Journal of Big Data

SURVEY PAPER

Open Access

Machine learning concepts for correlated Big Data privacy

Sreemoyee Biswas*, Nilay Khare, Pragati Agrawal and Priyank Jain

*Correspondence:
shonai.biswas@yahoo.in
Department of Computer Science, Maulana Azad National Institute of Technology, Bhopal, India

Abstract

With data becoming a salient asset worldwide, dependence amongst data kept on growing. Hence the real-world datasets that one works upon in today's time are highly correlated. Since the past few years, researchers have given attention to this aspect of data privacy and found a correlation among data. The existing data privacy guarantees cannot assure the expected data privacy algorithms. The privacy guarantees provided by existing algorithms were enough when there existed no relation between data in the datasets. Hence, by keeping the existence of data correlation into account, there is a dire need to reconsider the privacy algorithms. Some of the research has considered utilizing a well-known machine learning concept, i.e., Data Correlation Analysis, to understand the relationship between data in a better way. This concept has given some promising results as well. Though it is still concise, the researchers did a considerable amount of research on correlated data privacy. Researchers have provided solutions using probabilistic models, behavioral analysis, sensitivity analysis, information theory models, statistical correlation analysis, exhaustive combination analysis, temporal privacy leakages, and weighted hierarchical graphs. Nevertheless, researchers are doing work upon the real-world datasets that are often large (technologically termed big data) and house a high amount of data correlation. Firstly, the data correlation in big data must be studied. Researchers are exploring different analysis techniques to find the best suitable. Then, they might suggest a measure to guarantee privacy for correlated big data. This survey paper presents a detailed survey of the methods proposed by different researchers to deal with the problem of correlated data privacy and correlated big data privacy and highlights the future scope in this area. The quantitative analysis of the reviewed articles suggests that data correlation is a significant threat to data privacy. This threat further gets magnified with big data. While considering and analyzing data correlation, then parameters such as Maximum queries executed, Mean average error values show better results when compared with other methods. Hence, there is a grave need to understand and propose solutions for correlated big data privacy.

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

Introduction

Data Privacy is the appropriate use of data available with any individual or organization, unlike data security that guarantees confidentiality, integrity, and data availability. Some of the well-known data privacy preservation methods are k-anonymization [1, 2], l-diversity [2, 3], t-closeness [4], and differential privacy [5]. Table 1 summarizes these algorithms along with their pros and cons. Among these, k-anonymization, l-diversity, t-closeness fall under the category of De-identification algorithms. k-anonymization [1, 2] works on the principle of making each record identical to at least k-1 records, over a set of attributes called quasi-identifiers. It protects against linkage attacks by making k records indistinguishable from each other. The larger the value of k, the higher the privacy, and consequently, the data utility is lower. Despite introducing a larger value of k, there may be cases where the sensitive data in the equivalence class do not exhibit diversity. In that case, the dataset becomes vulnerable to homogeneity attacks. In order to prevent it, l-diversity [2, 3] ensures that the equivalence class should contain l well-represented values for each sensitive attribute. t-closeness [4] is an improved version of l-diversity where it ensures that the distance between the distribution of a sensitive attribute in the equivalence class and the distribution of the attribute in the whole table is not more than a threshold value t. Those mentioned above could prevent distribution attacks on datasets to a large extent. Data analysts and researchers used another mechanism, i,e., differential privacy [5] to obtain helpful information from the database and also ensured that its privacy is maintained when publishing it. The basic principle behind it is to introduce distortion in the database before publishing. Nevertheless, the distortion should be large enough to provide high privacy, and at the same time, it should be small enough to ensure data utility after data publishing [6, 7]. Hence to calculate the optimum amount of distortion, researchers apply many metrics. The traditional DP used global sensitivity [5]. Later researchers have also suggested other methods for it, in order to provide privacy can along with data utility [8–10]. Table 1 gives a brief desciption of the above mentioned mechanisms. Due to the widespread use of DP, the performing of rigorous research works were as done on it. Along with some other drawbacks,

Table 1 Different existing privacy techniques

S. no	Privacy measure	Definition	Limitations
1	k-anonymity	This is a mechanism for testing algorithms so that published data restricts what can be disclosed about the properties of entities that are to be secured [1, 2]	Homogeneity-attack, background knowledge [2]
2	l-diversity	The equivalence class is also said to l-diversity if the sensitive attribute has at least "l-well-represented" values [2, 3]	Insufficient to prevent attribute disclosure [2]
3	t-closeness	The table is said to have t-closeness if the distribution of the entire attribute in the whole table is no more than a threshold t [4]	Complex distribution of a sensitive attribute [4]
4	Differential privacy	It is a mechanism to provide an interface between user and database, which protects individual data with the highest mathematical guarantee [5]	In the case of data diversity, it includes too much noise, which reduces the data utility [5]

researchers identified a vital drawback that could threaten data privacy. It was the existence of correlation in the datasets [9, 11, 12].

Initial researches around DP ignored its existence and regarded data as IID. However, later researches showed that real datasets often had a high correlation among them, and hence researches including data correlation became primarily significant [10, 13, 14]. Many researchers have studied the adverse effects of data correlation on data privacy. Different researchers in their published works have proposed various approaches to deal with the issue of privacy threats due to data correlation in datasets. For this, many researchers have used a well-known machine learning technique, termed Data Correlation [15]. Along with ensuring data privacy in data correlation, another big concern is maintaining data utility, for which various algorithms also have been proposed.

The above-stated problem of Data Correlation as a data privacy threat further gets strengthened when the data is of large volume and involves more complexities, and higher dimensionalities [16]. The technical term for such data is "Big Data." So our concern increases more with the presence of big data in most of the fields. Due to the exhaustive data generation, big data is almost present everywhere, and with general data approaching big data, the magnitude of data correlation among the datasets also increases proportionally. Hence, the need to work towards the privacy of correlated big data increases manifold. In its later section, this paper also describes the structure and organization of big data, which is fundamental to big data privacy and correlated big data privacy.

This work outlines all the related works where data correlation has threatened data privacy and big data privacy. This work will be helpful to all those who wish to explore the same area further as it provides a detailed study of the critical points, the proposed methodology, the conclusions, and the limitations of the related works.

Machine learning for data privacy

Machine learning holds multidimensional capacity and witnesses its applications in varied fields such as speech recognition, online fraud detection, image recognition, product recognition, etc. [17]. Sensor machine learning is a comparatively new application of machine learning that uses various machine learning algorithms to work with sensor data. It is of great industrial importance as well. Through the usage of accurate sensor data and machine learning algorithms, the failure of heavy machines can be predicted well before time to minimize loss [18]. Also, several research papers have suggested algorithms to facilitate machine learning algorithms for sensor data [19, 20]. Another application of machine learning that is paving its way is in big data privacy. A large amount of data correlation poses a threat to their privacy. The researchers can ensure data privacy guarantees by utilizing machine learning concepts. By using one such concept of ML is Data correlation, they can analyze the relationship among data and later prevent it from being a threat. [13] uses MIC, one of the existing data correlation analysis techniques for this purpose. Similarly, other ML algorithms can also be analyzed and applied. The works discussed in [21, 22] have explored various correlation analysis techniques.

Fundamentals of Big Data

"A new generation of technologies and architectures built to economically separate value from massive volumes of a wide variety of data by allowing high-velocity collection, discovery, and analysis," according to the definition of big data [23]. According to this concept, the 3Vs reflect the properties of big data, which are volume, velocity, and variety. Later research has shown that the 3Vs definition is inadequate to understand the current state of big data. In order to correctly understand the definition of big data, the additions of veracity, validity, value, variability, venue, vocabulary, and vagueness led to the making of some complement descriptions of big data. The fact that big data can include text, audio, image, video, and other types of data is a common theme. Variety represents the various qualities of data. The processing and analysis of data that is too huge or complicated for traditional database systems are handled by a big data architecture, as described in Fig. 1. Batch processing of gigantic data sources at rest, real-time processing of big data in motion, interactive exploration of big data, predictive analytics, and machine learning are everyday responsibilities in big data management.

Various frameworks have been established in recent years to ensure 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. The algorithm builds on existing privacy-preserving data techniques, resulting in a new model that incorporates a new Enhanced Secured Map Reduce (ESMR) layer [24] of privacy on the Big Data architecture map reduces phase. As the data passes through the map-reduce process, this new layer applies the protection algorithms to each specific piece of data. When data processing occurs via this latest proposed ESMR layer of Big Data, it can be safe and secured.

In today's digital world, where they store lots of information in big data, the analysis of the databases can provide opportunities to solve big problems of society like Healthcare, Sensors, Satellites, Share Market, Election, Crop prediction, and others. Another field that has emerged as a great application of big data is remote sensing. Remote sensing holds large scope in newer applications such as a build-up of strategies for reduction of resources consumption, timely disaster forecast, monitoring global changes, etc. [25]. Some conventional remote application areas are sensing image classification [26], Crop Classification [27], Land Cover and Land Use Classification [28], Satellite Image Classification [28] and many more. Research papers [25, 29–31] presents a detailed review

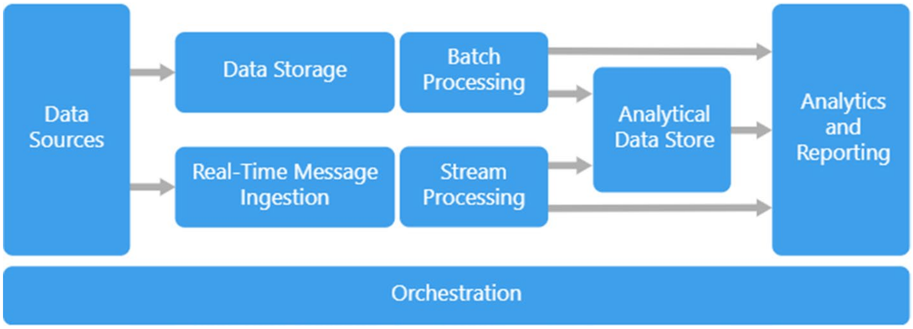

Fig. 1 Architecture of Big Data for batch and stream processing of data

on the need and use of Machine Learning in remote sensing. Authors of [25] also used examples of commercial players that are using remote sensing for Earth observation to advocate the need for machine learning in remote sensing. Remote sensing in itself holds the great capacity to be explored. It has a plethora of applications in varied fields, and the application of machine learning in the field of remote sensing can be a great boon.

Big Data is often challenged with many privacy issues [32] Its application in varied fields calls for stringent privacy measures [33]. One of the recent privacy threats is the presence of data correlation in big datasets. The researchers discussed it in detail in this paper.

Organisation of the paper

This paper presents a detailed review of research works from 2011 to 2019 that stated that existing algorithms are not enough to deal with the problem and proposed algorithms to resolve this issue. After the introduction of the paper, the following section describes, in brief, the basics of big data, which is fundamental to the topic of this paper. The following section is about the literature review, which in compact states the research papers and highlights the importance of each of them. The following section discusses the research papers that were the pioneers in understanding the problem of correlated data and classified it as a threat to data privacy. The upcoming section throws light and discusses in detail the research papers which have tried to provide some solution to the stated problem. The division of section is into several subsections as per the domain of the methodology proposed. The coming section deals with the experimental results of the experiments conducted to compare various proposed methods and analyze them based on few selected parameters. The last section states the conclusion and the open research areas and describes them in detail. Table 2 is a list of abbreviations used in this work.

