Privacy Preservation of High-dimensional Data using Manifold Learning

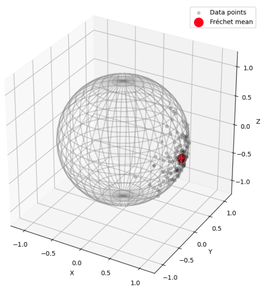
Sonakshi Garg, PhD Student, Umeå University
Department of Computing Science
Supervisors: Prof. Vicenç Torra

Motivation & Research Goals

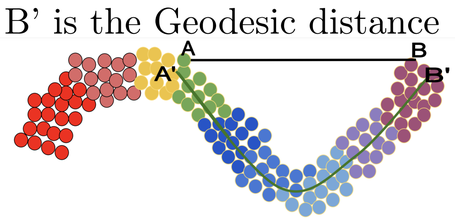
With an increase in abundant amount of data in recent years, an increase in complexity of data from scalars to highly structured and non-linear data can also be seen. Such data may contain confidential information and must be protected from disclosure and essential to safeguard GDPR policies. This work aims to provide privacy preserving model to anonymize high-dimensional data maintaining the manifold structure of the data. Manifold Learning hypothesize that real-world data lie on a low-dimensional manifold embedded in a higher-dimensional space. As the protection of high-dimensional data is important, similarly the protection of statistical summaries of data is equally important which could reveal a lot of information. Fréchet mean is one such operation which can be performed on a metric space, and is meaningful in the manifold setting. We show that to find a trade-off between utility and privacy it is important to preserve the structure of data which is achieved by using Manifold Learning.

Methods

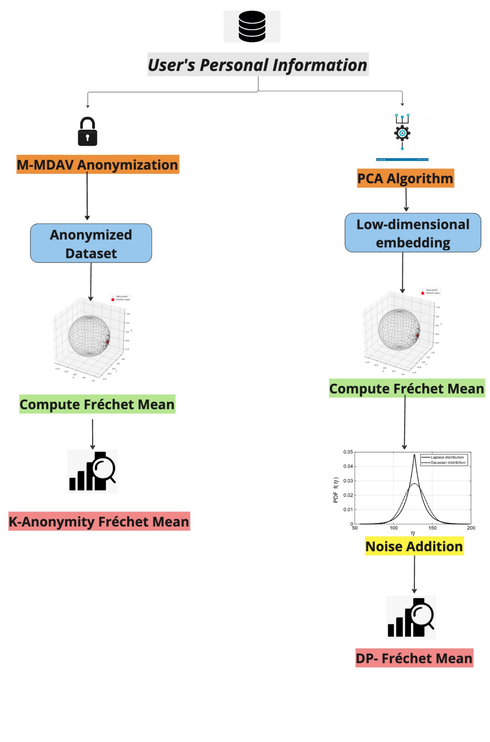
Fréchet Mean on the manifold surface



A→ B is the Euclidean distance vs A'→ B' is the Geodesic distance



Statistical summaries such as mean provides a lot of information about a data set. It is a measure of central tendency. But when the data possess manifold structure, arithmetic or geometric mean are not capable of providing such information. Thus, Fréchet mean generalises the concept of centroids for any metric space.



- We considered the problem of estimating the anonymized Fréchet mean lying on a manifold. We used K-Anonymity [1] and Differential Privacy models that are capable of anonymization while preserving the manifold structure of data, and provided a comparative analysis.
- To emphasize the importance for preserving the manifold structure of data we performed additional experiments. We used two manifold learning techniques such as ISOMAP and LLE to preserve the manifold structure of data and then anonymized using K-Anonymity . We compared this with the approach that directly anonymizes the data using K-Anonymity without preserving the manifold structure[2].

