 ISBN: 9789073592421

CPS Paper

Formal privacy on a subset of dataset variables

Author: Ms Pin Lin Tan

Submission ID: 503

Reference Number: 503

Presentation File

abstracts/ottawa-2023_4d2effe4e3d35b76857fc4db2f365ce7.pdf

Files/Uploads

CPS0082_PinLin

Brief Description

Differential privacy (DP) is a mathematical definition of privacy that has many attractive properties.

However, it requires all outputs from the dataset to be noisy, even those that only involve variables that are not considered sensitive or identifying.

In this paper, we introduce privacy definitions that only protect a subset of the variables in a dataset, thus allowing statistics only involving unprotected variables to be released accurately.

We explore their relation to DP and Pufferfish privacy, their composition properties and how it can be used to assign an epsilon value to each variable.

Abstract

Differential privacy (DP) is a mathematical definition of privacy that has many attractive properties. This includes allowing data providers to quantify the privacy risk of a collection of data releases from a dataset. As DP is defined for mechanisms acting on the dataset, the privacy guarantee of DP is at the level of the entire dataset and does not require the determination of which variables are sensitive or identifying. While this provides privacy guarantees that do not depend on the often-subjective classification of variables, this also leads to needing all outputs from the dataset to be noisy, even those that only involve variables that are not considered sensitive or identifying.

In this paper, we introduce privacy definitions that only protect a subset of the variables in a dataset, thus allowing statistics only involving unprotected variables to be released accurately. We show that the definitions are a generalisation and relaxation of DP and can be defined using the Pufferfish framework for privacy definitions, as well as provide an algorithm for generating privacy-protected count tables that are consistent with the counts that can be released accurately under the proposed privacy definition. We explore the properties of the definition in terms of composition and show how the definitions can be used to assign an epsilon value to each subset of variables. This contrasts with DP, which only assigns an epsilon value to the entire dataset. The epsilon value for a single variable is at most that of the entire dataset and can be smaller, reflecting that different variables in a dataset may be protected to different degrees in a data release. This may be useful for data providers who wish to protect some variables, for example more sensitive ones, more than others.