Benchmarking Differential Privacy and Existing Anonymization or De-identification Guidance

Office of the Privacy Commissioner of Canada Contributions Program 2023-24

Professor Rafal Kulik, University of Ottawa
Department of Mathematics and Statistics
150 Louis Pasteur Priv. , K1N 6N5
STM564
(613) 562 5800 ext. 6266

uOttawa

Table of Contents

Applicant and Team

Principal investigator: Rafal Kulik, PhD, is a Professor of Mathematics and Statistics, University of Ottawa. In the last four years, he has worked in the area of differential privacy and disclosure risk metrics. In this area, he has partnered with Privacy Analytics through two Mitacs grants: Privacy Guarantees and Risk Identification: Statistical Framework and Methodology (link), and Statistical Framework and Methodology for Risk and Privacy in Complex and High-Dimensional Data (link). This research collaboration thus far includes two PhD students and one master's student. Professor Kulik also specializes in theoretical statistics, extreme value theory and time series. He published two books (*Long Memory Processes*, *Heavy Tailed Time Series*). He is an author of 50 publications in top journals in mathematics and statistics. He serves as an Associate Editor in top four journals in the field. He has supervised three post-doctoral fellows, six doctoral students and numerous masters students.

Co-investigator: Teresa Scassa, PhD, is the Canada Research Chair in Information Law and Policy at the University of Ottawa, Faculty of Law. She is the author or co-author of several books, including Canadian Trademark Law (2d edition, LexisNexis 2015), and Electronic Commerce and Internet Law in Canada, (CCH Canadian Ltd. 2012) (winner of the 2013 Walter Owen Book Prize). She is a past member of the External Advisory Committee of the Office of the Privacy Commissioner of Canada, and of the Canadian Government Advisory Committee on Open Government. She is a member of the GEOTHINK research partnership and has written widely in the areas of intellectual property law, law and technology, and privacy.

Industry support: Privacy Analytics provided advisory support and coordinate access to relevant data and computing infrastructure for students to run experiments. Privacy Analytics will provide in-kind support only, to facilitate research.

- Luk Arbuckle is Chief Methodologist and Privacy Officer at Privacy Analytics. He is a recognized author on the subject of anonymization, and an industry supervisor for Mitacs students in collaboration with Professor Kulik.
- Devyani Biswal is Methodology Architect and AI Scientist at Privacy Analytics. She received Mitacs support for her PhD thesis on differential privacy and disclosure risk metrics, under the academic supervision of Professor Kulik.

Students: Two science students (Heidi Barriault and Patrick Fogaing Koumao) were funded to conduct literature reviews, gap analyses, run experiments, and summarize findings. One law student (Jasper Ross) was funded to conduct literature reviews, gap analyses, construct surveys and facilitate outreach to experts and civil society groups and assist in finalizing guidance and recommendations.

Project Title

Benchmarking Differential Privacy and Existing Anonymization or De-identification Guidance

Basic Information

Organization name: Department of Mathematics and Statistics, University of Ottawa
Address: 150 Louis-Pasteur Pvt, STEM Complex, room 336
Ottawa, ON K1N 6N5
Billing address: <same>
Fax number: N/A
Principal Investigator: Rafal Kulik
Email address: rkulik@uottawa.ca
Senior Administrative Office: Martine Bertrand-Bourgeois <martineb@uottawa.ca>

Executive Summary

The project consisted of three phases:

1) **Phase 1: Practical Assessment Framework of Differential Privacy**
2) **Phase 2: Experimentation Phase**
3) **Phase 3: Scope of Policies**

We have also conducted a survey of students, on their understanding and their attitude to different privacy concepts. The results of the survey can be found in the Appendices section.

Our work sparked some interest among policy makers. We are holding an online seminar with the Government of Manitoba.

Executive Summary – Phase 1

The first phase focuses on technical aspects related to differential privacy. We were able to discover the inconsistencies in the definitions of differential privacy and its parameters. We were also able to identify the basic techniques and the limitations of differential privacy. These discoveries then allowed us to determine suitable language and identify opportunities to overcome challenges of implementing differential privacy and better optimize its use within existing frameworks.

We chose 25 papers to read for this phase, they consisted in policy/legal publications as well as non-technical. To achieve our goals for this phase, we created a spreadsheet that had a combination of all the definitions of differential privacy and its parameters as well as the basic techniques and the limitations.

The definitions for differential privacy varied from one paper to the next but we were able to group them into 4 themes that we named noise injection, bounding information, hiding individuals and miscellaneous. The group entitled 'noise injection' was the most common one with 45% of the definitions being grouped into this one. The second most common group was 'bounding information', with 23% of the definitions grouped in it. The group which had 18% of the definitions was the miscellaneous group. This group did not have a definition but was rather a group of definitions that could not be linked to each other or other groups. Finally, the last group, with 14% of the definitions in it, was the group entitled 'hiding individuals'. Furthermore, we were able to rank the level of technicality for each paper by using the number of definitions that were mentioned. For example, if a paper only mentioned 1/6 definitions they are categorized as being little to not technical. Thus, from these results we were able to conclude that there is an important lack in consistency and consensus on a proper definition for differential privacy.

This phase allowed us to discover the inconsistencies in the definition of differential privacy itself and of its parameters as well. The two main parameters of differential privacy are the privacy budget and the sensitivity.[1] Considering that these parameters heavily influence the amount of noise that a dataset will be injected with, it is surprising that they are rarely mentioned in academic papers or guidance and, when they are, their definitions are inconsistent. The privacy budget was only referenced in 14 papers out of the 25 that we read and there was no consistency on what to call it and how they defined it. The DP-sensitivity, on the other hand, was defined in 3

[1] In order to avoid confusion between "sensitivity" that appears in privacy laws and "sensitivity" that appears in the context of differential privacy, we will call the latter as "DP-sensitivity".

of the 25 papers. The gaps in the parameters of differential privacy are even greater than those in the definition itself which could contribute to the complexity of its implementation.

From the research done in this phase we were also able to identify the two basic techniques of differential privacy. The first one is called query-based differential privacy and this technique is defined as when noise is added to the query before being returned to an untrusted aggregator. The second technique is called database-based differential privacy and it is defined as when noise is added directly to the individual before being sent to an untrusted aggregator.

Finally, phase 1 of the proposal allowed us to identify the main limitations of differential privacy. These were grouped into 3 themes: technical implementation, privacy and utility trade off, and disclosure risks.

Executive Summary – Phase 2

In the second, exploratory phase we conducted 5 experiments to delve deeper into the intricacies of differential privacy. Phase 2 allowed us to answer fundamental questions about the relationship between the main techniques (pre-processing and post-processing) of differential privacy, differential privacy in a data privacy and data utility context, the relationship and combination of k-anonymity and differential privacy, and the comparison of privatized data to original data. We also concluded that some standard statistical techniques cannot be used in the context of assessing data privacy and data utility.

The first experiment provided us with new insights on the privatization of data employing the two main methods of differential privacy, pre-processing and post-processing. The main takeaway from this experiment was that pre-processing was more conservative on data privacy compared to post-processing. In order words, for the same level of privacy, post-processing has a better data utility.

The second experiment allowed us to determine whether it was feasible to combine the two main methods of data privatization, namely k-anonymity and differential privacy. We observe that a combination of two methods enhances data protection.

The third experiment was done to compare the main methods of differential privacy in a data utility context. We performed the experiment on three common queries which allowed us to get a better understanding of the influence of the privacy budget and how it works using different methods of differential privacy.

The fourth experiment was conducted to see if it is feasible to compare two dependent distributions using a KS-test. In other words, we wanted to see if it was possible to compare the structure of two datasets: the original dataset and its privatized counterpart. The key outcome of this experiment was that the KS-test does not work in this scenario. We expect that there will be a similar issue with any classical statistical test that compares two distributions, one for original dataset, one for privatized dataset, which uses the assumption of independence.

Finally, **the last experiment** helped us determine the feasibility of comparing both k-anonymity and differential privacy via a data utility point of view. The purpose of this experiment was to get a clear understanding of a possible relationship between k-anonymity and differential privacy, particularly when k is equivalent to the privacy budget. We concluded that a comparison with respect to data utility between differential privacy and k-anonymity does not seem feasible.

In conclusion, the second phase experiments coupled with the insights gained from the exploratory phase, provided valuable understanding into the intricate interplay between privacy budgets, data utility, and the choice of differential privacy techniques. The findings underline the importance of careful consideration in selecting either pre-processing or post-processing differential privacy. This nuanced understanding contributes to refining the practical implications of implementing differential privacy, ensuring a balanced approach to preserving privacy while maintaining data utility.

Executive Summary – Phase 3

Although there has been considerable development of privacy enhancing technologies that go beyond anonymization, their relationship to the concept of anonymization in data protection law is not always clear. **Currently, there are no clear guidelines that explain how differential privacy may be aligned with the concept of anonymization in privacy law** or how it might relate to the relative approach to anonymization developed in Canadian case law. **This third part** of this project will examine how differential privacy can be integrated with legal requirements in PIPEDA and in the proposed Bill C-27.

In Canada's case law, the courts apply threshold test and assess "serious possibility". Whether something meets the "serious possibility" threshold (or other similar threshold developed under comparable legislation), will depend on the circumstances of each case. The relative approach to anonymization has been the dominant approach in Canadian case law, although the Privacy Commissioner of Canada has expressed support for an absolute standard of anonymization

We do not suggest that differential privacy is a technique that is superior to other identification techniques such that it should replace them. Rather, it should be a viable tool in the deidentification toolbox.

We provide the following set of guidance:

1) Anonymization can and should allow for the use of differential privacy techniques.
2) Guidance on anonymization should be clear and should allow for the selection of different tools or approaches.
3) Differential privacy should be clearly defined. As noted in Part I of this project, differential privacy is sometimes discussed in the literature in imprecise or problematic ways. Clarity regarding this technique should begin with a careful definition.
4) Differential privacy can be an anonymization technique or it can be used in conjunction with other privacy protective measures.

Project Phase I – Practical Assessment Framework of Differential Privacy

Vocabulary and definitions

Differential Privacy: A technical privacy model that protects individuals by limiting the information that can be contributed to an analytical output by any one individual.

- It randomizes the results of queries or to the database before results are shared.
- It randomizes by adding noise to a query or dataset so that it is impossible to reverse-engineer individual inputs (up to a known information limit).

Privacy Budget:[2] The level of protection in a given dataset or statistic.

- Can be interpreted as a tuning parameter that trades privacy for accuracy.

Sensitivity (DP-Sensitivity): A function that measures the maximum potential change in the output.

Global (DP)-Sensitivity: The maximum difference between the values that a function may take on a pair of datasets that differ in only one element.

Local (DP)-Sensitivity: The maximum difference between the values that a function may take on a dataset and the same dataset that differs in only one element.

De-identification:[3] A potential means of facilitating the use of personally identifiable information (PII) in a way that does not identify or otherwise compromise the privacy of an individual or group of individuals.

Anonymization:[4] A process that removes the link between the identifying dataset and the data subject.

Pseudonymization:[2] Techniques that involve replacing a data subject's identifier (or identifiers) with indirect identifiers created specifically for each data subject.

---

[2] Drechsler, Joerg. Differential Privacy for Government Agencies – Are We There Yet? 2, 2021. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.2102.08847.

[2] Information Security, Cybersecurity and Privacy Protection - Privacy Enhancing Data de-Identification Framework. ISO, Nov. 2022.

[3] Garfinkel, Simson. De-Identifying Government Data Sets. NIST SP 800-188 3pd, National Institute of Standards and Technology, 2022, p. NIST SP 800-188 3pd. DOI.org (Crossref), https://doi.org/10.6028/NIST.SP.800-188.3pd.

[4] Technical Guidelines for the Development of Small Hydropower Plants — Part 1: Vocabulary. ISO, Dec. 2019.

[5] Health Informatics - Pseudonymization. ISO, Jan. 2017.

[6] Privacy Enhancing Data De-Identification Terminology and Classification of Techniques. ISO, Nov. 2018.

Untrusted aggregator: A company or entity that does not have access to the original dataset.

Trusted curator: A company or entity who has access to the original dataset and can apply differentially private mechanisms to the dataset.

Query: A request for information that is calculated automatically from a dataset.

Database:[5] A collection of stored data.

Data subjects:[6] Person to whom data relate.

Noise injection:[7] A technique that modifies a dataset by adding random values to the values of a selected attribute.

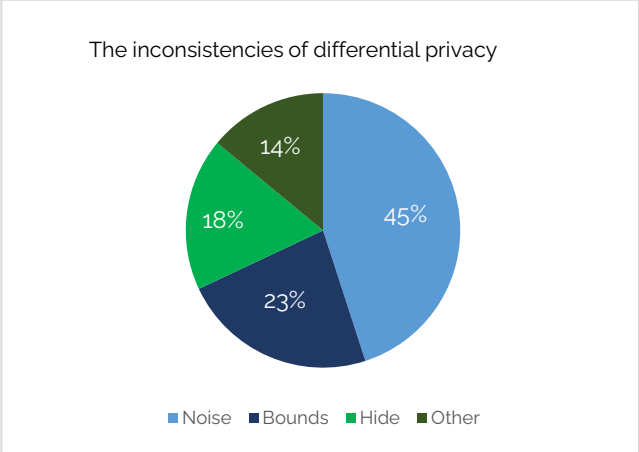
Overview

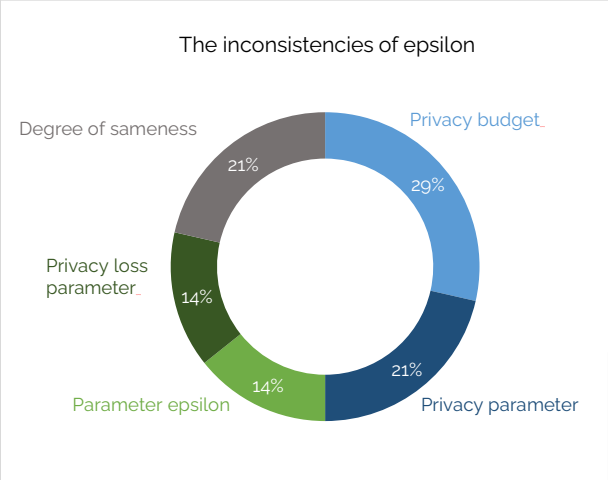
There are multiple variations of differential privacy that are being used at the moment, such as approximate differential privacy, zero-concentrated differential privacy, and others. Although they differ due to their parameters, they are all known as differential privacy because they are all based on the same concept of indistinguishability from one individual to another, and they all meet the same mathematical definition of privacy. In this report we will only be exploring (ε)-differential privacy.

Discrepancy

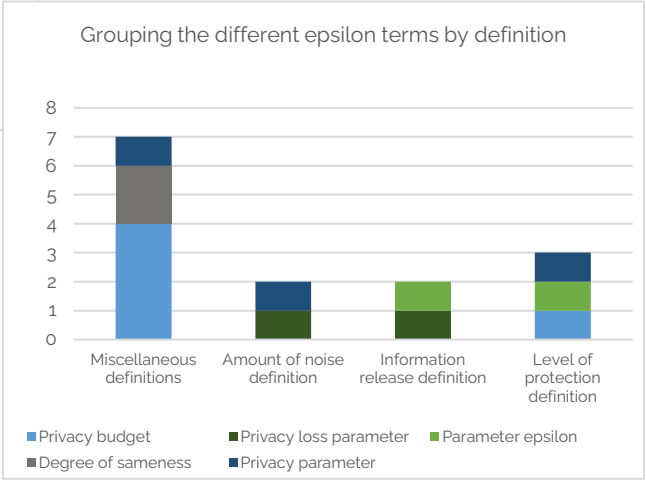
Phase 1 relies mainly on the excel file attached. The analysis done in this phase is based off the detailed survey that was done to create the excel file. We want to clarify that a limitation of this phase is that we have not read all the literature on differential privacy and have only read the selected 25 papers that we thought to be the most useful.

This phase focuses on conducting a review of current interpretations of differential privacy and its parameters to find suitable language. To accomplish this, we read through 25 publications, a combination of policy/legal and non-technical, and found inconsistencies throughout the definitions of differential privacy and its parameters. To get a clearer picture of the definition of differential privacy, we grouped it into four main themes: noise injection (noise), bounding information (bounds), hiding individuals (hide), and miscellaneous (other). As it can be seen in the image, **the most common definition for differential privacy is one that was based off noise injection,** with 45% of the definitions grouped in this theme. The definition that englobes this theme is: "Differential privacy adds noise to a dataset to protect the information". The second common theme, with 23% of the definitions grouped in this theme, is bounding information. This theme's definition is: "Differential privacy bounds the amount of information that can be revealed". The theme named miscellaneous was third with 18% of the definitions fitting in it. This theme is a combination of definitions that could not be placed in the other themes, nor could they be grouped together under one same theme, other than miscellaneous. Finally, the last group, hiding individuals, is composed of 14% of the definitions and has the following definition: "Differential privacy hides the presence or absence of an individual in a dataset".

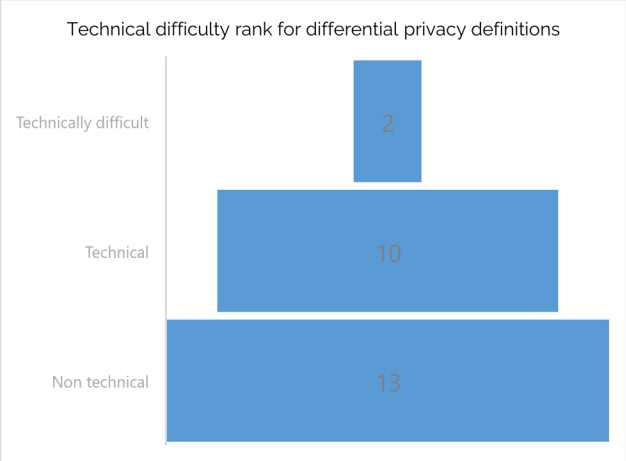




Not only there are gaps in the definition of differential privacy, as seen above, but in its parameters as well. The two main parameters of differential privacy are the privacy budget and the DP-sensitivity. Knowing that these parameters are what determines the amount of privacy given to a dataset, they are not mentioned frequently, and their definitions are not consistent. Starting with the privacy budget, it was only mentioned in 14 papers out of the 25 that we read. Furthermore, there was no consistency in what they called this parameter and how they defined it. Even if they had the same name, they did not necessarily have the same definition. However, one consistent aspect of the privacy budget in the papers was the effects of a large and small privacy budget; a large privacy budget results in better data utility, with less noise added and so low



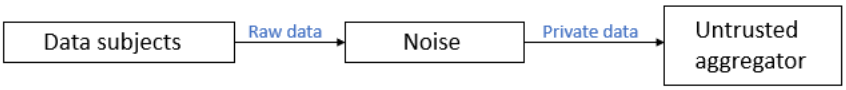
privacy and vice versa for a small privacy budget. The DP-sensitivity, on the other hand, was defined in 3 of the 25 papers. The DP-sensitivity, unlike the privacy budget, is not chosen by the user but greatly influences the amount of privacy provided by a differentially private algorithm, nonetheless. There are two different types of DP-sensitivity named global and local DP-sensitivity. Note that they have very similar titles to the different techniques of differential privacy and are not to be confused. Global DP-sensitivity is theoretical and uses an imaginary population to compute it. For example, if we want to calculate the global DP-sensitivity of age, we know that the minimum value it could ever be is 0 and we can assume that no one will be older than 200 years old. Thus, the DP-sensitivity function will use these values to calculate its value. Local DP-sensitivity, on the other hand, is experimental and uses the values from the dataset to compute it. So, if a dataset has a minimum age of 20 and a maximum of 65, they the DP-sensitivity function will use these values to calculate its value. The gaps in the parameters of differential privacy are even greater than those in the definition itself and makes their implementation complex.



