AN INTRODUCTION TO DE-INDENTIFICATION

INTRODUCTION

Realising the full potential of data often relies on the ability to share it with a wider audience, be it other organisations or the public. However, upholding the information privacy rights of individuals is equally important, and required by various privacy laws. De-identification is often regarded as a way to protect the privacy of personal information when sharing or releasing data.

This resource provides a high-level overview of what de-identification is, with a focus on unit-level data that relates to specific individuals (rather than aggregated data). It looks at the meaning of de-identification, how it works, and the risks and challenges of de-identification in relation to privacy.

The resource speaks to de-identification in a broad context, where de-identified data is made freely available through public release of data sets or other “open data” programs. It does *not* look at de-identification that occurs in a controlled environment, such as the work done by the Victorian Centre for Data Insights (**VCDI**) under the *Victorian Data Sharing Act 2017*.

WHAT IS DE-IDENTIFICATION?

The term ‘de-identification’ can have different meanings. It can refer to the process of removing or altering information, e.g. deleting information that directly identifies individuals, such as names and addresses or dates of birth. Alternatively, it can mean reaching a state where individuals can no longer be ‘reasonably identified’ from the information.

Under the *Privacy and Data Protection Act 2014* (**PDP Act**), personal information is considered to be de-identified when it no longer relates to an identifiable individual or an individual who can be reasonably identified from the information.[[1]](#endnote-1)

It is crucial to note that even where information has been de-identified, it may still be possible to reveal an individual’s identity. This can occur when there is other auxiliary information available that can be matched with the de-identified data, leading to re-identification. Re-identification is explored further below.

WHY DE-IDENTIFY?

There are a number of reasons why an organisation may need or want to de-identify personal information, including:

* **when required by law** – for example, the PDP Act requires Victorian government organisations to destroy or permanently de-identify personal information if it is no longer needed for any purpose;[[2]](#endnote-2)
* **for risk management purposes** – for example, an organisation may wish to use a de-identified dataset rather than the original, as a way of minimising the risk of a privacy breach occurring;
* **to promote transparency and accountability** – for example, publishing an organisation’s Gifts, Benefits and Hospitality Register while maintaining the privacy of those to whom the information relates; and
* **to enable information to be used in innovative ways**, such as for research purposes, or to improve policy and decision making.

HOW DOES DE-IDENTIFICATION WORK?

Generally, the de-identification process involves:

1. removing direct identifiers (e.g. name, address); and
2. removing or altering other identifying information (indirect or quasi-identifiers, e.g. date of birth, gender, profession).

There are numerous techniques for removing or altering indirect identifiers. Two of these methods include k-anonymity and differential privacy, which are briefly outlined below.

k-anonymity

*k*-anonymity is a way of grouping data to hide unit-level records that relate to specific individuals. The objective of *k*-anonymity is to transform a dataset so that the identity of individuals in that dataset cannot be determined.

To achieve *k*-anonymity, there needs to be zero or ‘*k*’ individuals in the dataset who share a set of each possible attribute. If each individual is part of a larger group with the same set of characteristics, then any of the records could belong to any person in that group.

For example, in the table below, the indirect identifiers are age and gender. The dataset has *k*-anonymity of *k*=2, as there are at least two individuals for every possible combination of age and gender: (25, F), (23, F) and (26, M).

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Age | Gender | Test score |
| \* | 25 | F | 98 |
| \* | 26 | M | 78 |
| \* | 23 | F | 93 |
| \* | 26 | M | 92 |
| \* | 23 | F | 98 |
| \* | 25 | F | 86 |

As the value of *k* increases, the less likely it becomes that an individual’s identity can be ascertained from a record.

There are a number of approaches to achieve *k*-anonymity, such as:

* **generalisation**: specific values are replaced with a broad range or category. For example, instead of a specific age (e.g. 25), the dataset uses an age range (25-29); and
* **suppression**: sensitive values are replaced or removed entirely. For example, names may be replaced with a symbol (as shown in the table above), or outliers in a dataset may be removed – if all the ages in a dataset are between 25 and 30, but one record contained an age of 50, then that outlier could be suppressed.

Differential privacy

Differential privacy is a mathematical model of privacy that allows for the analysis of a dataset as a whole, without revealing anything specific about a particular record.

Differential privacy involves adding random noise to a dataset, or ‘perturbation’. Adding random noise into the original dataset masks the impact that any one record can have on the outcome of an analysis.

