Journal of Big Data

RESEARCH Open Access

Enhancing correlated big data privacy using differential privacy and machine learning

Sreemoyee Biswas*, Anuja Fole, Nilay Khare and Pragati Agrawal

*Correspondence:
shonai.biswas@yahoo.in

Department of Computer Science, Maulana Azad National Institute of Technology, Bhopal, India

Abstract

Data are often correlated in real-world datasets. Existing data privacy algorithms did not consider data correlation an inherent property of datasets. This data correlation caused privacy leakages that most researchers left unnoticed. Such privacy leakages are often caused by homogeneity, background knowledge, and linkage attacks, and the probability of such attacks increases with the magnitude of correlation among data. This problem further got magnified by the large size of real-world datasets, and we refer to these large datasets as 'Big Data.' Several researchers proposed algorithms using machine learning models, correlation analysis, and data privacy algorithms to prevent privacy leakages due to correlation in large-sized data. The current proposed work first analyses the correlation among data. We studied the Mutual Information Correlation analysis technique and the distance correlation analysis technique for data correlation analysis. We found out distance correlation analysis technique to be more accurate for high-dimensional data. It then divides the data into blocks using the correlation computed earlier and applies the differential privacy algorithm to ensure the data privacy expectations. The results are derived based upon multiple parameters such as data utility, mean average error, variation with data size, and privacy budget values. The results showed that the proposed methodology provides better data utility when compared to the works of other researchers. Also, the data privacy commitments offered by the proposed method are comparable to the other results. Thus, the proposed methodology gives a better data utility while maintaining the required data privacy commitments.

Keywords: Big data privacy, Correlated datasets, Data correlation, Machine learning, Correlated big data, Data privacy threats, Data privacy algorithms

Introduction

The massive generation of data from our day-to-day life has led to large, voluminous, and heterogeneous data getting produced daily. Due to this reason, the real-world datasets are primarily large and possess high dimensionality. The traditional privacy algorithms are no longer sufficient to ensure the privacy of large-sized datasets, especially when data is highly correlated [1, 2]. Hence many researchers are working towards producing algorithms that can take care of these challenges. Our previous work [4] gave a detailed description of all the works of global researchers who shed light on this issue and proposed solutions to deal with it. Among all the other pieces, we identified the

work presented in [5] by authors Lv et al. as the most potential one and extended our research in the same direction. This paper presents a solution using Distance Correlation Analysis Technique and showcases our results. We compared our results with the results of [5] and subsequently established the supremacy of our method.

Data privacy protection

Data privacy has been a topic of concern for a long time. The concern grew among researchers with the data's increase in size and dimensionality. The classical data privacy algorithms are k-anonymity [6, 7], l-diversity [7, 8], t-closeness [9] and differential privacy [10, 11]. Figure 1 briefly describes the traditional data privacy algorithms. Most of the research regarding data privacy primarily revolves around DP. It is due to its widespread use in this field. Researchers have observed and studied many threats to data privacy over the years. Among them, the existence of data correlation within the data is one of the potential causes of privacy leakages [12–14].

Initial research around data privacy ignored data correlation and considered data as IID. A piece of data is said to be IID, i.e., Independent and Identically Distributed, when it does not hold any relation with other data of the dataset and its distribution is identical throughout the dataset. In other words, there exists no correlation among data within the dataset. But suppose a correlation exists among such data and during the application of data privacy algorithms. In that case, if it gets ignored, then such assumptions can lead to potential privacy leakages [5, 15, 16]. This threat increases with the size and dimensionality of data. Hence, one can conclude that data correlation is a more significant threat to big data [17]. Big data privacy often gets compromised by ignoring the data correlation within the data.

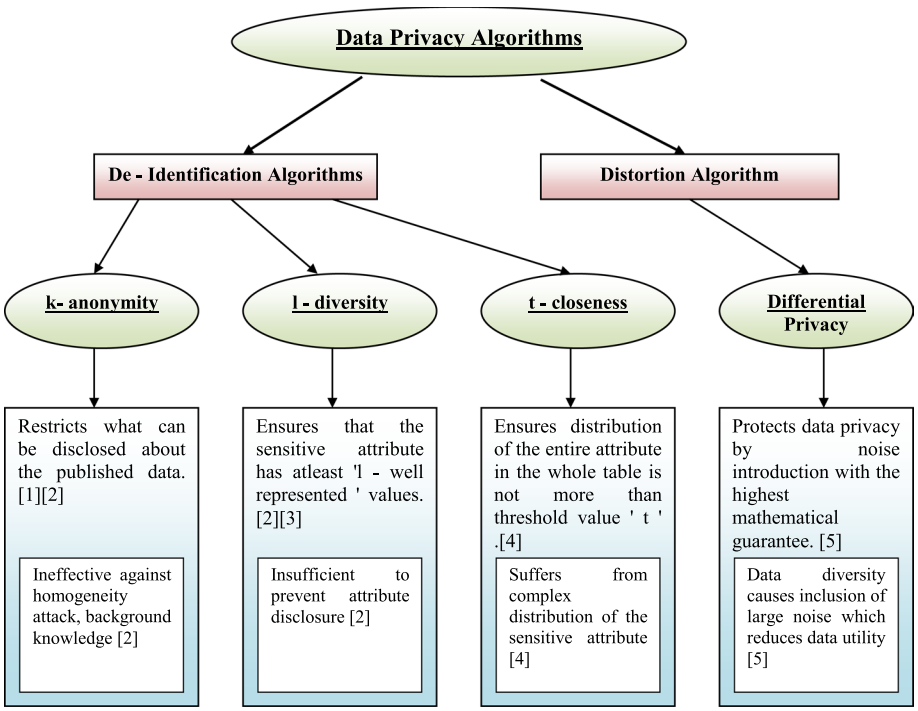

Fig. 1 Pros and cons of classical data privacy algorithms

The work presented in this research paper outlines all the related results where data correlation threatens data privacy and big data privacy. This work also suggests a methodology to deal with the mentioned problem. The proposed approach initially offers to realize the correlation amongst data using the Distance Correlation analysis method. Then, using this correlation as a parameter, clustering is performed over the dataset and divided into blocks. After that, data sensitivity gets calculated concerning the individual blocks instead of Global Sensitivity (GS). The last step is to use the calculated sensitivity to apply differential privacy for the data blocks. The Distance correlation analysis method applied at the first step ensures proper recognition and consideration of data correlation in the big dataset. The divide and conquer approach is adopted to handle the high dimensionality of the data. Calculating sensitivity ensures that noise of lower magnitude gets added to increase the data utility. The Pyspark technology is adopted to manage the processing of big data. Finally, we compare the observations extracted from the performed experiment with the results of [5] to establish its effectiveness.

Data privacy and data utility

Data utility means that we should maintain the availability of the data when conducting privacy preservation [18]. More clearly, data utility means the ease of using noisy data for data mining and other analysis, maintaining the correctness and accuracy of the analytical result drawn from the noisy data [18]. The most common metric to measure data utility is Information Loss. The more the information loss, the less the utility of data [18]. Data privacy also has a relationship with data utility. The privacy performance of the data decreases with the increase in data utility, and it creates a trade-off between data utility and data privacy. Figure 2 is a graphical representation of the same. To plot the graph between data privacy and data utility, we have calculated data privacy as the

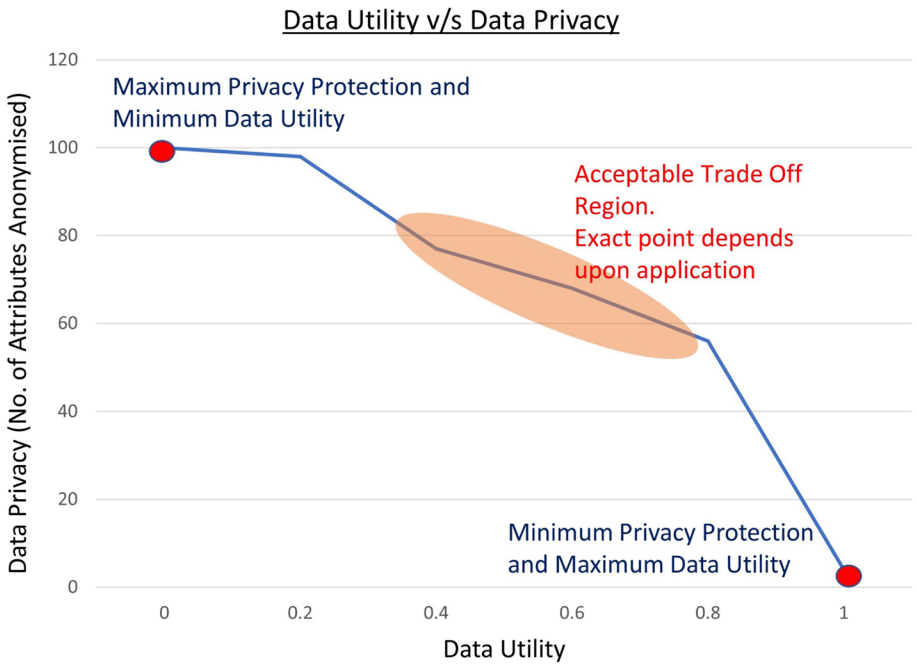

Fig. 2 Data utility and data privacy trade off

number of attributes that has been anonymized and to calculate data utility we have used Information Gain. In the presented work, we aim to provide enhanced data utility while maintaining the required data privacy levels. Thus, the measurement of data utility is crucial for the proposed methodology.

