Differential Privacy - foundations, techniques, policies

Rafał Kulik

Department of Mathematics and Statistics (University of Ottawa)

Government of Manitoba

Joint work with Luk Arbuckle (Privacy Analytics), Heidi Barriault (UofO), Devyani Biswal (PA and UofO), Chang Qu (UofO), Teresa Scassa (UofO)

18 June 2024

Plan

1. What is data privacy?
2. Very simple approach - k-anonymization
3. Query attack on a database
4. Differential privacy
5. Policies

What is data privacy?

Data privacy or information privacy is a branch of data security / computer science / statistics concerned with the proper handling of data – consent, notice, and regulatory obligations. More specifically, practical data privacy concerns often revolve around:

- How data is legally collected and/or stored.
- **How and if data is shared with third parties.**
- **Regulatory restrictions and policies (?)**

k-anonymization

The term k-anonymity was first introduced by Pierangela Samarati and Latanya Sweeney in the paper published in 1998, although the concept dates to a 1986 paper by Tore Dalenius.

Definition 1 (k-anonymity)

A database satisfies k-**anonymity** if each equivalence class of Quasi-Identifiers consist of at least k units.

In principle, k-anonymity should guarantee that the chance of re-identification is at most $1/k$.

k-anonymization

- **Explicit Identifiers (EI)** consist of identifying information (such as names) of the record holders.
- **Quasi-Identifiers (QI)** (such as date of birth, gender, and zip code) do not reveal identity, but can be used to link to a record holder or an explicit identity in some external sources.
- **Sensitive Attributes (SA)** consist of other person-specific but sensitive information (such as medication and DNA entries). In some risk disclosure literature, SAs are referred to as **target variables**.

Typically, the attacker wants to learn about Sensitive Attributes.

k-anonymization - original table

	EI	QI		SA
ID	Name	Gender	Year of Birth	Test Result
5	Alicia Freds	Female	1942	- ve
3	Alice Brown	Female	1955	- ve
11	Beverly McCulsky	Female	1964	- ve
7	Marie Kirkpatrick	Female	1966	Zero
13	Freda Shields	Female	1975	- ve
6	Gill Stringer	Female	1975	- ve
8	Leslie Hall	Female	1987	- ve
4	Hercules Green	Male	1959	- ve
12	Douglas Henry	Male	1959	+ ve
1	John Smith	Male	1959	+ ve
2	Alan Smith	Male	1962	- ve
14	Fred Thompson	Male	1967	- ve
9	Bill Nash	Male	1975	- ve
10	Albert Blackwell	Male	1978	- ve

k-anonymization - 2-anonymized table

	QI		SA
ID	Gender	Decade of Birth	Test Result
13	Female	1970-1979	-ve
6	Female	1970-1979	-ve
11	Female	1960-1969	-ve
7	Female	1960-1969	Zero
12	Male	1950-1959	+ve
1	Male	1950-1959	+ve
4	Male	1950-1959	-ve
2	Male	1960-1969	-ve
14	Male	1960-1969	-ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve

k-anonymization

- 2-anonymization is achieved with respect to DOB, not the rest result. If we know that Alice Brown is in the table, we can automatically guess her test result.
- Additional generalization is needed.
- Very inefficient if many variables are considered.

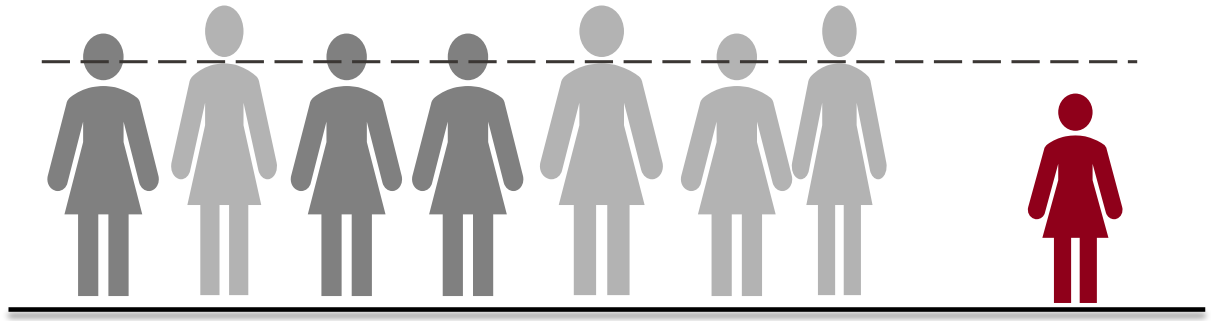
Query attack on a database - learning Sensitive Attributes

