Proceedings 64th ISI World Statistics Congress - Ottawa, Canada ISBN: 9789073592421

CPS Paper

Quantifying the contribution of individual records to the reidentification risk of (pseudo)anonymized datasets

Author: Dr Photis Stavropoulos

Coauthors: Vasiliki Daskalaki, Kimon Spiliopoulos, Konstantinos Spinakis, Gilbert Saporta, Michel Bera

Submission ID: 551

Reference Number: 551

Presentation File

abstracts/ottawa-2023_3a7332d6240cfd6fb54c54f8f7782d29.pdf

Brief Description

A measure of the risk of reidentifying individuals in a dataset and a method to estimate it have been proposed recently.

This paper presents various approaches to estimating the contribution of each record to the dataset's risk.
This serves various purposes.

The withdrawal of (possibly a few) records with large contribution may reduce the dataset's risk.

Furthermore, the contribution of a record can be considered as a proxy of the risk of identifying the individual the record corresponds to.

Abstract

The reidentification of individuals or business establishments in (pseudo)anonymized microdata may expose sensitive data and will lead to fines and reputational damage for the data's custodians. The QaR method (AFNOR, 2020) proposes a measure of the reidentification risk of a dataset, and a statistical technique, based on extreme-value theory, to estimate it. This risk has great value. It is a gauge of the effectiveness of whatever disclosure control the custodians apply to the data; it could be reported to regulatory authorities to demonstrate the custodians' level of care for the data subjects' privacy; it can be used to calculate an insurance premium against unauthorized disclosure or the amount of money that custodians need in their balance sheet to cover potential financial damages due to such disclosure.
The present paper deals with a particular aspect of the methodology: the quantification of the contribution of each record to the dataset's risk. It discusses its importance and its large computational demands in very large datasets, and proposes metrics that are faster to compute and could serve as proxies of record contribution. The results for some of these proxies are promises but more investigation is needed.