Furthermore, we can rank on their technical difficulty, where 1 is the most technically difficult and 3 is the least technically difficult. To do so, we count how many of the 6 definitions were mentioned in the paper. A more technical paper would be one that mentions 5-6 definitions whereas the least technical paper will have 1-2 definitions. Most of the publications were non-technical as they mentioned 1-2 definitions.

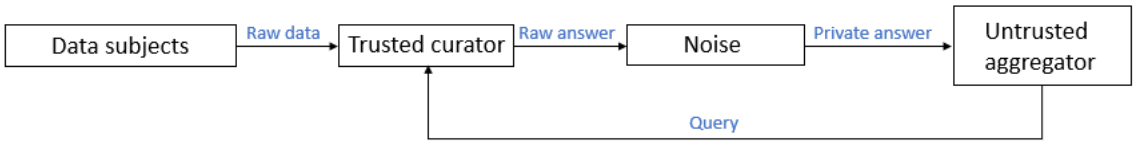
Basic techniques and parameters

Local differential privacy



Local differential privacy is when there is noise that is applied directly to the data subject's raw data. The data is then sent to an untrusted aggregator as private data. In simpler terms, there is noise that is added to the individuals' data before being received by an untrusted third party where they will use it in some way. This technique is done when the third party does not need to see the true data to perform their analysis. It also meets the requirements for differential privacy as it protects individuals by limiting the information that can be contributed to an analytical output by any one individual.

Global differential privacy



Global differential privacy is when the raw data from the data subjects is sent to a trusted curator. There is then an untrusted aggregator that asks queries to the trusted curator, the curator answers the query but before returning it they add noise to the answer. This technique is used when there is a person or entity that needs the real data. For example, hospitals need the true information on the patients but if a researcher were to ask them questions, they would then provide noisy data.

Combination

In addition to these two techniques there is a technique that combines the two. This technique is used when an untrusted aggregator can ask a query to see the entire database.

Absolute vs relative risk metric

Absolute and relative risk are used abundantly in health sciences as a way to measure risk. Absolute risk measures the difference between risks in two groups (risk of the second group minus risk of the first group), and relative risk measures a ratio that says how likely risk is to increase or decrease in two groups (risk of the second group divided by risk of the first group).

For example, let's compare the risk of developing skin cancer using sunscreen versus not using sunscreen. After having surveyed 200 000 individuals, 100 000 using sunscreen and 100 000 not using sunscreen, we see that the risk of developing skin cancer is twice as likely if you don't wear sunscreen. This means that the risk of developing skin cancer doubles when you do not apply sunscreen. However, we also see that the difference of developing skin cancer between both companies was 20 individuals per 100 000 individuals, or 0.0002%. The absolute risk tells us the risk of getting cancer between the two groups and the relative risk tells us that the risk of cancer increases if you don't wear sunscreen.

Another example of absolute and relative risk would be to consider a rare disease X, we want to measure the risk of developing this disease for the general population and for a more specific population, say plumbers. After analysing the results from the general population and those from the plumbers we discover that the risk of developing the disease in the general population is 0.2% more likely than the plumbers. Which means that the risk of developing the disease is much lower for the general population than it is for the plumbers. However, the difference of developing the diseases between both populations is 42 individuals per 100 individuals, or 42%. This means that for every 100 individuals in the general population, 42 fewer individuals are likely to develop the disease compared to the plumbers.

These examples show us the importance of measuring both absolute and relative risk:
"We therefore recommend to report both the relative risk and the absolute risk with their 95% confidence intervals, as together they provide a complete picture of the effect and its implications." Noordzij et al. (2017).

The Canadian guidance on anonymization uses a form of absolute risk to determine an acceptable level of disclosure. Two techniques used are k-anonymity and the threshold rule. For example, a group of twenty people on the same identifying information represents a risk of one over twenty, or 0.05, that their names can be randomly assigned. This approach makes it possible to compare the overall level of risk between two datasets. We can tie the absolute risk described here to the previous example. The higher the SPF (although it is not linear) is, the lower the risk of cancer and, in this case, the higher the number is in the group, the lower the risk of being identified.

Differential privacy provides a relative measure since it compares two databases, one that has all the individuals and one that has all but one individual. As a relative measure, differentially private outputs can only be compared on the database in which it is being applied. Two different databases will have different relative measures that are independent of one another. Unlike the techniques mentioned previously, differential privacy needs to be extended to include a relative risk metric.

"We argue that individuals should care about absolute disclosure risk and not relative risk alone." Hotz et al. (2022).

Data normalization is a common process used in statistics that reorganizes the data from the database that statistical tests can be run and have an improved data analysis. When using differential privacy, if we want to normalize the data, we must apply the same normalization procedure to the privacy budget and the DP-sensitivity that we did to the data. For example, if we multiplied the data by ten, we must also multiply the privacy budget and the DP-sensitivity by ten as well.

Current implementations of differential privacy

There are companies that have started to implement differential privacy such as Apple, Google, Uber, Amazon, US Census Bureau, and more. We will be delving into the techniques that the US Census Bureau and Apple uses.

The US Census uses local differential privacy, adding noise to an individual's contribution, to create a database assuming that only specific queries will be asked. In fact, the US Census set the DP-sensitivity to a specific value based on a square root function, and a specific privacy budget of 19.61. Furthermore, the US Census has implemented something called zero-concentrated differential privacy which meets the core definition of differential privacy with more parameters to consider.

Apple also uses local differential privacy but, in this case, they directly apply it to the user's device rather than to a dataset. They use a different privacy budget ranging from 2 to 8 depending on the data stream. They reset the privacy budget daily and limit the number of times an individual's data can be sent to Apple. This approach has been criticized because some experts believed that the privacy budget should be added across streams. Adding the privacy budget across streams results in a much higher overall privacy budget, in the range of 16 to 20, meaning less noise overall is added to the data. However, Apple explained that the data streams are segregated.

From these two examples alone, we can see the challenges in comparing the uses of differential privacy between Apple and US Census. Both are using different definitions of differential privacy, different DP-sensitivity functions, different parameters, and different assumptions. So, although we can compare the privacy budgets, they in fact mean very different things and cannot be compared directly.

Key challenges

We identify the most common challenges, pertaining to the implementation of differential privacy, to find opportunities to overcome them. Based on the literature, we can classify the challenges into three main themes: technical implementation, privacy and data utility trade-off, and disclosure risks.

The first theme, technical implementation challenges, refers to the challenges linked to the process of applying differential privacy to a database. The main challenges that are mentioned are the computational complexity of differential privacy [8] [9] [10], that its implementation requires an expert [3] [6] [11] [12] [13] [14] [15] [16] and that there are no ready-to-use tools to help with the process [3] [9] [15].

The second theme, privacy and data utility trade-off challenges, are the challenges that relate to the data privacy and the data utility of the noisy data. Here, noisy data refers to data that results after differential privacy is implemented. As the privacy of the database is directly related to its data utility, then the higher the privacy the worst the data utility and vice versa. Thus, finding the trade-off between privacy and data utility can be quite challenging. One reason the trade-off can be challenging to find is due to the absence of guidance or standards to determine the appropriate privacy [6] [9] [11] [12] [13] [15] [16] [17] [18] [19]. Another difficulty highlighted in the literature is interpreting the level of privacy and data utility [1] [10] [13] [16] after applying differential privacy techniques to a database. Finally,

---

[8] OPC blogger. "Privacy Enhancing Technologies for Businesses." Privacy Tech-Know Blog, 12 Apr. 2021, https://www.priv.gc.ca/en/blog/20210412/.

[9] Gandhi, Raina, and Amritha Jayanti. Technology Factsheet: Differential Privacy. 2020, https://www.belfercenter.org/sites/default/files/files/publication/diffprivacy-3.pdf.

[10] Cummings, Rachel, et al. "'I Need a Better Description': An Investigation Into User Expectations For Differential Privacy." Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2021, pp. 3037–52. DOI.org (Crossref), https://doi.org/10.1145/3460120.3485252

[11] Royal Society. From Privacy to Partnership. 2023, https://royalsociety.orgmedia/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf?la=en-GB&hash=4769FEB5C984089FAB52FE7E22F379D6.

[12] Fast-track action committee on advancing privacy-preserving data sharing and analytics networking and information technology research and development subcommittee of the national science and technology council. National Strategy to Advance Privacy-Preserving Data Sharing and Analytics. Office of the President, Mar. 2023, https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf.

[13] OECD. Emerging Privacy-Enhancing Technologies. OECD Digital Economy Papers, 354, 26 June 2023, p. 51. DOI.org (Crossref), https://doi.org/10.1787/c1faa51e-en

[14] Privacy-Enhancing Technologies (PETs). Information Commissioner's Office, June 2023, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/.

[15] Wood, Alexandra, et al. "Differential Privacy: A Primer for a Non-Technical Audience." SSRN Electronic Journal, 2018. DOI.org (Crossref), https://doi.org/10.2139/ssrn.3338027.

[16] Reiter, Jerome P. "Differential Privacy and Federal Data Releases." Annual Review of Statistics and Its Application, vol. 6, no. 1, Mar. 2019, pp. 85–101. DOI.org (Crossref), https://doi.org/10.1146/annurev-statistics-030718-105142.

[17] United Nations Committee of Experts on Big Data and Data Science for Official Statistics. The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics. 2023, https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf.

[18] Dwork, Cynthia, et al. "Differential Privacy in Practice: Expose Your Epsilons!" Journal of Privacy and Confidentiality, vol. 9, no. 2, Oct. 2019. DOI.org (Crossref), https://doi.org/10.29012/jpc.689.

[19] Garfinkel, Simson, et al. Issues Encountered Deploying Differential Privacy. 1, 2018. DOI.org (Datacite), https://doi.org/10.48550/ARXIV.1809.02201.

the addition of noise to data may inadvertently amplify any existing bias that is already in the database [12].

Finally, the last theme, disclosure risks of differential privacy, has grouped 4 of the challenges that are highlighted in the literature. These are the lack of a common description for differential privacy [10 20], the uniqueness of every dataset [8 9 11 16 21], that it cannot be used to study outliers [18], the fact that differential privacy does not necessarily result in anonymous information [10 14 21]. The lack of a standard or common description could potentially affect the accuracy and the functionality of the implementation of differential privacy. Furthermore, the aforementioned challenge as well as the uniqueness of the dataset makes it challenging to establish a standardized approach. In fact, implementation challenges tend to occur when the database is small or complex, i.e., categorical variables. Another limitation to differential privacy is that it cannot be used to study outliers in a database because it focuses on protecting the overall privacy of the dataset, thus hiding the outliers. Finally, if the differentially private algorithm is not configured properly it could lead to personal information leaks.

[20] Sparapani, Tim, et al. A Review of the Emerging Privacy Tech Sector. June 2021, https://fpf.org/wp-content/uploads/2021/06/FPF-PTA-Report_Digital.pdf.

[21] UNECE. Synthetic Data for Official Statistics: A Starter Guide. United Nations, 2023. Open WorldCat, https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf.

Bibliography (Phase 1)

1. (1) (2) (3) (4–21) (22) (23–26) (27,28) (29,30) (31)

1. Study Report: Reduced Exposure Study Using THS 2.2 Menthol with 5 Days in a Confinement Setting Followed by 86 Days in an Ambulatory Setting [Data file]. U.S. Food & Drug Administration; 2017 (updated 2018). Available from: https://www.fda.gov/tobacco-products/advertising-and-promotion/pmp-sa-module-7-scientific-studies-and-analyses

1. Sparapani T, Sherman J, Privacy Tech Alliance, Future of Privacy Forum. A Review of the Emerging Privacy Tech Sector. 2021 Jun; Available from: https://fpf.org/wp-content/uploads/2021/06/FPF-PTA-Report_Digital.pdf

2. Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries [Internet]. arXiv; 2015. Available from: http://arxiv.org/abs/1503.01214

3. Long G. Consistency of Data Products and Formal Methods for the 2020 Census [Internet]. MITRE Corporation; 2022 Jan p. 150. Report No.: JSR-21-02. Available from: https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-census-data-products-privacy-methods.pdf

4. Apple. Differential Privacy Overview [Internet]. Available from: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

5. Apple Differential Privacy Team. Machine Learning Research. 2017. Learning with Privacy at Scale. Available from: https://machinelearning.apple.com/research/learning-with-privacy-at-scale

6. Cummings R, Kaptchuk G, Redmiles EM. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security [Internet]. Virtual Event Republic of Korea: ACM; 2021 [cited 2023 Jul 6]. p. 3037–52. Available from: https://dl.acm.org/doi/10.1145/3460120.3485252

7. Dwork C, Kohli N, Mulligan D. Differential Privacy in Practice: Expose your Epsilons! JPC [Internet]. 2019 Oct 20;9(2). Available from: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/689

8. Elliot M, Mackey E, O'Hara K. The Anonymisation Decision-Making Framework 2nd Edition: European Practicioners' Guide [Internet]. UKAN; 118 p. Available from: https://ukanon.net/framework/

9. Fast-track action committee on advancing privacy-preserving data sharing and analytics networking and information technology research and development subcommittee of the national science and technology council. National Strategy to Advance Privacy-Preserving Data

Sharing and Analytics [Internet]. Office of the President; 2023. Available from: https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf

10. Garfinkel SL. De-identification of personal information [Internet]. National Institute of Standards and Technology; 2015 Oct [cited 2023 Jul 5] p. NIST IR 8053. Report No.: NIST IR 8053. Available from: https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

11. Garfinkel SL, Abowd JM, Powazek S. Issues Encountered Deploying Differential Privacy. 2018; Available from: https://arxiv.org/abs/1809.02201

12. Garfinkel S. De-Identifying Government Data Sets [Internet]. Gaithersburg, MD: National Institute of Standards and Technology; 2022 [cited 2023 Jul 5] p. NIST SP 800-188 3pd. Report No.: NIST SP 800-188 3pd. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-188.3pd.pdf

13. Nakanishi T, Hori S. Data Free Flow with Trust: Overcoming Barriers to Cross-Border Data Flows [Internet]. World Economic Forum; 2023. Available from: https://www3.weforum.org/docs/WEF_Data_Free_Flow_with_Trust_2022.pdf

14. Reiter JP. Differential Privacy and Federal Data Releases. Annu Rev Stat Appl. 2019 Mar 7;6(1):85–101.

15. UNECE. Synthetic data for official statistics: a starter guide [Internet]. Geneva: United Nations; 2023. 87 p. Available from: https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

16. United Nations Committee of Experts on Big Data and Data Science for Official Statistics. The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics [Internet]. 2023. Available from: https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf

17. Xiong A, Wang T, Li N, Jha S. Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension. 2020; Available from: https://arxiv.org/abs/2003.13922

18. Privacy enhancing data de-identification terminology and classification of techniques. ISO; 2018.

19. Technical guidelines for the development of small hydropower plants — Part 1: Vocabulary. ISO; 2019.

20. Information security, cybersecurity and privacy protection - Privacy enhancing data de-identification framework. ISO; 2022.

21. Privacy Implementation Notice 2023-01: De-identification [Internet]. Canada.ca; 2023. Available from: https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information-privacy-notices/2023-01-de-identification.html#app

22. Royal Society. From privacy to partnership [Internet]. 2023. 112 p. Available from: https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf?la=en-GB&hash=4769FEB5C984089FAB52FE7E22F379D6

23. Health information - Pseudonymization. ISO; 2017.

24. Gandhi R, Jayanti A. Technology Factsheet: Differential Privacy [Internet]. 2020. Available from: https://www.belfercenter.org/sites/default/files/files/publication/diffprivacy-3.pdf

25. Sparapani T, Sherman J. Privacy Tech Buyer Framework. 2022 Apr; Available from: https://fpf.org/wp-content/uploads/2022/04/FPF-Privacy-Tech-Buyer-Framework-R5-singles-1.pdf

26. Drechsler J. Differential Privacy for Government Agencies -- Are We There Yet? 2021; Available from: https://arxiv.org/abs/2102.08847

27. Privacy-enhancing technologies (PETs) [Internet]. Information Commissioner's Office; 2023. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/

28. Wood A, Altman M, Bembenek A, Bun M, Gaboardi M, Honaker J, et al. Differential Privacy: A Primer for a Non-Technical Audience. SSRN Journal [Internet]. 2018; Available from: https://www.ssrn.com/abstract=3338027

29. OPC blogger. Privacy Enhancing Technologies for Businesses [Internet]. Privacy Tech Know blog. 2021. Available from: https://www.priv.gc.ca/en/blog/20210412/

30. Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security [Internet]. 2014. p. 1054–67. Available from: http://arxiv.org/abs/1407.6981

31. Reimsbach-Kounatze C, Reynolds T. Emerging privacy-enhancing technologies: Current regulatory and policy approaches [Internet]. 2023 Mar. Report No.: 351. Available from: https://www.oecd-ilibrary.org/science-and-technology/emerging-privacy-enhancing-technologies_bf121be4-en

Project Phase 2 – Experimentation Phase

List of projects

1) Exploring Privacy Budget Effects: A Comparative Analysis of Pre- and Post-Processing Differential Privacy.
2) Comparative Analysis of Pre- and Post-Processing Differential Privacy: A Focus on Data Utility
3) Comparison of distributions: original vs. anonymized data
4) Exploring the Impact of Privacy Budget on Data Utility through Noisy Dataset Analysis
5) Experiment 5: Comparison of Differential Privacy and K-Anonymity

Experiment 1: Exploring Privacy Budget Effects: A Comparative Analysis of Pre- and Post-Processing Differential Privacy

Differential privacy can be achieved using either pre-processing (adding noise to the database) or post-processing (adding noise to a query) methods. Using these methods, we enhance data protection by minimizing the risk of disclosing sensitive information while still allowing for meaningful analyses.

This experiment was conducted to investigate the differences or similarities between using pre-processing and post-processing methods with different sensitivity measures: local and global. Furthermore, we wanted to attempt to find the most effective approach for introducing noise into a dataset to anonymize it, keeping in mind the differential privacy principles.

This experiment was done by using a dataset containing people's age. Furthermore, as there is a linear relationship for the mean between pre-processing and post-processing, we use this statistic to compare both methods.

The main findings of this experiment are that as the privacy budget increased (less noise was added), the variance of the noisy mean estimator tended to 0. However, the variance of pre-processing differential privacy was always bigger than that of post-processing differential privacy. This is in line with statistical theory. These results led us to conclude that pre-processing has a more conservative approach to data privatization than post-processing. In order words, for the same level of privacy, post-processing gives the better data utility. This conclusion is intuitive as post-processing is related to a particular query while pre-processing is supposed to anonymize against all possible queries.

Methodology

For both methods of differential privacy and both DP-sensitivity types, we performed the same steps. We applied differential privacy to the age variable. When we computed the pre-processing method, we added the noise to each variable prior to calculating the mean. On the other hand, when we computed the post-processing method, we calculated the mean prior to adding the noise. We repeated this step 1000 times to see what happened to the mean after each iteration. Finally, to analyse the tendency, we computed the variance for each iteration and plotted it.

Figure 1.1 displays the result for the comparison between pre- and post-processing with a global DP-sensitivity. Figure 1.2 displays the result of that same comparison but with a local DP-sensitivity. The result for each iteration of the experiment is displayed. We see that the variance of the mean estimator tends towards 0 as the value of the privacy budget increases. In fact, in both cases, post-processing tends to 0 faster than pre-processing. Furthermore, pre-processing tends to have a higher variance for the mean estimator compared to post-processing.

Conclusions

Thus, we conclude that data subjected to pre-processing has a higher level of protection than the data undergoing post-processing. We can conclude this as the variance for the mean estimator using post-processing starts off near 0 and tends closer to 0 as the privacy budget increases. This means that the mean query that is generated using post-processing is not likely to vary as the privacy budget increases. However, the pre-processing mean estimator generated will have

greater variability in its value as the privacy budget increases. Furthermore, these results may indicate that pre-processing is a more conservative approach as it holds a higher level of privacy than post-processing. Nevertheless, the question of data utility persists, raising concerns about whether this increased privatization comes at the expense of maintaining an acceptable level of data utility.

Summary: Pre-processing guarantees more privacy. This is intuitive, since pre-processing "protects against all possible queries", while post processing deals with one query only. Data utility measured by variance of the randomized queries.

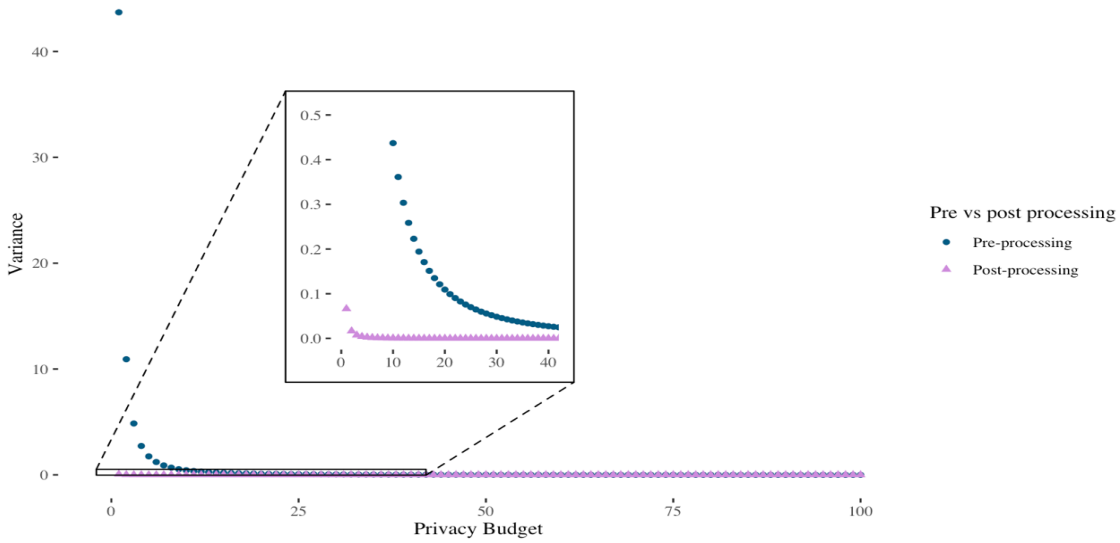


Figure 1.1 Variance of the mean estimator with global sensitivity as the privacy budget increases

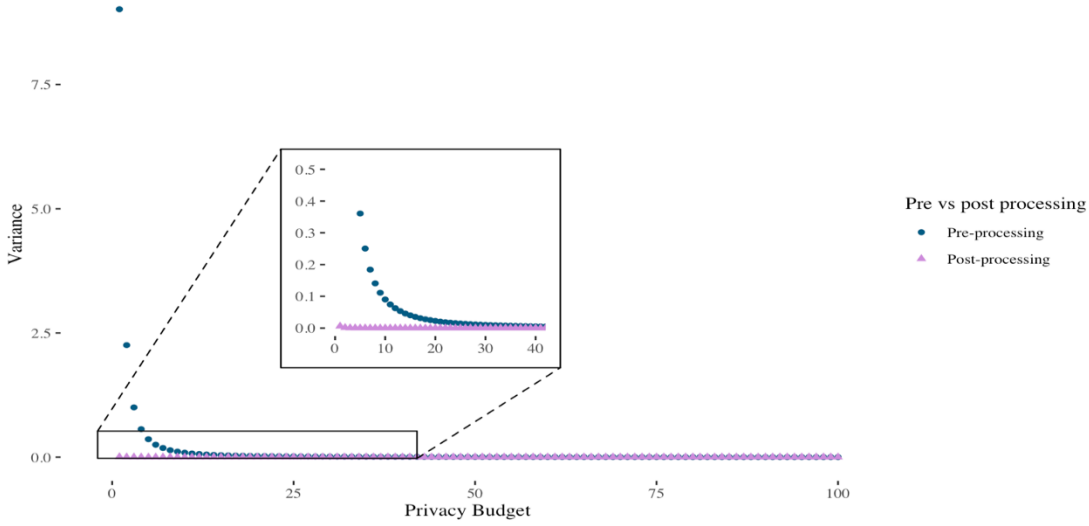


Figure 1.2 Variance of the mean estimator with local sensitivity as the privacy budget increases

Experiment 2: Comparative Analysis of Pre- and Post-Processing Differential Privacy: A Focus on Data Utility

When applying differential privacy techniques to a dataset, we must find a balance between data utility and data privacy. Although we all wish to have data that is privatized to the highest degree, in order to do this, we must compromise the utility of it. This is true for the reverse. Thus, it is important to be able to find a balance between the two to have useful but privatized data.

The purpose of experiment 3 was to compare the outcomes of the data utility of both pre- and post-processing differential privacy while using different queries and privacy budgets. From these comparisons, the goal was to get a better understanding of how data utility is achieved in different scenarios.

To conduct this experiment, we chose three basic and widely used queries: the mean, the median and the variance. We chose to do two scenarios, the first where both methods of differential privacy had the same privacy budget and the second where they had different privacy budgets.

The main takeaway from this experiment was that for the same level of privacy post-processing has better data utility.

Methodology:

We started the experiment by generating a dataset. This dataset was similar to the one used in Experiment 1 but contained age and BMI (body mass index) values. Using these continuous variables made it easier to follow the experimentation process. We performed the experiment under two separate scenarios. Each scenario was performed twice, once for pre-processing and the other for post-processing.

In the first scenario, we set the privacy budget to 1 for both pre-processing and post-processing. From here we calculated the mean, the median, and the variance and repeated this 1000 times. For every query, the outputs for post-processing are centered around the true query values, which can be seen in Figures 2.1-2.6. However, the outputs for pre-processing are widely spread out. For the mean and median, the queries acting on the privatized database are spread around the true parameter (mean or median, respectively), indicating that the privatized estimator is unbiased. On the other hand, the variance is not near the true value, which indicates bias. This shows us that the pre-processing variance is highly privatized and has low data utility compared to the mean and median.

In the second scenario, we set the privacy budget for pre-processing to the square root of the sample size and for post-processing to 1. We chose this value for pre-processing because the mean is a linear operation and so we can equate the data utility of the mean for both methods using that particular privacy budget. Thus, the privacy budget was around 7 for pre-processing. We observed that the data utility for both methods were similar with the exception of the variance, as the privatized mean and median for both were near their true respective values (Figures 2.7-2.12). The data utility for the variance for pre-processing was low compared to that of post-processing, as the variance for pre-processing was not as close as that of post-processing to the true variance value.

Conclusion

Thus, we can conclude that the data utility of post-processing is better when both privacy budgets are fixed. This can be seen through the graphs below and theoretically by comparing the variances of the mean and median query using both privacy budgets. Additionally, when computing the theoretical variance query, we saw that pre-processing induces a bias, which in turn means that post-processing has a better data utility.

Summary: Data utility is higher for post-processing than pre-processing for the mean, median and variance when using the same privacy budget. Data utility for the mean and the median for both methods are the same when the privacy budget for post-processing is 1 and the privacy budget for pre-processing is $\sqrt{n}$ times the privacy budget of post-processing (i.e., privacy budget of 1). The variance query for pre-processing varies far from its true value due to bias.

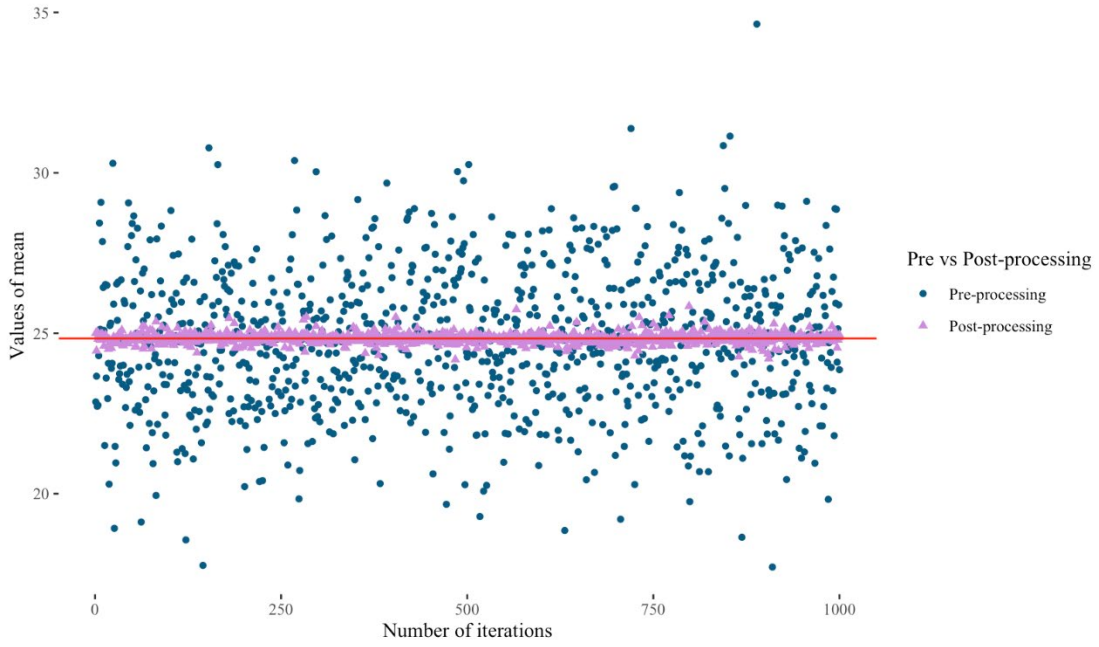


Figure 2.1 A comparison of the mean of BMI between pre– and post-processing DP, using a privacy budget of 1 for both

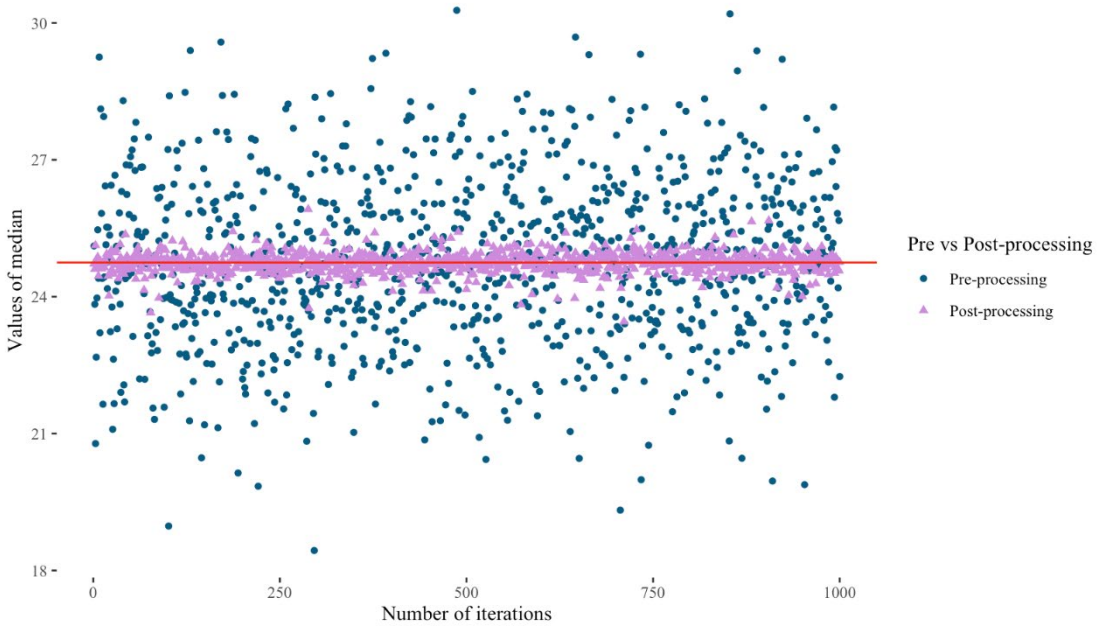


Figure 2.2 A comparison of the median of BMI between pre– and post-processing DP, using a privacy budget of 1 for both

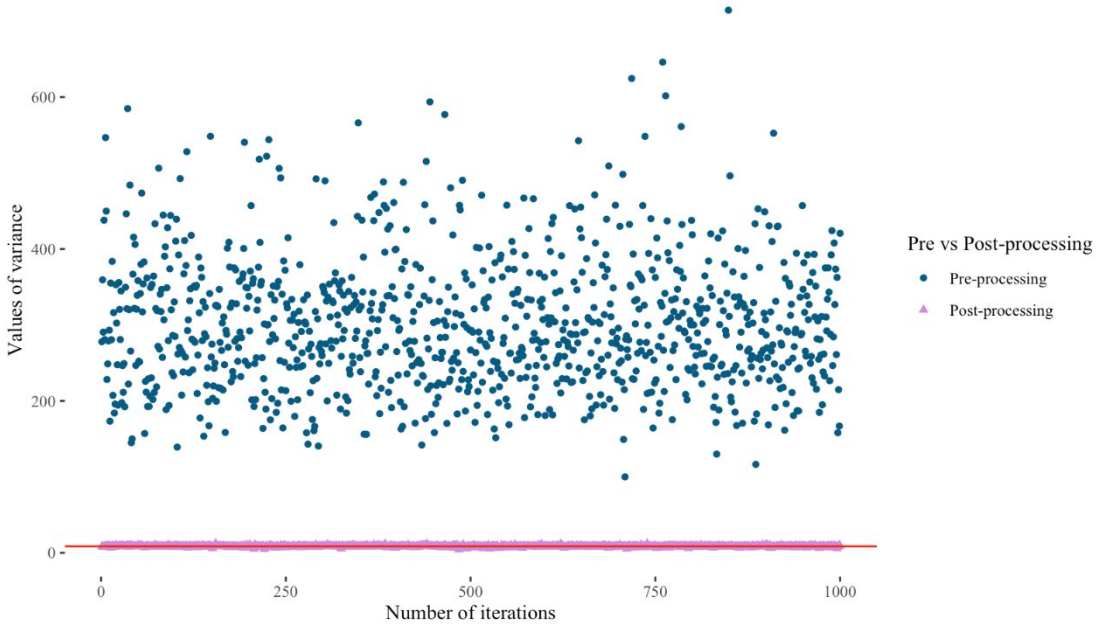


Figure 2.3 A comparison of the variance of BMI between pre– and post-processing DP, using a privacy budget of 1 for both

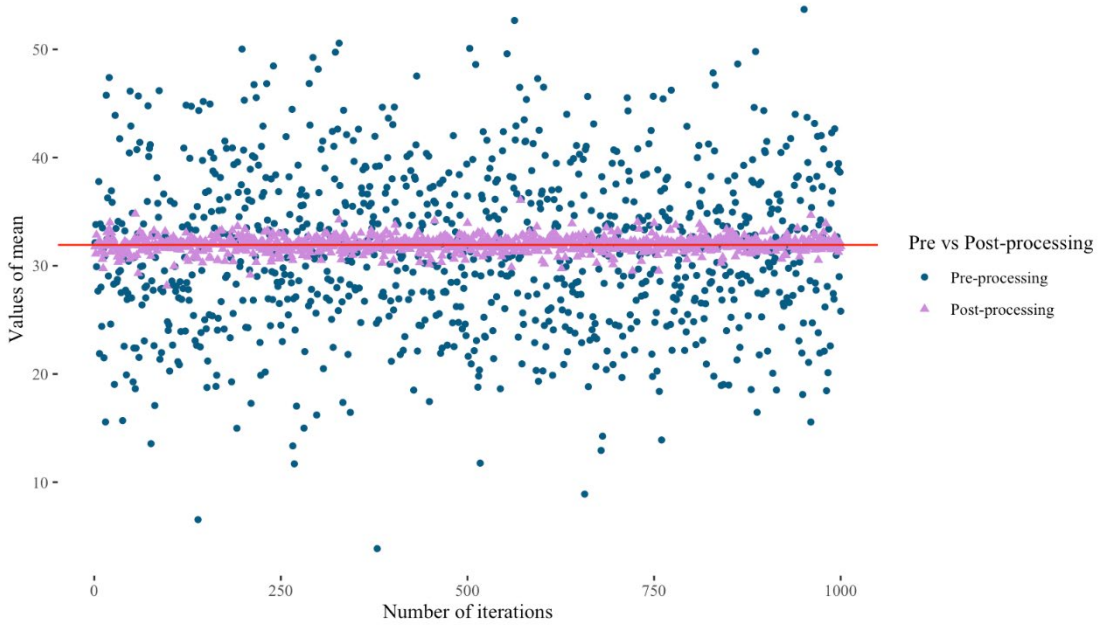


Figure 2.4 A comparison of the mean of age between pre– and post-processing DP, using a privacy budget of 1 for both

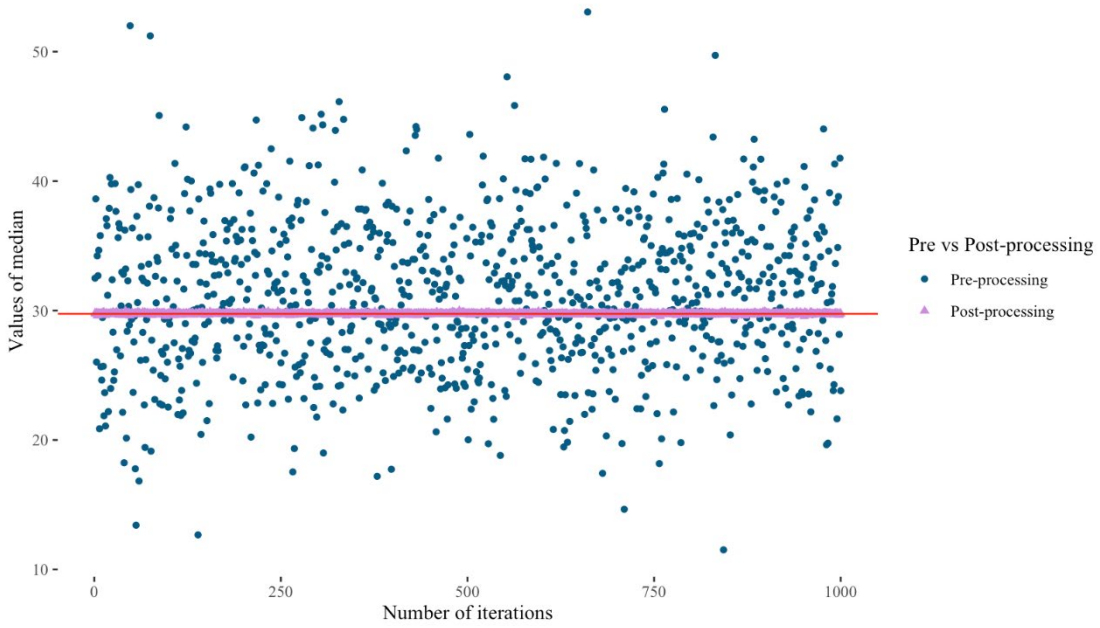


Figure 2.5 A comparison of the median of age between pre– and post-processing DP, using a privacy budget of 1 for both

