The Five Safes as a Privacy Context

James Bailie[1] & Ruobin Gong[2]

Department of Statistics
[1]Harvard University [2]Rutgers University

22 September 2023

The 5th Annual Symposium on Applications of Contextual Integrity

[1]jamesbailie@g.harvard.edu
[2]ruobin.gong@rutgers.edu

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

  A large family of technical standards (i.e. mathematical specifications).

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

  A large family of technical standards (i.e., mathematical specifications)

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

  A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

> A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

> A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

  > A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

> A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

> A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Motivations

- National Statistical Offices (NSOs) are modernizing their data protection.
- Example: US 2020 Census protected by differential privacy (DP).
- What is DP?

> A large family of technical standards (i.e. mathematical specifications) that conceptualizes the loss of privacy as a rate of change: the change in the (distribution of) the output statistics per unit change of an individual's information.

- Troubles in implementation: Context needs to be understood when choosing which DP standard to use.

Two main points

1. The Five Safes is a reparametrization of Contextual Integrity in the situation where the information flow is a statistical dissemination;
2. The Five Safes provides a context for Differential Privacy as a framework for controlling the disclosure risk of statistical dissemination.

The Five Safes[3]

Safe People

Safe Projects

Safe Settings

Safe Data

Safe Outputs

[3] See e.g. Australian Bureau of Statistics, Statistics New Zealand, Office for National Statistics (UK)

Two main points

1. The Five Safes is a reparametrization of Contextual Integrity in the situation where the information flow is a statistical dissemination;
2. The Five Safes provides a context for Differential Privacy as a framework for controlling the disclosure risk of statistical dissemination.

The Five Safes[3]

Safe People

Safe Projects

Safe Settings

Safe Data

Safe Outputs

[3] See e.g. Australian Bureau of Statistics, Statistics New Zealand, Office for National Statistics (UK)

Mapping the Five Safes to CI in statistical dissemination

The two types of information flow in statistical dissemination:

$$\textbf{data} \rightarrow \textbf{people} \text{ (researchers)} \tag{1}$$

$$\textbf{outputs} \rightarrow \textbf{people} \text{ (general public)} \tag{2}$$

Ex 1. Open Data: public use data files

Ex 2. Data enclaves:

- Physical: Federal Statistical Research Data Center (US); Canadian Research Data Centre Network (StatCan)
- Virtual: DataLab (Australian Bureau of Statistics); Real Time Remote Access (StatCan)

Ex 3. Synthetic data + validation server

e.g. U.S. Census Bureau Survey of Income and Program Participation (SIPP) Synthetic Beta + validation through Gold Standard File

Mapping the Five Safes to CI in statistical dissemination

The two types of information flow in statistical dissemination:

$$\textbf{data} \rightarrow \textbf{people} \text{ (researchers)} \tag{1}$$

$$\textbf{outputs} \rightarrow \textbf{people} \text{ (general public)} \tag{2}$$

Ex 1. Open Data: public use data files

Ex 2. Data enclaves:

- Physical: Federal Statistical Research Data Center (US); Canadian Research Data Centre Network (StatCan)
- Virtual: DataLab (Australian Bureau of Statistics); Real Time Remote Access (StatCan)

Ex 3. Synthetic data + validation server

e.g. U.S. Census Bureau Survey of Income and Program Participation (SIPP) Synthetic Beta + validation through Gold Standard File

Mapping the Five Safes to CI in statistical dissemination

The two types of information flow in statistical dissemination:

$$\textbf{data} \rightarrow \textbf{people} \text{ (researchers)} \tag{1}$$

$$\textbf{outputs} \rightarrow \textbf{people} \text{ (general public)} \tag{2}$$

Privacy norm parameters	Their meanings in statistical dissemination
sender	statistical agencies/NSOs/data custodians
recipient	**people**: researchers (1) and general public (2)
subject	is a component of **data** (1)
information type	is a component of **data** (1) and **outputs** (2)
transmission principles	encompass **projects**, **settings**, and more

DP in the context of the Five Safes

1. DP pertains to some aspects of Safe Outputs and Safe Data and is silent on other aspects.
2. DP does not purport to assess Safe People, Projects or Settings.
3. The Five Safes is a solution concept for implementing DP in a way that respects contextual integrity.

The Five Safes

Safe People

Safe Projects

Safe Settings

Safe Data

Safe Outputs

jamesbailie@g.harvard.edu
ruobin.gong@rutgers.edu

DP in the context of the Five Safes

1. DP pertains to some aspects of Safe Outputs and Safe Data and is silent on other aspects.
2. DP does not purport to assess Safe People, Projects or Settings.
3. The Five Safes is a solution concept for implementing DP in a way that respects contextual integrity.

The Five Safes

Safe People

Safe Projects

Safe Settings

Safe Data

Safe Outputs

jamesbailie@g.harvard.edu
ruobin.gong@rutgers.edu

DP in the context of the Five Safes

1. DP pertains to some aspects of Safe Outputs and Safe Data and is silent on other aspects.
2. DP does not purport to assess Safe People, Projects or Settings.
3. The Five Safes is a solution concept for implementing DP in a way that respects contextual integrity.

The Five Safes

Safe People

Safe Projects

Safe Settings

Safe Data

Safe Outputs

jamesbailie@g.harvard.edu
ruobin.gong@rutgers.edu

Five Components of DP → Safe Data & Outputs

- The protection domain (what can be protected?): as defined by the dataset space $\mathcal{X}$;
- The scope of protection (to where does the protection extend?): as instantiated by the data multiverse $\mathscr{D}$, which is a collection of data universes $\mathcal{D} \subset \mathcal{X}$;
- The protection units (who are the units for data perturbation?): as conceptualized by the divergence $d_{\mathcal{X}}$ on the dataset space $\mathcal{X}$;
- The standard of protection (how to measure the output variations?): as captured by the divergence $d_{\mathcal{T}}$ on (the probability distributions on) the output space $\mathcal{T}$;
- The intensity of protection (how much protection is afforded?): as quantified by the privacy-loss budget $\epsilon_{\mathcal{D}}$ for each data universe $\mathcal{D}$.