Assume that $\boldsymbol{x} = (x_1, \ldots, x_{10})$ is a database of income of 10 people (Sensitive Attribute). Let x_1 be the income of John Smith.
The attacker asks the question (**query**): *What is the average income*?
From an answer, the attacker cannot infer the income of John Smith. But, assume that the attacker has a big **privacy budget**. That is, the attacker can ask another question: *What is the average income of the first 9 people*?

Query attack on a database - protecting outliers

Calculating a statistic

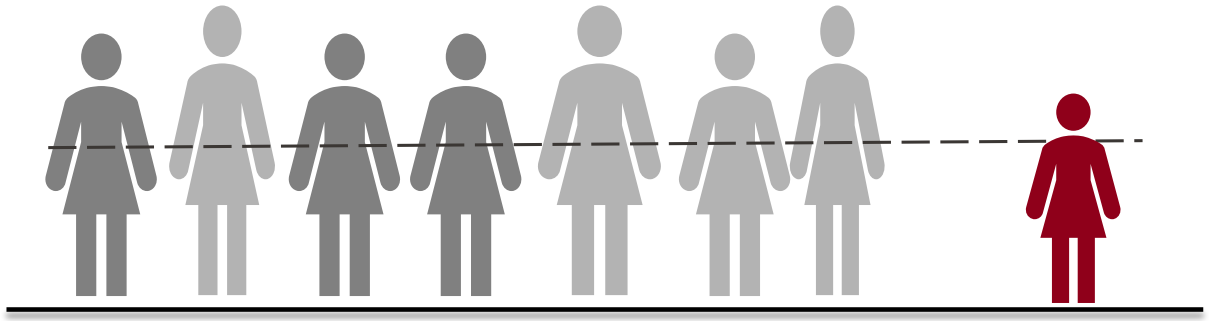
The database contains the heights of women



Query attack on a database - protecting outliers

Calculating a statistic (again)

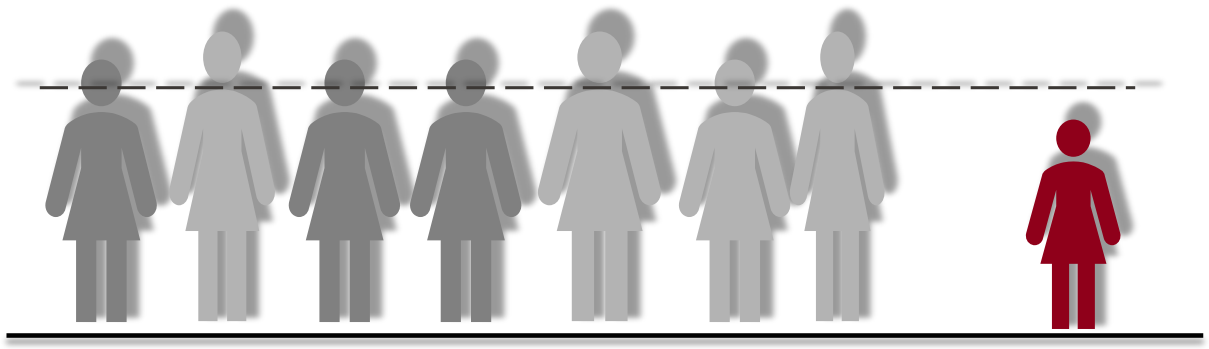
The database contains the heights of women



Query attack on a database - protecting outliers

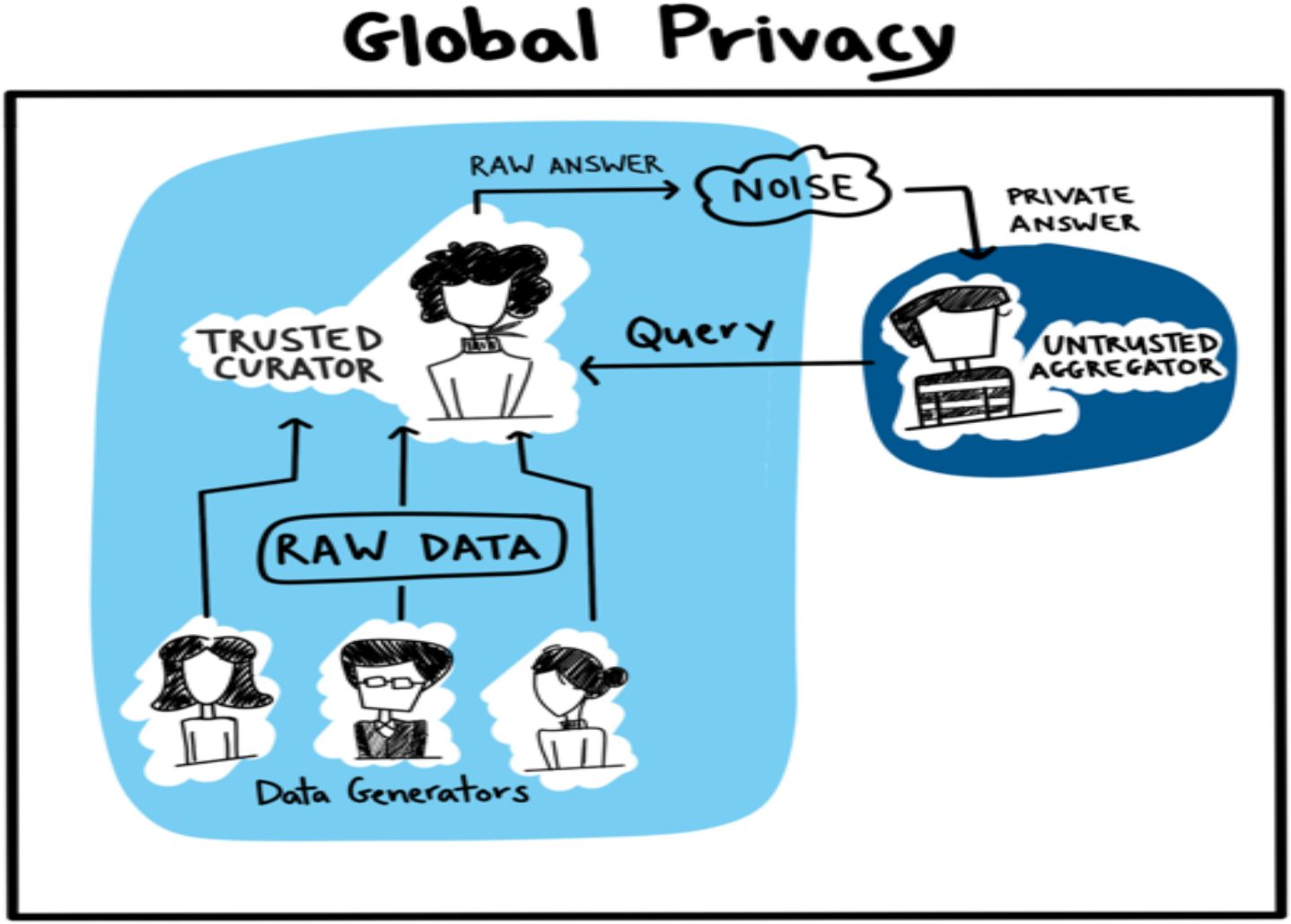
Calculating a statistic from randomized inputs

The database contains the heights of women

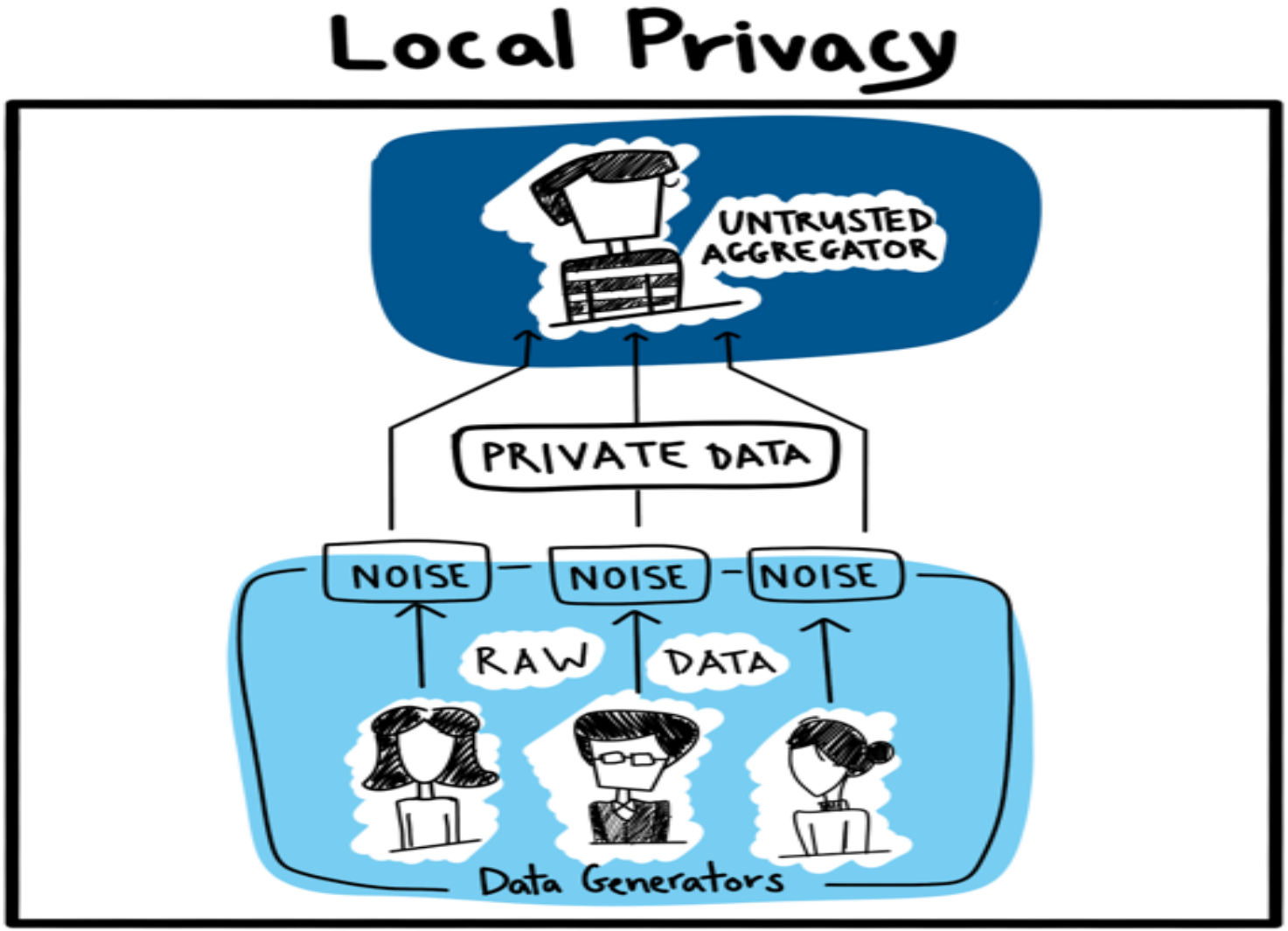


Calibration error

Query perturbation vs database perturbation



Query perturbation vs database perturbation



Differential privacy

Differential privacy is a **probabilistic guarantee** that the inclusion of **an individual in a database does not alter the outcome of a query on the database by more than a user chosen specified bound**. It represents a robust framework for quantifying and managing the privacy of individuals in databases that undergo analysis. Formally introduced by Cynthia Dwork in 2005, it has since become a popular implementation choice in the field of data privacy. It provides mathematical guarantees against identity inference and data re-identification attacks.