Data correlation and data privacy

Before 2011, when researchers talked about data privacy algorithms, they considered that there existed no correlation among data. But soon, in 2011, researchers started studying the potential of data correlation as a privacy threat to data and have given enough instances to support it [4]. Many privacy leakages became evident with the existence of data correlation among data. The privacy leakages were caused due to the homogeneity attack, background knowledge attack, and linkage attack. The main contributing factor to these attacks was unnoticed data correlation. If one ignored the existence of data correlation, these privacy leakages would also go unnoticed and cause a reduction in the privacy commitments of the data privacy algorithms. The pioneer in this was Gehrke et al. [19], who considered the case of social networks where users and their data are highly correlated, and even the strong privacy guarantee provided by Differential Privacy could not assure privacy for Social Network settings. Then in the subsequent years, researchers worked over the same lines and with various data to study the privacy threats associated with data correlation and proposed solutions. Table 2 gives a summary of all such works. Also, the following section provides a deeper insight into the same.

Another problem associated with the existence of data correlation is sensitivity. Correlated data causes a higher value of sensitivity of the data. While applying differential privacy algorithm, a higher sensitivity value will cause higher noise to be added to the original data. It adversely affects the data utility and causes its reduction. This is an undesirable effect and may render the privatized data useless.

Correlated big data privacy

As stated initially, data correlation poses a big threat to data privacy, and it causes privacy leakages that go unnoticed and causes unexpected compromises in data privacy. The real-world datasets are often large and accompanied by high dimensionality, which in turn causes high data correlation [4], which causes a potential threat to big data privacy. Given the massive amount of data and the combination of structured and unstructured data, some new Big Data models are a need to improve privacy and protection [4].

Organisation of the paper

This paper suggests a mechanism that deals with three main problems—Processing high dimensional big datasets, ensuring privacy protection of correlated data while using DP as the main privacy algorithm, and maintaining data utility while enhancing data privacy. The organization of the remaining paper is as follows: "Literature review" section presents a brief literature review of the topic. "Basic principles and theories" section gives a short description of the basic principles and theories used. Whereas "Proposed methodology" section discusses the proposed solution for the privacy protection of correlated big data and submits the model algorithm to implement it. "Experiment and analysis"

section describes the performed experiments and presents the analysis and results. Then the paper finally concludes. Table 1 provides a list of abbreviations used in this work.

Literature review

Differential privacy provides very robust privacy protection for data. It is dependent on pure mathematical theories. Researchers from 2011 have studied the potential of data correlation as a privacy threat to data and have given enough instances to support it [4]. Gehrke et al. [19], in the year 2011, considered the case of social networks where users and their data are highly correlated, and even the strong privacy guarantee provided by Differential Privacy could not assure privacy for Social Network settings. Kifer et al. [13] in 2011 gave initial arguments that the consideration of correlation between records is pivotal as the correlation between records or attributes can substantially decrease the privacy guarantee provided by any algorithm. Those mentioned above were the initial attempts to formalize Data Correlation as a general phenomenon for real-time datasets. These are considered pioneers in realizing the existence of data correlation in datasets and its potential as a privacy threat. Kifer et al. [21], in their successive work in the year 2014, proposed a privacy mechanism called Pufferfish. The mechanism helped develop privacy definitions for different data-sharing needs, studied existing privacy definitions, studied privacy compromise due to non-independent data records, and several other critical issues in terms of privacy. Since then, Yang et al. [22] in 2015, Wang et al. [23], Chen et al. [24] in 2017 proposed some solutions to it using Bayesian Networks, [21–24] proposed solutions using Probabilistic Models, Cao et al. in 2012 [25] and 2013 [26] proposed solutions using Behavioural and Similarity Analysis, Chen et al. [27] in 2013, and Liu et al. [15] in the year 2016 proposed modified perturbation mechanisms, Authors Kumar et al. [17] in 2018 have tried to offer solutions using Statistical Correlation Analysis Method, Lv et al. [5] in 2019 and Zhao et al. [15] proposed modifications to the DP algorithm, and the authors of [16, 28, 29] present the recent advancements regarding the same. Table 2 summarizes the notable works done by previous researchers and throws light on the limitations of the approaches proposed by them.

Table 1 List of abbreviations

S. no.	Abbreviation	Meaning
1	DP	Differential privacy
2	IID	Independent and identically distributed
3	GS	Global sensitivity
4	PM	Pufferfish mechanism
5	MIC	Mutual information correlation
6	LM	Laplace mechanism
7	DCov	Distance covariance
8	DCor	Distance correlation
9	DVar	Distance variance
10	MAE	Mean average error
11	CIS	Coupled item similarity

Table 2 Proposed solutions

S. no.	Privacy measure	Definition	Limitations
1	Pufferfish mechanism [21]	Formed strong foundation for other similar research	Could not satisfy differential privacy guarantee
2	Coupled behavior analysis (CBA) [25]	Good experimental results were obtained on real datasets	It does not furnish expected results for high dimensional data, other challenges of CBA are yet to be explored
3	CBA-HG and CBA-HC [31]	Experimental comparison showed that CBA-HG outperformed the mechanisms of [25]	Applicability on other datasets with different couplings is uncertain
4	Coupled item similarity (CIS) [26]	Proposed an effective mechanism to measure the non-IIDness	No solution to deal with the non-IIDness was proposed, without mentioning non-IIDness of data adversely affecting data privacy
5	Modified sensitivity calculation [27]	Multiplication of global sensitivity with the no. of correlated records for correlated datasets.	Data utility got highly degraded
6	Correlated sensitivity [12]	Noise reduced by an enormous amount and greater data utility as compared to [27]	Few parameters held trade-off with utility
7	Bayesian differential privacy [22]	Mechanism provided privacy for correlated data and against an adversary with partial background knowledge	Prior knowledge of probabilistic relationships is not possible
8	Dependent differential privacy [15]	High accuracy achieved	The estimation of value of ρ_{ij} is the key challenge
9	Pufferfish Wasserstein distance mechanism [6]	Mathematically proved the unnecessity to consider the correlation of distant nodes	When compared to the results of [17], it performed slightly worse for a particular range of values
10	Identity differential privacy [7]	Mechanism concluded that concepts of Information Theory are well suited to model the problems of dependent data in differential privacy	Practical implementation is not suggested as other privacy leakages were not studied
11	Bayesian network perturbation mechanism [8]	Proposed perturbation mechanism provided a decreased privacy budget and increased data utility	The requirement of modeling the Bayesian Network in advance may not be practically feasible
12	Statistical correlation analysis [9]	Enhanced accuracy by using correlation analysis techniques	The correlation analysis techniques and feature selection techniques used were not good enough to study complex relationships
13	Correlated differential privacy of big data publication [10]	Proposed use of Divide and conquer approach along with machine learning, Used correlated big datasets	Traditional correlation analysis technique used could not handle high dimensional data
14	Dependent differential privacy [13]	Proposed DDP and proved mathematically how it can be derived from DP	Lacks practical implementation
15	Temporal privacy leakage [14]	Temporal correlation along with the study of the relationship between data privacy and data utility	Other correlation models were not studied for temporal leakages
16	Weighted hierarchical graph mechanism [16]	Mechanism offers privacy guarantee in case of negative correlation as well	Not applicable to nonlinear queries
17	Temporal correlation mechanism [28]	Proposed w-event privacy using DP for location statistics and provided results regarding data utility	Correlation between other values was not studied

Table 2 (continued)

S. no.	Privacy measure	Definition	Limitations
18	Bayesian DP with Gaussian correlation model [29]	Proposed Bayesian DP model which used Gaussian correlation model to study data correlation	Approximation of accurate probabilistic values is a challenge
19	Social network data privacy analysis [30]	Analysed data privacy leakages in social network data using single and cross social extrapolation	They did not provide any solution framework for the problem

Among the discussed works, [5] is the most relevant to the proposed work. In [5], authors Lv et al. studied the data correlation among the dataset and then utilized the same to assure data privacy of the dataset using differential privacy. The main shortcomings of this paper were—(i) the Use of the Mutual Information Correlation analysis technique to calculate the data correlation among data, (ii) Its inefficiency in handling voluminous and high dimensional data, i.e., Big Data, (iii) Very few parameters were used for evaluation of the solution. In the presented work, we have proposed solutions to the shortcomings of [5].For the same, we used the Distance Correlation analysis method as it could handle high dimensional data. PySpark technology and the divide and conquer approach supported big data processing. We used parameters such as data utility, mean average error, information loss, variation with data size, and privacy budget values to evaluate the proposed method.

