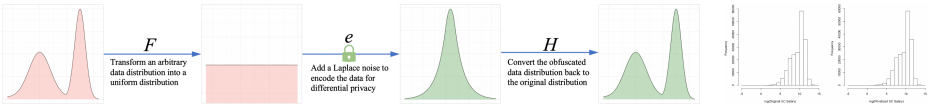
UNIVERSITY OF MINNESOTA
Driven to Discover℠

A data privatization method which maintains statistical accuracy

A novel data protection method that maintains the statistical accuracy of the data and also provides state-of-the-art data privacy.



IP Status: Provisional Patent Application Filed

Applications

- Data sharing with privacy

Key Benefits & Differentiators

- **Provides differential privacy** This method provides a rigorous form of data privacy while sharing data.
- **Maintains statistical accuracy:** Original data distribution is maintained, hence statistical accuracy is preserved, enabling better predictions from machine learning models trained on the data.

Technology Overview

Datasets that consist of uniquely identifiable sensitive information need a high level of privacy protection. Differential privacy, which enables data sharing by adding randomness and describing the patterns of groups within the dataset while withholding information about individuals in the dataset, is the most rigorous form that is currently available. However, the mechanisms that provide differential privacy transform the dataset such that the statistical accuracy and usefulness of the dataset are lost, and requires bounded support of the data distribution to guard against extreme events.

Profs. Xuan Bi and Xiaotong Shen at the University of Minnesota have developed a novel distribution-invariant privatization (DIP) mechanism to address the two aforementioned challenges for all types of univariate and multivariate data involving continuous, discrete, mixed, and categorical variables. Methodologically, DIP transforms each variable to a variable with bounded support, followed by perturbation to mask data, and a suitable mapping to the original scale. Since the data distribution is maintained, the downstream statistical accuracy is preserved. In real-world data analysis, with a given privacy budget, DIP improves existing benchmarks by 82% to 3170% in terms of accuracy across multiple datasets covering many mainstream statistical and machine-learning tasks. This method would be advantageous for companies across all industries that deal with uniquely identifiable or sensitive individual data. Organizations would no longer need to get a de-identification sign-off from the data owner and a time-consuming encryption process if this method is opted for. With this method, the

Technology ID

2021-044

Category

Software & IT/Algorithms
Software & IT/Artificial Intelligence
Software & IT/Data Mining

Learn more

transformed data can be shared easily since individual privacy is protected, yet the statistical properties of the data are maintained.

This method would be beneficial to companies where data protection is needed, such as in the financial, medical, and smart-manufacturing industries. Additionally, companies that work with the following types of data will benefit from this method: Genomics, Health Care, Internet of Things (IoT), Geolocation based, US Census based, Social Media, Search Engine, Travel, Online Shopping, etc.

Phase of Development

TRL: 5-6
The method has been demonstrated to produce accurate results with code written in Python with some core subroutines to be run in R.

Desired Partnerships

This technology is now available for:

- License
- Sponsored research
- Co-development

Please contact our office to share your business' needs and learn more.

Researchers

- **Xuan Bi** Assistant Professor, Information & Decision Sciences
- **Xiaotong Shen** Professor, School of Statistics

References

1. Bi, Xuan, and Xiaotong Shen(18 June 2022) , https://www.sciencedirect.com/science/article/pii/S030440762200121X, Journal of Econometrics