Differential privacy (DP)

Definition 2

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two neighbouring databases (that differ by one record). An output perturbation mechanism is ε-**differentially private**

$$\sup_{B \in \mathbb{R}^d} \frac{\mathbb{P}(Q(\boldsymbol{x}, Z) \in B)}{\mathbb{P}(Q(\boldsymbol{y}, Z) \in B)} \leq \mathrm{e}^{\varepsilon}.$$

- Typically, $Q(\boldsymbol{x}, Z) = \underbrace{f(\boldsymbol{x})}_{query} + \underbrace{Z}_{random\ noise}$ (output perturbation mechanism) - noise added to query.
- Alternatively, $Q(\boldsymbol{x}, Z) = f(\boldsymbol{x} + \underbrace{\boldsymbol{Z}}_{vector\ random\ noise})$ (sanitized responses mechanism) - noise added to the database.

Differential privacy (DP)

- ε is called the **privacy budget**. Bigger ε, less privacy. The query output from two distinct databases are very different.
- Typically, Z has **Laplace distribution** with the variance proportional to $1/\varepsilon^2$. Large ε, mean small variance, means little noise added, means little privacy (distinct databases remain distinct). This Laplace distribution depends not only on the privacy budget ε, but also on so-called sensitivity of the query. For example, if the query is the sample mean, then the sensitivity is equal to the range of data divided by the sample size. The bigger sensitivity, the bigger variance of the noise.
  There are technical issues how to calculate the sensitivity.
- Small ε means a lot of privacy, but poor **data utility**.
- Big variance of noise means a lot of privacy, means poor data utility.
- Academics recommend $\varepsilon = 1$. US census used $\varepsilon = 14$. Private company reset on daily basis.

Differential privacy (DP)

- DP is designed for continuous data. It has to be adjusted for count data.
- DP may lead to unreasonable outcomes. For example, negative income. This is addressed by *bounded* differential privacy.
- The original DP was designed for queries. Nowadays, noise added to databases. Thus, in principle, one needs to protect against all possible queries. Hence, a lot of noise has to be added.
- In general if we want to keep the same level of privacy, we need to add much more noise to the database than to a query. This is intuitively clear. When we want to release a randomized database, we need to protect against all possible queries.

DP in PIPEDA and Bill C-27?

We considered the concept of DP 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 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. The threshold test for identifiability in information that is the **serious possibility** test from Gordon v. Canada. According to that test, information is personal information "if *there is a serious possibility*" that an individual could be identified.

DP in PIPEDA and Bill C-27?

Privacy 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 (**data utility**).[1]
The **serious possibility** of re-identification and the need for **data utility** have to be balanced through necessity and proportionality approach. That is, *we need enough privacy, but not too much privacy*.
In the context of DP, ε should be chosen *not too big, not too small*.

[1] https://www.priv.gc.ca/en/opc-news/speeches/2023/sp-d20230525/

DP in policies

- In the current legislation, there is no mention of particular anonymization techniques.
- It is not obvious what does *serious possibility* (of re-identification) mean? We surveyed 150 students in Statistics, and they gave the range 50%-90% for the serious possibility.
- *Serious possibility* depends on the context.
- At this moment, there is no mention of DP in policies/legislation. It is hard to link ε to *serious possibility*.

Policies - general guidance

- **Anonymization can and should allow for the use of differential privacy techniques.** It should be clear to those seeking to anonymize data that the appropriate use of differential privacy techniques qualifies as anonymization for the purposes of the interpretation and application of PIPEDA and, if it is passed, Bill C-27.
- **Guidance on anonymization should be clear and should allow for the selection of different tools or approaches.** The evolution of different privacy enhancing techniques such as differential privacy make it clear that there is no one-size-fits-all approach.
- **Quebec's draft anonymization regulations offer an interesting model for the development of general anonymization guidelines.** In particular, the regulations require a preliminary risk assessment for the data set. Based on this assessment, the appropriate anonymization techniques must be selected. Once anonymization techniques have been applied, reidentification risk is re-assessed on the anonymized dataset. The regulations also require periodic reassessment of the dataset on the basis that as circumstances change so might the risks of reidentification.

Thank you!!!!