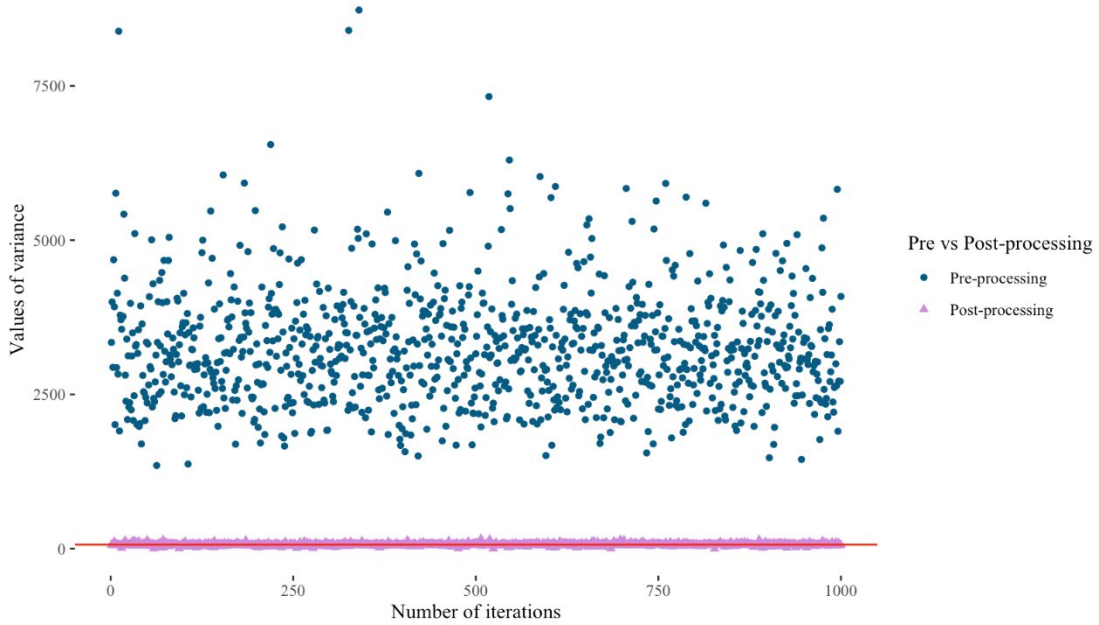


Figure 2.6 A comparison of the variance of age between pre– and post-processing DP, using a privacy budget of 1 for both

Figure 2.7 A comparison of the mean of BMI between pre– and post-processing DP, each using a different privacy budget

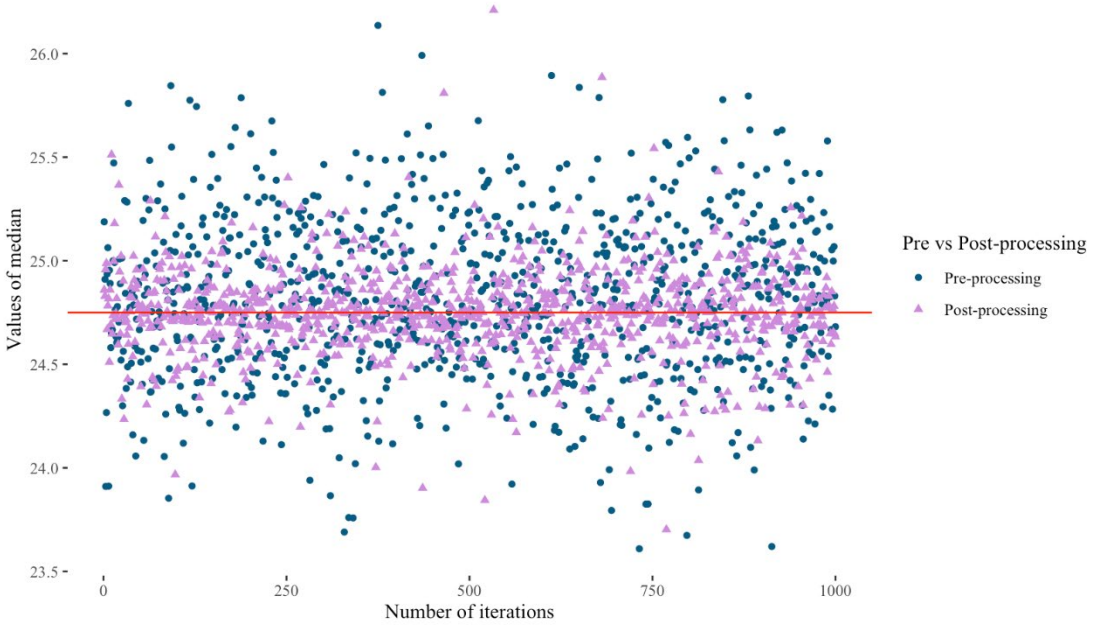


Figure 2.8 A comparison of the median of BMI between pre– and post-processing DP, each using a different privacy budget

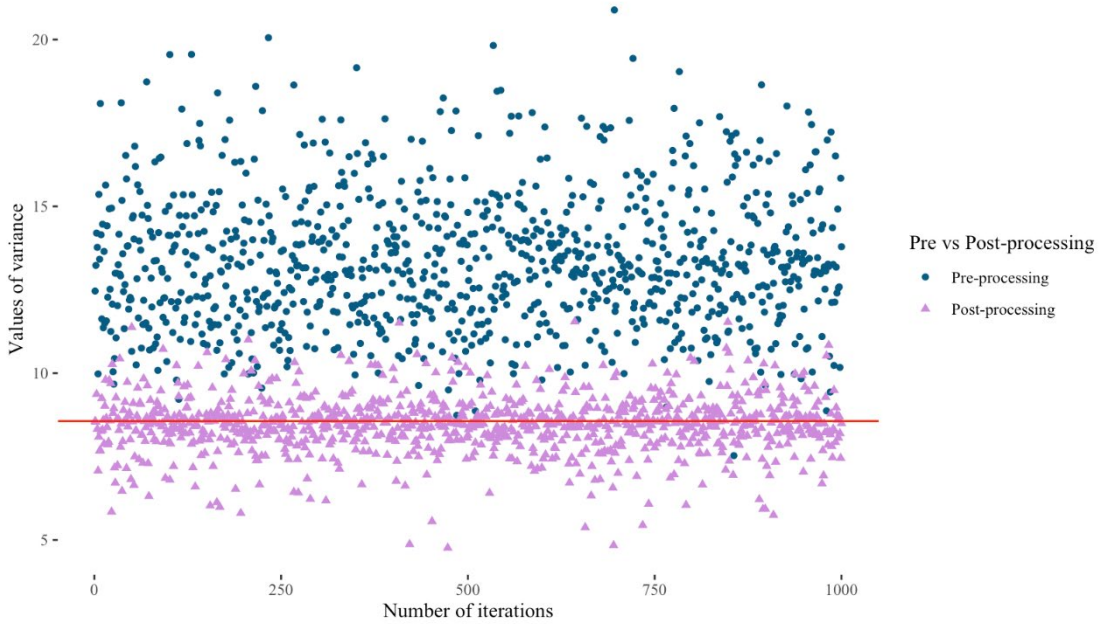


Figure 2.9 A comparison of the variance of BMI between pre– and post-processing DP, each using a different privacy budget

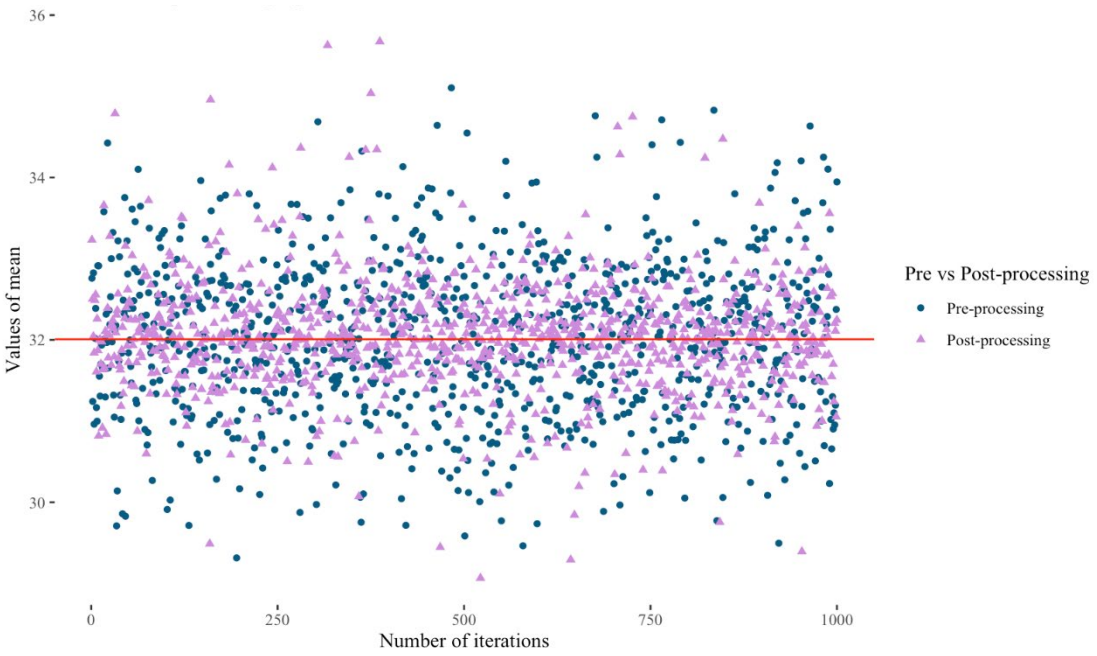


Figure 2.10 A comparison of the mean of age between pre– and post-processing DP, each using a different privacy budget

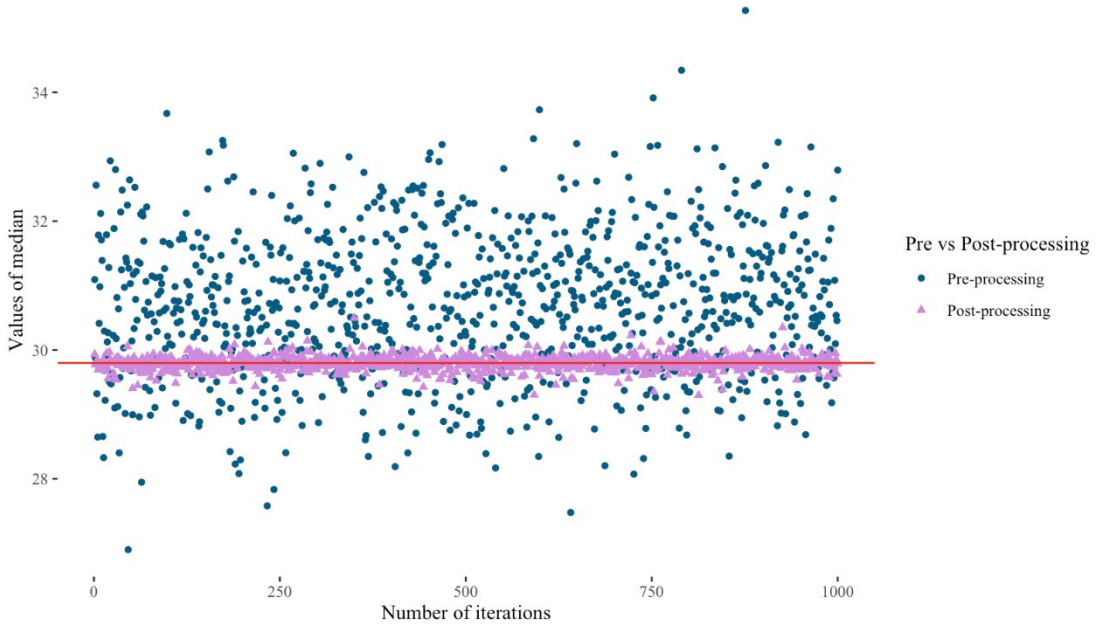


Figure 2.11 A comparison of the median of age between pre– and post-processing DP, each using a different privacy budget

Figure 2.12 A comparison of the variance of age between pre– and post-processing DP, each using a different privacy budget

Experiment 3: Comparison of distributions: original vs. anonymized data

The comparative analysis of two distributions is a technique used in numerous disciplines spanning through different fields such as statistics, data science, economics, and social sciences. This method is popular as it aims to uncover similarities and differences between the variables of two distinct datasets. This experiment was conducted to conceptualize how this process can be adapted to differential privacy concerns. We wanted to explore the impact of the privacy budget on the distribution of the data.

We use the Kolmogorov-Smirnov (KS) test to compare two distributions and explore the limitations or challenges with its application to differential privacy. This experiment was computed to conceptualize the break in the structure of the dataset when noise is added.

This experiment was conducted in four steps. The first two steps, which can be found in the annex, were done to understand what a rejection probability was and how the KS-test can determine the distribution of a dataset. The latter two parts were done using two datasets. One test was performed using two independent datasets and we compared their distributions using the KS-test. This test was done to be able to visualize how the KS-test works with independent samples, for which we have the existing statistical theory. The second test was performed using two dependent datasets. To make the data dependent, we generated the first dataset and then added noise to it to generate the second one. The goal of this test was to visualize how the KS-test works with dependent data obtained via pre-processing differential privacy when we vary the privacy budget.

The main finding of our analysis is that the KS-test does not work with two dependent datasets.

Methodology and conclusions:

Kolmogorov-Smirnov test – independent case

In this part of the experiment, we created two independent datasets of size 100. Both datasets were composed of data generated from a normal distribution with mean 0 and variance 1. We applied the KS-test to the distributions to determine whether they came from the same distribution. We retained the value of the test statistics (called D). Then, we repeated this experiment 1000 times, hence we obtained 1000 values of test statistics. The graph below shows the histogram of the obtained values of KS-test statistics. The shape of the histogram is predicted by the classical theory.

Two sample Kolmogorov-Smirnov test – dependent case

For this part of the experiment, we created a dataset of size 100 that was normally distributed with a mean of 0 and a variance of 1. The second dataset was the first one that underwent pre-processing differential privacy. Thus, the datasets are dependent. As in the independent case, we calculated KS statistics. We repeated this experiment 1000 times. The results are displayed in Figure 3.2. There, we plotted the histogram for the independent case (as above; in blue) and overlayed with the histogram for dependent case stemming from differential privacy (in pink). We note a big difference between the two histograms. This means that the p-values and the rejection probabilities will be calculated incorrectly when applied to anonymized and original data.

Summary: The classical KS-test (and similar tests of goodness of fit) cannot be applied in the context of differential privacy.

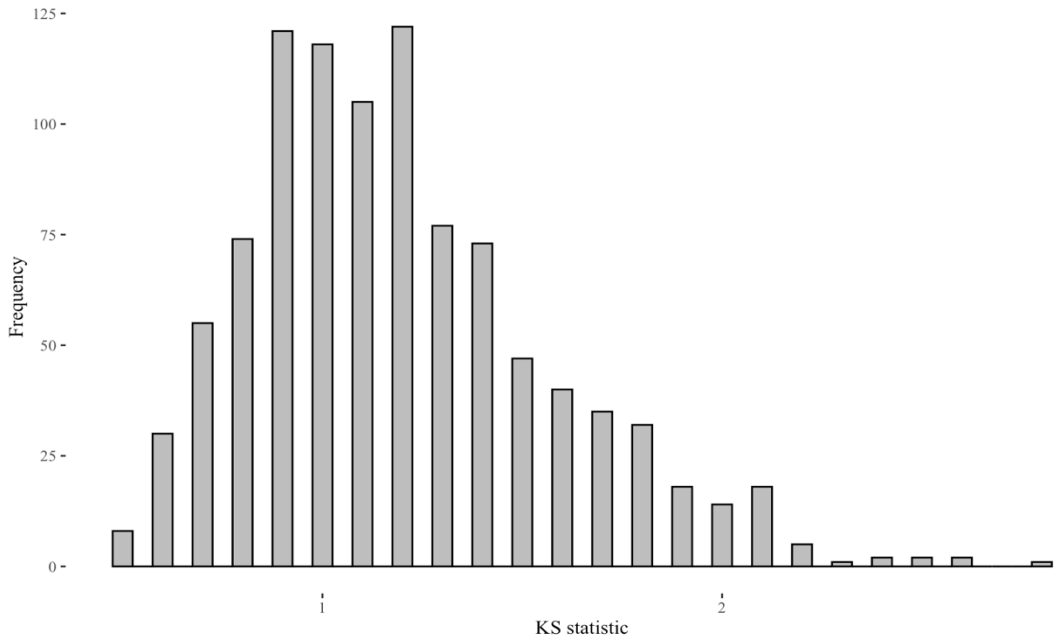


Figure 3.1 Histogram of the KS statistic for each iteration; independent case

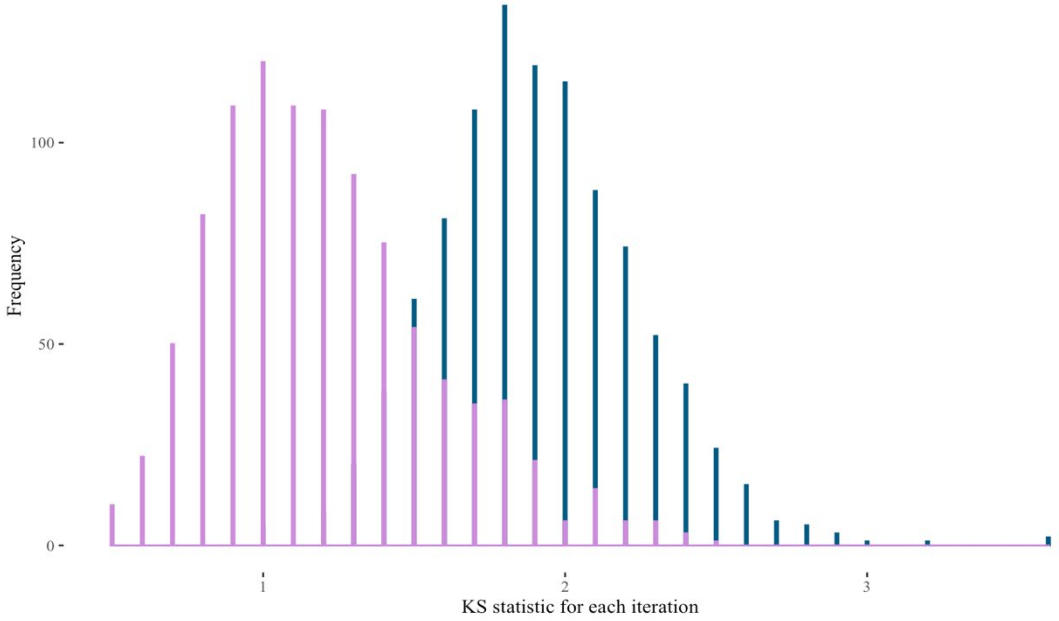


Figure 3.2 Histogram of the KS statistic for each iteration: independent (blue) and dependent (pink) case

Experiment 4: Exploring the Impact of Privacy Budget on Data Utility through Noisy Dataset Analysis

K-anonymity and differential privacy both have their own benefits regarding data privatization. As such, we wanted to see if it was possible to combine both methods to be able to utilize the benefits of both. For example, by combining both approaches, we would be able to "generalize" data while also having perturbed (noisy) data entries.

This experiment was done to see if it was feasible to utilize both privatization methods simultaneously to achieve heightened data protection while preserving adequate utility. Thus, we wanted to determine the effects of noise addition combined with group clustering applied to a dataset in the context of data utility and protection.

To be able to reach a conclusion, we compared the distributions of the original dataset and the dataset on which we applied both differential privacy and k-anonymity. Experiment 2 was performed using the variable "Age" from the same dataset used in Experiment 1.

The main findings are that as we increased the privacy budget for the noise addition, the data resembled the original grouping, thus increasing data utility. Furthermore, we determined at which point the privacy budget no longer influenced the data utility for a specific dataset.

Methodology and conclusions

To properly combine k-anonymity and differential privacy, we applied k-anonymity, with a k of 3, to the dataset. Thus, we grouped the data into three main groups: Age 10-25, Age 26-40, and Age 41-65. Following the grouping, we introduced noise into each group, using pre-processing differential privacy, to try and enhance data privacy while maintaining a decent data utility.