Basic principles and theories

Differential privacy definition

The foundation of the differential privacy mechanism is based on the concept of adjacent data sets. Let's consider two datasets, D_1 and D_2, which differ in one record, denoted by $| D_1 \Delta D_2 | = 1$; then these datasets are termed as adjacent datasets. The conventional definition of Differential Privacy is as follows—Suppose D_1 and D_2 are adjacent datasets, A is a privacy mechanism and if A satisfies ϵ -differential privacy for any output $A(D_1) \rightarrow R$, $S \in R$, then

$$Pr[A(D_1) \in S] \leq e^{\epsilon} Pr[A(D_2) \in S] \tag{1}$$

Often, such privacy mechanisms are realized using global sensitivity (GS). It is the measure of change in other records due to modification of another record. Let D_1 and D_2 be adjacent datasets; the global sensitivity can be defined as follow:

$$GS = max_{D_1,D_2} ||f(D_1) - f(D_2)||_1 \tag{2}$$

where f is the query function, $||.||_1$ is the L1-norm.

The differential privacy can be realized by adding noise 'e' to the output.

$$A(D) = f(D) + e \tag{3}$$

Definition of laplace mechanism

The Laplace mechanism will compute the function and perturb each coordinate with noise drawn from the LM distribution [32]. The noise scale will get adjusted to the sensitivity of the function (divided by ϵ). LM is used when the output is numerical [32].

Given the query function f and dataset D and the parameter ϵ, the Laplace mechanism A satisfies the following equation:

$$A(D) = f(D) + Laplace\left(\frac{GS}{\epsilon}\right) \tag{4}$$

where $Laplace\left(\frac{GS}{\epsilon}\right)$ is Laplace distribution, and parameter ϵ is called privacy budget. In practical applications, the inverse cumulative distribution function usually realizes the noise e.

Distance correlation

In the world of statistics, distance correlation means a measure of dependence between a pair of random vectors not necessarily having equal dimensions. The coefficient value ranges from [0,1], where the distance correlation coefficient is zero if and only if the random vectors are independent. It has the power to measure both linear and non-linear associations between the pair of random vectors, whereas the Pearson correlation can only capture the linear association between them. [33] shows how distance correlation gets estimated from the given data samples. A practical inference is that, by using two matrices, one can calculate the distance correlation. One of the matrices contains the pairwise distances between observations from X, and the other matrix contains observations from Y. We say X and Y are highly correlated if the items in these matrices co-varies together; otherwise, they have a meager correlation value.

Distance covariance The equation for Pearson covariance (Cov) between X and Y is given below:

$$Cov(x, y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)(y_i - y_j) \tag{5}$$

The terms $(x_i - x_j)$ and $(y_i - y_j)$ can be considered as a signed distance between i^{th} and j^{th} sample in one-dimension. These have been modified to centered Euclidean distances $D(x_i, x_j)$ in order to define distance covariance as given below:

$$DCov(x, y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}D(x_i, x_j).D(y_i, y_j) \tag{6}$$

Following are the properties for distance covariance:

I. If X and Y are independent, then DCov (X, Y) = 0
II. DCov (X, Y)≥ 0 and DCov (X, Y)≤1
III. $DCov^2(a_1 + b_1C_1X, a_2 + b_2C_2Y) =\mid b_1b_2 \mid DCov^2(X, Y)$ for all scalars b_1, b_2, constant vectors a_1, a_2 and orthonormal matrices C_1, C_2.
IV. This is also defined for random variables in different dimensions, as we can compute the distance between observations in any dimension.
V. If the random vectors $(X_1, Y_1)and(X_2, Y_2)$ are independent then $DCov(X_1 + X_2, Y_1 + Y_2) \leq DCov(X_1, Y_1) + DCov(X_2, Y_2)$

Distance variance It is a special case of distance covariance where the two variables are the same

$$DVar(x) = DCov(x, x) \tag{7}$$

Properties of distance variance are:

I. DVar(X) = 0 if and only if X = E[X], where E is the Euclidean Distance.
II. DVar(X) = 0 if and only every observation is the same.
III. DVar(A+bCX) = | b | DVar(x) for all constant orthonormal matrices C, constant vectors A, and scalars B
IV. DVar(X+Y) ≤ DVar(X) + DVar(Y) if and only if X and Y are independent.

Distance correlation One can obtain it for two random variables by dividing distance covariance by the product of their distance standard deviation. It is given by:

$$DCor(x, y) = \frac{DCov(x, y)}{\sqrt{DVar(x) * DVar(y)}} \tag{8}$$

Properties:

I. 0≤ DCor (X, Y) ≤ 1 for all X and Y; This is different from Pearson Correlation, where correlation can even be negative
II. DCor (X, Y) = 0 if and only if X and Y are independent
III. It defines a statistical test for dependence with a permutation test.

Information gain

Information gain is a popular metric that is used to measure Data Utility. It measures the amount of information that can be derived from the given dataset. To compute it, we have used Information Loss [34]. Its value ranges between 0 to 1. One can compute Information Loss using the:

$$Information\ Loss\ in\ each\ field\ = \frac{|(Value_{original} - Value_{modified})|}{(Value_{original} + Value_{modified})} \tag{9}$$

where $Value_{original}$ is the original value of the attribute and $Value_{modified}$ is the attribute's value after applying the final methodology. Information Gain gets calculated by complementing the value of Information Loss [34].

$$Information\ Gain\ of\ each\ field\ = (1 - Information\ Loss) \tag{10}$$

Proposed methodology

As per the literature review and detailed research on the topic, one can extract the following:

1. Data Correlation must be studied, analyzed, and considered to ensure data privacy needs of data. When left unnoticed, these can cause severe privacy leakages, which

may lead to the degradation of the privacy commitments of the applied data privacy algorithm.

2. When there is an increase in the volume, size, and dimensionality of data, the correlation among data also increases. Either the number of correlated data gets increased, or the magnitude of data correlation increases. In some cases, both may happen. Thus, correlated big data has great potential for data privacy leakages.
3. The implementation of differential privacy mechanism usually consumes too many computing resources and a privacy budget. This problem further gets magnified when the data is big data.
4. The increasing magnitude of data correlation, size, and dimensionality of data causes an increase in the overall sensitivity of data. Increased sensitivity will result in the addition of larger noise when applying differential privacy algorithm. Consequently, this results in a highly polluted dataset that can be of very little use after data publication.

Based on the above observations, this paper proposes a novel mechanism to study data correlation using the Distance Correlation Analysis Method, efficiently handle big data using the divide and conquer approach and finally apply differential privacy over the correlated big dataset.

The proposed mechanism can be summarised as a sequence of steps in the following manner:

Step 1—Calculation of Data Correlation—The standard approaches to study data correlation are—Statistical Methods, Mutual Information Correlation Method, and Distance Correlation Analysis Method [5]. The statistical methods of data correlation analysis could only study the linear relationship among data. The MIC could explore the non-linear relationship of data along with the linear relationship, but it was inefficient in handling high dimensional data [4, 5]. But the distance correlation analysis could efficiently handle linear, non-linear, and high dimensional data. Thus in the presented work, we have used the distance correlation analysis method to generate the Data Correlation Matrix instead of the MIC method, which was used in [5].

Step 2—Division into smaller data blocks—The enormous size of big data leads to increased overheads. The proposed mechanism divides the data into smaller data blocks to overcome this setback. It is done using a combination of the k-means clustering algorithm and the distance correlation matrix (generated in Step 1). The k-means clustering algorithm, along with the distance correlation matrix, divides the large dataset D into multiple smaller data blocks, i.e., $D_1, D_2, \ldots, D_n$ such that $D_1 U D_2 U \ldots U D_n = D$. In other words, $D_i \Delta D_j = \phi$ where $i \neq j$.

Step 3—Computing Correlation Sensitivity—This paper uses the Correlation Sensitivity method proposed in [5] as it is very appropriate. Correlation sensitivity states

that sensitivity must be calculated of any query f only over D_i instead of the entire dataset D. This helps reduce the noise added for distortion. It also helps to improve data utility.

Step 4—Noise Addition—The noise gets calculated using the correlation sensitivity (as calculated in Step 3) and using the Laplace Mechanism. This step solves the problems discussed in observations 3 and 4. It is a crucial step as the working and effectiveness of the differential privacy algorithm primarily depends on the sensitivity parameter. By reducing the magnitude of noise to be added, one can increase the data utility of the protected dataset. The calculated noise is added to the record value to generate a noisy value.

$$\mathrm{N(D)} = \mathrm{f(D)} + \mathrm{noise\ (e)}$$

where N(D) is the noisy data, f is the query executed over data D, and noise is the noise value calculated using correlation sensitivity and the Laplace mechanism.

Step 5—Implementing Differential Privacy—In our proposed work, we used the parallel combination of differential privacy to apply the differential privacy algorithm over the partitioned dataset. This step is a solution to reduce the overheads associated with high-dimensional datasets. This step must be applied to all the formed data blocks to ensure the overall privacy protection of the big data.

Figures 3, 4, 5, and 6 give a schematic description of the steps of the proposed methodology.

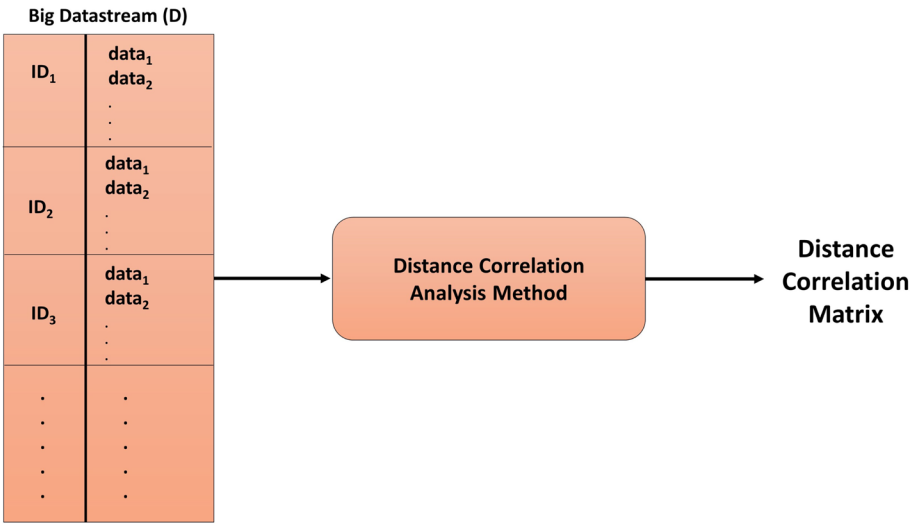

Fig. 3 Step 1 of proposed methodology

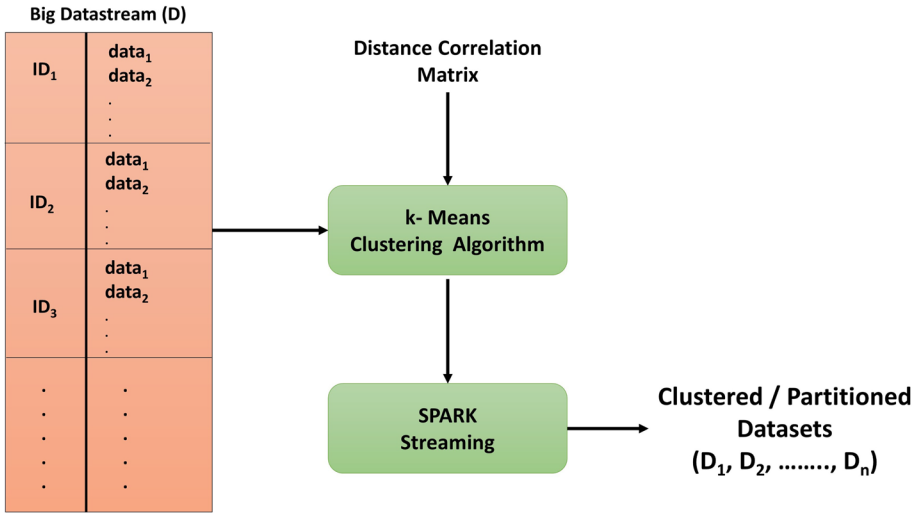

Fig. 4 Step 2 of proposed methodology

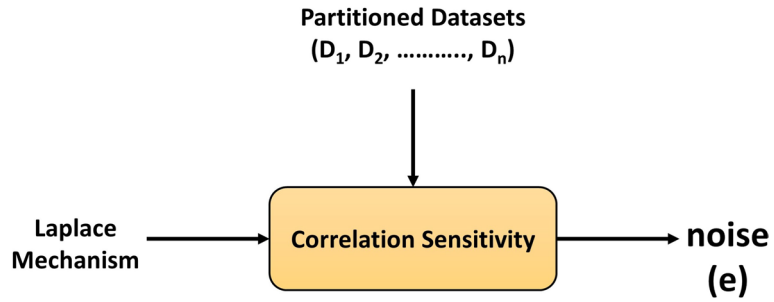

Fig. 5 Step 4 of proposed methodology

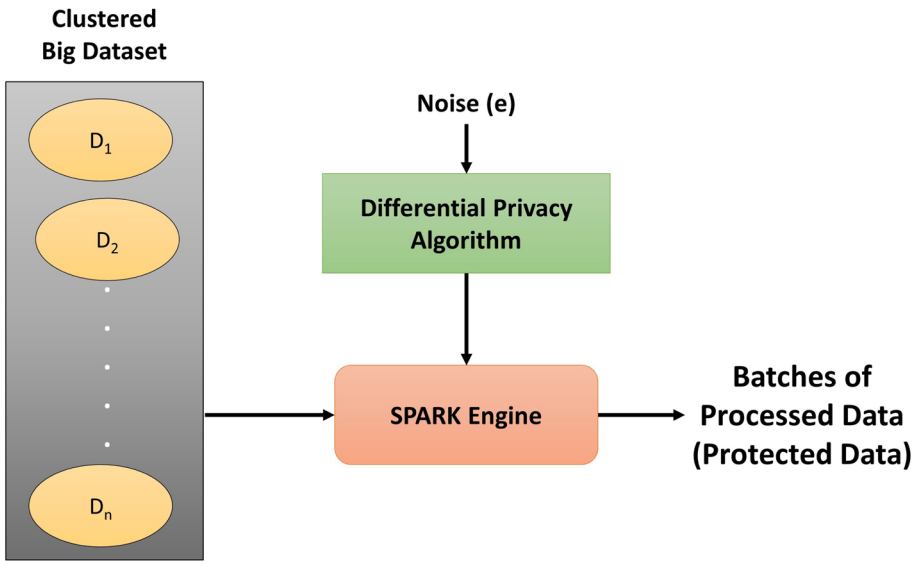

Fig. 6 Step 5 of proposed methodology

Algorithm 1 Algorithm to compute Distance Correlation

DistanceCorrelation($\mathbf{X}, \mathbf{Y}$)
    Inputs to the algorithm are two two dimensional arrays X and Y
    (They will have dimension = (n*m)*2)
if No.OfRows(X) == No.OfRows(Y) **then**
    Proceed to next step
else
    Raise Error
end if
$N = No.\ of\ rows(X)$
Store the matrix in variable a, $a = pairwiseDistance(X)$
Store the matrix in variable b, $b = pairwiseDistance(Y)$
A = a − mean of "a" along column − mean of "a" along row + mean of complete "a"
B = b - mean of "b" along column − mean of "b" along row + mean of complete "b"
$\mathbf{DCov2_{xy}} = \mathbf{sum(A * B)/(n * n)}$
$\mathbf{DCov2_{xx}} = \mathbf{sum(A * A)/(n * n)}$
$\mathbf{DCov2_{yy}} = \mathbf{sum(B * B)/(n * n)}$
$\mathbf{DCor} = \frac{\sqrt{(\mathbf{DCov2_{xy}})}}{\sqrt{(\sqrt{(\mathbf{DCov2_{xx} * DCov2_{yy}})})}}$
Return DCor

Algorithm 2 Clustering Algorithm

Input: Given a large dataset D containing n records and p features, i.e., D of size n*p
Output: B dataset containing k independent blocks from dataset D
ClusteringAlgorithm()
    Initialize k randomly chosen records from D as initial cluster centers.
        Choose a clustering algorithm.
        Choose distance evaluation metric for clustering algorithm i.e. MIC or distance correlation
        Choose number of iterations as a stopping criterion.
    D = $r_1, r_2, r_3, ...r_i, \ldots r_n | 1 \leq i \leq n$, where r_i is a record of length p.
        InitialCentroid = $c_1, c_2, \ldots ...c_k | 1 \leq k \leq n$
    Repeat till given number of iterations
        for i = $\mathbf{1, 2, 3, ..., n}$ **do**
            for j = $\mathbf{1, 2, ..., k}$ **do**
                $\mathbf{dist = DistanceEvaluationMetric(D[i], InitialCentroid[i])}$
                Assign the record to the cluster where the distance has the maximum value
        Calculate the mean of each cluster, update the centroid of a cluster with the cluster's mean.
        Group all the records based on clusters and form a new dataset B.
    Output is new Dataset B with k independent blocks, $B = d_1, d_2, d_3, ..., d_k$, where $d_1 U d_2 U d_3 \ldots\ldots U d_k = D$

Experiment and analysis

Experimental setup

The experimental platform is a laptop with Intel® Core. i5-8250U CPU @ 1.6 GHz 1.80 GHz processor, 8 GB RAM, 64-bit operating system, x64-based processor, Windows 10. We used the MIC and Distance Correlation Analysis method to calculate the correlation between the records and performed a comparative study. We can present the algorithms adopted as Algorithm 1 and 2. Then we compiled the experiments and results and implemented the methodologies using the TensorFlow environment on Google Colabotary, with 13 GB RAM and 108 GB disk. The following subsection gives a detailed description of the datasets used.

Dataset description

In the present work, we utilized the following datasets. The main reason for considering these datasets is the existence of an adequate amount of data correlation and their privacy needs. Table 3 gives brief descriptions of the datasets, and Tables 4, 5, and 6 present the statistical features of the respective datasets. During the experiments and observations, these datasets get referred to as Dataset a, Dataset b, and Dataset c.

1. New-York-City trip data 2016: This dataset contains 19 attributes providing information about taxi trips in 2016.
2. Chicago Crime data: This dataset contains 22 attributes providing information about crime incidents. The researchers obtained the crime data of Chicago from the city of Chicago data portal.
3. New-York-City trip data 2013: This dataset contains 26 attributes providing information about taxi trips in 2013.

Analysis and results

We used data correlation analysis, epsilon values, data size, mean average error, and data utility to analyze the proposed methodology results. For ease, we have named the

Table 3 Dataset description

S. no.	Dataset	Attributes	Attributes considered for experimental purpose
1	New-York-City trip data-green Taxi-2016	VendorID, lpep pickup datetime, Lpep dropoff datetime, Store and fwd flag, RateCodeID, Pickup longitude, Pickup latitude, Dropoff Longitude, Dropoff latitude, Passenger count, Trip distance, Fare Amount, Extra, MTA tax, Tip amount, Tolls amount, Ehail fee, improvement surcharge, Total amount, Payment type, Trip type	VendorID, RateCode, Pickup longitude, Pickup latitude, Dropoff Longitude, Dropoff Latitude, Passenger count, Trip distance, Fare Amount, Extra, MTA tax, Tip amount, Tolls amount, improvement surcharge, Total amount, Payment type, Trip type
2	Chicago Crime Data	Id, caseNo, date, block, IUCR, Primarytype, Description, location, arrest, domestic, beat, district, ward, community area, FBIcode, xcoordinate, ycoordinate, year, updated, latitude, longitude, location	Id, caseNo, date, block, IUCR, arrest, domestic, beat, district, community area, FBIcode, xcoordinate, ycoordinate, year, updated, latitude, longitude, location
3	New-York-City trip data-2013	Store and fwd flag, rate code, Dropoff latitude, Passenger count, Trip distance, Fare amount, Extra, MTA tax, Trip amount, Tolls amount, Ehail fee, Total payment, Trip type, Pickup Longitude, Pickup Latitude, Droffoff longitude, Vendor id, pickup hour, pickup day, Pickup month, pickup year, pickup minute, Dropoff hour, Dropoff day, Dropoff month, Dropoff year, Dropoff minute	Store and fwd flag ,rate code, Dropoff latitude, Passenger count, Trip distance, Fare amount, Extra, MTA tax, Trip amount, Tolls amount, Total payment, Pickup Longitude, Pickup Latitude, Droffoff longitude, Vendor id, pickup hour, pickup day, Pickup month, pickup year, pickup minute, Dropoff hour, Dropoff day, Dropoff month, Dropoff year, Dropoff minute

Table 4 Statistical features of the dataset a

Features → Attributes ↓	Count	Mean	SD	Min	Max
VendorID	9999	1.8638863	0.34292631	1.0	9998.0
RateCodeID	9999	1.0482048	0.4243966	1.0	5.0
Pickup longitude	9999	− 73.899834	1.65348802	− 74.073845	0.0
Pickup latitude	9999	40.7194723	0.91256991	0.0	40.92078
DropOff longitude	9999	− 73.8976524	1.653648802	− 74.07387	0.0
DroffOff latitude	9999	40.72007131	0.91274002	0.0	41.178898
Passenger count	9999	1.4497449	1.1375009	0.0	6.0
Trip distance	9999	3.05243424	2.83776974	0.0	61.7
Fare amount	9999	12.485061506	8.74345086	70.0	222.5
Extra	9999	0.49219921	0.07028614	− 0.5	0.5
MTA tax	9999	1.1725532	2.09252519	0.0	33.37
Tip amount	9999	1.17255325	2.09252194	0.0	58.0
Toll amount	9999	0.0592209	0.5719925	0.0	41.178898
Improvement surcharge	9999	0.2957995	0.03912370	1.0	5.0
Total amount	9999	14.99873372	9.860037110	1.0	229.34
Payment type	9999	1.58855885	0.51321125	1.0	5.0
Trip type	9999	1.010601	0.1024193	1.0	2.0

Table 5 Statistical features of the dataset b

Features → Attributes ↓	Count	Mean	SD	Min	Max
ID	7472435	6785755.704	3364129.889	1000	9999999
Case number	7472431	305161.8421	134932.5016	JB299184	ZZZ199957
Date	7472435	Null	Null	01/01/2001	12/31/2021
Block	7472435	Null	Null	0000X E 100 PL	XX unknown
IUCR	7472435	1124.011711	813.08648	0110	9901
Primary type	7472435	Null	Null	ARSON	Weapons violation
Description	7472435	Null	Null	300 and under	WireRoom/sports
Location desc	7463560	Null	Null	CTA L platform	YMCA
Arrest	7472435	Null	Null	False	True
Domestic	7472435	Null	Null	False	True
Beat	7472435	1187.31387	702.93333	0111	2535
District	7472388	11.294631	6.9475874	001	16
Ward	6857594	22.72791258	13.8358939	1	9
Community area	6858956	37.53571709	21.540734015	0	9
FBI code	7472435	12.1414402	7.3373007	01A	26
X co-ordinate	7396684	1164568.996	16850.99545	0	1205119
Y co-ordinate	7396684	1885737.64226	32278.566004	0	1951622
Year	7472435	2009.4327	5.855461479	2001	2021
Updated on	7472435	Null	Null	01/01/2007	12/31/2021
Latitude	7396684	41.84206024	0.088805891	36.6194463	42.0229103
Longitude	7396684	− 87.67161265	0.06109258	− 87.5245293	− 91.68656568
Location	7396684	Null	Null	19446395, − 9.	2910333, − 8.