Literature review

Initially, while people have dwelled upon data privacy, they inherently assumed no relations among data exist. Furthermore, all the proposed mechanisms for data privacy had an implicit assumption that they will be working upon independent data sets. Gradually,

Table 2 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

researchers and scholars around the globe started noticing that getting an independent data set has negligible chances in a real-life scenario. Most of the data sets upon which researchers and scholars were working had Data correlation as it was the realistic approach to understand the data and the working of various proposed mechanisms. Over the past few years, researchers have been working on analyzing the effects of Data Correlation over data privacy. Some of them had also proposed algorithms that would modify the existing privacy techniques for cinching privacy for correlated data.

Gehrke et al. [34] in the year 2011 and Kifer et al. [11] in 2011 gave initial arguments that considering the correlation between records is pivotal as the correlation between records or attributes can substantially decrease the privacy guarantee provided by any algorithm. These were the earliest works that realized the existence of data correlation within datasets and identified them as a threat to data privacy. They were the pioneers in this but severely lacked practical application. [35–38] deal with a special case of correlated data i.e. behavior analysis of the stock market. It is due to that the data of various investors and traders are often not mutually independent. The behavioral coupling between investors–traders and investors–investors is frequently noticed.

Cao et al. [35] in 2012 used an example of a stock market to study Coupled Behaviour Analysis. Cao et al. [35] were the first to do some research in Coupled Behaviour Analysis. This work also discussed Multivariate time series, sequence analysis, and Coupled Hidden Markov Model (CHMM) to study the abnormal behavior of databases such as a trading database. The proposed algorithms in this work were tested on a real dataset of the Asian Stock Exchange from 1st June 2004 to 31st December 2005. Longbing Cao [39] in 2013 presented another study about non-IIDness of data which analyzed the IIDness of several classical algorithms and showed how they might not work correctly in the case of non-IID data. To measure the non-IIDness, authors in this paper have used a measure termed Coupled Item Similarity (CIS). This work suggested a method to study the dependency among data but did not suggest any practical implementation to ensure data privacy.

Kifer et al. [40] in their successive work in the year 2014, proposed a privacy mechanism called Pufferfish Mechanism. The works of other researchers have used the proposed Pufferfish mechanism as a part of their work towards privacy. However, the proposed framework did not satisfy the Differential Privacy, and hence privacy preservation for Correlated Data still called for further investigation. Yang et al. [41] in 2015, proposed another mechanism in which they used Bayesian principles to help Differential Privacy cope up with correlated data. The mechanism proposed in this was very efficient, but it had the same disadvantages as a Bayesian Model. Wang et al. [6] in the year 2016, used Pufferfish Mechanism (PM) along with Wasserstein Distance to ensure privacy for correlated data with the help of two examples-Physical activity monitoring and Flu status. They proposed that the added noise must be proportional to Wasserstein Distance. The authors concluded that the proposed mechanism performs satisfactorily for a fixed range of values and fails for the remaining range. Chen et al. [7] in 2017, proposed a perturbation mechanism for Mobile Crowd Sensing (MCS) data using Bayesian Network that tries to add noise with the increased utility of published data. The proposed mechanism requires full knowledge of the probabilistic relationship among records which is not always practically feasible.

Chen et al. [8] in 2013, with the help of Social Network data, proposed multiplication of global sensitivity with the no. of correlated records to deal with the issue. The proposed method sets the major drawback of high data utility loss. Zhu et al. [9] in 2015, shown in this work that how the solution proposed in [8] induced an enormous magnitude of noise which significantly damaged the utility of data. They also proposed a different sensitivity analysis method called Correlated Sensitivity which reduced the amount of noise compared to global sensitivity. In the later sections of this paper, one can find the description of correlated sensitivity. Liu et al. [10] in the year 2016, proposed a perturbation mechanism for dependent data called Dependent Differential Privacy (DDP) to deal with the problem of correlated data. Liu et al. have performed an inference attack over a real-world dataset to demonstrate how an adversary can infer a user's sensitive information by using the noise added query outputs and exploiting the user's social relationships, thus violating the Differential Privacy guarantees. The proposed perturbation mechanism has added an extra parameter called the dependence coefficient to measure the dependence relationship between tuples minutely. It involves the estimation of ρ_{ij} value. The key challenge of the proposed approach is to accurately compute the value of ρ_{ij} as it relies on probabilistic models of statistical data.

Wu et al. [12] in the year 2017 proposed the use of foundations of Information Theory to study the relationship between Information Theory Principles and Differential Privacy. The authors concluded that concepts of Information Theory are well suited to model the problems of dependent data in Differential Privacy. This work did not study other privacy leakages, i.e., whether using concepts of Information theory induces any other privacy leakages or not. In order to practically suggest its application, one might study other privacy leakages.

Kumar et al. [42] in 2018 used Pearson's Product Moment correlation method to study the relationship between data. The datasets used were-Depressive disorder symptom dataset for evaluating depression severity, the Local weather dataset for classifying depression severity, and the Physiological sensor dataset for emotion detection. However, the authors of [42] must use a better correlation analysis technique as Pearson's Product Moment could not correctly define relationships among independent variables and other complex relationships.

Lv et al. [13] in 2019, studied data correlation among a correlated big data set to ensure its data privacy. They have used the MIC (Mutual information correlation) algorithm to analyze data correlation based on Information Theory and Mutual Information Theory. The authors also proposed two Correlated Differential Privacy models for Big Data Privacy Protection: (1) k-Correlated Record Differential Privacy and (2) r- Correlated Block Differential Privacy. For the experimental analysis, they have chosen the National Air Quality Data as time-series data are usually highly correlated. For examination purposes, the proposed mechanism is compared to mechanisms proposed in [8, 9] and [10] by setting different parameters to suitable values. Nevertheless, the proposed approach has a few shortcomings, which provide an open scope for further research in this field.

Zhao et al. [10] in 2019, considered and did all feasible combinations by combining tuple correlations and the query responses to provide a slight modification to differential privacy to solve the above-stated problem.

Cao et al. [43] in the same year took up temporal data to further understand the problem. The main objectives of this work were-Analysing temporal privacy leakage and quantify temporal privacy leakage. The researchers used the Concepts of Backward privacy leakage (BPL) and Forward privacy leakage (FPL) for this purpose. The future scope of this work lies in studying privacy leaks under temporal correlations combined with other types of correlation models.

Hemkumar et al. [45] in the year 2021 took up the same lines and studied temporal correlation. They proposed w-event privacy to deal with the problem. The proposed methodology lacks in the study of the correlation between other values.

Li et al. [14] later in the same year, took small examples to explain cases of positive correlation, negative correlation, and no correlation, along with how an adversary can use them with little or full prior knowledge. Mechanisms proposed in [46–48], and [41] based on the Bayesian Inference method can only be applied in case of positive correlation, unlike the proposed Prior Differential Privacy (PDP) mechanism. The authors have initially used the Weighted Hierarchical Graph technique as a solution, but later, they have pointed out its disadvantages and have applied Multi-Variate Gaussian Model. The paper further describes the observations and the results derived. Nevertheless, the proposed analysis was applied only to linear queries in the current work.

Further in the same year, in the work [49], the proposed perturbation mechanism is based on the minimization of Laplace noise using the Bayesian network model. Subsequently, the authors Chen et al. conducted simulations to evaluate the proposed algorithms and demonstrated their effectiveness. They also proved the influence of the maximum correlated group on the Bayesian differential privacy mechanism by using the Gaussian correlation model.

Figure 2 gives an illustration of the categories of the proposed solution. In this, they have categorized based on the technique used. Table 3 summarizes the proposed methodologies and their pros and cons in a tabular format for easy and better understanding.

Realization of correlated data as a threat

Gehrke et al. [34] 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. It was not a direct analysis of correlated data and its effects, but it was an indirect study of the effects of data correlation. This work also emphasized achieving a Zero-Knowledge Definition of Privacy along the lines of Dalenius. It states that an adversary must have Zero additional Knowledge by accessing the proposed mechanism. Authors have also formalized the notion of 'Aggregated Information,' which would be acceptable to release to attain the desired utility level without compromising privacy. The proposed Zero-Knowledge Privacy mechanism makes use of the Aggregated Information parameter. In the proposed work, mechanisms, adversaries, and simulators are simply randomized algorithms that play certain roles in our definitions. Let San be a mechanism that operates on databases in $D_1 andD$ be the class of all databases. For any database $D_1 \in D$, any adversary A, and any $z \in \{0,1\}^*$, let $OutA(A(z) \Leftrightarrow San(D1))$ denote the random variable representing the output of A on input z after interacting with the mechanism San operating on the database D_1. Note that San can be interactive or

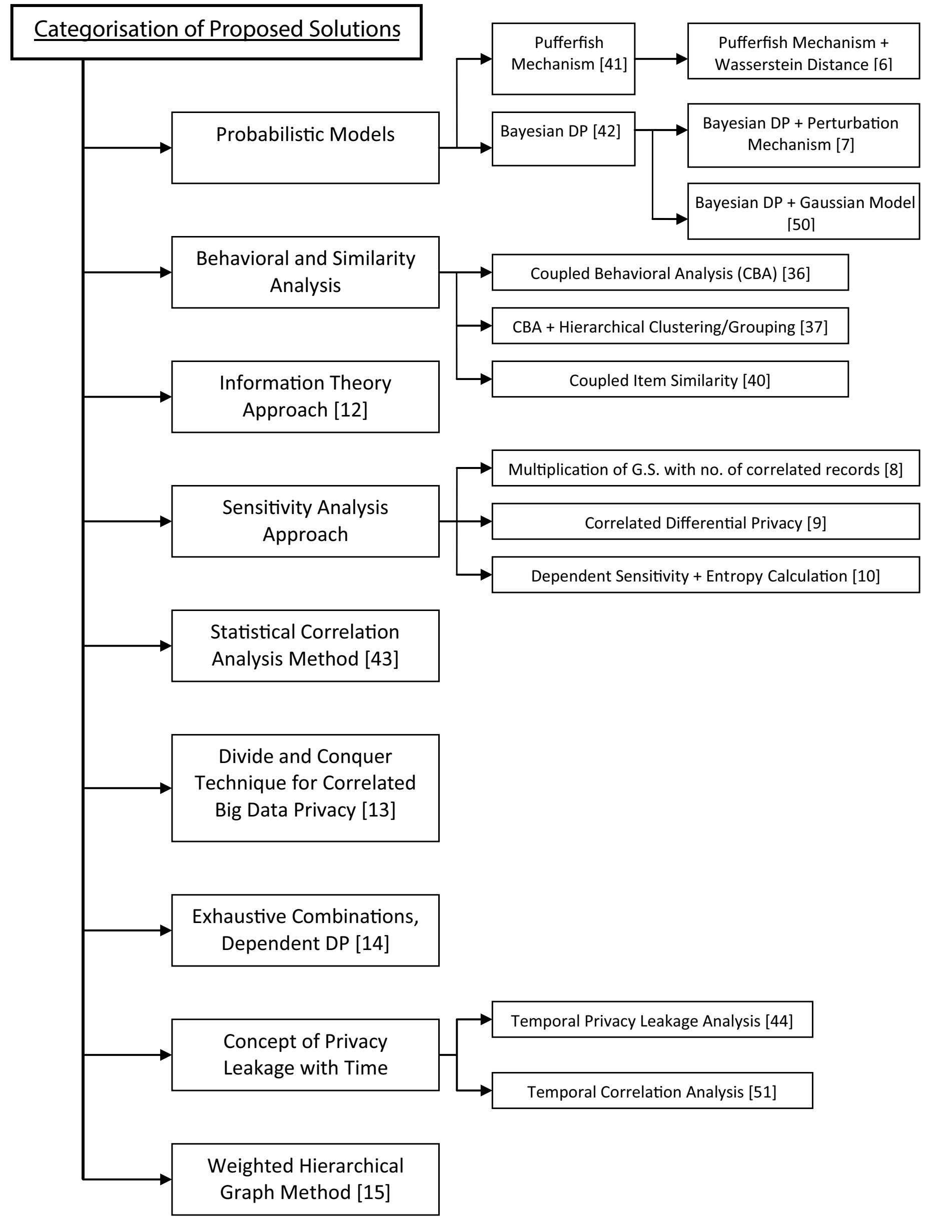
Categorisation of Proposed Solutions
Probabilistic Models
Pufferfish Mechanism [41]
Pufferfish Mechanism + Wasserstein Distance [6]
Bayesian DP [42]
Bayesian DP + Perturbation Mechanism [7]
Bayesian DP + Gaussian Model [50]
Behavioral and Similarity Analysis
Coupled Behavioral Analysis (CBA) [36]
CBA + Hierarchical Clustering/Grouping [37]
Coupled Item Similarity [40]
Information Theory Approach [12]
Sensitivity Analysis Approach
Multiplication of G.S. with no. of correlated records [8]
Correlated Differential Privacy [9]
Dependent Sensitivity + Entropy Calculation [10]
Statistical Correlation Analysis Method [43]
Divide and Conquer Technique for Correlated Big Data Privacy [13]
Exhaustive Combinations, Dependent DP [14]
Concept of Privacy Leakage with Time
Temporal Privacy Leakage Analysis [44]
Temporal Correlation Analysis [51]
Weighted Hierarchical Graph Method [15]