Figure 2.1 portrays the distribution of the three groups created with k-anonymity without the addition of noise. Subsequently, we introduced a significant amount of noise, differential privacy with a privacy budget of 1, which can be observed in Figure 2.2. The distribution of the noisy groups differs significantly from the initial distribution due to the amount of noise added, which in turn highlights the consequences of overly perturbing the data. Following this result, we gradually increased the privacy budget, which signifies adding less noise, until achieving a distribution similar to the initial one. The privacy budget that we had reached when the original distribution matched the noisy distribution was 17. To find this match, we compared the distribution as well as looked at the composition of the groups (number of values in each group) and chose the most similar ones.

As such, we concluded that the combination of k-anonymity and differential privacy can enhance data protection.

Summary: Combining k-anonymity with a small noise perturbation may enhance data privacy without significantly impacting its utility.

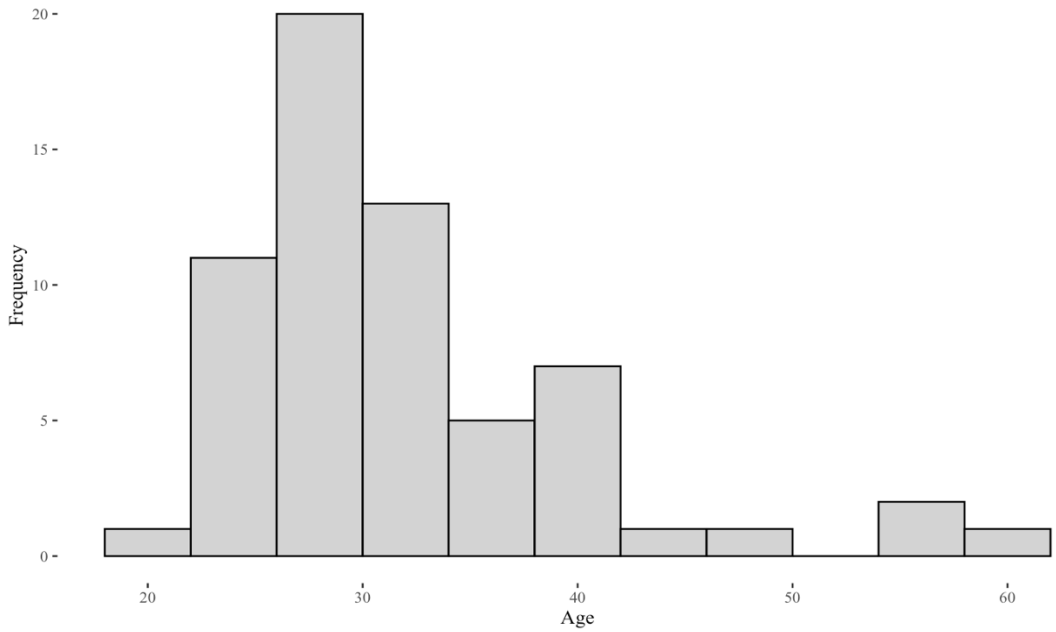


Figure 4.0 Dataset distribution

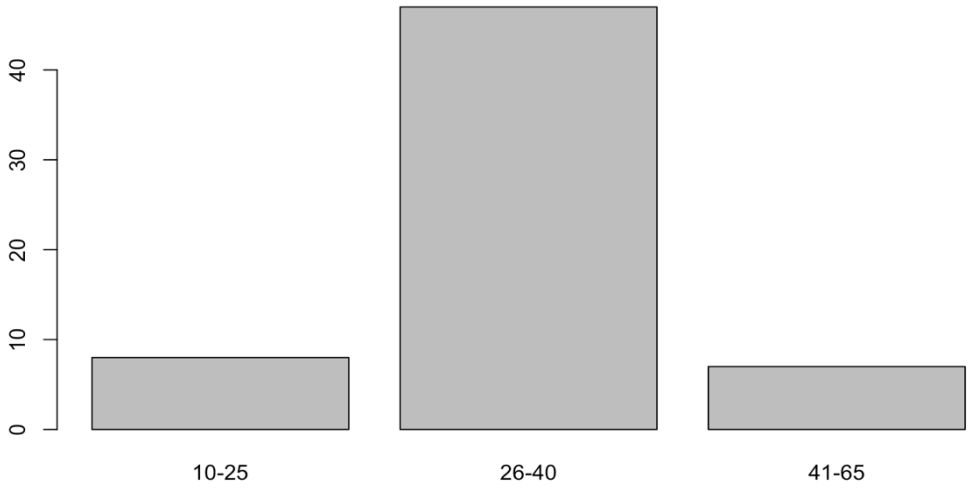


Figure 4.1 Dataset with k-anonymity

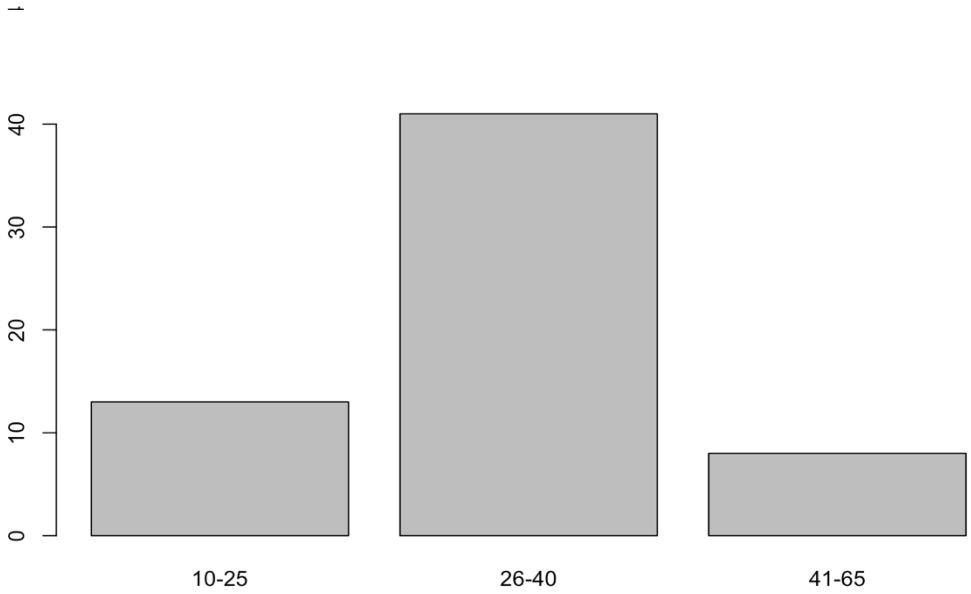


Figure 4.3 Dataset with k-anonymity and differential privacy (privacy budget=17)

Experiment 5: Comparison of Differential Privacy and K-Anonymity

Within the realm of data protection and privacy, two fundamental concepts have garnered increasing significance over time: differential privacy and k-anonymity. Both approaches are geared towards maintaining the confidentiality of data while enabling their utilization for statistical analyses and related operations.

At the core of our experiment lies the inquiry into potential resemblances or shared aspects between these two methodologies, although they may appear to have distinct goals and techniques to achieve privacy. To analyze the resemblances or shared aspects, we sought to equate the privacy budget with the k parameter of k-anonymity.

As k-anonymity uses a different approach to treat discrete (i.e., count data) and continuous data, we conducted the experiment using both. For both data types, we generated a dataset and its anonymized counterpart twice, once using differential privacy and the other using k-anonymity. For both the original and the anonymized datasets, we computed basic statistics such as the mean and the median. We then compared the statistics of the original and anonymized datasets to observe when the anonymization methods led to the same data utility. This allowed us to conclude whether or not both techniques can be compared via a data utility point of view.

After performing the experiment with both types of data, we concluded that a comparison with respect to data utility between differential privacy and k-anonymity does not seem feasible. K-anonymity does not seem to influence data utility in the context of its definition used in this experiment**.**

Methodology and conclusions:

Discrete data:

For this experiment, we generated a dataset using a Poisson distribution. Subsequently, we created various datasets based on the original by applying both anonymization methods.

Starting with k-anonymity, we need to find a group representative. These representatives are used to make the data private using k-anonymity. For the sake of the experiment, we used four different representatives in separate experiments: the group mean, a randomly chosen representative, the minimum and the maximum values for the group. Then, we found the mean squared error (MSE) for each representative. This measures the error by using the average squared difference between the original and anonymized dataset. Afterwards, we repeated these steps for the differentially private dataset but only calculating the mean and median. We then repeated this with different values of K or privacy budget.

The results of this experiment are shown on Figures 5.1 and 5.2. We can discern from these graphs that k-anonymity has a minimal effect on the MSE as it has relatively little variation when k increases. In contrary, when using differentially private data, we observed a gradual decrease in MSE as the privacy budget values increase. This observation aligns with the notion that a larger privacy budget leads to a smaller amount of noise added to the original data.

However, the key takeaway from this test is the challenge in comparing the two results and establishing a direct equivalence between both methods remains a complex task.

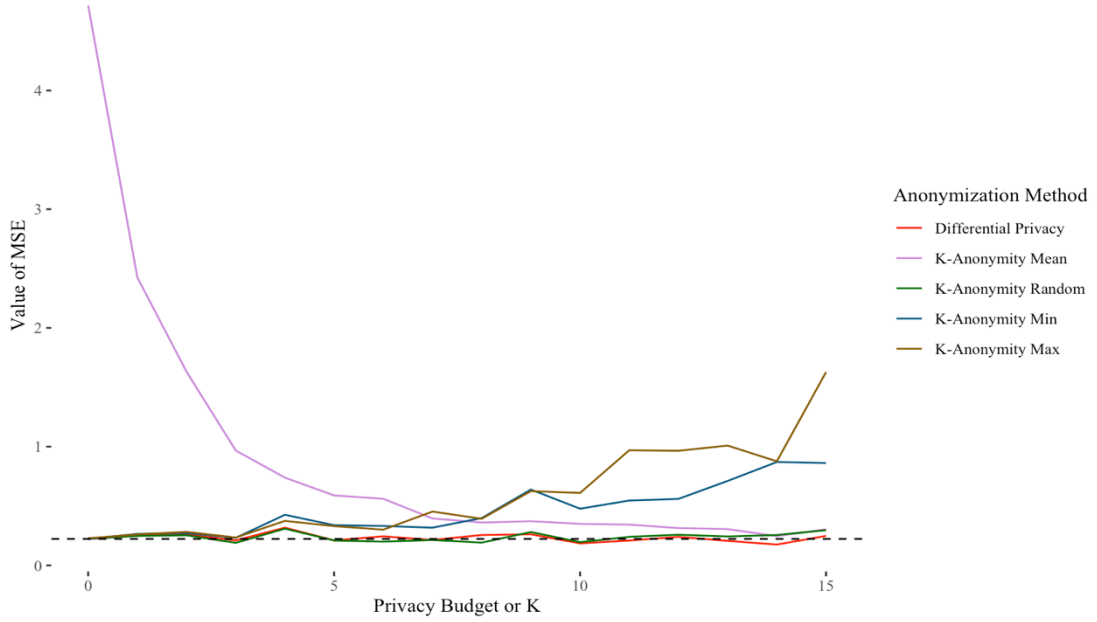


Figure 5.1: MSE for mean and different anonymization methods (four methods of k-anonymization and DP) using a poisson distribution

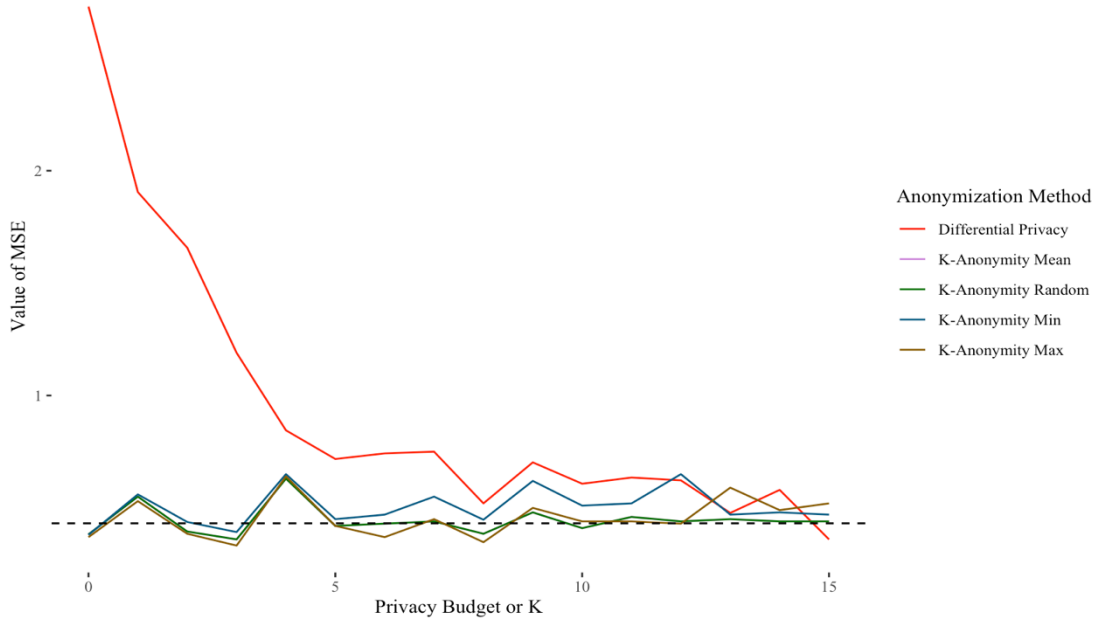


Figure 5.2: MSE for median and different anonymization methods (four methods of k-anonymization and DP) using a poisson distribution

Continuous data

In this scenario, instead of Poisson, we used a normal distribution to generate a dataset. The result are shown below. Our conclusion aligns with that of the discrete case that it is rather infeasable to compare differential privacy and k-anonymity. However, it is important to observe that the Mean Squared Error (MSE) exhibits adverse behaviour in relation to the K-anonymity, when the representaitve of each group is chosen as minimum or maximum. This is due to bias.

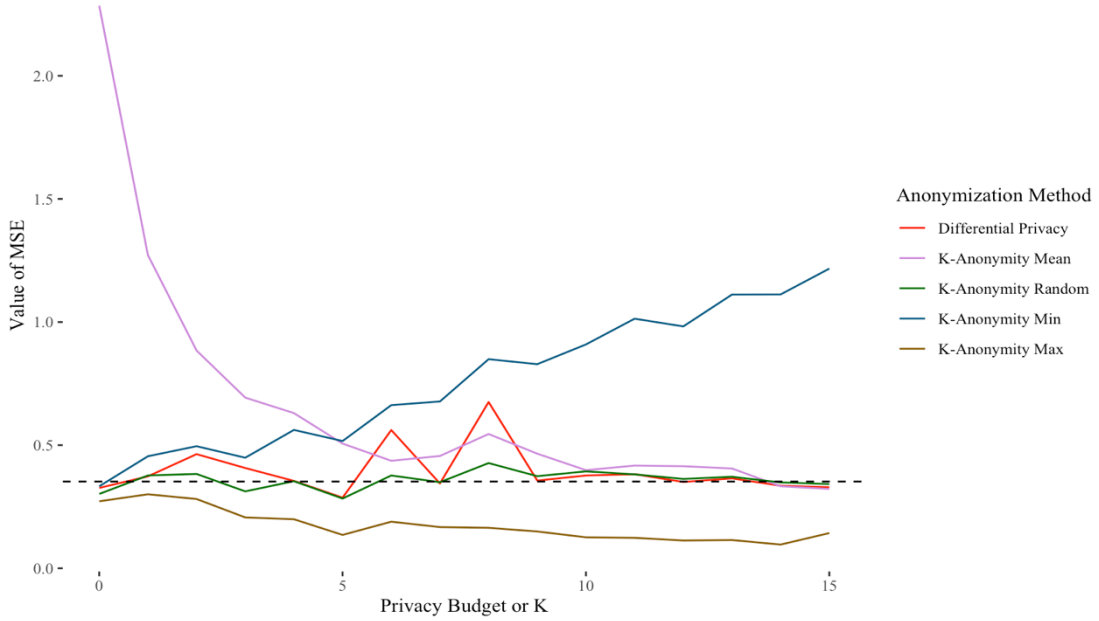


Figure 5.3: MSE for median and different anonymization methods (four methods of k-anonymization and DP) using a normal distribution

Summary: Comparing differential privacy and k-anonymity does not seem feasible and K-anonymity does not seem to influence data utility.

Annex: Experiment 4:

I. One sample parametric test

In general, we have a set of *n* observations coming from an unknown distribution with a mean and a variance and want to test if the mean is equal to a certain number or is different from it. If it is different, then we reject the hypothesis that the mean was equal to a certain value. For example, if we simulate from a normal with mean 0 and variance 1 and test for a mean of 0, the rejection probability should be around 0.05. Otherwise, if we simulate from a different mean, the rejection probability will be higher than 0.05, yielding the power curve.

In this experiment, we simulate n=100 observations from a normal distribution with different means. To demonstrate how this test works, we test to see if our distribution has a mean of 0 and record whether the test rejects it or not. As such, we can see in these graphs that when we test for mean 0, we get a rejection probability of around 0.05. Furthermore, we conclude that for any distribution, the test statistic follows a normal distribution, when the original hypothesis is true.

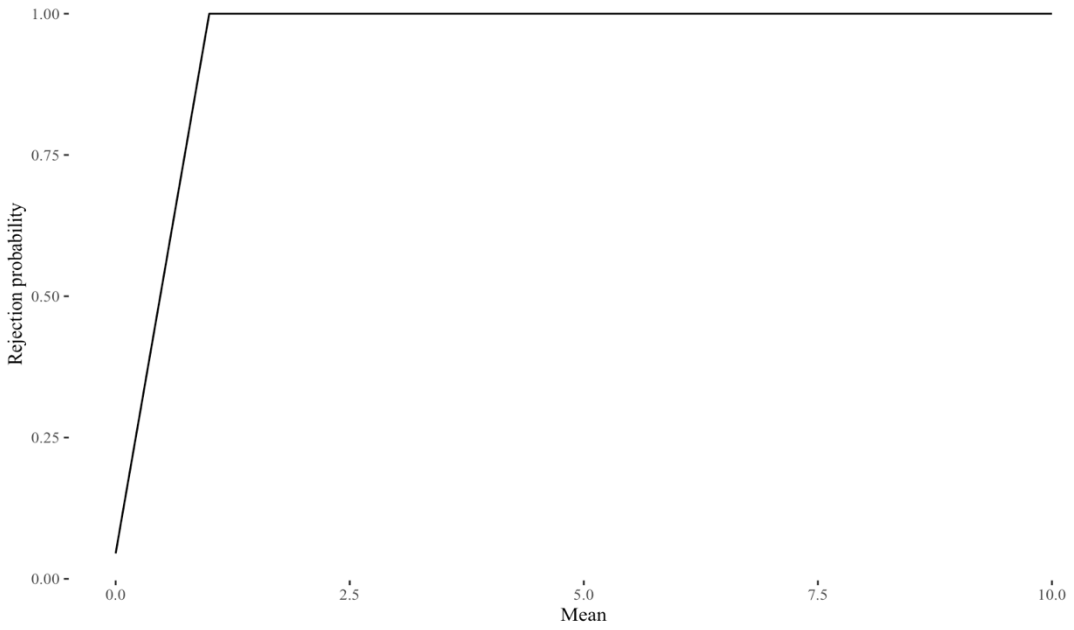


Figure 6.1 Rejection probability graph

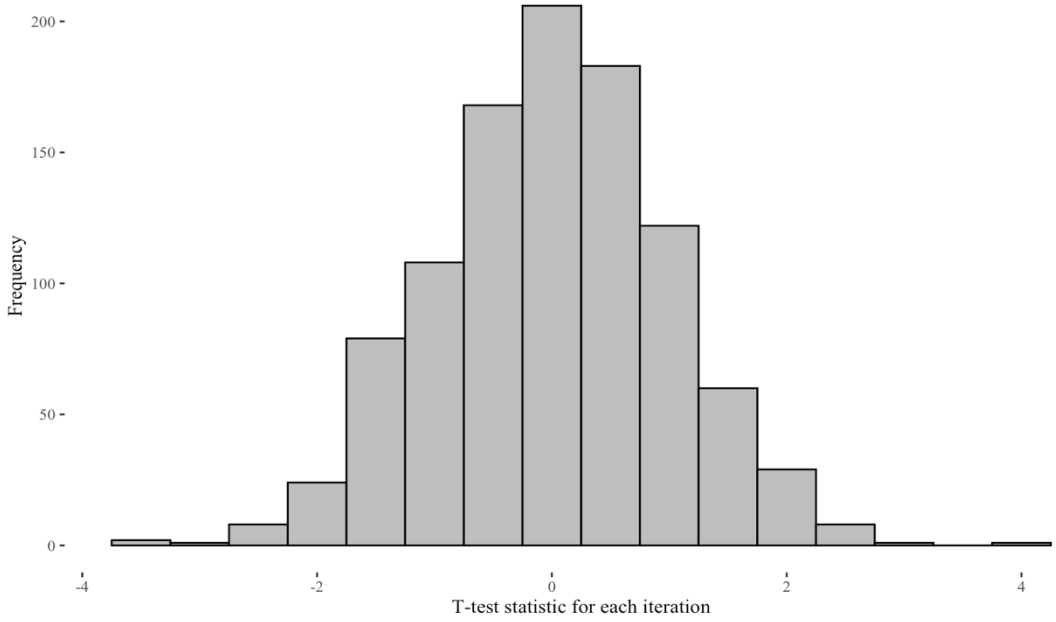


Figure 6.2 Histogram of the KS statistic for each iteration

II. One sample KS-test

We simulated n=100 observations from normal distribution with different means. We tested for whether the distribution has a mean of 0 and whether the test rejects the hypothesis or not.