Table 6 Statistical features of the dataset c

Features → Attributes ↓	Count	Mean	SD	Min	Max
Store and Fwd flag	7472431	1.48811223	0.58421322	0.0	1.0
DroffOff latitude	6857594	40.72007131	0.91274002	0.0	41.178898
PassengerCount	7472435	3.222114	2.8712441	1.0	6.0
Trip distance	7472435	3.05243424	2.83776974	0.0	68.7
Fare amount	7472435	14.485061506	9.74345086	72.0	254.42
MTA tax	7396684	3.1725532	2.88417522	1.24	58.37
Trip amount	7472435	1.17255325	2.09252194	0.0	58.0
Tolls amount	7463560	5.44124560	3.88891100	0.0	41.178898
Total payment	7472435	15.99873372	9.860037110	5.0	302.34
Pickup longitude	6857594	− 73.899834	1.65348802	− 74.073845	0.0
Pickup latitude	6857594	40.7194723	0.91256991	0.0	40.92078
Drop off longitude	6857594	− 73.89765	1.653648802	− 74.07387	0.0
Vendor ID	7472435	12.1414402	7.3373007	01A	26BZ
Pickup month	7472435	5.4433841	4.3200112	1	12
Pickup year	7472435	2013	0.2211384	2013	2013
DropOff month	7396684	4.7770122	4.0998411	1	12
DropOff year	7396684	2013	0.06109258	2013	2013

methodology proposed in [5] as cdp-method and the method proposed in the current work as d-method.

Data correlation

There are numerous methods to study data correlation among data. Our proposed methodology used distance correlation, and the method proposed by [5] used mutual information correlation (MIC) for the same. Figures 7 and 8 present the correlation coefficients calculated using distance correlation and mutual information correlation for Dataset a. Datasets b and c adopted the same methodology. One can observe that Distance correlation analysis can measure the data correlation better than MIC.

Epsilon values

Epsilon is an important parameter that affects differential privacy protection, also known as the Privacy Budget. The lower the value of epsilon, the higher the level of privacy protection, which lowers the utility of the data. Higher perturbation results in a higher loss of original data values, but at the same time, it provides higher privacy protection levels. The observations from the experimental analysis suggest that the proposed d-method, when compared with the cdp method [5], provides nearly equal privacy performance. We depicted the above using Fig. 11, where we compared both the methods for privacy performance using three different datasets.

Datasize

As per the traditional assumptions of differential privacy, the privacy protection levels do not vary with the amount of data. This is considered one of the advantages of differential privacy. The comparative analysis of the cdp method [5] and proposed

d-method shows that both the methods have a meager impact on privacy performance with increasing data size and have nearly identical curves, as shown in Fig. 12.

Number of clusters

Figure 9 shows the variation of privacy performance epsilon with changing values of r, i.e., the number of data subsets formed from the large datasets. Observations show that a small value of r results in a more significant value of r-MAE, which results in a more unsatisfactory privacy protection performance. Nevertheless, after a threshold value of r, the value of r-MAE becomes stable, and one can obtain optimal privacy protection performance. This result is similar to the development of [5]. Hence, one can say that in terms of variation of the value of r, the proposed method is equally efficient to the cdp-method [5].

$$r - MAE = \sum_{i=1,\epsilon}^{r} \frac{MAE_{i,\epsilon}}{r} \tag{11}$$

Data utility

Data utility and Data Privacy have an inherent trade-off nature. So when we see an incline in the data utility, it implies a decline in data privacy and vice versa. Due to this, the measurement of data utility becomes very important. Various metrics have been used to measure the data utility, which measures the data privacy level. We have used information gain as the metric for measuring data utility. The more the information gained more is the data utility. Table 7 states the information gain values of different clusters using the cdp-method [5] and d-method, respectively. Also, Fig. 10 is the same graphical representation. After comparison of data utility values between the cdp-method [5] and the d-method, we observed that the d-method incurs greater Information Gain values for all the clusters implying a better data utility. Then both the methods are compared against the conventional Differential Privacy Algorithm, and one can observe that the conventional DP provides the least data utility, followed by the cdp-method [5]. The d-method offers the highest Data Utility. Table 8 depicts these values.

Table 7 Information gain values for different clusters using cdp-method [5] and the proposed d-method

Cluster no.	Dataset a		Dataset b		Dataset c	
	Info gain values using cdp-method	Info gain values using d-method	Info gain values using cdp-method	Info gain values using d-method	Info gain values using cdp-method	Info gain values using d-method
Cluster 1	0.0	0.163	0.038	0.144	0.019	0.078
Cluster 2	0.020	1.0	0.055	0.312	0.176	0.221
Cluster 3	0.026	0.149	0.070	0.971	0.036	0.349
Cluster 4	0.039	0.731	0.085	0.712	0.069	0.918
Cluster 5	0.084	0.096	0.091	0.196	0.088	0.159

The values are normalised for easy interpretation

Table 8 Information gain values for conventional DP, cdp-method [5] and d-method

Info gain values when using Conventional DP	Info gain values when using cdp-method	Info gain values when using d-method
0.00	0.313	1.00

The values are normalised for easy interpretation

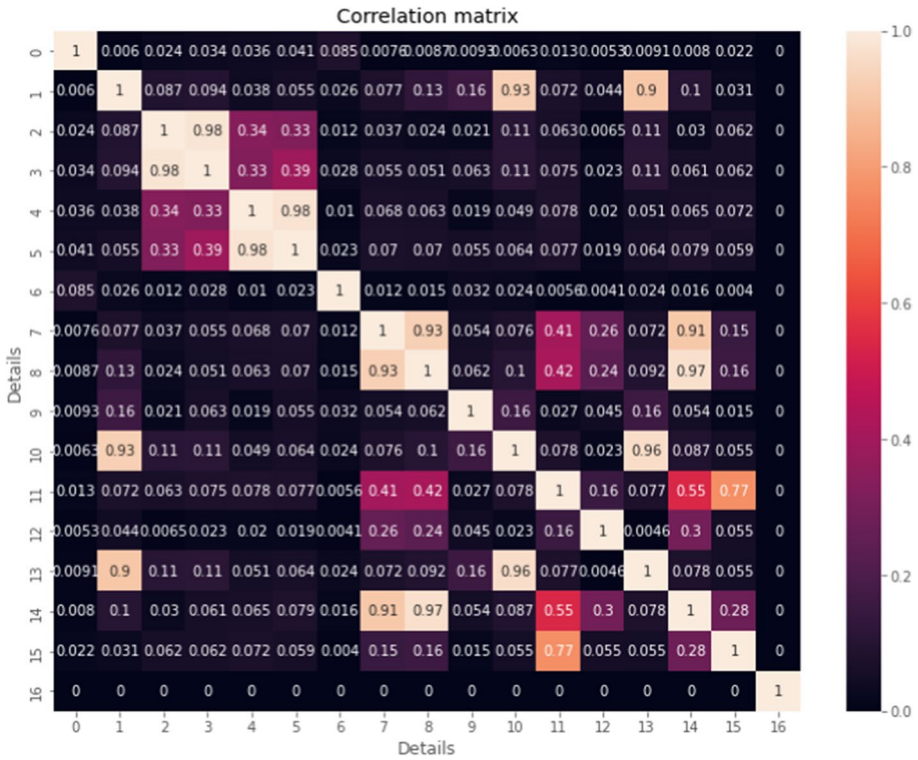

Fig. 7 Distance correlation matrix

Conclusion

The experimental analysis showed satisfactory results of the proposed mechanism. This paper initially studied how the existence of correlation among data can adversely affect the privacy guarantees of any privacy algorithms. We looked at the above with the help of an extensive literature survey and other experimental analyses. One can observe a noticeable difference when data were clustered: (i) on a general basis and (ii) based on the existing correlation. The clusters formed were very different for the two cases. It further strengthens the notion that correlation dramatically impacts how data gets interpreted and is considered for privacy mechanisms. The proposed mechanism used the distance correlation analysis technique to study the correlation among data in real-world datasets. This correlation analysis technique is selected because it can handle high-dimension, linear, and non-linear data, as most real-world datasets are high-dimensional and non-linear by nature. To address the large size of the real-world dataset, the proposed mechanism has used the cdp-method, i.e., the divide and conquer approach, further combined with the parallel combination of Differential Privacy Protection Mechanism and PySpark technology to ensure the privacy of data. Results showed that the

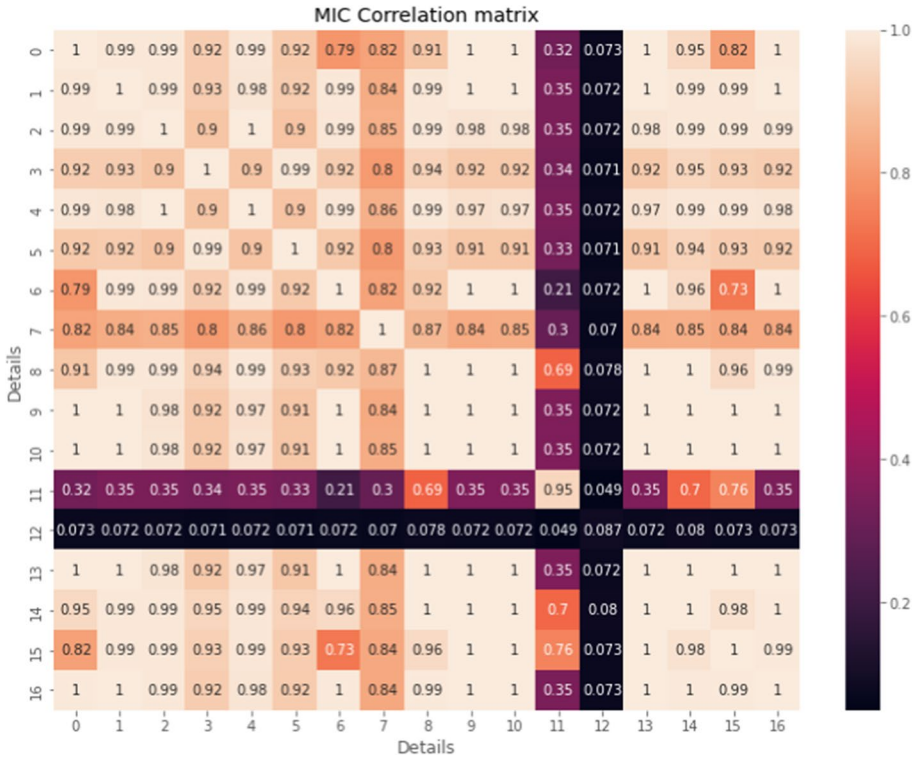

Fig. 8 MIC correlation matrix

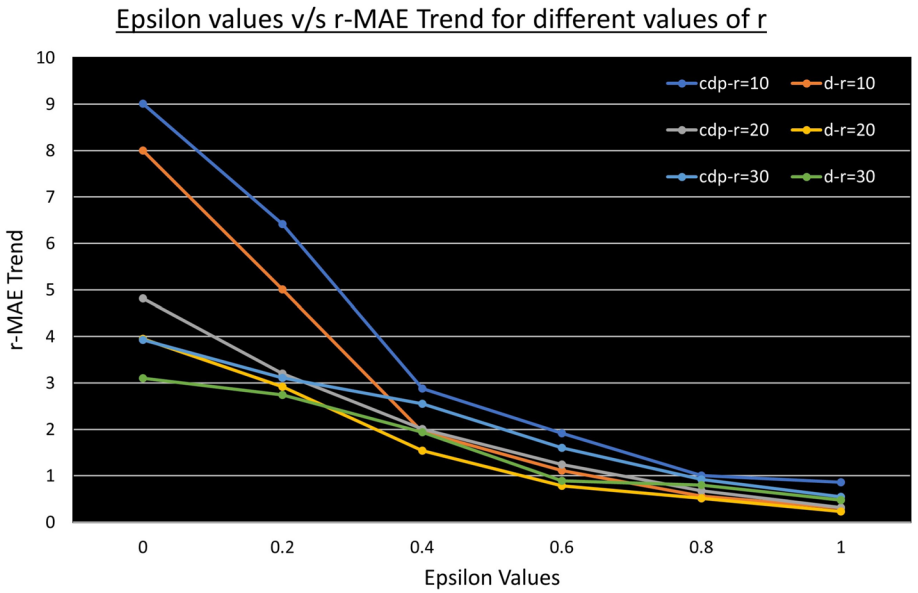

Fig. 9 Epsilon versus r-MAE trend for the different datasets

distance correlation analysis method is better than the MIC method in correlation analysis and data utility. Other results were similar to the results of the cdp-method, which proved that the proposed methodology provides better data utility while maintaining the data privacy levels offered by the cdp-method. We can shortly summarize these observations as the following:

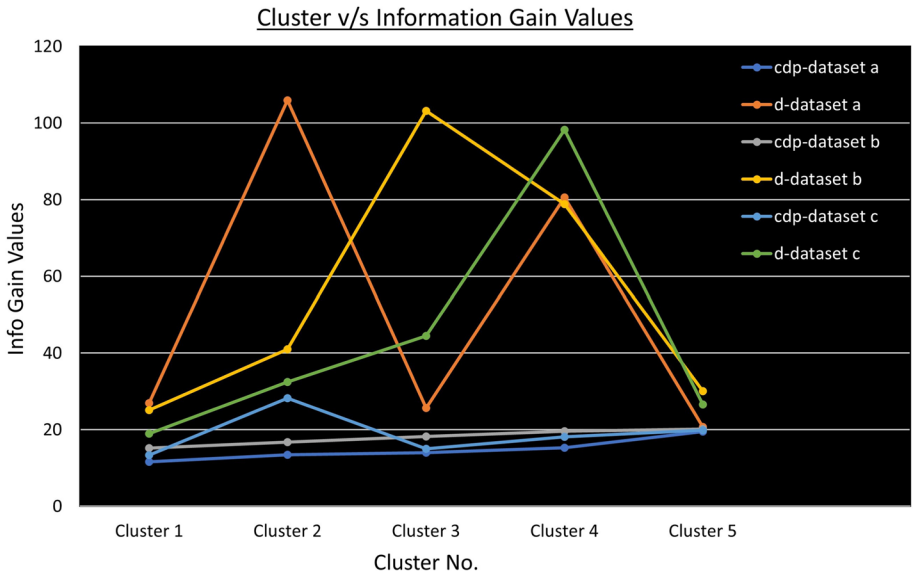

Fig. 10 Information gain values for cdp method [5] versus proposed d method

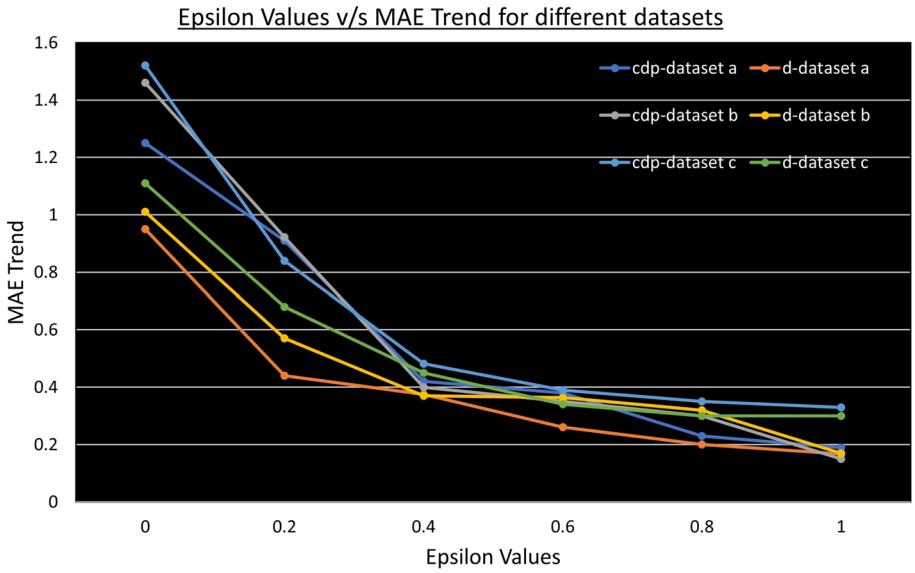

Fig. 11 Epsilon versus MAE Trend for the different datasets

1. This paper studied the adverse effect of data correlation in real-world datasets.
2. We used the distance correlation analysis technique to study the correlation among data. It could efficiently handle high-dimensional data, unlike the traditional MIC analysis method.
3. Divide and conquer methodology is used to handle the big dataset, and a parallel combination of Differential Privacy Mechanism ensures privacy protection for the entire dataset. This is similar to the cdp-method.