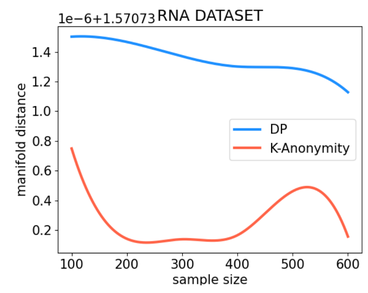
References

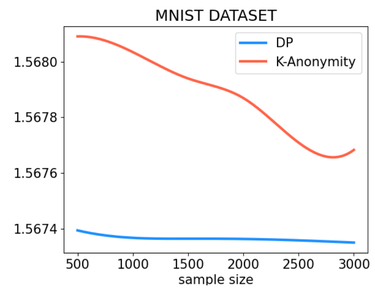
[1] Protecting privacy when disclosing information: k-anonymity
P. Samarati and L. Sweeney
Technical report at SRI International in 1998

[2] K-Anonymous Privacy Preserving Manifold Learning
S. Garg and V. Torra
2023 SECRYPT

Selected Results

Sample size vs Manifold distance for RNA Dataset having dimensions of (800 × 20531) and MNIST Dataset having dimensions of (60000 × 28 × 28)

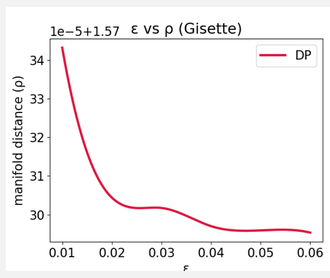




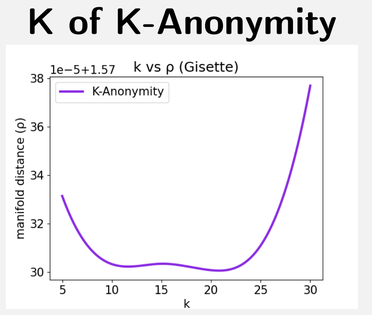
The Manifold distance is calculated as the geodesic distance between original Fréchet Mean and anonymized Fréchet Mean, which is computed by minimizing the objective function:

$$Z = \arg\min_{p \in M} \sum_{i=1}^{n} d(p, x_i)^2$$

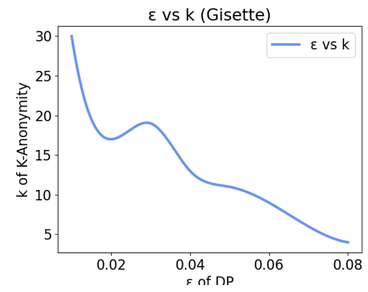
Epsilon of Differential Privacy

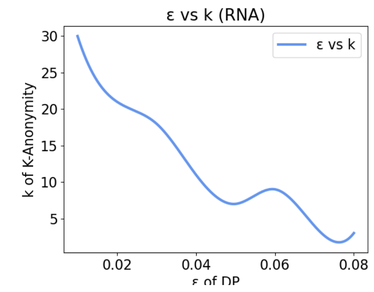


K of K-Anonymity



Relationship between different privacy models: Epsilon of Differential privacy vs K of K-Anonymity





Evaluation to understand the importance of Manifold Learning

Dataset (D)	D(n,p)	Algorithm	Accuracy	Precision	Recall	K-Stress
RNA	800 × 20531	M-ISOMDAV	**99.17**	**99.18**	**99.17**	**0.43**
		M-LLEMDAV	58.12	59.3	58.13	0.73
		M-MDAV	90.10	90.12	90.11	–
Gisette	6000 × 5000	M-ISOMDAV	77.79	76.82	77.78	0.69
		M-LLEMDAV	**85.13**	**86.10**	**85.14**	**0.64**
		M-MDAV	69.21	69.87	69.18	–
SPAM	5272 × 5055	M-ISOMDAV	**85.20**	**84.34**	**85.21**	**0.45**
		M-LLEMDAV	42.61	43.13	42.59	0.89
		M-MDAV	39.56	40.10	39.81	–

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM