Data Anonymization and De-Identification: Challenges and Options
August 2019

Executive Summary

This whitepaper is intended to create a cohesive understanding of data anonymization and de-identification concepts, describe the risks and challenges associated with processing personal data that has been anonymized or de-identified, as well as briefly outline options for managing privacy risks related to re-identification. In sharing this paper among members of the University of Washington (UW) community, the UW Privacy Office aspires to raise awareness, accountability, and stewardship practices among those responsible for processing personal data at the UW.

For decades, those using data sets containing personal information have implemented a variety of techniques intended to "anonymize" and/or to de-identify such data, in an effort to protect privacy and prevent future re-identification of the individuals. The combination of emerging technological capabilities and evolving cultural practices is causing many to conclude that data anonymization/de-identification is no longer possible and that related risks of re-identification of individuals will continue to grow. The associated privacy risks of re-identification are defined in various ways and often contemplate an organizations' values and privacy principles, trusted relationship with individuals, compliance obligations, reputation, and financial well-being. Privacy risks include the possibility of objective or subjective harms to individuals, including: loss of liberty/opportunity, economic loss, social detriment, (unconscious or conscious) behavioral changes and/or psychological dangers[1]. When personal data is processed "unexpectedly," it may result in one or more objective or subjective harms to individuals' or result in other corresponding violations of privacy.

As a University, we must think strategically about the paradox, and the challenges and possibilities associated with the important question of "How might the UW continue to protect the privacy of personal data and appropriately process such data to create innovative research, technologies, learning methods, products, and solutions?"

Background

The University of Washington (UW) collects, creates, and processes many types of personal data. In general, personal data is any information that identifies or can identify an individual, either on its own or in conjunction with other information. The processing of personal data can take various forms such as collecting, recording, using, sharing, adapting, altering, or storing of structured, un-structured, or meta data. As stewards of personal data, all members of UW who process personal data have a role in helping protect the privacy of personal information.

In order to manage privacy risks associated with processing of personal data, the users of personal data (e.g. analysts, researchers, data scientist, fiscal or computing specialists) use a variety of techniques to manipulate the data elements, including redaction, pseudonymization, de-identification and/or anonymization. These techniques differentially modify or impact the personal data elements included, elements which can be classified by the levels of "identifiability" implicit in each along an identifiability continuum[3]:

- Direct identifiers serve to "uniquely" identify an individual, and include data elements such as Social Security Numbers, Tax ID numbers, Passport numbers, full names or addresses.
- Indirect identifiers, while not unique to an individual, can be combined with other indirect identifiers to identify an individual among a set of individuals. Indirect identifiers include items such as zip code, birth date, IP address, etc.
- Other personal data elements may be associated with multiple individuals, such as level of education, area of study, or communication preferences, and within a single data set a combination of such elements often does not allow the identification of a single individual.
- When data have been appropriately manipulated, combined or aggregated (perhaps in census data or survey results) they typically can no longer be linked to any individual, and are considered anonymized.
- Finally, some data elements (such as weather) are simply not related to individuals, and would not be considered personal information.

Techniques

Data sets containing either direct or indirect identifiers are generally perceived to be more useful for research or analytics, and typically present greater risks to individual privacy. Historically, in order to reduce such privacy risks, the following techniques[3], described in a simplified manner, have been used:

1. Deletion, redaction or obfuscation:
   Direct identifiers are covered, eliminated, removed or hidden. These techniques are difficult to accomplish well, particularly on unstructured data, and use of unsophisticated techniques may enable easy re-identification.
   Example: Jane Doe – DOB 8/15/1970 – St. Louis → [redacted] – DOB 8/15/1970 – St. Louis

2. Pseudonymization:
   Information from which direct identifiers have been eliminated, transformed or replaced by pseudonyms, but indirect identifiers remain intact. Re-identification may occur where there is failure to secure the pseudonymization method or key used, and/or when reverse engineering is successful.
   Example: Jane Doe – DOB 8/15/1970 – St. Louis → ID:TRXD 8/15/1970 St. Louis

3. De-identification:

Direct and known indirect identifiers (perhaps contextually identified by a particular law or regulation, i.e. HIPAA) have been removed or mathematically manipulated to break the linkage to identities.
Example: Jane Doe – DOB 8/15/1970 – St. Louis → Female 1970 Missouri

4. Anonymization:
   Direct and indirect identifies are removed or manipulated together with mathematical and technical guarantees, often through aggregation, in order to prevent re-identification. Anonymization is intended to be irreversible.
   Example: Jane Doe – DOB 8/15/1970 – St. Louis → Female Adult Missouri

   Note that encryption is sometimes inaccurately thought of as an obfuscation or de-identification technique. However, it is not a such a technique, but rather is a security measure intended to protect the personal data that may contain any combination of identifiable data elements.