A mechanism or algorithm is considered to be differentially private if the probability of getting any particular outcome from an analysis on the dataset is essentially the same regardless of the absence or presence of any single unit-level record.

For example, consider two databases. They are identical, except one contains a particular record that the other does not. If successfully applied, differential privacy requires that the probability of a query producing a given result is similar, regardless of whether it is conducted on the first or second database.

RISKS AND CHALLENGES

Limitations of privacy enhancing techniques

Each technique for removing or altering information as part of a de-identification process has its own limitations and criticisms – differential privacy and *k*-anonymity are no exceptions.[[3]](#endnote-3) Where unit-level data is concerned, some techniques used to treat the data can deliver better privacy protections than others, however, no method can guarantee perfect privacy.

Balancing utility of data and privacy

One of the challenges of de-identification is balancing the utility of data with robust privacy protections. A dataset that has been altered significantly in a de-identification process will likely have a higher degree of privacy protection. For example, in relation to differential privacy, the more that the data is altered (that is, the more noise added), the more that privacy is preserved, yet this comes at a high cost to data utility.[[4]](#endnote-4)

Re-identification

Another challenge is the possibility of re-identification. All de-identified data carries a risk that it may be matched with other available auxiliary information, regardless of whether it is released publicly or shared in a controlled environment. However, this risk becomes more significant in an open data context, where restrictions on how the data is used or accessed or combined may be impossible to anticipate.

Essentially, re-identification is when an individual’s identity has been determined from seemingly de-identified data.

Re-identification generally occurs when a de-identified dataset is linked with another dataset that contains identifying information. In the context of “open data”, this auxiliary information can come from anywhere, including publicly available sources.

If the de-identified and the auxiliary datasets have related records (for example, they both contain individuals’ age, gender and postcode), then a combination of features that uniquely identify an individual should be the same in both datasets, therefore revealing that individual’s identity.

Assessing and managing the risk of re-identification

For unit-level data that is released publicly, the question to consider is: what auxiliary information will likely be available for re-identification? This can change over time as more information is generated and becomes available. In this context, assessing the risk of re-identification becomes difficult, if not impossible, as it would require knowledge of all the different data types that exist, and continue to be created.

Managing the risk of re-identification may therefore require more than reliance on de-identification techniques alone, particularly for unit-level data which carries a higher risk of re-identification. It may be appropriate to use access controls and other safeguards to help mitigate the risk of re-identification and provide an additional layer of privacy protection. In some cases, restricting access to the data may be the only practicable option.

There are a number of different controls and safeguards that can be adopted as an alternative to “open data”, such as:

* **electronic access controls** – for example, user authentication, or encrypting the data;
* **physical access controls** – this may include locking buildings in which data analytics work is done, or proper supervision;
* **personnel access controls** – for example, ensuring that only authorised officers have access to the data, or vetting individuals who may potentially access it;
* **only allowing access in a controlled environment**, such as a secure research environment or data lab;
* **enabling analysis of the data without providing access to it** – for example, running an analysis and providing the result to those who need to work with it, but not the raw data itself; and
* **releasing aggregate statistics** rather than detailed unit-level record data.

Conclusion

When done robustly in a controlled environment, the application of de-identification techniques to unit-level data can indeed enhance privacy. However, in circumstances where data is released to the public through an “open data” policy, de-identification can be risky. The growing availability of new and more information increases the risk that de-identified data can be matched with other auxiliary information, leading to the re-identification of seemingly de-identified data.

While de-identification can have many benefits, it is not a panacea for protecting the privacy of personal information in unit-level records. Given the risk of re-identification, it may not be appropriate to publicly release de-identified unit-level data. Victorian public sector organisations should consider working with expert data analytics bodies such as the VCDI when dealing with unit-level datasets.

1. Section 3 of the PDP Act. [↑](#endnote-ref-1)
2. Information Privacy Principle 4.2. [↑](#endnote-ref-2)
3. For more information about *k*-anonymity and differential privacy, see *The limitations of de-identification* available at: <https://ovic.vic.gov.au/privacy/privacy-guidance-for-organisations/>. [↑](#endnote-ref-3)
4. Data utility refers to the value of a given dataset as an analytical resource. It comprises the analytical completeness and validity of the data.

   **Disclaimer:** The information in this document is general in nature and does not constitute legal advice. [↑](#endnote-ref-4)