Fig. 2 Categorisation of proposed solutions

non-interactive. If San is non-interactive, then $San(D_1)$ sends information (e.g., a sanitized database) to A and then halts immediately; the adversary A then tries to breach the privacy of some individual in the database D_1. Let agg be any class of randomized algorithms that provide aggregate information to simulators.

Definition 1 Zero-Knowledge Privacy with aggregate- We say that San is ϵ -zero-knowledge private with respect to agg if there exists a $T \in agg$ such that for every adversary A, there exists a simulator S such that for every database $D_1 \in X_n$, every $z \in \{0, 1\}^*$, every integer $i \in [n]$, and every $W \in \{0, 1\}^*$, the following hold:

Table 3 Proposed solutions

S. no	Privacy measure	Definition	Limitations
1	Pufferfish Mechanism [40]	Formed strong foundation for other similar research	Could not satisfy differential privacy guarantee
2	Coupled Behaviour Analysis (CBA) [35]	Good experimenatl 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 [36]	Experimental comparison showed that CBA-HG outperformed the mechanisms of [35]	Applicability on other datasets with different couplings is uncertain
4	Coupled Item Similarity (CIS) [39]	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 [8]	Multiplication of global sensitivity with the no. of correlated records for correlated datasets	Data utility highly degraded attribute
6	Correlated Sensitivity [9]	Noise reduced by an enormous amount and greater data utility as compared to [8]	Few parameters held trade-off with utility
7	Bayesian Differential Privacy [41]	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 [10]	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 [44], it performed slightly worse for a particular range of values
10	Identity Differential Privacy [12]	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 [7]	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 [42]	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 [13]	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 [10]	Proposed DDP and proved mathematically how it can be derived from DP	Lacks practical implementation
15	They study Temporal Privacy Leakage [43]	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 [14]	Mechanism offers privacy guarantee in case of negative correlation as well	Not applicable to nonlinear queries
17	Temporal Correlation Mechanism [45]	Proposed w-event privacy using DP for location statistics and provided results regarding data utility	Correlation between other values was not studied

Table 3 (continued)

S. no	Privacy measure	Definition	Limitations
18	Bayesian DP with Gaussian Correlation Model [49]	Proposed Bayesian DP model which used Gaussian Correlation Model to study data correlation	Approximation of accurate probabilistic values is a challenge

$$Pr[OutA(A(z) \Leftrightarrow San(D_1)) \in W] \leq e.Pr[S(z, T(D_i, \perp), i, n) \in W] \tag{1}$$

$$Pr[S(z, T(D_i, \perp), i, n) \in W] \leq e.Pr[OutA(A(z) \Leftrightarrow San(D)) \in W] \tag{2}$$

The probabilities are over the random coins of San and A and T and S, respectively. Equations (1) and (2) means that whatever adversary can compute without accessing the mechanism, the same can be accessed by accessing the mechanism but with certain aggregate information. By properly setting the value of this parameter, the researchers might achieve its optimum value. Also, Gehrke et al. [34] have proved mathematically that Differential Privacy is a particular case of the proposed Zero-Knowledge Privacy mechanism. This work was among the initial research papers that tried to understand the dependencies of data though not directly but with the use of Social networks and their analysis. Some limitations of the proposed work are lack of application as the approach was purely theoretical, and there is no evidence of data utility preservation.

Kifer et al. [11] in 2011, gave initial arguments that considering the correlation between records is pivotal as the correlation between records or attributes can substantially decrease the privacy guarantee provided by any algorithm. This was the first attempt to formalize the term Data Correlation as a general phenomenon for real-time datasets. Later this work formed a significant foundation for further research regarding correlated datasets. This work identified the existence of data correlation and its adverse effects on data privacy but did not provide enough solutions to deal with it.

Various solutions proposed by other researchers

Proposed solutions using probabilistic models

Kifer et al. [40] in their successive work in the year 2014, proposed a privacy mechanism called Pufferfish, which helped in developing privacy definitions for different data sharing needs, study existing privacy definitions, study privacy compromise due to non-independent data records, and the number of other issues which are critical in terms of privacy. The proposed Pufferfish mechanism mainly depends on three foundation components. They are: (a) Set of Potential Secrets—What is to be protected? (b) The Set of Discriminative Pairs (Si, Sj)—The attacker should not be able to distinguish between Si and Sj, which is a desirable privacy guarantee. (c) The Evolution Scenario- This includes knowledge of how the data has evolved and the attackers' potential. The authors of [40] are discussing these three components in detail with the help of detailed examples in [40]. They also express the primary guarantees of the proposed Pufferfish mechanism in terms of Odds and Odds ratios. As per their definitions in this work, if E1 and E2 are two mutually exclusive events, then the fraction $\frac{P(E1)}{P(E2)}$ is their Prior Odds. If the ratio is equal to α (say), that means E1 is α times as likely as E2. If A is the piece of information available then,$\frac{P(E1|A)}{P(E2|A)}$ is the Posterior Odds. And the ratio $\frac{P(E1|A)}{P(E2|A)} \div \frac{(P(E1)}{(P(E2)}$ is the Odds Ratio and if $\frac{(P(E1|A))/(P(E2|A))}{(P(E1))/(P(E2))} \approx 1$ then one can say that A did not furnish any information that facilitated the differentiation between E1 and E2. Pufferfish provides this semantic guarantee. In this work, the concept of Hedging Privacy is used, which provides good privacy levels. It is strengthening privacy algorithms weaker than Differential Privacy. Specifically, it strengthens single prior privacy to protect such data publishers who have poor data models as prior belief. The set of potential secrets factors and set of discriminative pair factors remain the same. However, the evolution scenario factor changes by including conditional probabilities and other fixed-valued parameters containing all conditional probabilities. They use this for histogram publishing with privacy guarantees of Pufferfish in the proposed work. Also, in this work, Kifer et al. [40] mathematically proved that the Pufferfish framework provides the following with the help of various Lemma, Theorems, an: (a) Protection of Continuous attributes (b) Protection of Secrets that are aggregate properties of data (c) Protection of Distributional Privacy. This work also discussed other points related to privacy and Differential Privacy. As we have observed in the works of other researchers, they have used the proposed Pufferfish mechanism as a part of their work towards privacy. However, the proposed framework did not satisfy the Differential Privacy, and hence privacy preservation for Correlated Data still called for further investigation.

Yang et al. [41] in 2015, proposed another mechanism in which they used Bayesian principles to help Differential Privacy cope up with correlated data. The proposed Bayesian Differential Privacy provides privacy as per formal commitments of Differential Privacy. The above also provided privacy for correlated data and against an adversary with partial background knowledge. They did this by constructing a Gaussian Correlation Model to describe correlated data with complex correlations using the proposed algorithm, which efficiently prevented it from being intractable.

Definition 5 Bayesian Differential Privacy. Let A = A(i,K) be an adversary $(K \subseteq [n]/\{i\})$ and $M(x) = Pr(r \in S \mid x)$ be a randomized perturbation mechanism on database x.

The Bayesian differential privacy leakage of M w.r.t. A is

$$BDPLA(M) = sup_{xi,xi',xk,S} log \frac{(Pr(r \in S \mid x_i, x_k))}{(Pr(r \in S \mid x'_i, x_k))} \tag{3}$$

We say M satisfies ϵ-Bayesian differential privacy, or ϵ-BDP, if $sup_A BDPLA(M) \leq \epsilon$. Further, we could calculate the values of BDPLA(M) for discrete and continuous values of database x using ([41], Eqs. 6, and 7, respectively).

Wang et al. [6] in the year 2016, used Pufferfish Mechanism (PM) along with Wasserstein Distance to ensure privacy for correlated data because Pufferfish is capable of hiding personal values against correlation among multiple entries, unlike Differential Privacy and PM is capable of providing better utility in situations where large no. of entries are correlated [40]. For the demonstration, the authors have used two examples—Physical activity monitoring and Flu status. Using conventional definitions of global sensitivity and the Pufferfish privacy mechanism, they proposed that the added noise must be proportional to Wasserstein distance.

Definition 6 p-Wasserstein Distance. Let (X, d) be a Radon space, and u,v be two probability measures on X with finite p-th moment. Then the p-th Wasserstein distance between u and v is defined as:

$$Wp(u,v) = (inf \int_{\gamma \in \Gamma(u,v)} d(x,y)^p d\gamma(x,y))^{1/p} = (infE_{\gamma \in \Gamma(u,v)}[d(X,Y)^p])^{1/p} \tag{4}$$

where Γ(u,v) is the set of all couplings over u and v and each $\gamma \in \Gamma(u,v)$ can be regarded as a way to shift probability mass between u and v the cost of a shift is $E(X,Y) \approx [d(X,Y)^p]^{1/p}$. The cost of the min-cost shift is the Wasserstein distance. Initially, the researchers applied the proposed version of Pufferfish using Wasserstein distance to the examples mentioned above. They observed that the amount of noise added would be less than the amount of noise added if they use global sensitivity. Hence utility was enhanced. Then this result was generalized using mathematical proofs supported by a no. of theorems. Nevertheless, the proposed mechanism could have been computationally costly. Therefore, the authors have used a more restricted setting using a Bayesian network to describe the dependence between Markov Quilt Mechanism entries. It stated that if nodes X_i and X_j are far apart in a graph g, then X_j is largely independent of X_i and the amount of noise needed to obscure X_i will only be proportional to the local nodes around X_i. This work further explained how to select which nodes are local and add small effects concerning the distant nodes. Case studies of earlier mentioned examples initially infer the better utility of the proposed mechanism and later were supported by mathematical proofs. In order to show that the proposed Markov Quilt Mechanism also offered considerable privacy, the researchers used simulation results compared them to [44]. Ultimately, the authors concluded that the proposed Markov Quilt Mechanism performs significantly better than the method proposed in [44] for a certain range of parameter values and slightly worse for the remaining range.

