Machine Learning Concepts for Correlated Big Data Privacy

SREEMOYEE BISWAS (✉ shonai.biswas@yahoo.in)

MANIT: Maulana Azad National Institute of Technology <https://orcid.org/0000-0002-7097-8018>

Nilay Khare

MANIT: Maulana Azad National Institute of Technology

Pragati Agrawal

MANIT: Maulana Azad National Institute of Technology

Priyank Jain

IIIT-Bhopal

Survey paper

Keywords: Big Data Privacy, Correlated Datasets, Data Correlation, Machine Learning, Correlated Big Data, Data Privacy Threats, Data Privacy Algorithms

Posted Date: April 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-388753/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Journal of Big Data on December 1st, 2021.
See the published version at <https://doi.org/10.1186/s40537-021-00530-x>.

RESEARCH

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
Full list of author information is
available at the end of the article

Abstract

With data becoming a salient asset worldwide, dependence within 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 that where there exists a correlation among data, the existing privacy guarantees could not be assured with existing 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 have considered to utilize a well known machine learning concept, i.e., Data Correlation Analysis to understand the relationship between data in a better way. This has given some promising results as well. Though its less but still a considerable amount of research has been done for correlated data privacy. But correlated big data privacy is very less explored. The real world datasets that are worked upon, are often large in size (technologically termed as big data) and house a high amount of data correlation. 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 any organization, unlike data security that guarantees confidentiality, integrity, availability of data. 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. Larger the value of k, higher is the privacy and consequently lower is the data utility. In spite of 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. 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 . This could prevent distribution attacks on datasets to a large extent. Therefore data analysts and researchers used differential privacy [5] to obtain useful 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. But the distortion should be large 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, many metrics have been applied by researchers. The traditional DP used global sensitivity [5]. Later researchers have also suggested other methods for it so that privacy can be provided along with data utility [8][9][10]. Due to the widespread use of DP, rigorous research works were also done on it and along with some other drawbacks, researchers identified an important drawback that could be a potential threat to data privacy. It was the existence of correlation in the datasets [11][9][12]. Initial researches around DP ignored its existence and regarded data as IID but later researches showed that real datasets often had high correlation among them and hence researches including data correlation became primarily significant [13][14][15]. Many researchers have studied the adverse effects of data correlation on data privacy and various approaches have been proposed by different researchers in their published works to deal with the issue of privacy threat due to the existence of data correlation in datasets. For this, many researchers have used a well known machine learning technique, termed as Data Correlation. Along with ensuring data privacy in presence of data correlation, another big concern is to maintain data utility for which various algorithms have also been proposed. This paper presents a detailed review of research works from the year 2011 to 2019 that stated that existing algorithms aren't enough and proposed algorithms to resolve this issue.

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.	Homogeneity-attack, background knowledge.
2	l-diversity	The equivalence class is also said to l-diversity if the sensitive attribute has at least "l-well-represented" values.	Insufficient to prevent attribute disclosure.
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 .	Complex distribution of a sensitive attribute.
4	Differential Privacy	It is a mechanism to provide an interface between user and database, which protects individual data with the highest mathematical guarantee.	In the case of data diversity, it includes too much noise, which reduces the data utility.

Literature Review

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

data sets. Gradually with time, researchers and scholars around the globe started noticing that in a real-life scenario, getting an independent data set has negligible chances. 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 the privacy of data and some of them had also proposed algorithms which would modify the existing privacy techniques so that privacy for correlated data could be cinched. The first subsection of the Literature Review cites the research papers which were the pioneers in identifying the existence of Data Correlation as a threat to data privacy. The last subsection discusses about the various constraints related to data correlation. Rest of the subsections discuss about various techniques proposed by different researchers as a solution to the above mentioned threat. Figure 6 gives illustration of the categories of the proposed solution. Categorization is done on the basis of the technique used. Table 2 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 [16] 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 to achieve 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 the 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 and D be the class of all databases. For any database $D_1 \in D$, any adversary A , and any $z \in \{0, 1\}^*$, let $\text{Out}A(A(z) \Leftrightarrow \text{San}(D_1))$ 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 non - interactive. If San is non - interactive, then $\text{San}(D_1)$ 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 $T \in \text{agg}$ 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:

$$\Pr[\text{Out}A(A(z) \Leftrightarrow \text{San}(D_1)) \in W] \leq e \cdot \Pr[S(z, T(D_i, \perp), i, n) \in W] \quad (1)$$

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

The probabilities are over the random coins of San and A, and T and S, respectively. Eq. (1) and eq.(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, its optimum value may be achieved. Also Gehrke et al.[16] have proved mathematically that Differential Privacy is a special 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 its analysis. Some major limitations of the proposed work are lack of application as the approach was purely theoretical and also there is no evidence of the presence 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 major 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 solution to deal with it.

Proposed solutions using Probabilistic Models

Kifer et al. [17] 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. This is the desirable privacy guarantee. (c) The Evolution Scenario- This includes knowledge of how the data has evolved and about the potential of the attackers. These three components are discussed in detail with the help of detailed examples in [17]. The primary guarantees of the proposed Pufferfish mechanism are expressed 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. Then 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 we can say that A did not furnish any such information that facilitated to differentiate between E1 and E2. Pufferfish provides this semantic guarantee. In this work, the concept of Hedging Privacy is used which provided good privacy levels. It is strengthening of privacy algorithms weaker than Differential Privacy. In specific it is strengthening of single

prior privacy to protect such data publishers who have poor models of data as prior belief. The set of potential secrets factor and set of discriminative pair factors remain the same but the evolution scenario factor gets changed with the inclusion of conditional probabilities and some other fixed valued parameters within which all conditional probabilities are contained. This is used for histogram publishing with privacy guarantees of Pufferfish in the proposed work. Also in this work, Kifer et al. [17] mathematically proved that the following are provided by Pufferfish framework with the help of various Lemma, Theorems, etc. - (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've observed in works of other researchers, they have used the proposed Pufferfish mechanism as a part of their work towards privacy. But the proposed framework did not satisfy the Differential Privacy and hence privacy preservation for Correlated Data still called for further investigation.

