <www.theguardian.com>.



There may be different reasons for preserving data which may be identifiable, or could be reverse engineered back to an identifiable state on an as need basis. This might be to support internal business or legal requirements where identifiable records need to be maintained for audit purposes. However, in many cases companies do not need the data to be identifiable to further use or share it for other secondary purposes.

De-identified data provides flexibility to preserve privacy depending on the purposes for which data is needed. It allows a risk-based approach to personal data management.

Assessing privacy risks for de-identification requirements

The risk approach in data management is attractive. Quantifying the risk related to the use of data we could: a) make a decision about particular use(s) of the data and b) particular methods of mitigating that risk. However, as yet, there is no consensus or understanding about how this should be quantified or whether a general approach is appropriate for all scenarios.

There is a very high bar for 'anonymisation'[10] and an ambiguous continuum for which information may be considered, or required to be, 'de-identified'. Understanding whether personal information meets either of these classifications, and to what extent, is highly contextual and largely depends on the dataset in question, the data ecosystem and IT environment which the organisation controls and manages, current technologies and the effectiveness of de-identification methods.

The following table outlines common factors which may be relevant for stronger or weaker de-identification standards: [11]

Factors which may require stronger de-identification	**Factors which may allow weaker de-identification**
• Sensitivity of data • Likely exposure or availability of data to other users (internal or external)	• Limited users who may have access • Short data retention periods • Strong security mechanisms implemented, including use of

[10] Considered as almost unattainable by some researchers, e.g. Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye "Estimating the success of re-identifications in incomplete datasets using generative models" (2019) 10 Nat Commun 3069.

[11] Some of the factors are discussed in Rubinstein and Hartzog, above n 6, at 741 ff.

• Vulnerability of individual(s) or groups with similar characteristics (e.g. ethnic minorities) • Uniqueness of data elements • Volume of data collected • Matching risk of data relating to same individuals controlled or managed by the same agency (current and future) • De-identification method • Availability of other data sources for matching • Individuals' or public expectations (which also may depend on the type of data use, e.g. data for Covid-19 response)	modern one-way encryption, user access controls or tiered access, data segregation, policies and other governance controls • Whether data is, or expected to be, reasonably public • Risk of privacy harm relating to personal data if exposed or identified to individuals or specific groups • Legitimate interests for legal defence, dispute or grievances.

There is no single set of rules for how these factors could be taken into account to devise particular de-identification methods. A useful guidance on the methods of anonymisation can be found on the websites of privacy and data protection authorities.[12] However, one thing is certain, data use needs to be limited to the purpose of its collection and internal policy safeguards and contractual protections should still apply. As the quantity, complexity and sprawl of data continues to increase across the public and private domains, organisations and individuals need to be clear about what level of data protection and privacy protections are afforded to them when they share and use data. The risks to individuals and organisations have to be properly managed by implementing appropriate data minimisation and disclosure practices.[13] This begins with agreeing and understanding what key terms such as "de-identified" and "anonymised" means, the potential risks associated with these, and having clear responsibilities for organisations that accurately reflect current social and data risks to protect individuals' interests and rights.

[12] See Article 29 Working Party *Opinion 05/2014 on Anonymisation Techniques* (0829/14/EN WP216 2014); Datatilsynet *A guide to the anonymisation of personal data* (2015); Singapore Personal Data Protection Commission *Guide to basic data anonymisation techniques* (2018).

[13] See for example Privacy Commissioner's recommendation for the Ministry of Health, Privacy Commissioner, above n 3, at 4.