Chen et al. [7] in 2017 initially showed how correlated data could be a threat to privacy in Mobile Crowd Sensing (MCS). They proposed a perturbation mechanism for the same using Bayesian Network that tries to add noise with the increased utility of published data. They presumed that the aggregate server has complete knowledge of the probabilistic relationship among records. They have applied the proposed perturbation mechanism to multiple available datasets like the ADULT dataset from the UCI Machine learning repository, NLTCS dataset from StatLib. Search logs generated by interpolating Google Trends data and American Online Search Logs and compared with [8] and [9] and concluded that the proposed mechanism provided much higher utility and also provided a decreased privacy budget. Another product of the proposed mechanism is the study of the influence of the size of the maximum correlated data groups over the noise generated for perturbation. It showed how they could exclude it to produce less noise when it has a low impact on statistical results. They assumed that the aggregate server has complete knowledge of the probabilistic relationship among records. However, this may not always be practically feasible. Nevertheless, to its relief, one may often not require to model the Bayesian Network over the entire dataset as a relationship between only correlated data is rudimentary to model. Fortunately, they are very few.

In [49], authors Chen et al. firstly investigated the influence of sensing data correlation on differential privacy protection. Due to the complex and diverse relationship among sensing data, different correlation models were explored by them, and accordingly, perturbation mechanisms are proposed from the attacker's point of view and the data owner's point of view. The proposed perturbation mechanism is based on the minimization of Laplace noise using the Bayesian network model. Subsequently, they conducted simulations to evaluate the proposed algorithms and demonstrated their effectiveness. They also proved the influence of the maximum correlated group on the Bayesian differential privacy mechanism by using the Gaussian correlation model.

Proposed solutions using behavioural and similarity analysis

Cao et al. [35] in 2012 used an example of a stock market to study Coupled Behaviour Analysis. The works of [36–38] also threw light on the same particular type of Correlated data, i.e., coupling between different users and their behaviors, by using the same example of the stock market. When the authors of [35] correlated activities of one or more actors with each other, then such activities are termed Coupled Behaviour(CB), and their analysis as Coupled Behaviour Analysis (CBA). This work studies CB and CBA and how they can be a challenge. Cao et al. [35] were the first to do some research in Coupled Behaviour Analysis. They described A behavior (IB) as a four-ingredient tuple $IB = (\epsilon, O, C, R)$ where actor ϵ is the one who does a behavior or behavior is imposed upon the actor. Operation O is what an actor does to accomplish the goal. Context C is the environment where the behavior occurs, and Relationship R is a tuple that defines complex interactions between multiple actors. [35] Also, in [35], they have defined a Behavior-Feature Matrix FM(IB) as follows:

$$FM(IB) = \begin{bmatrix} IB_{11} & IB_{12} & .... & IB_{1jmax} \\ IB_{21} & IB_{22} & .... & IB_{2jmax} \\ . & & & \\ . & & & \\ IB_{I1} & IB_{I2} & .... & IB_{Ijmax} \end{bmatrix}$$

Here it is assumed that there are I actors (customers) $\epsilon_1, \epsilon_2, ..., \epsilon_I$, an actor ϵ_i undertakes J_i behaviors $IB_{i1}, IB_{i2}, ..., IB_{iJ}$, actor $\epsilon_i' s j^{th}$ behaviour IB_{ij} is a K-variable vector, it's variable $[p_{ij}]_k$ reflects the k^{th} behavior property [35]. Others have earlier studied multivariate time series-based analysis, which is close to CBA, but technically they are different. This work also discussed Multivariate time series, sequence analysis, and Coupled Hidden Markov Model (CHMM) as they are related to the proposed work. CHMM is a collection of Multiple HMMs. The researchers have used this work to study the abnormal behavior of a database such as a trading database. Adapting the CHMM to changes was time-consuming and hence subsequently, the researchers developed an adaptive CHMM by adding an automatically adaptive mechanism to improve its efficiency. The key stages of CHMM/ACHMM based abnormal behavior are described in Fig. 3.

Experiments performed showed that using CHMM will outperform the use of a single HMM. Moreover, ACHMM outperformed CHMM. The researchers tested proposed

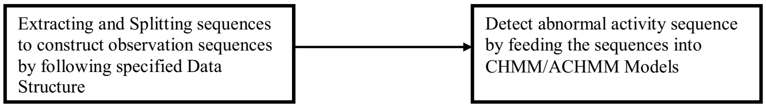
Extracting and Splitting sequences to construct observation sequences by following specified Data Structure
Detect abnormal activity sequence by feeding the sequences into CHMM/ACHMM Models

Fig. 3 Steps of ACHMM

algorithms in this work on a real dataset of the Asian Stock Exchange from 1st June 2004 to 31st December 2005. The metrics used for the evaluation of performances were

a. $Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
b. $Precision = \frac{TP}{(TP+FP)}$
c. $Recall = \frac{TP}{(TP+FN)}$
d. $Specificity = \frac{TN}{(TN+FP)}$

where TP stands for True Positive, which means that it belonged to the positive class. It is also classified as positive by algorithm; FP stands for False Positive, which means that it belonged to negative class but misclassified as positive by algorithm; TN stands for True Negative, which means that it belonged to negative class. It is also classified as negative by the algorithm. FN stands for False Negative, which means that it belonged to the positive class but was misclassified as negative by an algorithm. Tables 4 and 5 gives the accuracy values and recall values for CHMM and ACHMM, respectively, obtained from the experiments performed. The tables record values for winsize=20 minutes, and P-Num stands for the number of detected abnormal activity sequences. The values of other technical parameters for both the methods can be found in [35] with their detailed analysis.

This work gave an initial insight into CB and CBA, proposed CHMM and ACHMM, and showed their performance using experimental analysis of a real dataset. The major drawback was that when applied to datasets of high dimensionality, the proposed mechanisms could not furnish expected results. Hence, the proposed mechanisms need more

Table 4 Accuracy values for CHMM and ACHMM

P-Num	CHMM	ACHMM
20	0.88	0.90
30	0.85	0.88
40	0.83	0.84
50	0.80	0.82
60	0.77	0.79

Table 5 Recall Values for CHMM and ACHMM

P-Num	CHMM	ACHMM
20	0.32	0.41
30	0.36	0.45
40	0.45	0.50
50	0.45	0.50
60	0.50	0.55

exploration regarding their integration with more sophisticated domain knowledge. For future research, CBA has excellent potential, challenges, and opportunities open.

As stated above [35–38] deal with a special case of correlated data i.e. behavior analysis of the stock market. Due to the data of various investors and traders are often not mutually independent, the behavioral coupling between investors–traders and investors–investors are frequently noticed. The behavioral coupling between one investor and other investors is undesirable as they together try to manipulate and earn extra higher profit and influence the trend. Hence financial market regulators are interested in detecting such couplings. The two proposed frameworks in [36] consist of the major three stages, as described in Fig. 4.

In this work, the researchers proposed algorithms for all the above three stages using multiple variants of hierarchical grouping (HG) and hierarchical clustering (HC) techniques. They were compared with the algorithms proposed in [35]. The same metrics were used to evaluate the performance in both the works and algorithms of [36] provided better results for anomaly detection. Tables 6 and 7 gives the accuracy values and recall values for the three methods, respectively, obtained from the experiment formed. The tables record values for winsize=20 minutes, and P-Num stands for the number of detected abnormal activity sequences. The values of other technical parameters for both the methods can be found in [36] with their detailed analysis.

When they compared the two proposed methods to each other, then it was observed that the CBA-HG framework obtained better performance in terms of all the used

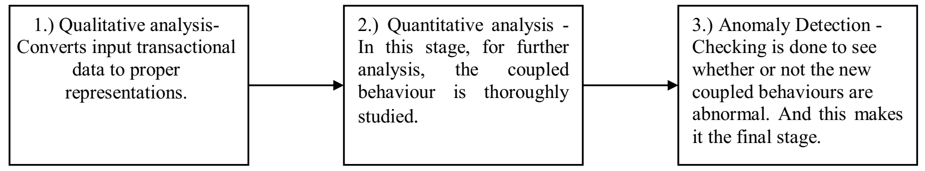
1.) Qualitative analysis- Converts input transactional data to proper representations.
2.) Quantitative analysis - In this stage, for further analysis, the coupled behaviour is thoroughly studied.
3.) Anomaly Detection - Checking is done to see whether or not the new coupled behaviours are abnormal. And this makes it the final stage.

Fig. 4 Steps of the proposed framework

Table 6 Accuracy values for ACHMM,HC,HG

P-Num	ACHMM	CBA-HC	CBA-HG
20	0.90	0.92	0.93
30	0.88	0.89	0.91
40	0.84	0.86	0.89
50	0.82	0.84	0.87
60	0.79	0.82	0.85

Table 7 Recall Values for ACHMM,HC,HG

P-Num	ACHMM	CBA-HC	CBA-HG
20	0.32	0.41	0.43
30	0.36	0.45	0.46
40	0.45	0.50	0.52
50	0.45	0.50	0.55
60	0.5	0.55	0.58

metrics. Nevertheless, none of the mechanisms performed satisfactorily when applied to datasets of high dimensionality. Hence, more research regarding their varied applications is needed.

Longbing Cao [39] in 2013 presented another study about non-IIDness of data which is usually present in the form of couplings and heterogeneity between objects, values, and attributes. This work initially analyzed the IIDness of several classical algorithms and showed how they might not work properly in the case of non-IID data. To measure the non-IIDness, authors in this paper have used a measure termed Coupled Item Similarity (CIS). CIS between categorical items X and Y are defined as follows:

$$CIS = \Sigma_{(j=1)}^{n} \delta A_j(Xj, Yj) \tag{5}$$

where X_j and Y_j are the values of item feature j for X and Y, respectively and δA_j is coupled attribute value similarity (CAVS). δ_j^A is dependent upon the intra-coupled attribute value similarity (IaAVS) δ_j^{Ia} and inter-coupled attribute value similarity (IeAVS) δ_j^{Ie} as proposed in this work.

$$\delta A_j(X_j, Y_j) = \delta_j^{Ia}(X_j, Y_j) + \delta_j^{Ie}(X_j, Y_j) \tag{6}$$

The values of δ_j^{Ia} and δ_j^{Ie} can be calculated using [1], Eqs. (4.2), (4.7) respectively. Also, in this, Longbing Cao has shown in detail that how exploring the non-IIDness allows us to comprehensively, systematically, and deeply explore couplings, correlation, heterogeneity between values, attributes, and objects. This subsequently results in more robust algorithms such as learning algorithms, pattern matching algorithms, and many more. This also profoundly explains the need to consider data as non-independent with classic examples, but it does not state its relation with its privacy. It does not device any mechanism to improvise data privacy in the case of non-independent data.

Proposed solutions using the sensitivity analysis approach

Chen et al. [8] in 2013, analyzed how correlated data can lead to unexpected privacy loss and proposed multiplication of global sensitivity with the no. of correlated records to deal with the issue. For this, they analyzed Social network data as there was a high correlation among data records. The proposed method successfully achieved the required privacy in the correlated dataset. However, as per the privacy and data utility trade-off, it set the major drawback of high data utility loss. Hence, the authors of [8] were yet to achieve an optimum balance of data privacy and data utility in the case of correlated data. As used in this work,

Definition 2 Global sensitivity. For any function $f : D \rightarrow R^d$, the global sensitivity of f is

$$GS(f) = max_{\{D1,D2\}} \parallel f(D1) - f(D2) \parallel_1 \tag{7}$$

for all D1, D2 s. t. $\mid D1 \bigtriangleup D2 \mid = 1$ i.e. Two databases, D1 and D2, are neighbors if they differ on at most one record, denoted by$\mid D1 \bigtriangleup D2 \mid = 1$.

$$ProposedSensitivitycalculation = no.ofcorrelatedrecordsGlobalsensitivity = k * GS(f) \quad (8)$$

Then in the year 2014, the Pufferfish mechanism was proposed in [40] to give a solution to the problem of privacy loss due to correlated data, and it has already been discussed earlier in this paper.

In the consecutive years, there were also many proposed mechanisms to address the same threat. Zhu et al. [9] in 2015, intricately described problems that arose due to correlated datasets and how they compromised the privacy guarantee provided by conventional Differential Privacy, and they coined the problem as Correlated Differential Privacy. Traditional Differential Privacy never considered the relationship among records, and it greatly and adversely influenced the privacy guarantees. In this work, the researchers also explained the drawbacks of the naive solution of multiplying the global sensitivity with the no. of correlated records to ensure privacy level as proposed in [8] in detail. They clearly showed in this work that how the solution proposed in [8] induced an immense magnitude of noise which significantly damaged the utility of data. In this, Zhu et al. have categorized correlated record analysis methods. Also, in the same work, they have proposed a different sensitivity analysis method called Correlated Sensitivity which reduced the amount of noise compared to global sensitivity.

Definition 3 Correlated Sensitivity: For a query Q, the correlated sensitivity is determined by the maximal record sensitivity,

$$CSq = max_{i \in q}(CS_i) \quad (9)$$

where q is a record set of all records responding to a query Q.

Definition 4 Record Sensitivity: For a given δ and a query Q, the record sensitivity of r_i is

$$CS_i = \Sigma_{j=0}^{n} \mid \delta_{ij} \mid (\parallel Q(D_1) - Q(D_2) \parallel_1) \quad (10)$$

where $\delta_{ij} \in \triangle$. The record sensitivity measures the effect on all records in D when deleting a record r_i. δ is the correlated degree analysis matrix that holds the degree of correlation between all sets of records. Along with it, the authors have also proposed a mechanism that satisfies Differential Privacy in the case of correlated datasets and also preserves utility for further application. Nevertheless, during the practical implementation of the proposed mechanism, it was observed that it depends on a no. of parameters. One such parameter held a trade-off with the accuracy of the proposed mechanism, and hence its application could not be suggested.

Liu et al. [10] in the year 2016, initially stated how Differential Privacy could not give a privacy guarantee in case of correlated data and later in their work have proposed a perturbation mechanism for dependent data called the Dependent Differential Privacy (DDP) to reduce such threats. To show vulnerable threats to correlated data, Liu et al. have performed inference attack over a real-world dataset. This facilitated demonstrating how an adversary can infer sensitive information such as user location information by using the noise added query outputs and exploiting user's social relationships, thus violating the Differential Privacy guarantees. The authors [10] measured the performance of the Inference attacks by using the following two metrics:

$$1.) InferenceError = \frac{1}{n}\Sigma_{i=1}^{n} Dist(d_i, D_i) \tag{11}$$

$$2.) LeakedInformation = \frac{1}{n}\Sigma_{i=1}^{n} H(Di) - H(Di \parallel \mu D_{-i}) \tag{12}$$

Dist(.) is defined in [[10], equation 5] and H(.) denotes the entropy information of a random variable [50]. H(Di) measures the entropy of prior probability and measures the adversary's prior information for Di without considering the social relationships.$(H(Di) - H(Di \parallel \mu D_{-i})$ is the entropy of posterior probability and measures the adversary's posterior information after the inference attack. By evaluating the Leaked Information, one can measure the privacy breaches in terms of change in the adversary's before posterior beliefs. The above metrics applied over real-world datasets clearly showed that conventional Differential Privacy applied to a dependent dataset would disclose more sensitive information than expected and hence is a serious privacy violation issue. The proposed perturbation mechanism called DDP has added an extra parameter called the dependence coefficient to measure the dependence relationship between tuples minutely. The sensitivity under dependence relationship between two tuples D_i and D_j is proposed as:

$$Sensitivityunderdependencerelationship = \Delta D_i + \rho_{ij} \Delta D_j \tag{13}$$

where $\rho_{ij} \in [0, 1]$ and is a metric to measure dependence between two tuples D_i and D_j. Rest they theoretically and experimentally proved the utility and privacy superiority of the proposed DDP mechanism in the current work. Furthermore, they have stated in this work that the DPP mechanism is more accurate in computing k-means clustering centroids, SVM classifiers, publishing degree distance of large-scale social networks. They have also proved in this work that the proposed mechanism is also resilient to inference attacks in the case of dependent tuples. Even conventional Differential Privacy fails to provide privacy in such a case. The key challenge of the proposed approach is to accurately compute the value of ρ_{ij} as it relies on probabilistic models of statistical data. Hence the probabilistic models must be known beforehand to calculate ρ_{ij}. The overestimated value of ρ_{ij} still guarantees the privacy commitments of the proposed DDP, but its underestimation leads to degradation of the privacy commitments.

Proposed solution using information theory models

Wu et al. [12] in the year 2017 proposed Identity Differential Privacy to capture the weakness of conventional Differential Privacy when dealing with dependent data by using the foundations of Information Theory in their published work. This work also studied the connection between Information theory and Differential Privacy. With the help of mathematical proofs, they showed how concepts of information theory could well explain ϵ-Differential Privacy and how using the proposed Identity Differential Privacy one can show that mutual information between dependent records will stay less than or equal to ϵ. As given in this work,

Lemma 1 *If the mechanism M satisfies ϵ-identity differential privacy, then for any source X_i, any $r \in R$, and any record $t \in \chi_i$, there are*

$$\frac{(Pr[X_i = t])}{(Pr[X_i = t \mid Y = r])} \leq e^{\epsilon} \quad and \quad \frac{(Pr[X_i = t \mid Y = r])}{(Pr[Xi = t])} \leq e^{\epsilon} \tag{14}$$

We can find the mathematical proof of Lemma 1 in detail in [12]. In information theory, the relative entropy is used to measure the distance between two probability distributions [51]. The relative entropy of X_i and $(X_i|Y = r)$, denoted as $D(X_i||(X_i|Y = r))$, has the following result

Theorem 1 Let the mechanism M satisfies ϵ-identity differential privacy and let Y be the output random variable of the mechanism. We have

$$D(Xi||(Xi|Y = r)) \leq \epsilon \tag{15}$$

Using Lemma 1, we have

$$\begin{aligned} &D(Xi||(Xi|Y = r)) \\ &\quad = \Sigma Pr[Xi = t] log \frac{(Pr[Xi = t])}{(Pr[Xi = t|Y = r])} \leq \Sigma Pr[Xi = t] loge^{\epsilon} = \epsilon \end{aligned} \tag{16}$$

Hence, the proof is complete.

Theorem 2 Let the mechanism M satisfies ϵ-identity differential privacy and let Y be the output random variable of the mechanism. We have

$$I(Xi; Y) \leq \epsilon \tag{17}$$

Finally, in the current work, it was concluded that concepts of Information Theory are well suited to model the problems of dependent data in Differential Privacy. This work did not study about other privacy leakages i.e., whether using concepts of Information theory induces any other privacy leakages or not. Without studying about other privacy leakages its application may not be practically suggested.

Proposed solution using statistical correlation analysis method

Kumar et al. [42] in 2018 have tried to do correlation analysis over some selected datasets to enhance the accuracies when some classification problem arises. The datasets used are-Depressive disorder symptom dataset for evaluating depression severity, the Local weather dataset for classifying depression severity, and the Physiological sensor dataset for emotion detection. To study the correlation analysis, the authors have used Pearson's Product Moment correlation method. Consider two variables, A and B. Then the Pearson's correlation coefficient can be calculated using the following formula:

$$C_{A,B} = \frac{Covariance(A,B)}{\sigma_A \sigma_B} \tag{18}$$

where $C_{A,B}$ is the correlation coefficient, Covariance(A,B) is the covariance, and $\sigma_A \sigma_B$ are the standard deviations of A and B, respectively. In the case of a dataset involving two sets, {a1, a2,....,an} and {b1, b2,..., bn}, the correlation coefficient can be calculated as:

$$C = \frac{\Sigma_{i=1}^{n}(a_i - a')(b_i - b')}{\sqrt{\Sigma_{i=1}^{n}(a_i - a')^2}\sqrt{\Sigma_{i=1}^{n}(b_i - b')^2}} \tag{19}$$

where n is the sample size, ai and bi are the ith data values, and a', b' are the mean values. The value of the coefficient C ranges between −1 and + 1. Values close to + 1 show a strong positive correlation, those close to − 1 show a strong negative correlation, and those closest to 0 show no relation. Then for feature selection, the stepwise selection algorithm backward elimination was used and performed by Weka Tool using the Merit parameter in this work. The classification algorithms used for classification purposes are—Random forest algorithm, Multinomial Logistic Regression, Logit Boost, and SVMs. And then accuracy is calculated by taking different numbers of features in each iteration using the Accuracy formula as in [35]. And then, an optimum number of features was selected as per the assigned ranks. The researchers must use a better correlation analysis technique as Pearson's Product Moment could not properly define relationships among independent variables and other complex relationships. Also, more study is required to find an appropriate feature selection method.

Proposed solution using divide and conquer technique

Lv, et al. [13] in 2019, aimed to achieve a Differential Privacy mechanism for correlated big data. Their initial step was to find the correlation between data. Moreover, for the same, they have used MIC (Mutual information correlation) algorithm, based on Information Theory and Mutual Information Theory. The main reason behind selecting MIC Algorithm is, it can measure both linear and nonlinear correlation between variables [21]. The authors then proposed two Correlated Differential Privacy models for Big Data Privacy Protection (1) k-Correlated Record Differential Privacy and (2) r-Correlated Block Differential Privacy.

Definition 7 Adjacent Data Set. Let D_1 be a big data set with n as total no. of records, including l correlated records ($l \ll n$). Suppose when a record is modified, k-1 records are changed to get dataset D2 then D1 and D2 are adjacent data sets represented as $\mid D_1 \triangle D_2 \mid = k$ where $1 \leq k \leq l$.

Definition 8 k-Correlated Record Differential Privacy (k-CRDP). Let D_1 and D_2 be two adjacent data sets, i.e. $|D_1 \triangle D_2| = k$, A be the privacy mechanism and f be the query function. For any output $S \in R$, k-CRDP is as follows-

$$CRDP(A) = \left| ln\frac{Pr[A(f(D_1)) \in S}{Pr[A(f(D_2)) \in S} \right| \leq \epsilon \tag{20}$$

Definition 9 r-block division. Let D be a big data set if there exists a set B $= \{D_1, D_2, ..., D_r\}$ and $D_1, D_2, ..., D_r$ are independent of each other such that $D_1 UD_2 U.........UD_r = D$, then B is called a r-block division of big data set D. In order to calculate sensitivity, the researchers used the new technique in this work, i.e., the Machine learning approach to establish dependency of correlated data records. Clearly, the algorithms used the divide and conquered approach. For the experimental analysis, they have chosen the National Air Quality Data as time-series data are usually highly correlated. For examination purposes, the proposed mechanism is compared to mechanisms proposed in [8–10] by setting different parameters to suitable values wherein they MAE as a performance evaluation function in this. An obstacle in this was that Big data is often accompanied by enormous computing overhead, but the authors proposed the r-Correlated Block Differential Privacy (r-CBDP) protection, model. This approach gave a solution to a large extent to the problem and improve efficiency through data partitioning and parallel computing. Concurrently, the proposed MIC k-means algorithm improved efficiency in large distance calculation time for the k-means clustering algorithm (Fig. 5).

The study of the influence of the block parameter r on privacy protection performance revealed that with a decrease in value of r, the amount of data correlation between sub-datasets increased, and privacy protection performance decreased. Hence a large value of r would result in high privacy protection performance. However, increasing the value of r will also result in increased overhead time, and thus, parameter r must be adjusted to balance time overhead and privacy protection performance. In this work, the researchers calculated the time complexity of the proposed mechanism to be $O(n^2)$ where n is the size of the data block.

The simulation results showed that: (a) Privacy protection performance of the proposed mechanism is better than the other three mechanisms proposed in [8–10]. (b) As per the curves drawn using experimental results, the influence of different privacy budget parameters ϵ is not stable for mechanisms provided in [8, 9]. However, its influence on the proposed method in the current work is relatively stable, and also, the mechanism of [10] is close to the mechanism proposed in this. Therefore, the performance advantage of the mechanism proposed is more than the mechanisms of [8, 9] and [10].

Some of the shortcomings of this work upon which one may further work are: (a) Examination of external correlation of datasets. Only internal correlations have been considered. (b)The Traditional MIC algorithm used in this cannot handle high dimensional data, and hence improved algorithms may be used for future works. (c) k value selection is complex in the k-means algorithm. Other clustering algorithms may be explored and used for improved efficiency.

Proposed solution using exhaustive combinations

Zhao et al. [10] in 2019 proposed a slight modification to Differential Privacy to give a privacy guarantee for correlated data. To do so, they had considered all viable combinations by combining tuple correlations and the query responses. After that, the authors of [10] concluded by stating that the modification done to Differential Privacy, termed as Dependent Differential Privacy (DDP), facilitated privacy guarantee to any data

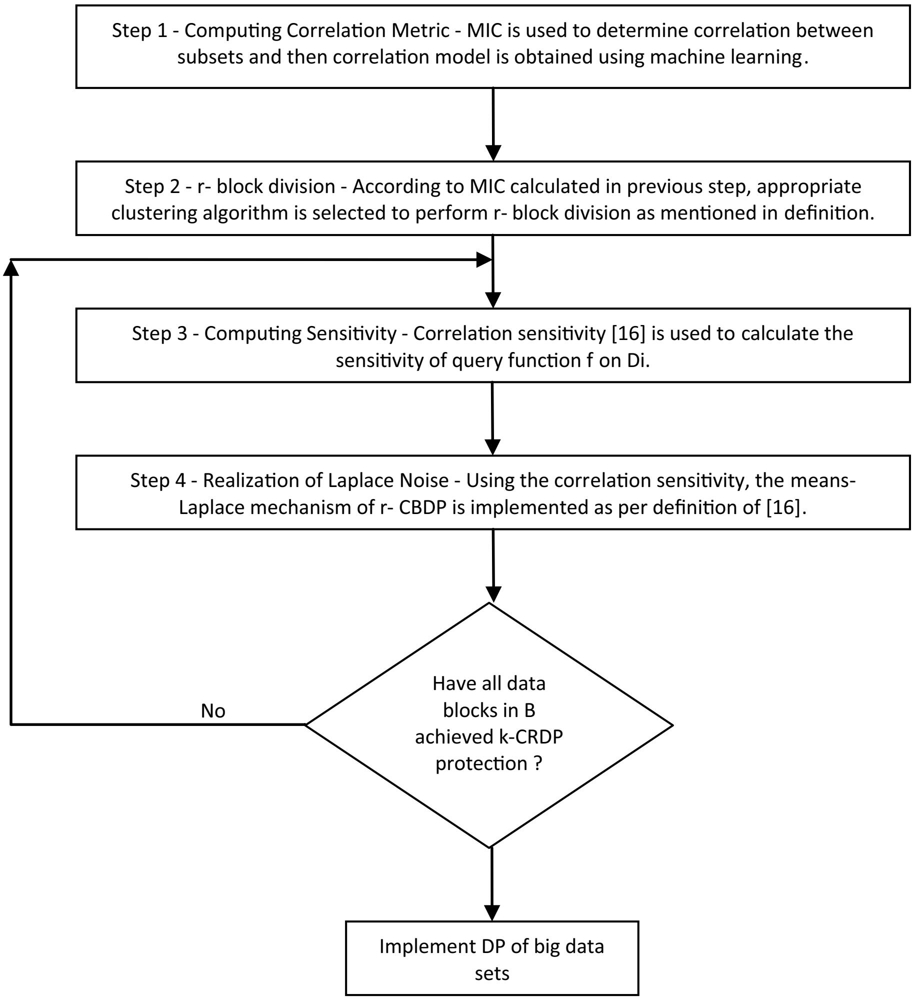
Step 1 - Computing Correlation Metric - MIC is used to determine correlation between subsets and then correlation model is obtained using machine learning.
Step 2 - r- block division - According to MIC calculated in previous step, appropriate clustering algorithm is selected to perform r- block division as mentioned in definition.
Step 3 - Computing Sensitivity - Correlation sensitivity [16] is used to calculate the sensitivity of query function f on Di.
Step 4 - Realization of Laplace Noise - Using the correlation sensitivity, the means-Laplace mechanism of r- CBDP is implemented as per definition of [16].
Have all data blocks in B achieved k-CRDP protection ?
No
Implement DP of big data sets

Fig. 5 Flowchart of r-CBDP [13]

correlation model. Moreover, the adversary's knowledge and authors also showed how quantitatively DDP could be deduced from DP using a more robust privacy parameter.

Proposed solution using the concept of privacy leakage with time

Cao et al. [43] midway in the year 2019, tried to address the issue of privacy loss of temporal data release in their work. If data is recorded concerning time, then such a correlated dataset is said to be a temporally correlated dataset, or simply a temporal dataset [43]. The main objectives of this work were—Analysing temporal privacy leakage and quantify temporal privacy leakage. To analyze Temporal Correlations, they have used the concept of the Markov Chain. The parameter of the Markov Chain is a transition matrix. They have accordingly used transition matrices PBi and PFi, which are called backward temporal correlation and forward temporal correlation, respectively, in this work. In this, the researchers also considered it rational that the adversary may know the matrices.

Backward Privacy Leakage, BPL—The privacy leakage of M_t caused by $r_1, ..., r_t$ w.r.t. A_i^T is called backward privacy leakage, defined as follows:

$$BPL(A_i^T, M_t) = sup_{l_i^t, l_i^{t1}, r^1, ..., r^t} log \frac{Pr(r^1, ..., r^t \mid l_i^t, D_k^t)}{Pr(r^1, ..., r^t \mid l_i^{t'}, D_k^t)} \tag{21}$$

$$BPL(M_t) = max_{\forall A_i^T, i \in U} \; BPL(A_i^T, M_t) \tag{22}$$

Forward Privacy Leakage, FPL-The privacy leakage of M_t caused by $r^t, ..., r^T$ w.r.t. A_i^T is called forward privacy leakage, defined by follows:

$$FPL(A_i^T, M_t) = sup_{l_i^t, l_i^{t1}, r^1, ..., r^t} log \frac{Pr(r^1, ..., r^t \mid l_i^t, D_k^t)}{Pr(r^1, ..., r^t \mid l_i^{t'}, D_k^t)} \tag{23}$$

$$FPL(M_t) = max_{\forall A_i^T, i \in U} \; FPL(A_i^T, M_t) \tag{24}$$

In this work, it was already shown how TPL can be calculated using the formula-

$$TPL = BPL + FPL - PL_0(A_i^T, M_t) \tag{25}$$

Substituting the values of BPL and FPL as calculated in (16) and (18) into equation (20):

$$TPL(A_i^T, M_t) = BPL(A_i^T, M_t) + FPL(A_i^T, M_t) - PL_0(A_i^T, M_t) \tag{26}$$

Since the worst-case privacy leakage must be considered among all users in the database, therefore by substituting (18) and (20) in (22):

$$TPL(M_t) = BPL(M_t) + FPL(M_t) - PL_0(A_i^T, M_t) \tag{27}$$

After quantifying TPL, α -DPT i.e. α -DP under temporal correlations was proposed. Using Eq. (23):

$$\alpha = (\alpha_t^B + \alpha_t^F - \epsilon^t)$$

where α_t^B denotes the Backward Privacy leakage, α_t^F denotes the Forward Privacy leakage and ϵ^t is the privacy leakage due to conventional DP. Therefore,

$$\alpha - DPT = (\alpha_t^B + \alpha_t^F - \epsilon^t) - DP \tag{28}$$

The current work has consequently evinced the sequential composition theorem of α-DPT. It showed how temporal correlations only affect event-level privacy (i.e., privacy at a particular time t) and do not affect user-level privacy (i.e., the privacy of each individual during the whole timeline). Then further, to calculate the mathematical values of BPL and FPL to find TPL mathematically, Cao et al. [43] have formalized the objective function and constraints. The [43] authors also showed that they could find their optimum values using the Simplex Optimization Algorithm as the objective function is a form of Linear- Fractional Programming. The overall time complexity using Simplex Algorithm is $O(n^2 2^n)$ as calculated in this work. Also, the discussed [43] authors several different methods to calculate the value of BPL, FPL, and TPL, such as Polynomial Algorithm, Quasi-Quadratic Algorithm, Problem Formulation, Sub-Linear Algorithm. The relationship between α and the run time of these algorithms was also studied. They have also studied the relationship between data privacy and data utility by applying the proposed mechanism for different degrees of correlations and temporal data. The future scope of work lies in studying privacy leaks under temporal correlations combined with other types of correlation models.