Re-identification Motivations, Methods, and Myths

Motivations for re-identification of individual data subjects can vary widely:

- Threat actors may want to re-identify in order to conduct identity theft and fraud or social engineering of individuals.
- Statisticians, data analysts, data scientists or other users of personal data may want to re-identify when challenged to prove data cannot (or that it can) be re-identified.
- Scientific researchers may attempt to re-identify in order to further test related hypotheses.
- Capitalistic individuals and organizations increasingly want to re-identify in order to profile individuals, monetize personal data, or use personal data in ways that may not be expected, anticipated, or desired by the individuals. This is exemplified by data brokers, marketing and advertising organizations, social media firms, and so many other types of organizations today.

Among the prevalent methods used when attempting re-identification of anonymized or de-identified data:

- "Reverse" redaction (as seen in the movie "Hidden Figures" by way of exposing a manually redacted document to intense light source, or as in the technological glitch in redaction of the Manafort legal documents, where cutting and pasting redacted text into a new document rendered the text visible once again);
- "Reverse" pseudonymization – uncovering the methods, accessing the key, or reverse-engineering the pseudonymization techniques implemented;
- And increasingly by way of combining or linking data with other data sets available either publicly or for purchase. As eloquently phrased by Boris Lubarsky **"The proliferation of publicly available information online, combined with increasingly powerful computer hardware, has made it possible to re-identify "anonymized" data."[4]**

Finally, there are many myths about de-identification methods that simply must be refuted[5]:

- *Myth 1: Only highly knowledgeable data scientists can re-identify individuals within anonymized or de-identified data sets*. Readers of this paper are encouraged to read the resources linked below for many recent examples where students and/or junior analysts successfully re-identify individuals within data sets.
- *Myth 2: De-identified data can be used for any purpose*. To limit resulting privacy risks to organizations and individuals, stewards of personal data should limit personal data use to the original purpose for data collection/creation.
- *Myth 3: Once data is de-identified, it can be given to any recipient*. Corresponding privacy risk of re-identification will likely increase with each subsequent release of de-identified data.

Basic Privacy by Design Steps to Help Protect Personal Data

Privacy by design, in general, is the protection of personal data by embedding privacy practices into University operations, business processes, information systems, and technologies, including: at the earliest design stage when initially determined that data processing will involve personal data; during data processing; and at the conclusion of the information lifecycle when personal data is no longer needed for the purpose it was collected or created by the University. The following privacy by design steps may help address the data anonymization and de-identification challenges and options.

1. Clearly articulate the purpose for processing any personal data.
2. Only collect what is required for the specific purpose.
3. Plan how to protect personal data and identities before you collect or create data.
4. Protect data as required under all applicable laws (see UW Privacy Laws webpage). Note that personal data protection requirements often differ substantially, at times in conflict, under the large number of state, federal, international, and federation-level privacy laws.
5. Anticipate privacy risks, and identify possible consequences of related harms should the data be compromised, in order to ensure the benefits of personal data collection outweigh possible costs.
6. Provide notice (or seek consent) about data collection and purpose.
7. Be intentional about the technique used to redact or obfuscate, de-identify, pseudonymize, or anonymize personal data to ensure compliance with relevant laws or regulations.
8. Control and monitor access to details of de-identification process and/or keys.
9. If you need to de-identify or anonymize data elements, engage with colleagues to discuss options and issues, and to identify reliable sources and strategies for de-identification.
10. If sharing de-identified data, then specify with others that data will not be re-identified.
11. Be mindful of sustaining data integrity, and include practices to periodically refresh personal data.
12. Manage privacy across the entire data lifecycle from collection/creation to data destruction consistent with records retention schedules, once data purpose has been achieved.
13. Acknowledge (and wherever possible, offset) increasing likelihood of re-identification privacy risks for the UW and individuals.
14. Be prepared to respond to data subjects' requests to exercise their various privacy rights under evolving laws.

Future Anonymization/De-identification Challenges

The many challenges associated with use of anonymized or de-identified personal data are expected to grow as a result of a combination of shifting factors:

- New privacy legislation has been and is expected to be introduced at all levels of privacy law;
- Existing privacy laws are expected to be reviewed, revised, updated or pre-empted;
- De-identification and anonymization techniques, as well as re-identification techniques, will continue to be revised, enhanced, and invented; and
- Technological improvements in computing capacity will continue**.**

…(T)he gathering evidence shows that all of the ("de-identifying") methods are inadequate, said Dr. de Montjoye. "We need to move beyond de-identification," he said. "Anonymity is not a property of a data set, but is a property of how you use it."[6]

Adopting broad "Privacy by Design" practices throughout UW helps ensure continued stewardship and protection of the vast personal data under UW's care.

Citations:

[1] R. Jason Cronk**, Strategic Privacy by Design**, International Association of Privacy Professionals (IAPP) 2018, https://iapp.org/ (search for print or electronic copy of this book), pages 154-155. The privacy harms content within Cronk's book is based upon **The Boundaries of Privacy Harm**, Ryan Calo, Indiana Law Journal, Vol. 86, No. 3, 2011, written July 16, 2010 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1641487

[2] Boris Lubarsky, **Re-Identification of "Anonymized" Data**, Georgetown Law Technology Review, https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/, pages 203-204

[3] Barbara Sondag, Elimu Jajunju, and Elena Elkina, Privacy.Security.Risk.2017 conference presentation entitled: **Global Technological and Legal Effects of De-Identification and Anonymization**, (no longer available online), slide 2.

[4] Boris Lubarsky, **Re-Identification of "Anonymized" Data**, Georgetown Law Technology Review, https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/, page 203

[5] Barbara Sondag, Elimu Jajunju, and Elena Elkina, Privacy.Security.Risk.2017 conference presentation entitled: **Global Technological and Legal Effects of De-Identification and Anonymization**, (no longer available online), slide 5.

[6] Gina Kolata, "**Your Data Were 'Anonymized'? These Scientists Can Still Identify You**," New York Times, July 23, 2019. https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html

---

Many additional resources were reviewed to inform overall content. The following were of particular help:

Ryan Calo, "**The Boundaries of Privacy Harm**," Indiana Law Journal, Vol. 86, No. 3, 2011, written July 16, 2010, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1641487.

Mark Elliot, Elaine Mackey, Kieron O'Hara and Caroline Tudor, **The Anonymisation Decision-Making Framework**, United Kingdom Anonymisation Network (UKAN), University of Manchester, Oxford Road Manchester, M139PL, 2016, https://ukanon.net/ukan-resources/ukan-decision-making-framework/.

Kelsey Finch, **A Visual Guide to Practical Data De-Identification**, Produced by the Future of Privacy Forum (FPF), https://fpf.org/2016/04/25/a-visual-guide-to-practical-data-de-identification/.

Personal Data Protection Commission, Singapore, "**Guide to Basic Data Anonymisation Techniques**," Published 25 January 2018, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf.

Jules Polonetsky, Omer Tene, Kelsey Finch, "**Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification**," Vol. 56, Number 3 Santa Clara Law Review, Article 3 (6-17-2016). Available at: http://digitalcommons.law.scu.edu/lawreview/vol56/iss3/3.

Balaji Raghunathan, **The Complete Book of Data Anonymization: From Planning to Implementation**, Boca Raton: CRC Press, Taylor & Francis Group, 2013

Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, "**Estimating the success of re-identifications in incomplete datasets using generative models**," Nature Communications 10, Article number: 3029 (23 July 2019). Available at Nature Communications: https://www.nature.com/articles/s41467-019-10933-3

Ira Rubinstein, Woodrow Hartzog, "**Anonymization and Risk**," 91 Washington Law Review 703 (2016); NYU School of Law, Public Law Research Paper No. 15-36, (August 17, 2015). Available at SSRN: https://ssrn.com/abstract=2646185

Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke & Salil Vadhan (Harvard University Privacy Tools Project), "**Differential Privacy: A Primer for a Non-technical Audience**," Vanderbilt Journal of Entertainment & Technology Law (JETLaw), Vol. 21, Issue 1, pages 209 – 276, http://www.jetlaw.org/journal-archives/volume-21/volume-21-issue-1/differential-privacy-a-primer-for-a-non-technical-audience/.