Following this step, we performed a goodness of fit test. This test is used to verify whether the data follows a certain distribution or not. The associated test statistics should follow so-called Kolmogorov-Smirnov distribution, displayed on Figure 6.5.

In our case, the test obtained a rejection probability close to 0.05 when it was assumed that the data came from a normal distribution with mean 0 and variance 1. However, when it was assumed that it came from a normal distribution with mean 1 and variance 1, it had a rejection probability of 1. Finally, the test works well as its distribution follows the Kolmogorov-Smirnov distribution. Indeed, the histogram of the values of KS statistics resembles the theoretical histogram. This is supported by the classical statistical theory.

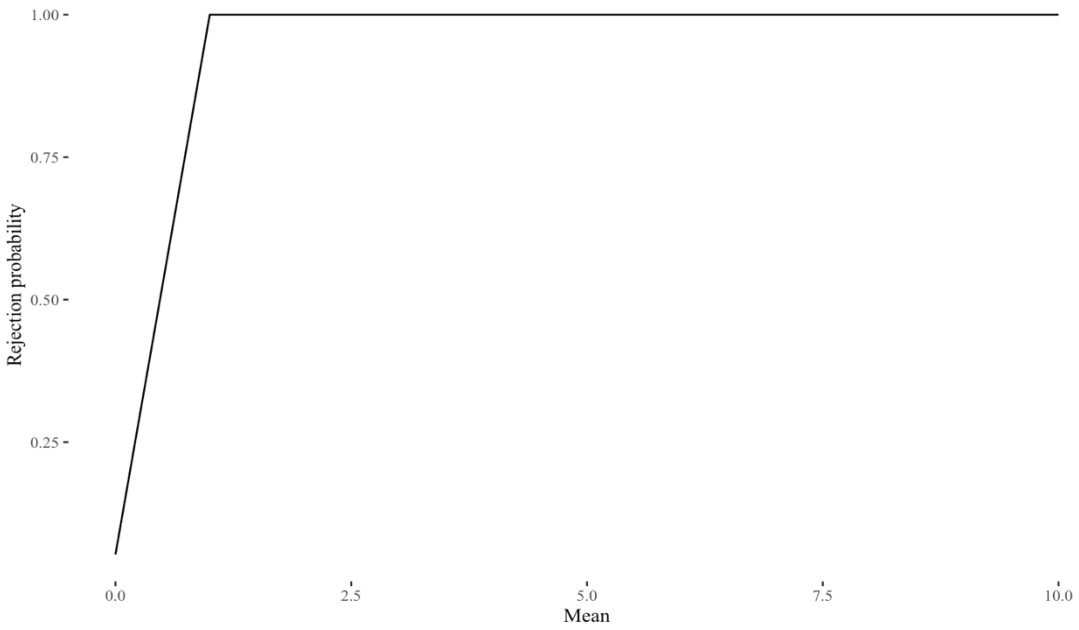


Figure 6.3 Rejection probability

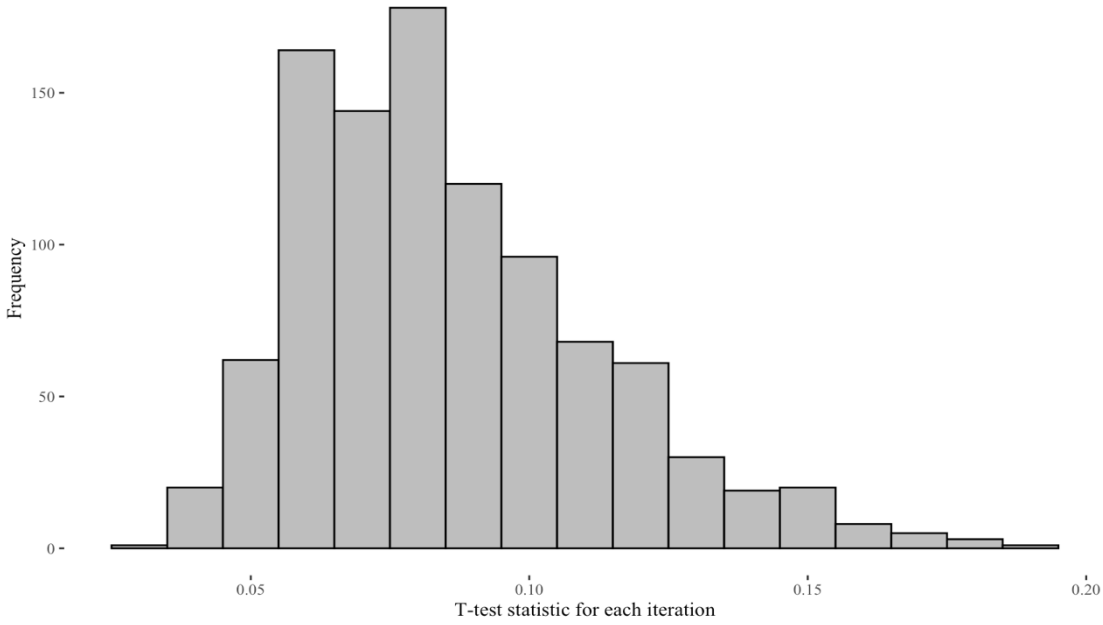


Figure 6.4 Histogram of the T-test statistic for each iteration

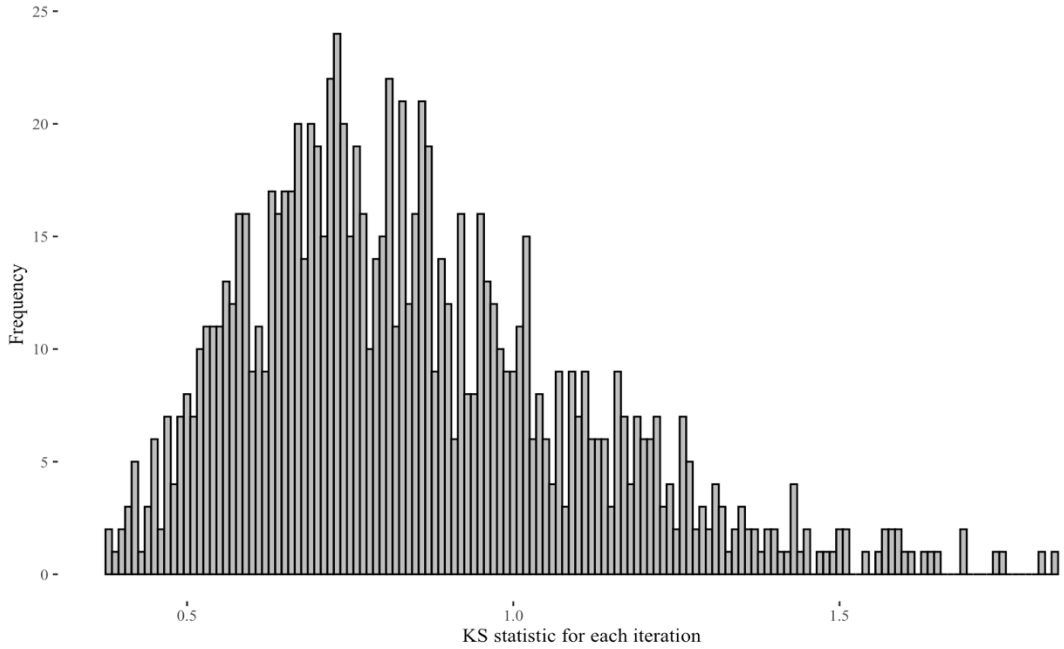


Figure 6.5 Histogram for the goodness of fit of the distribution

Project Phase 3 – Scope of Policies

Overview

The *Personal Information Protection and Electronic Documents Act* (PIPEDA) does not define anonymization, although the definition of "personal information" ("information about an identifiable individual") suggests that properly anonymized data is no longer considered to be "personal information". Bill C-27 – a bill to reform PIPEDA that is currently before Parliament – will make it clear that properly anonymized data will fall outside the scope of the legislation.

Anonymization – particularly k-anonymization – has become an important means of anonymizing data to enable its reuse in a range of contexts. Although there has been considerable development of privacy enhancing technologies that go beyond anonymization, their relationship to the concept of anonymization in data protection law is not always clear. A lack of certainty can lead to underutilization – or improper utilization – of techniques and approaches that have real value in protecting privacy.

This project considers the concept of "differential privacy" and its relationship to PIPEDA and to the Consumer Privacy Protection Act in Bill C-27. Currently, there are no clear guidelines that explain how differential privacy may be aligned with the concept of anonymization in privacy law or how it might relate to the relative approach to anonymization developed in Canadian case law. The first part of this project has explored definitions and techniques of differential privacy, while the second part has used experimentation to test the boundaries of differential privacy in different scenarios, compare differential privacy to other techniques and to establish a framework to compare different outcomes against established disclosure control methods and formal privacy models. This third part of this project will examine how differential privacy can be integrated with legal requirements in PIPEDA and in the proposed Bill C-27.

Anonymization

Although PIPEDA does not expressly refer to anonymization, as noted above, the definition of "personal information" can be interpreted to mean that if individuals are not identifiable in data, the data is not personal information and falls outside the scope of the legislation.[22] The threshold test for identifiability in information that is "not, on its face, personal information"[23] is the "serious possibility" test from *Gordon v. Canada*.[24] According to that test, information is personal information "if "there is a serious possibility" that an individual could be identified through the use of that information, alone or in combination with other available information."[25] The Federal Court has defined a "serious possibility" as "a possibility that is greater than speculation or a 'mere possibility', but does not need to reach the level of 'more likely than not'".[26] This interpretation originally only applied in a freedom of information

[22] Cadillac Fairview re MAC addresses.
[23] Cadillac Fairview, at para 143.
[24]
[25] CF at para 143.
[26] *Canada (Information Commissioner) v. Canada (Public Safety and Emergency Preparedness),* 2019 FC 1279, at para 53.

request, but it's often quoted. The OPC described this as "more than a frivolous chance, less than a statistical probability."

Courts that have interpreted the threshold for "personal information" have made it clear that whether something meets the "serious possibility" threshold (or other similar threshold developed under comparable legislation),[27] will depend on the circumstances of each case. In *Cain v. Canada*,[28] for example, Justice Pentney emphasized the relationship between sensitivity of information and reidentification risk, noting that "the type of personal information in question is a central concern for this type of analysis".[29] Thus, for highly sensitive information, it is important to reduce reidentification risk "as much as is feasible".[30] Other factors that might be taken into consideration could include the difficulty and cost of reidentification as well as the motivation for doing so.[31] In *Cain*, the federal Privacy Commissioner noted the enhanced risks of reidentification that come with "(a)dvancements in technology combined with the proliferation of public or quasi-public data sources magnify the potential for re-identification of datasets unless sufficient precautions are taken" (Intervener's Factum, para 14)."[32] A further important consideration is that, as noted by the OPC, "risk of re-identification is not a static consideration and may increase over time with the improvement of re-identification techniques and the availability of additional resources and data that may be linked to the de-identified dataset."[33]

One of the reasons why the legal test for anonymization is a relative one and not absolute is that "true" anonymization, in the sense of achieving zero risk, is (increasingly) difficult to achieve. Another is that achieving "true" anonymization might lead to a degradation in data quality that seriously undermines the usefulness of the resulting data. These considerations are also part of a balancing approach to privacy rights that considers privacy in relation to other relevant interests. In the access to information context (where tests such as the one in *Gordon* have been developed), the right to privacy is balanced against a right to access information in the hands of government – a right that has some relationship to the constitutionally protected freedom of expression.[34] In this context, the right to privacy is balanced against a range of public or private interests in the use of data. Under PIPEDA, these interests may be commercial in nature. This has sparked some discussion about the nature of this balance, with the current and former federal Privacy Commissioners calling for privacy rights to take precedence over commercial interests in the balancing approach.[35] Nevertheless, privacy commissioners have recognized and supported the public and other interests in the legitimate use of data, and Commissioner Dufresne has stated: "Privacy

[27] Eg in Ontario the test is set out in …and is…
[28]
[29] *Ibid*. at para 107.
[30] *Ibid*. at para 108.
[31] Literature. See also: https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-federal-institutions/2022-23/pa_20230529_phac/ at para 15.
[32] Cain at para 95.
[33] https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-federal-institutions/2022-23/pa_20230529_phac/ at para 18.
[34] Ontario (Public Safety and Security) v. Criminal Lawyers' Association (2010), 319 D.L.R. (4th) 385; 2010 SCC 23
[35]

supporting the public interest and Canada's innovation means that it is not a zero-sum game between privacy rights and public and private interests."[36]

The relative approach to anonymization has been the dominant approach in Canadian case law, both with respect to interpretations of federal and provincial legislation, and it is widely accepted.[37] Quebec's new privacy legislation adopts the relative approach, stating that: "information concerning a natural person is anonymized if it is, at all times, *reasonably foreseeable in the circumstances* that it irreversibly no longer allows the person to be identified directly or indirectly".[38] Bill C-27's proposed *Consumer Privacy Protection Act* (which will reform PIPEDA), currently defines anonymization according to an absolute standard, retaining a relative standard for the term "deidentification".[39] This has caused some debate and confusion. The Canadian Anonymization Network has argued for the adoption of a relative standard in the definition of "anonymization",[40] although the Privacy Commissioner of Canada has expressed support for an absolute standard of anonymization.[41] Clearly, the standard adopted for the definition of anonymization will have a significant impact on what anonymization techniques will suffice to meet this legislative standard.

Anonymization and Differential Privacy

Traditional approaches to anonymization of data have tended to rely upon techniques that involve masking, generalization, and suppression of data.[42] These include k-anonymity[43] and data aggregation. With such techniques, all directly identifiable data is removed from data sets as part of the anonymization process. In addition, there is some modification of the remaining data so that individuals cannot be reidentified based upon particular features or variables in the data. Such techniques may not be an absolute barrier to re-identification, but as noted earlier, the legal test for reidentification risk has been a relative one.

Differential privacy can be distinguished from techniques such as aggregation and k-anonymization because, as discussed in Phase 1, in addition to removing unique identifiers, it adds noise to the data (rather than simply removing or generalizing existing data).

Depending upon the amount of noise introduced in the data, the risk of re-identification may be lowered – even more so than with standard anonymization techniques. However, the amount of noise added to the data may impact its quality and its fitness for some purposes. There is a trade-off between the degree of privacy and the utility of the data. How significant this trade-off is may depend upon variables such as the intended use of the data, since an

[36] https://www.priv.gc.ca/en/opc-news/speeches/2023/sp-d_20230525/
[37] https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf at 2.
[38] Emphasis added. *Act respecting the protection of personal information in the private sector*, CQLR c P-39.1, s. 23.
[39]
[40]
[41]
[42] https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf at 3.
[43]

assessment of fitness for purpose will obviously rely very much how the data are intended to be used.

Necessity and Proportionality

The necessity and proportionality approach common in the human rights context has been relied upon by the federal Privacy Commissioner (and by some provincial privacy commissioners in Canada)[44] as a means of balancing the fundamental right to privacy with the need of a data user to collect, use or disclose personal information. This approach is also used in the European Union.[45] In our view, the necessity and proportionality framework can also serve as a useful guide to thinking about the noise/utility ratio in differential privacy.

Adding noise enhances privacy while adversely impacting data quality. The more noise that is introduced, the more the quality is affected. A necessity and proportionality analysis will examine the necessity of the proposed use of the data and balance it against proportionality considerations. In this context, these could include the connection between the data in question and the desired objective. It might then consider whether the right to privacy is minimally impaired by what is proposed. If the usability of the data requires only a small amount of noise to be introduced, causing greater risks to privacy, it would be reasonable to consider whether other privacy enhancing techniques should be used to better protect privacy. This could include, for example, limited sharing under a restrictive data sharing agreement. Although such data might not be considered "anonymized" (in the sense that they are no longer personal information), they might still be capable of use in the circumstances. By contrast, if the purposes for the use of the data can be achieved with the introduction of more noise such that the reidentification risk is substantially lowered, the data might be considered suitably anonymized. A necessity and proportionality approach assess keeps the right to privacy at the forefront and assesses the application of differential privacy in terms of its ability to properly balance privacy with the fitness for purpose of data.

Guidance

We do not suggest that differential privacy is a technique that is superior to other identification techniques such that it should replace them. Rather, it should be a viable tool in the deidentification toolbox. In our view, general guidance on anonymization would be valuable, and should make clear that different techniques or approaches are available.

At the same time, we believe that there would be value in providing more information regarding differential privacy and how it relates to anonymization under the law. To this end, and based upon our research, we have several recommendations for information specific to differential privacy that could assist organizations looking to anonymize data to choose the appropriate tool, and, if differential privacy is chosen, to understand how to use it appropriately.

---

[44]

[45] https://www.edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en

We have the following recommendations:

I. General guidance

1) **Anonymization can and should allow for the use of differential privacy techniques when incorporated into appropriate risk measures**. It should be clear to those seeking to anonymize data that the appropriate use of differential privacy techniques, as a relative risk measure, qualifies as anonymization for the purposes of the interpretation and application of PIPEDA and, if it is passed, Bill C-27.
2) **Guidance on anonymization should be clear and should allow for the selection of different tools or approaches**. The evolution of privacy enhancing techniques such as differential privacy make it clear that there is no one-size-fits-all approach. Guidance on anonymization should provide for the selection and use of a technique or approach that meets the objectives and that is consistent with best practices, taking into consideration the distinction between relative and absolute measures and how they can be use appropriately to achieve effective anonymization.
3) Quebec's draft anonymization regulation offers an interesting model for the development of general anonymization guidelines. In particular, the regulation combines EU GDPR (no individualization) with EU guidance (no correlation or inference to an individual) with US HIPAA Expert Determination (re-identification risk analysis by an expert to demonstrate very low risk, with regular updates to ensure it remains very low risk. There are also documentation requirements.[46] General guidance of this kind for anonymization, incorporating differential privacy, over the lifecycle of a data could be beneficial.

II. Guidance/information specific to differential privacy

1) Differential privacy should be clearly defined, as a relative risk measure, with an explanation of the relationship between the privacy budget and DP-sensitivity. As noted in Part I of this project, differential privacy is sometimes discussed in the literature in imprecise or problematic ways. Clarity regarding this technique should begin with a careful definition.

2) Differential privacy can be an anonymization technique or it can be used in conjunction with other privacy protective measures. The relationship between relative and absolute risk measures— evaluating the concepts of individualization, correlation and inference to an individual—needs to be further developed and incorporated into guidance. This would provide necessary guardrails for how differential privacy and other disclosure control techniques are incoroproated into responsible data sharing.

Differential privacy inherently involves a balance between privacy and data quality. In fact, the clear trade-off between the two is an advantage of this technique, as it

46

requires the party carrying out the anonymization to understand and factor in the data quality required for a given task. Where tasks do not require highly accurate data, more noise can be introduced, raising the level of privacy accordingly. Where tasks require precise and high-quality data, it may be impossible to maintain quality while sufficiently protecting privacy using differential privacy. Where the noise/utility ratio will leave a certain level of reidentification risk, this could also be a signal that anonymization is not possible and that other privacy enhancing approaches should concurrently be adopted to properly protect the data. This could mean, for example, that in some cases, even where differential privacy tools are used, the resulting data sets are unsuitable for broader distribution and should only be shared under a carefully drafted data sharing agreement.

3) **Calibrating the noise/utility ratio to the factors for reidentification risk could be a useful part of guidance**.

Reidentification risk can be a function of several factors, many of which have been identified by courts. The sensitivity of the data at issue is a key factor. The more sensitive the data, the higher the risk of reidentification – both in terms of the likelihood that an adversary will attempt reidentification and in terms of the impact on individuals if reidentification occurs. Ideally, these factors could be identified and articulated in general guidance on anonymization.

In the case of differential privacy, highly sensitive data may require the introduction of higher levels of noise. The introduction of noise will impact the quality of the data which could be a problem for proper use of the data. For example, health-related data are highly sensitive, but accuracy may also be very important to the quality of research output. An assessment of noise/utility ratio using differential privacy could be helpful in determining whether such data can be fully anonymized in the legal sense of the term, or whether their use must necessarily be accompanied by other privacy-enhancing techniques including data sharing agreements.

Bibliography (Phase 3)

Teresa Scassa, "Court decision explores reidentification risk in access to information request", February 7,2023. https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=368:court-decision-explores-reidentification-risk-in-access-to-information-request&Itemid=80

Office of the Privacy Commissioner of Canada, "PIPEDA fair information principles", May 2019 https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/

Office of the Privacy Commissioner of Canada, Gordon v. Canada (Health), 2008 FC 258 Case Summary, June 2014 https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-privacy-act/summaries-of-leading-privacy-act-federal-court-cases/lpac/lpac_019/

Gazette Officielle Quebec, Anonymization of personal information. (2023). Part II,Vol. 155, No. 51 144(4), https://www.publicationsduquebec.gouv.qc.ca/fileadmin/gazette/pdf_encrypte/lois_reglements/2023A/106606.pdf

Office of the Privacy Commissioner of Canada, "Summary of privacy laws in Canada", January 2018, https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/

European Data Protection Supervisor, "Necessity & Proportionality", https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en

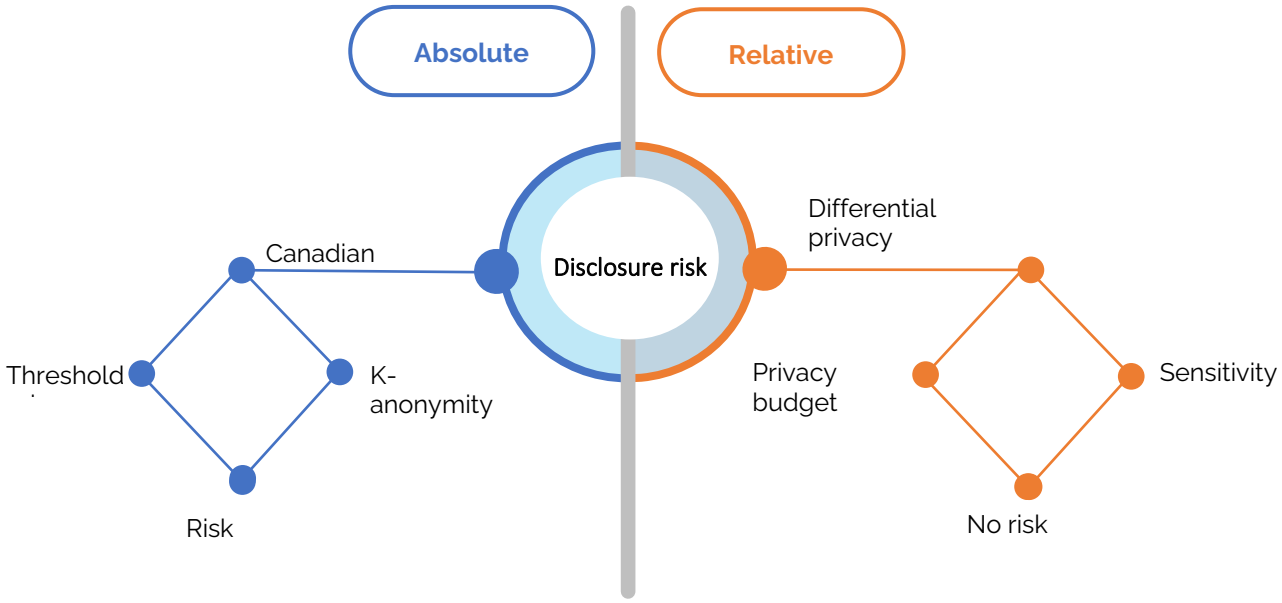
Canadian Anonymization Network, "Proposed amendments to the de-identification and anonymization provisions in the Digital Charter Implementation Act, 2022 (Bill C-27)",December 7, 2022 https://deidentify.ca/wp-content/uploads/2022/12/CANON-Proposed-Amendments-to-Bill-C-27-Dec-7-2022.pdf

Appendices

1) Appendix A: Absolute and Relative Risk
2) Appendix B: Differential Privacy
3) Appendix C: Canadian guidance to anonymization
4) Appendix D; Some legal definitions
5) Appendix E: Tables with definitions related to Differential Privacy
6) Appendix F: Survey

Appendix A: Absolute and Relative Risk

What is the difference between absolute and relative risk?
Absolute and relative risk are used abundantly in health sciences as a way to measure risk. Absolute risk difference measures the difference between risks in two groups (risk of the first group minus risk of the second group), and relative risk measures a ratio that says how likely risk is to increase or decrease in two groups (risk of the first group divided by risk of the second group). For example, let's compare the risk of developing skin cancer using sunscreen versus not using sunscreen. After having surveyed 200 000 individuals, 100 000 using sunscreen and 100 000 not using sunscreen, we see that the risk of developing skin cancer is twice as likely if you don't wear sunscreen. This means that the risk of developing skin cancer doubles when you do not apply sunscreen. However, we also see that the difference of developing skin cancer between applying it and not was 20 Individuals per 100 000 individuals, or 0.0002%. The absolute risk tells us the risk of getting cancer between the two groups and the relative risk tells us that the risk of cancer increases if you don't wear sunscreen.



What is absolute and relative risk in anonymized data?

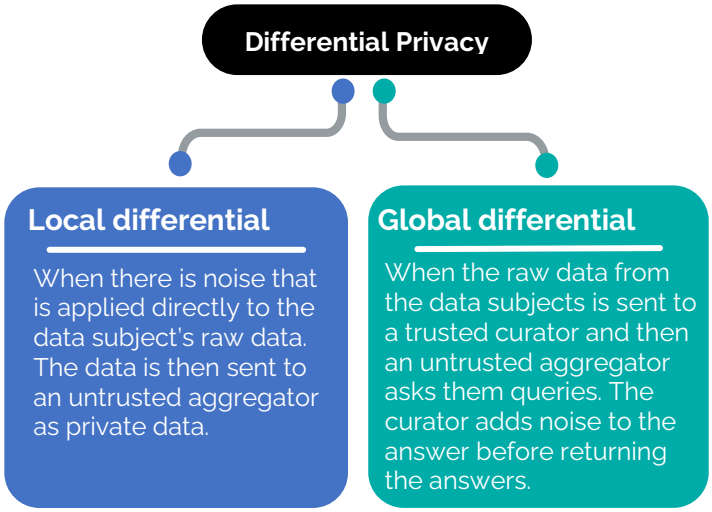
The Canadian guidance on anonymization uses a form of absolute risk to determine an acceptable level of disclosure. For example, a group of twenty people on the same identifying information represents a risk of one over twenty, or 0.05, that their names can be randomly assigned. This approach makes it possible to compare the overall level of risk between two datasets. Differential privacy[47] provides a relative measure since it compares two databases, one that has all the individuals and one that has all but one individual. As a relative measure, differentially private outputs can only be compared on the database in which it is being applied. Two different databases will have different relative measures that are independent of one another. Unlike the techniques mentioned previously, differential privacy needs to be extended to include a relative risk metric.

[47] Differential privacy is a technical privacy model that protects individuals requiring that the information contributed by any individual does not significantly affect the output.

Appendix B: Differential Privacy

The purpose of this Appendix is to showcase the different parameters of differential privacy and its techniques. It is also to show that you cannot have a clear understanding of the outputs of differential privacy using only the privacy budget. Instead, you must know the technique that is used, the privacy budget and the sensitivity function.

DEFINITIONS – Differential Privacy and Parameters
Differential privacy: Differential privacy is a technical privacy model that protects individuals by requiring that the information contributed by any individual does not significantly affect the output.
Privacy budget: The privacy budget is used to specify the level of protection in a given dataset or statistic. The privacy budget can be interpreted as a tuning parameter that trades privacy for accuracy.
Sensitivity (DP-Sensitivity): The sensitivity function measures the maximum potential change in output.



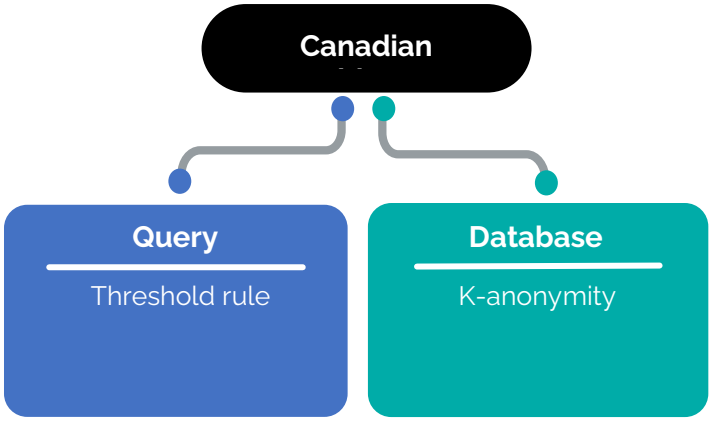
Considerations:

- Differential privacy is a relative risk as we are comparing a dataset with all the individuals to one that has a missing individual. However, at this moment, it cannot become a relative risk metric. As such, we cannot compare two differentially private datasets or queries to one another.
- There are different variations of differential privacy that are used. Although they differ by the parameters that each variation uses, they are all based off the same concept of indistinguishability between individuals and the mathematical definition of privacy.
- The noise cannot be solely determined by the privacy budget or the sensitivity it is the combination of the two that will determine the noise added.

Appendix C: Canadian guidance to anonymization

The purpose of this one-pager is to show what the Canadian guidance uses for data anonymization. It also shows the different parameters of the techniques used in the guidance and to clarify the understanding the basic aspects of these techniques.

<table>
<tr><th colspan="4">DEFINITIONS – K-Anonymity , Threshold Rule and their parameters</th></tr>
<tr><td colspan="4">K-anonymity: K-anonymity is a formal privacy measurement model that ensures that for each identifier there is a corresponding equivalence class containing at least K records</td></tr>
<tr><td colspan="4">Threshold rule: A minimum number of data principals in a selected attribute is defined by a threshold, n, below which the number of data principals in the selected attribute is deemed sensitive.</td></tr>
<tr><td>Maximum risk:
the maximum level of identifiability for a single data principal is taken, measured across all data principals in the shared or released data</td><td>Average risk:
the average level of identifiability for a single data principal is taken, measured across all data principal in the shared or released data</td><td>Population attack:
an adversary knows the targeted entity is in a defined population in the data made available</td><td>Sample attack:
an adversary does not, or cannot, know if the targeted entity is in the data being made available, most</td></tr>
</table>



Appendix C: Some legal definitions Interpretations

This Appendix is related to Phase 3

CANON definitions of Anonymization and De Identification and applicable provisions

Anonymization: to irreversibly and permanently modify personal information, in accordance with generally accepted best practices, to ensure that no individual can be identified from the information, whether directly or indirectly, by any means.

De Identify: to modify personal information — or create information from personal information — by using technical processes to ensure so that the information does not identify an individual or could not be used in reasonably foreseeable circumstances, alone or in combination with other information, to identify an directly identified from it, though a risk of the individual."

"This Act does not apply in respect of personal information that has been anonymized."

"An organization may use an individual's personal information without their knowledge or consent for the organization's internal research, analysis and development purposes, if the information is de-identified before it is used."

"A person who carries out any regulated activity and who processes or makes available for use anonymized data in the course of that activity must, in accordance with the regulations, establish measures with respect to (a) the manner in which data is anonymized; and (b) the use or management of anonymized data."

Applicable sections of Quebec Draft Regulation: respecting the anonymization of personal information

"**correlation criterion**" means the inability to connect datasets concerning the same person;

"**individualization criterion**" means the inability to isolate or distinguish a person within a dataset;

"**inference criterion**" means the inability to infer personal information from other available information;

5. At the beginning of a process of anonymization, a body must remove from the information it intends to anonymize all personal information that allows the person concerned to be directly identified. The body must then conduct a preliminary analysis of
the re-identification risks considering in particular the individualization criterion, the correlation criterion and the inference criterion, as well as the risks of other information available, in particular in the public space, being used to identify a person directly or indirectly.

6. On the basis of the re-identification risks determined in accordance with the second paragraph of section 5, a body must establish the anonymization techniques to be used, which must be consistent with generally accepted best practices. The body must also establish protection and security measures to reduce re-identification risks.

7. After implementing the anonymization techniques established for the process of anonymization and the protection and security measures in accordance with section 6, a body must conduct an analysis of the re-identification risks. The results of the analysis must show that it is, At all times, reasonably foreseeable in the circumstances that the information produced further to a process of anonymization irreversibly no longer allows the person to be identified directly or indirectly.

For the purposes of the second paragraph, it is not necessary to demonstrate that zero risk exists. However, taking into account the following elements, the results of the analysis must show that the residual risk of re-identification is very low: (1) the circumstances related to the anonymization of personal information, in particular the purposes for which the body intends to use the anonymized information; (2) the nature of the information; (3) the individualization criterion, the correlation criterion and the inference criterion; (4) the risks of other information available, in particular in the public space, being used to identify a person directly or indirectly; and (5) the measures required to re-identify the persons, taking into account the efforts, resources and expertise required to implement those measures.

8. A body must regularly assess the information it has anonymized to ensure that it remains anonymized. For that purpose, the body must update the analysis of the re-identification risks it conducted under section 7. The update must consider, in particular, technological advancements that may contribute to the re-identification of a person. The results of the analysis must be consistent with the second paragraph of section 7. If they are not, the information is no longer considered anonymized."

Definition of necessity and proportionality from European Data Protection Supervisor (EDPS):

1: **Necessity**: organizations should only pursue privacy-invasive activities and programs where it is demonstrated that they are necessary to achieve a pressing and substantial purpose.

2: **Proportionality**: organizations should only pursue privacy-invasive activities and programs where the intrusion is proportional to the benefits to be gained. Proportionality restricts authorities in the exercise of their powers by requiring them to strike a balance between the means used and the intended aim.

Appendix E: Tables with definitions

Table of definitions for differential privacy

ISO 20889	Differential privacy is a formal privacy measurement model that ensures that the probability distribution of the output from a statistical analysis differs by at most a specified value, whether any particular data principal is represented in the input dataset. It bounds the probability that the presence or absence of any particular data principal in the dataset is able to be inferred from the de-identified dataset or from system responses.
UKAN	Differential privacy is a system which limits the amount of information specific to any individual that can be revealed by an analysis. It is a guarantee, not a risk assessment model. The guarantee is a limit on the amount of information that is revealed by an analysis.
Harvard Kennedy School	Differential privacy is a safeguard used to protect an individual's data privacy. It allows for the collection and publication of data patterns and trends, while protecting the privacy of individuals captured in a dataset. Differential privacy is not a tool or method, but rather a criterion or a property that multiple methods can achieve. More specifically, it is a mathematical definition of privacy that quantifies privacy risk. It considers a maximum level of privacy loss, called the privacy loss parameter, and manipulates the content of a dataset in order to achieve that level of privacy, while maintaining the utility and accuracy of a dataset.
PET guide	Differential privacy provides an information-theoretic notion of Output Privacy. Its goal is to quantify the maximum amount of information about individual records in a database that could be leaked by releasing the result of any computation on that database. It specifies a property that a data analysis algorithm must satisfy to protect the privacy of its inputs. In this sense, DP is a privacy standard, rather than a single tool or algorithm. The DP property is stated in terms of an alternate world where the input of a particular individual has been removed from or added to a database. DP requires that the outputs produced by the algorithm in the real and alternate world are statistically indistinguishable. Differential privacy offers a mathematical guarantee to individuals contributing sensitive data to a database on which certain queries will be performed.
NIST (Published)	Differential privacy is a set of technique based on a mathematical definition of identity disclosure and information leakage from operations on a dataset. Differential privacy prevents disclosure by adding non-deterministic noise (usually some random values) to the results of mathematical operation before the results are reported. Differential privacy's mathematical model holds that the result of an analysis of a dataset should be roughly the same before and after the addition or removal of a single data record (which is usually taken to be the data from a single individual).
OPC	Differential privacy offers organizations a formal method for preserving a certain amount of privacy. At its core, differential privacy involves adding a mathematically defined amount of "noise" – or fake data - to a dataset. The noise is added using an equation that makes it very difficult, if not impossible, to tell who or what was in the original dataset.

UK Royal Society	Differential privacy is a security definition which means that, when a statistic is released, it should not give much more information about a particular individual than if that individual had not been included in the dataset. Differential privacy also allows for risk to be quantified as the probability of reidentification, allowing the controller to 'dial up or down' and adjust for performance-privacy trade-offs by referring to a set 'privacy budget' or how much data is determined acceptable to be leaked from the site. Differential privacy could be used to add 'noise', to make any one true datapoint more difficult to trace to a real individual. The resulting 'noisy' dataset can then be shared more safely.
US National Science and Technology Council	Differential privacy, a data perturbation approach, adds noise to the original data in such a way that an adversary cannot tell whether any individual's data was or was not included in the original dataset.
OECD	These techniques make small changes (add noise) to the raw data to mask the details of individual inputs, while maintaining the explanatory power of the data. The idea is that small changes to individual records can securely de-identify the inputs without having a significant impact on the aggregated results. Differential privacy is relevant as a PET because it provides data subjects with some protection of deniability in cases where someone attempts to re-identify released data. Noise introduced into the dataset should not alter any large-scale analysis but makes any individual data less reliable and protective for the data subjects.
World Economic Forum	Differential privacy is when noise is added to a dataset so that it is impossible to reverse-engineer the individual inputs.
UN Economic Commission for EU	Differential privacy is neither a method nor an algorithm, but a definition supporting a mathematical disclosure-control framework. Thus, despite its name, the intended use of DP in official statistics is used to prevent disclosure when releasing statistical information rather than to address privacy concerns when gathering personal data from individuals.
Future Privacy Forum (*Emerging Privacy Tech Sector*)	Differential privacy used to assess mathematical guarantees of disclosure control for a particular privacy model.
Future Privacy Forum (*Buying Privacy Tech*)	Type of privacy: Organization releasing statistics or derived information – generally an organization that holds a large amount of data
Differential Privacy for Government Agencies – Are we there yet?	DP is a query response system. A system of this kind accepts specific queries as input – a query for the mean of a variable, for example - and then returns a noisy answer to the query, with the noise calibrated to ensure that the requirements of DP are met. Differential privacy guarantees that the influence that any record in a database can have on the reported output is strictly limited. This ensures that the information that can be learned about any individual in the database is also limited. These guarantees are made by bounding the difference of the probability distribution of the query response when changing

	one record in the data
Differential privacy in practice - expose your epsilons!	Differential privacy hides the presence or absence of any individual, or small group of individuals, in a dataset, in the sense that, for each individual, any conclusion reached from the analysis would be essentially as likely to have been reached, whether the given individual joined, or refrained from joining, the dataset. Differential privacy is a mathematical definition of privacy tailored to statistical data analysis.
Differential privacy: a primer for non-technical audience	Differential privacy is a formal mathematical framework for quantifying and managing privacy risks. It provides provable privacy protection against a wide range of potential attacks, including those currently unforeseen. Differential privacy mathematically guarantees that anyone viewing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.
Towards effective differential privacy communication for users' data sharing decision and comprehension	Differential privacy protects an individual's privacy by perturbing data on an aggregated level (DP) or individual level (LDP). To protect data privacy and ensure utility in the context of data publishing, the concept of differential privacy has been proposed, which adds noise to the aggregated result such that the amount of revealed information for any individual is bounded.
Issues Encountered Deploying Differential Privacy	Differential privacy provides a mathematical definition for privacy loss to individuals associated with the publishing of statistics based on their confidential data. Today the differential privacy literature provides numerous mechanisms for privacy preserving data publishing and privacy preserving data mining while limiting the resulting privacy loss to mathematically provable bounds
Differential Privacy and Federal Data Releases	It promises to protect attackers from learning whether or noy any individual is in a database regardless of the background information held by the attackers. As such, it provides a strong guarantee of Privacy & Confidentiality protection, even against the worst scenarios. The definition involves probability bounds.
ICO	Differential privacy generates anonymous statistics. This is usually done by randomising the computation process that adds noise to the output. Differential privacy is a property of a dataset or database, providing a formal mathematical guarantee about people's indistinguishability. It is based on the randomised injection of noise.

Table of definitions for epsilon

ISO 20889	**Privacy budget** is a design choice, not a straightforward process. If the value is large: there is a smaller standard deviation, typically spending more of the privacy budget when answers are provided to users, but also carries a greater privacy risk because smaller noise values are more likely to be added to the actual results. If the value is small: it increases magnitude of the standard deviation, thus increasing the likelihood that larger noise values are added to the actual results, providing greater privacy protection
Harvard Kennedy School	**Privacy loss parameter** determines the upper bound for privacy loss. It represents a trade-off between privacy and accuracy, since the parameter determines how much noise will be added to the dataset. A smaller value means greater privacy protection, but less accurate output. A larger value results in a more useful analysis, but less privacy protection.
PET guide	The adversary's inability of determining the presence of a record in the database is measured in terms of the similarity between the probability distributions over outputs when the record is either present or missing in the database. This similarity measure is parametrized numerically (typically represented by Greek letters epsilon and delta), with smaller values of these parameters representing a stronger privacy protection. Furthermore, privacy budgets are typically maintained by a technical component called a privacy accountant. These budgets take into account the previous queries made and how information from these queries can compound with one another to leak a greater level of information then each individually in isolation.
NIST (Draft)	**Privacy loss:** A measure of the extent to which a data release may reveal information that is specific to an individual. **The degree of sameness** is defined by the parameter epsilon. The smaller the parameter, the more noise is added, and the more difficult it is to distinguish the contribution of a single individual. The result is increased privacy for all individuals – both those in the sample and those in the population from which the sample is drawn who are not present in the dataset
NIST (Published)	**The degree of sameness** is defined by the parameter epsilon. The smaller the parameter, the more noise is added, and the more difficult it is to distinguish the contribution of a single record. The result is increased privacy for all of the data subjects.
UK Royal Society	**Privacy budget:** a quantitative measure of the change in confidence of an individual having a given attribute.
US National Science and Technology Council	**Privacy parameter** is used control the strength of the privacy guarantee while optimizing for accurate analytic results.
UN Economic Commission for EU	**Privacy parameter** – whose value is set by a data custodian - determines the degree of disclosure protection a DP-epsilon compliant method *M* is offering by means of the upper limit it imposes on the amount of person-level information *M* might be disclosing. More explicitly, the data custodian controls through epsilon the amount of suitable random noise used by *M* to produce its outputs. The larger the value set to epsilon the less noisy *M's* outputs become and, the greater the disclosure risk they pose and the greater their utility.
"I need a better	The **parameters epsilon** and delta control the maximum amount of information that can leak about any individual entry in the dataset.

description": An Investigation Into User Expectations For Differential Privacy	
Differential Privacy for Government Agencies - Are we there yet?	The **parameter epsilon** can be used to specify the level of protection. Larger values of epsilon allow for larger differences in the output distribution between two neighbouring databases, thus offering lower levels of privacy. However, larger values of epsilon will typically increase the level of accuracy of the reported output. Epsilon can be seen as a tuning parameter that trades privacy for accuracy of the estimate obtained
Differential privacy in practice - expose your epsilons!	The **privacy parameter,** typically called epsilon, provides a technical measure of privacy loss, with smaller epsilon corresponding to less privacy loss.
Differential privacy: a primer for a non-technical audience	An essential component of a differentially private computation is the **privacy loss parameter**, which determines how well each individual's information needs to be hidden and, consequently, how much noise needs to be introduced. It can be thought as a tuning knob for balancing privacy and accuracy.
Differential Privacy and Federal Data Releases	The epsilon, also known as the **privacy budget**, controls the degree of privacy offered by A, with lower values implying greater privacy guarantees.
ICO	Epsilon determines the level of added noise. Epsilon is also known as the "privacy budget" or "privacy parameter". Epsilon represents the worst-case amount of information inferable from the result by any third party about someone, including whether or not they participated in the input. Epsilon is the maximum distance between a query on a database (real-world computation) and the same query on a database with a single entry added or removed. Small values of ε provide very similar outputs when given similar inputs, and therefore provide higher levels of privacy as more noise is added. Therefore, it is more difficult to distinguish whether a person's information is present in the database. Large values of ε allow less similarity in the outputs, as less noise is added and therefore it is easier to distinguish between different records in the database.

Table of definitions for DP-sensitivity

ISO 20889	The sensitivity, S, of a given query or function describes the worst case scenario of how much the answer to that query or that function can change if a single data principal is removed from the database.
Differential Privacy for Government Agencies – Are we there yet?	Sensitivity in this context measures how much the statistic changes if one record in the data is altered.
Differential Privacy and Federal Data Releases	The global sensitivity of f, is the maximum L1 distance of the outputs of the function f between any two neighboring databases.

Table of definitions for de-identification

ISO 20889	De-identification refers to a process that removes the association between a set of data attributes and the data principal which they concern
ISO 27559	De-identification is one potential means for facilitating the use of personally identifiable information (PII) in a way that does not identify or otherwise compromise the privacy of an individual or a group of individuals.
UKAN	The removal or masking of direct identifiers within a dataset
NIST (Draft)	De-identification is a process that is applied to a dataset with the goal of preventing or limiting informational risks to individuals, protected groups, and establishments while still allowing for the production of aggregate statistics.
NIST (Published)	De-identification removes identifying information from a dataset so that individual data cannot be linked with specific individuals. De-identification is a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations. De-identification: General term for any process of removing the association between a set of identifying data and the data subject.
US National Science and Technology Council	HIPAA also defines a de-identification standard for protected information that requires there be no reasonable basis to believe the information can identify an individual.
OECD	De-identification means a process by which a set of personal health data is altered, so that the resulting information cannot be readily associated with particular individuals
Future Privacy Forum (*Emerging Privacy Tech Sector*)	US HIPPA: "the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual"
CANON	Canada 2016: de-identification is the process of removing personal information from a record or data set. HIPAA standard for de-identification: "Health information that does not identify and individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information"
Treasury Secretariat Board of Canada	Personal information which can be modified through a process to remove or alter identifiers to a degree that is appropriate in the circumstances

Table of definitions for anonymization

UKAN	A complex process to transform identifiable data into non-identifiable (anonymous) data. This usually requires that identifiers be removed, obscured, aggregated and/or altered in some way. It may also involve restrictions on the data environment
NIST (Draft)	Anonymization is a process that removes the association between the identifying dataset and the data subject.
NIST (Published)	Anonymization: "process that removes the association between the identifying dataset and the data subject."
US National Science and Technology Council	Data anonymization techniques address privacy risks in publishing data by transforming the original data to limit the disclosure of sensitive information or prevent the re-identification of individuals or groups represented in the data.
OECD	Anonymization is the process of rendering personal data impossible to link with an identified or identifiable natural person, even through matching them with other data.
UN Economic Commission for EU	Anonymization process is one where identifying information is modified or suppressed to avoid identification of individual entities in a data file.
PET guide	Anonymization of dataset is done by altering input data either by masking some fields or perturbing values up to a point where records can no longer be identified in the altered data (perturbation method).
Treasury Board Secretariat of Canada	Personal information that has been de-identified to the point that there is no serious possibility of re-identification, by any person or body using any additional data or technology at this point in time.

Table of definitions for pseudonymization

ISO 20889	Pseudonymization refers to a category of de-identification techniques that involve replacing a data principal's identifier (or identifiers) with a pseudonym in order to hide the identity of that data principal.
NIST (Draft)	Pseudonymization is a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.
NIST (Published)	Pseudonymization is a particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.
UKAN	Pseudonymization is a term defined in GDPR as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is subject to technical and organisational measures to keep it separate.
OECD	Pseudonymization involves removing potentially identifiable information from the data to reduce the risk of identification of the data subject, although some residual risk remains.
Treasury Board Secretariat of Canada	Pseudonymization is a process of masking direct identifiers. Pseudonymization is very similar to nulling and field suppression but where the direct identifiers are replaced with aliases and that same alias is used consistently across datasets.

Table of definitions for basic techniques

	Query based	Database based
NIST (draft)	The definition (of DP) is usually satisfied by adding noise to the result of a query ensuring that the added noise masks the contribution of any individual.	
Future Privacy Forum (*Buying Privacy Tech*)	To deploy differential privacy, they need software to calculate the statistical summaries they will release, add carefully calibrated noise, and fulfill the formal guaranteed of differential privacy.	
Harvard Kennedy School	In a curator model, a database administrator (the "curator") has access to a database which includes private data. This administrator uses database to generate differentially private data summaries. This means that the database itself does not satisfy differential privacy, but differentially private analyses run on the data yield differentially private output.	A local model ensures differential privacy at the point of data collection.
PET guide	If the curator is trusted, individuals may send their information directly to them for the purpose of running a differentially private data analysis algorithm whose output is released. *Some libraries, offering open-source implementations of the main differentially private primitives, focus on differentially private ML model training. The mechanisms there are based on appropriate modification of stochastic gradient descent, wherein the gradient is clipped, to limit its dependence on individual points and it is also perturbed by noise addition*	Two well-known applications of DP are its use in Google Chrome and Apple's iOS/OSX to collect usage statistics in a privacy-preserving way. These applications follow the local model of DP, where each individual user privatizes their own data before sending it to a centralized server for analysis
OECD	Noise can be added at the central location before the data are released (centralised)	Noise can be added at the time of data collection (distributed)
ISO (20889)	Random noise is added to the outputs provided by the differentially private system to an analyst (server model). Mechanisms that follow the "server model" for differential privacy typically preserve data in unmodified form in a secure database. In order to preserve privacy, responses to queries are only able to be obtained through a software component or "middleware", known as the "curator". The curator takes queries from system users, or from reporting software, and obtains the	Random noise is added at the user device to inputs from each data principal (local model). The local model is useful when the entity receiving the data is not necessarily trusted by the data principals, or if the entity receiving the data is looking to reduce risk and practice data minimization. In this model, data belonging to a single data principal, or the results of computations on these data, are first randomized before they

	correct, noise-free answer from the database. However, before responding to the user or reporting software, the curator adds random noise whose magnitude is inversely proportional to the privacy loss implied by the query.	are transferred to, and stored on, a server
"I need a better description": An Investigation Into User Expectations For Differential Privacy	In the central model users share their sensitive information directly, and the curator is trusted to perturb results that are released.	In the local model, users randomly perturb their information (with the help of the collection mechanism, e.g., their device) before sending it to a central entity in charge of analysis, called the curator
Differential Privacy for Government Agencies - Are we there yet?	A popular context taken to illustrate the underpinnings of DP is a query response system. A system of this kind accepts specific queries as input—a query for the mean of a variable, for example—and then returns a noisy answer to the query, with the noise calibrated to ensure that the requirements of DP are met.	
Differential privacy: a primer for a non-technical audience	In addition, some tools rely on a curator model, in which a database administrator has access to and uses private data to generate differentially private data summaries.	Others rely on a local model, which does not require individuals to share their private data with a trusted third party, but rather requires individuals to answer questions about their own data in a differentially private manner.
Towards effective differential privacy communication for users' data sharing decision and comprehension	Differential privacy: server has access to the true sensitive values of the users	Local DP: aggregator does not see the actual private data of each individual – users send randomized information to the aggregator who infers the data distribution based on that
Issues Encountered Deploying Differential Privacy	The Census Bureau operates as a trusted curator, which collects sensitive data from individuals, performs statistical tabulations, and publishes them.	Apple and Microsoft use the local model of differential privacy: randomization is performed by software running on the individual's computer.
Differential Privacy and Federal Data Releases	In the non-interactive setting, the agency releases a data product D constructed from differentially private algorithms; for example, D could be a set of summary statistics or a synthetic data set generated to satisfy ϵ-DP.	In the interactive setting, users query the confidential database D repeatedly for noisy answers to arbitrary statistical questions. These questions are determined adaptively by the user, not by the agency.

ICO	global differential privacy adds noise during aggregation; It involves an "aggregator" having access to the real data. Each user of the system that differential privacy is being used in sends information to the aggregator without noise. The aggregator then applies a differentially private mechanism by adding noise to the output (eg a response to a database query or the noise is embedded in the entire dataset). The noise is added during computation of the final result before it is shared with the third party. the global model leads to more accurate results with the same level of privacy protection, as less noise is added; the global model provides deniability of people's non-participation (ie you cannot prove whether a person's information was in the dataset);	local differential privacy is where each user adds noise to individual records before aggregation. It has the user of the system (or a trusted third party on a person's behalf) applying the mechanism before they send anything to the aggregator. Noise is added to the individual (input) data points. The aggregator receives "noisy" data – this addresses the trust risk of global differential privacy as the real data is not shared with the aggregator. the local model provides deniability of a person's record content, but not record association; the local model is not necessarily suitable for producing anonymous information (eg statistics). However, you can use it to mitigate sensitive attribute inference

Appendix F: Survey

In 2018, HBR Magazine published a survey to understand how people assign a probability to words often associated with random chance, such as "usually", "probably", "maybe", and "possibly".[48] The probabilities associated with words varied quite a bit, with some interested trends in general. These could be seen as a means to interpret the language used by courts and regulators when attempting to calculate the likelihood of events, such as the risk of re-identification or disclosure risk more generally. We incorporated a small sample of this experiment in surveys to capture the language used in this context, as part of a larger survey to explore how people interpret various notions of risk.

In March 2024, we conducted a survey of 3rd and 4th year students from the University of Ottawa. These students are enrolled in such programs as Statistics, and Financial Mathematics and Economics. As such, these students have a background in probability, statistics and risk computation. 98 students responded to this survey, which is a decent sample size and some statistical conclusions can be drawn.

The main conclusions are as follows:

- The students assign 80%-100% as "serious possibility".
- Half of the students do not know the notion of relative risk (even though in Q2 75% of them are confident they know the difference, the next questions proves otherwise).
- Understanding of this difference is important to students.
- Media do not report properly on risk.

In conclusion, this survey indicates an importance of education on risk. Relative vs. absolute risk concepts can be easily taught in high school!

Question 1:[49] Please assign the probability value for being identified in a statistical database that you associate with the following list of words or phrases, on a scale from 0 to 100%,

Frivolous chance
Possibility
Serious possibility
Statistical probability
Very low
Very small

For the "serious possibility", most of the responses identified the range 80%-100%. However, some students responded 50%-100% or similar. We had some students who answered 0%-50%. It is not clear what is the reasoning behind such answer. All in all, it seems that **80%-100%** is a reasonable range for "serious possibility".

[48] If You Say Something Is "Likely," How Likely Do People Think It Is? https://hbr.org/2018/07/if-you-say-something-is-likely-how-likely-do-people-think-it-is

[49] The terms selected are from OPC, Federal Court of Canada, Quebec legislation, and US HIPAA legislation.

Question 2: This question deals with absolute vs. relative risk. If your answer is YES, please TRUE, otherwise choose FALSE. Do you know the difference between absolute and relative risk?

Answers: TRUE 75%, FALSE 25%

Question 3: If a medication reduces the risk of a disease from 2% to 1%, do you consider this a 50% reduction in relative risk? (Please answer TRUE if your answer is YES, otherwise, select FALSE).

Answers: TRUE 50%, FALSE 50%

Question 4: Can an increase in relative risk be misleading without knowing the absolute risk? (Please answer TRUE if your answer is YES, otherwise, select FALSE)

Answers: TRUE 85%, FALSE 15%

Question 5: Do you think media reports usually specify whether a reported risk change is absolute or relative? (Please answer TRUE if your answer is YES, otherwise, select FALSE)

Answers: TRUE 18%, FALSE 82%

Question 6: Would knowing both absolute and relative risk figures influence your health-related decisions? (Please answer TRUE if your answer is YES, otherwise, select FALSE)

Answers: TRUE 92%, FALSE 8%

Question 7: On a scale from 0 (Not confident) to 10 (Very confident), how confident are you in your ability to explain the difference between absolute and relative risk?

Value	% of responders
0	0%
1	1.02%
2	2.04%
3	10.20%
4	7.14%
5	20.41%
6	12.24%
7	25.51%
8	13.27%
9	2.04%
10	6.12%

Question 8: How important is it for you to know the absolute risk when making health-related decisions? Rate from 0 (Not important) to 10 (Very important)

Value	% of responders
0	1%
1	0.00%
2	2.04%
3	2.04%
4	0.00%
5	12.24%
6	6.12%
7	14.29%
8	29.59%
9	14.29%
10	18.37%

Question 9: On a scale from 0 to 10, how effectively do you think the media communicates risks in health news? (0 - not effective at all, 10 - very effective)

Value	% of responders
0	0%
1	4.08%
2	11.22%
3	19.39%
4	13.27%
5	19.39%
6	16.33%
7	5.10%
8	8.16%
9	1.02%
10	2.04%

Question 10: To what extent does knowing the relative risk without absolute risk numbers change your perception of a health risk? Rate from 0 (Not at all) to 10 (Significantly).

Value	% of responders
0	0%
1	0.00%
2	2.04%
3	5.10%
4	11.22%
5	21.43%
6	13.27%
7	19.39%
8	15.31%
9	7.14%
10	5.10%

Question 11: If you read a news article stating a treatment reduces a risk by a certain percentage, how likely are you to look for more information about the absolute risk? Rate from 0 to 10 (10 - extremely likely)

Value	% of responders
0	0%
1	5.10%
2	8.16%
3	8.16%
4	7.14%
5	6.12%
6	10.20%
7	19.39%
8	18.37%
9	9.18%
10	8.16%