In 2021, authors Hemkumar et al., in their work [45] have evaluated the correlation between timestamps and location statistics and their ability to pose a threat to the data privacy guarantees. They have termed this type of correlation as temporal correlation. They studied the difference between the current timestamp and the previous timestamp and used this value to introduce distortion. As privacy algorithm, they have used differential privacy algorithm. Conventional differential privacy assumes that the location statistics are independent, but this is different from the nature of real-world datasets. Thus, the use of differential privacy for such datasets causes more privacy leakages than expected. The solution proposed in [45] adopts w-event privacy for continuously releasing location timestamp data. The proposed methodology introduces distortion in the privacy budget depending on the correlation (similarity and dissimilarity) between current and previous timestamps. The paper also evaluated the data utility values for real and synthetic data. This work lacks in the study of the correlation between other types of attributes. It is limited only to temporal correlation study, whereas the correlation between other values is also an important factor for data privacy guarantees.

Proposed solution using weighted hierarchical graph technique

Li et al. [14] in the year 2019, initially took small examples to demonstrate the effect of prior knowledge of an adversary, correlations between data, and sensitivity analysis. Then using the same examples, they have explained cases of positive correlation, negative correlation, and no correlation, along with how an adversary can use them with little or full prior knowledge. Mechanisms proposed in [46–48], and [41] based on the Bayesian Inference method can only be applied in case of positive correlation. The proposed Prior Differential Privacy (PDP) mechanism considers a set of distributions, unlike Bayesian Differential Privacy (BDP) mechanism, which considers a single distribution. As per the proposed mechanism, the privacy leakage caused by the adversary with the help of data correlations must be less than or equal to ϵ, which is the maximal privacy leakage caused by adversaries with public distributions. Then, to demonstrate various effects, they have used a weighted hierarchical graph (WHG) and calculated their node and edge values using the formulae specified in the current work.

Few key points became evident in finding the edge values and node values of WHG: (a) The weakest adversary can cause privacy leakage when there is a positive correlation. (b) When WHG has both positive and negative edges, the whole WHG will have to be traversed to compute the correct privacy leakage. To traverse the full WHG, they have initially used a complete space searching algorithm with time complexity $O(n^4 2^{n-1})$ where n is the number of tuples in a database. In order to reduce this computational time complexity, a fast searching algorithm on a subspace of the whole original space the researchers proposed with little sacrifice of accuracy and

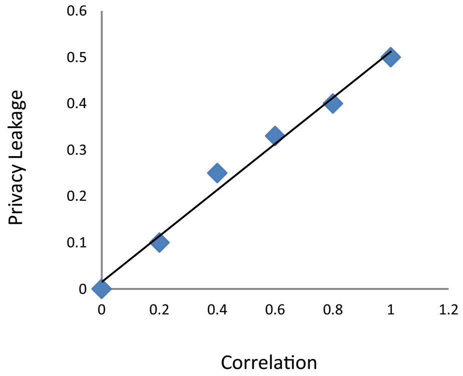

Fig. 6 Increasing privacy leakage with increasing correlation

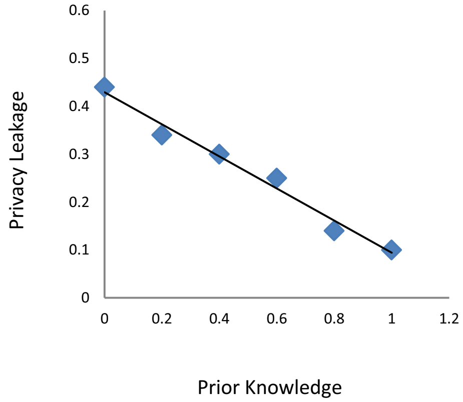

Fig. 7 Decreasing privacy leakage with increasing Prior Knowledge

time complexity of $O(n^4)$. To analyze the impact of the factors mentioned above on continuous-valued data, Li et al. [19] described why the WHG method would not be appropriate and instead have used the Multi-Variate Gaussian Model. The results are consistent with the results of discrete-valued data analysis.

Finally, with the help of numerical simulation, the following observations were noticed: (1) (a) No effect on privacy leakage when there is strongest prior knowledge. (b) In case of less strong prior knowledge as correlation increases, privacy leakage generally increases (for the positive and negative correlation). (2) Privacy leakage decreases with an increase in prior knowledge for positive correlation in discrete-valued and positive and negative correlation in continuous-valued data. (3) The efficiency of the fast searching algorithm is much more than the full space search algorithm. Also, the accuracy of the fast searching algorithm is quite close to the full space searching algorithm; hence in practical applications, the fast searching algorithm is better suited. Figures 6 and 7 shows observation 1(b) and 2 of the current proposed work, respectively to describe them. The proposed analysis was applied only to linear queries in the current work. However, it should be applied to nonlinear queries as well to evaluate its performance better.

Study of correlation constraints

Hyma et al. [52] by the end of 2015, studied considering the needed various correlation constraints while ensuring privacy to a correlated dataset. They are:

a. Simple Correlation Constraints: The entire database is analyzed to identify correlations among data. It can be the correlation either between records or between attributes.
b. Value-based Correlation Constraints: Based on the value of data, they scrutinized the privacy level. The value decides the impact of the correlation on the privacy level and mechanism.
c. Attribute-based Correlation Constraints: In this, they examined the associations between multiple attributes. If there is any association, then the chances are high that knowing the values of all attributes of one record will help the adversary correctly guess the value of attributes of another record if some attribute values are known, and some attribute values are unknown. Hence this holds a noticeable impact on the privacy level.
d. Event-based Correlation Constraints: Events in the case of database mining often refer to query occurrence. The correlation between successive events or queries holds great potential to challenge the privacy level decided for a non-correlated system.
e. Personalized v/s Universal Correlation Constraints: Personalised privacy is the privacy level achieved as per the record owner, and universal privacy refers to the privacy applied by the publisher after the owner to the publisher hands it over. A proper balance between the two is required to achieve the optimistic privacy of data.

These constraints were necessary to study before understanding correlation and its effect on data privacy. In this work, the researchers firmly concluded that the privacy mechanisms applied to correlated data and non-correlated data are very different. Moreover, a precise correlation constraint mechanism is required to achieve the required level of privacy. However, this work has given theoretical study of the above-stated facts over real-time examples, and also, it lacks practical implementations of the same.

Experiment and analysis

Experimental setup

In order to analyze the correlation between the data, the researchers used Time-series data. We chose “Air Quality Data” daily from various stations across several Indian cities. The data file contains nearly 30k records with a total of 13 columns, including city, date, air quality parameters AQI and AQI bucket classifying the air quality to an appropriate level. We chose to consider the data from 2015 to 2020, excluding the missing data. After removing the missing records, we chose to keep the city, date AQI and AQI bucket and few other air quality parameters except PM10, NH3, and Xylene. After pre-processing, the experimental dataset used had nearly 17k records with a record length of 17. Since the experimental dataset contains city records, there is a potential correlation between the dataset. We used the MIC method to calculate the correlation existing between the records. They compiled experiments and results and implemented using the TensorFlow

Table 8 Statistical features of the dataset

Features	PM2.5	NO	NO2	NOx	CO	SO2	O3	Benzene	Toluene	AQI
Mean	67.45	17.57	28.56	32.31	2.25	14.53	34.49	3.28	8.70	166.46
Std	64.66	22.78	24.47	31.65	6.96	18.13	21.69	15.81	19.97	140.69
Min	0.04	0.02	0.01	0	0	0.01	0.01	0	0	13
Max	949.99	390.68	362.21	467.63	175.81	193.86	257.73	455.03	454.85	2049

Table 9 Quantative analysis

S.No	Privacy Measure	ϵ-value	ϵ-Available Value Interval	Max Queries (Step=0.1)	Max Queries (Step=0.01)
1	Method A	0.52	[0.52,1]	7	54
2	Method B	0.44	[0.44,1]	8	68
3	Method C	0.28	[0.28,1]	9	80
4	Method D	0.17	[0.17,1]	10	89

environment on Google Colabotary, with 13GB RAM and 108GB disk. The experimental platform is a laptop with Intel®Core^{TM} i5-8250U CPU @ 1.6GHz 1.80GHz processor, 8GB RAM, 64-bit operating system, x64-based processor, Windows 10. Also, Table 8 states the statistical features of the dataset.

Analysis and results

The quantitative analysis between the various proposed methods in [8–10] and [13] show that the methodology proposed in [13] outperforms the other methods with respect to various parameters mentioned in Table 9. For convenience, methodology of [8] has been referred as Method A, [9] as Method B, [10] as Method C, and [13] as Method D.

The results of the various methods, as mentioned earlier, have been evaluated with the help of parameters such as ϵ values, ϵ value interval, maximum queries (for different step sizes), and Mean Average Error values.

The Mean Average Error (MAE) is fixed at 0.5 to evaluate the methods' performances, as aforementioned. As it is clear from the data of Table 9 that under identical conditions and $\Delta\,\epsilon = 0.01$, Method D can provide 89 queries. In contrast, Method A, Method B, and Method C can service 54, 68 and 80 queries. Also, when they compared MAE values for the four methods, Method D outperformed the other three as stated in Table 10.

Method D, as proposed in [13] explicitly considered Big Data instead of standard data. Hence, this paper attracted much of our attention, and we did a deeper study and analysis of the paper. Therefore, the quantitative analysis contains results concerning [13].

Discussion

This paper presents a review of all such works that identified data correlation as a privacy threat and tried to maintain data privacy guarantee by considering data correlation as an inherent property of real-world datasets. Also, some of the proposed algorithms reduce the amount of noise to be added so that the data utility could be maintained at the required level and ensured data privacy. We have also presented a few graphs and tables to describe

Table 10 Comparative analysis of MAE Values

S.No	Privacy measure	MAE Value
1	Method A	2.671
2	Method B	1.227
3	Method C	0.722
4	Method D	0.432

results in a better way. Table 3 provides a discussion about all the reviewed articles and their methodology in a tabular format for quick understanding.

Few pieces of research have utilized Mutual Information Coefficient (MIC) to study the relationship between data in a better way. MIC is a data correlation analysis technique and a machine learning concept, which has also given some promising results. Results show that the method using correlation analysis of data, i.e., MIC, yielded better results than methods that did not consider data correlation. Exploring other data correlation analysis techniques to provide better data privacy can be an up-and-coming solution to the problem.

Correlated Big Data privacy

Various frameworks have been established in recent years to ensure 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. One more observation is that most of the works have considered standard data to study the above-mentioned issue. In contrast, real-world datasets are often large and technologically termed big data. With the increase in data size, several other parameters also change, and the correlation among data also increases. A key feature of big data is its high dimensionality which makes the application of various algorithms complex and yields unexpected results. High dimensionality also results in high data correlation. If gone unnoticed, this can be a significant privacy breach, and hence understanding and analyzing data correlation is a fundamental step towards ensuring data privacy guarantees for big data.

Machine learning for correlated Big Data privacy

Researchers have provided various measures to realize data correlation as a data privacy threat and have proposed various methodologies to deal with it. One of the promising methods suggested by the authors of [13] is the use of MIC to calculate data correlation within the dataset. Tables 9 and 10 show the experimental results and conclude that using MIC for correlated data privacy is superior to the other methods. This paved the way for considering other data correlation analysis techniques to do the same. Also, this made other researchers believe that other machine learning tools can help provide solutions to the problem mentioned above of data privacy threat.

Future scope

- Research regarding data correlation as a privacy threat for big data is comparatively less. Hence, there lies a dire need to study data correlation present in big datasets and explore it further to ensure its privacy.