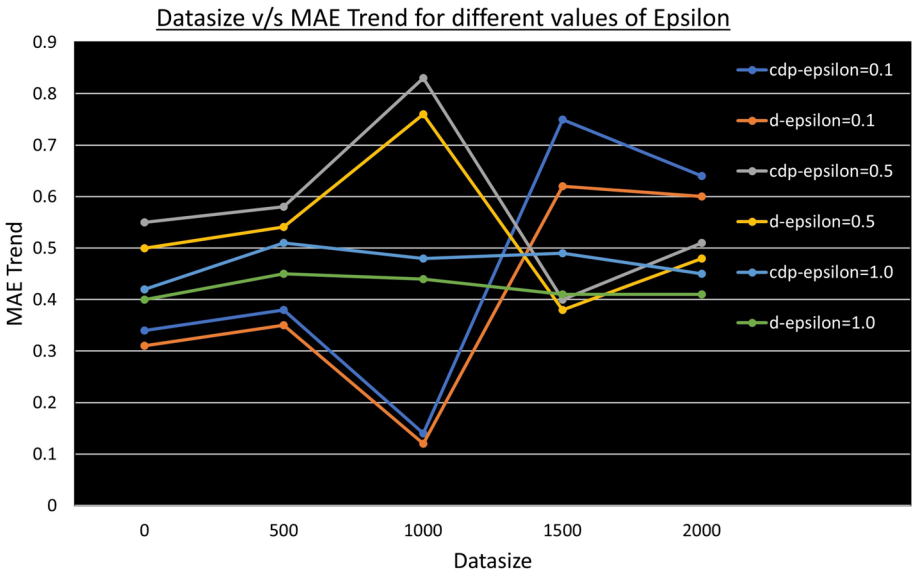

Fig. 12 Datasize versus MAE trend for the different datasets

4. Experimental results showed that distance correlation offered the required privacy protection of correlated data and maximum data utility.

$$DataUtility(ConventionalDP) < DataUtility(cdp - method) < DataUtility(d - method)$$

5. The proposed work utilized the big data framework, 'PySpark,' to handle Big Data efficiently. The researchers of the cdp-method used no such technology.

Acknowledgements
Not applicable.

Author contributions
SB conducted the literature review process, proposed the novel approach, wrote the manuscript, extracted the results of the experiments and arranged them in a presentable manner. AF helped with the presentation of the manuscript, performing the experiments and preparation of the figures. NK has helped with his technical expertise throughout the research work. He has been a constant guide from the beginning to the end. PA has helped with the final formatting of the paper and preparation of several tables. All the authors reviewed the manuscript.

Authors' information
Sreemoyee Biswas is currently pursuing a Ph.D. in Computer Science and Engineering from Maulana Azad National Institute of Technology, Bhopal, India. Her field of research is "Big Data Privacy." Other areas of specialization include Data Privacy, Information Security, and Machine Learning. She has about two years of experience as an Assistant Professor. Her Educational Qualification is M.Tech & B.E. in Computer Science and Engineering. Ms. Sreemoyee Biswas has publications in SCI indexed journals, Scopus indexed journals & National Conference.
Anuja Fole has completed her Mtech in Advanced Computing specialization under Computer Science branch from Maulana Azad National Institute of Technology.She is currently working as Data Scientist. Her areas of specialization are Machine Learning, Big data and Data privacy. Currently her interest of research is "Big Data". She has two years of industry experience as Associate Software Engineer. Her educational qualification is B.E in Information Technology.
Nilay Khare is working as Professor in MANIT Bhopal. He has more than 21 years of experience. His Educational Qualification is a Ph.D. in Computer Science & Engineering. Dr. Nilay Khare's areas of Specialization are Big Data, Big Data Privacy & Security, Wireless Networks, Theoretical computer science. He has publications in 54 International and National Conferences and International Journal. He is a Life Member of ISTE.
Pragati Agrawal is working as Assistant Professor in MANIT Bhopal. She has more than five years of experience. Her Educational Qualification is a Ph.D. in Computer Science & Engineering. Dr. Pragati Agrawal's areas of Specialization are Theoretical Computer Science, Energy Efficiency. Dr. Pragati Agrawal's publications are in International and National Conferences and International Journal. She is a Life Member of IEEE and ACM.

Funding
Not applicable.

Availability of data and materials
All relevant research data and materials are available with the authors.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors have given consent for publication of the matter.

Competing interests
The authors declare that they have no competing interests.

Received: 23 March 2022 Accepted: 23 February 2023
Published online: 07 March 2023

References

1. Liang JY, Feng CJ, Song P. A survey on correlation analysis of big data. J Soft. 2016;01(39):1–18.
2. Reshef D, Reshef Y, Finucane H, Grossman S, McVean G, Turnbaugh P, et al. Detecting novel associations in large data sets. Science (New York, NY). 2011;12(334):1518–24.
3. Abdalla HB. A brief survey on big data: technologies, terminologies and data-intensive applications. J Big Data. 2022;11:9.
4. Biswas S, Khare N, Agrawal P, Jain P. Machine learning concepts for correlated Big Data privacy. J Big Data. 2021;12:1.
5. Lv D, Zhu S. Achieving correlated differential privacy of big data publication. Comput Secur. 2019;05:82.
6. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data. 2007;1(1):3-es.
7. Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd international conference on data engineering; 2007. p. 106–15.
8. Dwork C. Differential privacy. In: 33rd international colloquium on automata, languages and programming, part II (ICALP 2006). vol. 4052 of lecture notes in computer science. New York: Springer; 2006. p. 1–12. Available from: https://www.microsoft.com/en-us/research/publication/differential-privacy/.
9. Yang X, Wang T, Ren X, Yu W. Survey on improving data utility in differentially private sequential data publishing. IEEE Trans Big Data. 2017;1:1.
10. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big Data. 2016;11:3.
11. Jain P, Gyanchandani M, Khare N. Differential privacy: its technological prescriptive using big data. J Big Data. 2018;04:5.
12. Zhu T, Xiong P, Li G, Zhou W. Correlated differential privacy: hiding information in non-IID data set. IEEE Trans Inf Forensics Secur. 2015;10(2):229–42.
13. Kifer D, Machanavajjhala A. No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data. SIGMOD'11. New York, NY, USA: Association for Computing Machinery; 2011. p. 193–204. Available from: https://doi.org/10.1145/1989323.1989345.
14. Wu G, Xia X, He Y. Extending differential privacy for treating dependent records via information theory; 2017, 03.
15. Zhao J, Zhang J, Poor HV. Dependent differential privacy for correlated data; 2017. p. 1–7.
16. Li Y, Ren X, Yang S, Yang X. Impact of prior knowledge and data correlation on privacy leakage: a unified analysis. IEEE Trans Inf Forensics Secur. 2019;14(9):2342–57.
17. Kumar S, Chong I. Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. Int J Environ Res Public Health. 2018;15(12):1. Available from: https://www.mdpi.com/1660-4601/15/12/2907.
18. Yang X, Wang Teng RXYW. Survey on improving data utility in differentially private sequential data publishing. IEEE Trans Big Data. 2017;1:1.
19. Gehrke J, Lui E, Pass R. Towards privacy for social networks: a zero-knowledge based definition of privacy. In: Ishai Y, editor. Theory of cryptography. Berlin: Springer; 2011. p. 432–49.
20. Belcastro CRMFe. Programming big data analysis: principles and solutions. J Big Data. 2022;01:9.
21. Kifer D, Machanavajjhala A. Pufferfish: a framework for mathematical privacy definitions. ACM Trans Datab Syst (TODS). 2014;01:39.
22. Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data; 2015.
23. Wang Y, Song S, Chaudhuri K. Privacy-preserving analysis of correlated data; 2016. arXiv:1603.03977.
24. Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. IEEE Trans Big Data. 2017;12(PP):1.
25. Cao L, Ou Y, Yu P. Coupled behavior analysis with applications. IEEE Trans Knowl Data Eng. 2012;08(24):1.
26. Cao L. Non-IIDness learning in behavioral and social data. Comput J. 2013;08(57):1358–70.
27. Chen R, Fung B, Yu P, Desai B. Correlated network data publication via differential privacy. VLDB J. 2014;08(23):653–76.

28. Hemkumar D, Ravichandra S, Somayajulu DVLN. Impact of data correlation on privacy budget allocation in continuous publication of location statistics. Peer-to-Peer Netw Appl. 2021;14(3):1650–65.
29. Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. IEEE Trans Big Data. 2021;7(04):1.
30. Cerruto CSDDEAF. Social network data analysis to highlight privacy threats in sharing data. J Big Data. 2022;02:9.
31. Song Y, Cao L, Wu X, Wei G, Ye W, Ding W. Coupled behavior analysis for capturing coupling relationships in group-based market manipulations. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining; 2012.
32. Dwork C, Roth A. The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci. 2014;9(3–4):211–407. Available from: https://doi.org/10.1561/0400000042.
33. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769–94. Available from: http://www.jstor.org/stable/25464608.
34. Jain P, Gyanchandani M, Khare N. Enhanced secured map reduce layer for big data privacy and security. J Big Data. 2019;03:6.

Publisher's Note