- More methods of calculating the correlation between data may be studied and applied to datasets, and their application to big datasets must be analyzed.
- More efficient mechanisms for providing correlated data privacy protection may be explored and applied.
- Other machine learning concepts must be explored and applied for ensuring data privacy for correlated datasets and correlated big datasets.

Conclusion

- This work is a study of all the articles and research papers that identified data correlation as a data privacy threat and attempted to provide solutions. Table 3 provides a discussion about all the reviewed articles and their methodology in a tabular format for quick understanding.
- MIC is a data correlation analysis technique and a machine learning concept. The researchers successfully used it to study and understand data correlation and consequently provide a solution for the data mentioned above privacy threat. We considered this as the most crucial finding because this introduced an effective way to calculate data correlation. Also, this is important as this paved the way for more research in using machine learning for data privacy.
- Real-world datasets are often large and high dimensional. Notwithstanding, most of the research is done on routine datasets. Hence, the researchers must explore data correlation as a threat to big data privacy in a better form.

Acknowledgements
Not applicable.

Authors' contributions
All authors have equally contributed to the building of this survey paper. All authors read and approved the final 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 Scopus Journals & National Conference.
Dr. 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.
Dr. 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.
Dr. Priyank Jain is working as an Assistant Professor in IIIT Bhopal. He has more than ten years of experience as an Assistant Professor and Research Scholar. His Ph.D. is in the "Big Data" field. He has experience from the Indian Institute of Management, Ahmedabad, India (IIM-A) in the research field. His Educational Qualification is M.Tech & BE in Information Technology. Mr. Priyank Jain's areas of specialization are Big Data, Big Data Privacy & Security, data mining, Privacy-Preserving, & Information Retrieval. Mr. Priyank Jain has publications in various International Conference, International SCI, SCIE, and Scopus Journals & National Conference. He is a member of HIMSS.

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: 2 April 2021 Accepted: 23 October 2021
Published online: 15 December 2021

References

1. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data. 2007;1(1):3. https://doi.org/10.1145/1217299.1217302.
2. 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. https://doi.org/10.1109/ICDE.2007.367856.
3. Dwork C. Differential privacy. In: 33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006). Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer
4. Yang X, Wang T, Ren X, Yu W. Survey on improving data utility in differentially private sequential data publishing. IEEE Trans Big Data. 2017. https://doi.org/10.1109/TBDATA.2017.2715334.
5. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big Data. 2016. https://doi.org/10.1186/s40537-016-0059-y.
6. Wang Y, Song S, Chaudhuri K. Privacy-preserving analysis of correlated data. ArXiv arXiv:abs/1603.03977 2016.
7. Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. IEEE Trans Big Data. 2017. https://doi.org/10.1109/TBDATA.2017.2777862.
8. Chen R, Fung B, Yu P, Desai B. Correlated network data publication via differential privacy. VLDB J. 2014;23:653–76. https://doi.org/10.1007/s00778-013-0344-8.
9. Zhu T, Xiong P, Li G, Zhou W. Correlated differential privacy: Hiding information in non-iid data set. IEEE Trans Inf Foren Security. 2015;10(2):229–42. https://doi.org/10.1109/TIFS.2014.2368363.
10. Zhao J, Zhang J, Poor HV. Dependent differential privacy for correlated data, 2017;pp. 1–7. https://doi.org/10.1109/GLOCOMW.2017.8269219
11. 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, pp. 193–204. Association for Computing Machinery, New York, NY, USA, 2011. https://doi.org/10.1145/1989323.1989345.
12. Wu G, Xia X, He Y. Extending differential privacy for treating dependent records via information theory, 2017.
13. Lv D, Zhu S. Achieving correlated differential privacy of big data publication. Computers Security. 2019. https://doi.org/10.1016/j.cose.2018.12.017.
14. Li Y, Ren X, Yang S, Yang X. Impact of prior knowledge and data correlation on privacy leakage: A unified analysis. IEEE Trans Inf For Sec. 2019;14(9):2342–57. https://doi.org/10.1109/TIFS.2019.2895970.
15. Sunil K, Iliyoung C. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. Int J Environ Res Public Health. 2018. https://doi.org/10.3390/ijerph15122907.
16. Reshef D, Reshef Y, Finucane H, Grossman S, McVean G, Turnbaugh P, Lander E, Mitzenmacher M, Sabeti P. Detecting novel associations in large data sets. Science (New York, NY). 2011;334:1518–24. https://doi.org/10.1126/science.1205438.
17. Pandey R, Dhoundiyal M, Kumar A. Correlation analysis of big data to support machine learning. Big Data. 2015. https://doi.org/10.1109/CSNT.2015.32.
18. Moraru A, Pesko M, Porcius M, Fortuna C, Mladenić D. Using machine learning on sensor data. CIT. 2010. https://doi.org/10.2498/cit.1001913.
19. Namuduri S, Narayanan BN, Davuluru VSP, Burton L, Bhansali S. Deep learning methods for sensor based predictive maintenance and future perspectives for electrochemical sensors. J Electrochem Soc. 2020;167(3):037552. https://doi.org/10.1149/1945-7111/ab67a8.
20. Moraru A, Pesko M, Porcius M, Fortuna C, Mladenic D. Using machine learning on sensor data. In: Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, 2010;pp. 573–578.
21. Liang J-Y, Feng C-J, Song P. A survey on correlation analysis of big data, Big Data. 2016;**39**, 1–18. https://doi.org/10.11897/SP.J.1016.2016.00001
22. MC Kennel. A survey on correlation analysis of big data. Big Data. 2016; 39, 1–18. https://doi.org/10.11897/SP.J.1016.2016.00001
23. Priyank J, Manasi G, Nilay K. Big data privacy: a technological perspective and review. J Big Data. 2016. https://doi.org/10.1186/s40537-016-0059-y.
24. Priyank J, Manasi G, Nilay K. Enhanced secured map reduce layer for big data privacy and security. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0193-4.
25. Zhu XX, Tuia D, Mou L, Xia G-S, Zhang L, Xu F, Fraundorfer F. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci Rem Sens Magazine. 2017;5(4):8–36. https://doi.org/10.1109/MGRS.2017.2762307.
26. Maggiori E, Tarabalka Y, Charpiat G, Alliez P. Convolutional neural networks for large-scale remote sensing image classification. IEEE Trans Geosci Remote Sens. 2017;55:645–57. https://doi.org/10.1109/tgrs.2016.2612821.

27. Zhong L, Hu L, Zhou H. Deep learning based multi-temporal crop classification. Remote Sens Environ. 2019;221:430–43. https://doi.org/10.1016/j.rse.2018.11.032.
28. Ce Zhang XP. Isabel Sargent: Joint Deep Learning for land cover and land use classification. Rem Sens Environ. 2019;221:173–87. https://doi.org/10.1016/j.rse.2018.11.014.
29. Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J Photogrammetry Remote Sens. 2019;152:166–77. https://doi.org/10.1016/j.isprsjprs.2019.04.015.
30. Liu X, Han F, Ghazali KH, Mohamed II, Zhao Y. A review of convolutional neural networks in remote sensing image. In: Proceedings of the 2019 8th International Conference on Software and Computer Applications. ICSCA '19, vol. 5, pp. 263–267. Association for Computing Machinery, New York, NY, USA, 2019. https://doi.org/10.1145/3316615.3316712.
31. Youssef R, Aniss M, Jamal C. Machine learning and deep learning in remote sensing and urban application: A systematic review and meta-analysis. In: Proceedings of the 4th Edition of International Conference on Geo-IT and Water Resources 2020, Geo-IT and Water Resources 2020. GEOIT4W-2020, p. 5. Association for Computing Machinery, New York, NY, USA, 2020. https://doi.org/10.1145/3399205.3399224.
32. Kantarcioglu M, Ferrari E. Research challenges at the intersection of big data, security and privacy. Front Big Data. 2019;2:1. https://doi.org/10.3389/fdata.2019.00001.
33. Haina Ye MY, Xinzhou C. A survey of security and privacy in big data. Big Data. 2016. https://doi.org/10.1109/ISCIT.2016.7751634.
34. Gehrke J, Lui E, Pass R. Towards privacy for social networks: A zero-knowledge based definition of privacy. In: Ishai, Y. (ed.) Theory of Cryptography, 2011;pp. 432–449.
35. Cao L, Ou Y, Yu P. Coupled behavior analysis with applications. Knowledge Data Eng IEEE Trans. 2012;24:1–1. https://doi.org/10.1109/TKDE.2011.129.
36. 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. https://doi.org/10.1145/2339530.2339683.
37. Brand M, Oliver N, Pentland A. Coupled hidden markov models for complex action recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition 0, 994, 1997. https://doi.org/10.1109/CVPR.1997.609450
38. Ghosh A, Kleinberg R. Inferential privacy guarantees for differentially private mechanisms. CoRR, 2016. arXiv:1603.01508.
39. Cao L. Non-iidness learning in behavioral and social data. Computer J. 2013;57:1358–70. https://doi.org/10.1093/comjnl/bxt084.
40. Kifer D, Machanavajjhala A. Pufferfish: A framework for mathematical privacy definitions. ACM Trans Database Syst. 2014. https://doi.org/10.1145/2514689.
41. Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. 2015.
42. Kumar S, Chong I. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. International Journal of Environmental Research and Public Health, 2018;**15**(12). https://doi.org/10.3390/ijerph15122907
43. Cao Y, Yoshikawa M, Xiao Y, Xiong L. Quantifying differential privacy under temporal correlations. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), 2017;pp. 821–832. https://doi.org/10.1109/ICDE.2017.132
44. Li N, Qardaji W, Su D, Wu Y, Yang W. Membership privacy: A unifying framework for privacy definitions. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer &; Communications Security. CCS '13, pp. 889–900. Association for Computing Machinery, New York, NY, USA, 2013. https://doi.org/10.1145/2508859.2516686.
45. Hemkumar D, Ravichandra S, Somayajulu DVLN. Impact of data correlation on privacy budget allocation in continuous publication of location statistics. Peer-to-Peer Network Appl. 2021;14(3):1650–65. https://doi.org/10.1007/s12083-021-01078-6.
46. Kifer D, Machanavajjhala A. A rigorous and customizable framework for privacy. In: Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. PODS '12, pp. 77–88. Association for Computing Machinery, New York, NY, USA, 2012. https://doi.org/10.1145/2213556.2213571.
47. Lee J, Clifton C. Differential identifiability. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12, pp. 1041–1049. Association for Computing Machinery, New York, NY, USA, 2012. https://doi.org/10.1145/2339530.2339695.
48. Cover TM, Thomas JA. Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). New York: Wiley-Interscience; 2006.
49. Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. IEEE Trans Big Data. 2021;7:4. https://doi.org/10.1109/TBDATA.2017.2777862.
50. Cover TM, Thomas JA. Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing,2006). Wiley-Interscience
51. Wang C, Cao L, Wang M, Li J, Wei W, Ou Y. Coupled nominal similarity in unsupervised learning. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11, pp. 973–978. Association for Computing Machinery, New York, NY, USA, 2011. https://doi.org/10.1145/2063576.2063715.
52. Janapana H, Prasad PVGD, Damodaram A. A study of correlation impact on privacy preserving data mining. Int J Computer Appl. 2015;129:22–5. https://doi.org/10.5120/ijca2015907152.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.