Yang et al. [18] in 2015, proposed another mechanism in which they used Bayesian principles to help Differential Privacy cope up with correlated data. In it, the proposed Bayesian Differential Privacy provides privacy as per traditional commitments of Differential Privacy and along with it, it also provided privacy for correlated data and against 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 is efficient and also prevented it from being intractable.

Definition 5 - Bayesian Differential Privacy - Let $A = A(i, K)$ be an adversary ($K \subseteq [n] \setminus \{i\}$) 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

$$BDPLA(M) = \sup_{x_i, x_{i'}, x_k, S} \log \frac{\Pr(r \in S \mid x_i, x_k)}{\Pr(r \in S \mid x_{i'}, x_k)} \quad (3)$$

We say M satisfies ϵ - Bayesian differential privacy, or ϵ - BDP, if $\sup_A BDPLA(M) \leq \epsilon$. Further, the values of $BDPLA(M)$ with respect to discrete and continuous values of database x can be calculated using [[18], equation 6, and equation 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 private 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 [17]. For the demonstration, two examples have been used - Physical activity monitoring and Flu status. By 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:

$$Wp(u, v) = \left(\inf_{\gamma \in \Gamma(u, v)} \int d(x, y)^p d\gamma(x, y) \right)^{1/p} = \left(\inf_{\gamma \in \Gamma(u, v)} E_{\gamma}[d(X, Y)^p] \right)^{1/p} \quad (4)$$

where $\Gamma(u,v)$ 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 $E(X,Y) \approx [d(X,Y)^p]^{1/p}$, and the cost of the min-cost shift is the Wasserstein distance. Initially, the proposed version of Pufferfish using Wasserstein distance was applied to the above-mentioned examples and it was observed that the amount of noise added will be less than the amount of noise added if global sensitivity was used. Hence utility was enhanced. Then this result was generalized using mathematical proofs supported by a no. of theorems. But the proposed mechanism could have been computationally very expensive, therefore, the authors have used a more restricted setting by using a Bayesian network to describe the dependence between entries called Markov Quilt Mechanism. 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 how to add small effects with respect to the distant nodes. The better utility of the proposed mechanism is initially inferred by case studies of earlier mentioned examples and later were supported by mathematical proofs. To show that the proposed Markov Quilt Mechanism also offered considerable privacy, simulation results were used where it was compared to [19]. Ultimately, the authors concluded that the proposed Markov Quilt Mechanism performs significantly better than the method proposed in [19] 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 can 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 full knowledge of the probabilistic relationship among records. They have applied the proposed perturbation mechanism to multiple available datasets like ADULT dataset from UCI Machine learning repository, NLTCs dataset from StatLib and SearchLogs 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 it can be excluded to produce less noise when it has a low impact on statistical results. It was assumed that the aggregate server has full knowledge of the probabilistic relationship among records. But this may not always be practically feasible. But 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 and fortunately, they are very few in number.

Proposed solutions using Behavioural and Similarity Analysis

Cao et al. [20] in 2012 used an example of a stock market to study about Coupled Behaviour Analysis. The works of [21],[22],[23] also threw light on the same special type of Correlated data i.e. coupling between different users and their behaviors by using the same example of the stock market. When activities of one or more

actors are correlated with each other then such activities are termed as Coupled Behaviour(CB) and their analysis is termed as Coupled Behaviour Analysis (CBA). In this work, a study about CB and CBA is done and how they can be a challenge is also studied. Cao et al.[20] were the first to do some research in Coupled Behaviour Analysis. A behavior (IB) is described 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 behaviour takes place and Relationship R is a tuple that defines complex interactions between multiple actors.[20] Also in [20], 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) $\epsilon_1, \epsilon_2, \dots, \epsilon_I$, an actor ϵ_i undertakes J_i behaviors $IB_{i1}, IB_{i2}, \dots, IB_{iJ}$, 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 [20]. 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. In this work, it is used to study the abnormal behaviour of 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 Figure 1 -

Experiments performed showed that using CHMM will outperform the use of single HMM. And also ACHMM outperformed CHMM. 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. The metrics used for evaluation of performances were

- $Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
- $Precision = \frac{TP}{(TP+FP)}$
- $Recall = \frac{TP}{(TP+FN)}$
- $Specificity = \frac{TN}{(TN+FP)}$

where TP stands for True Positive which means that it belonged to positive class

and 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 and also classified as negative by the algorithm and FN stands for False Negative which means that it belonged to positive class but was misclassified as negative by an algorithm. This work gave an initial insight of CB and CBA and proposed CHMM and ACHMM and showed their performance using experimental analysis of a real dataset. For future research, CBA has great potential, challenges, and opportunities open.

As stated above [20],[23],[21],[22] deal with a special case of correlated data i.e. behavior analysis of the stock market. It is because the data of various investors and traders are often not mutually independent and, 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 to detect such couplings. The two proposed frameworks in [23] consist of the major three stages, as described in Figure 2 -

In this work, algorithms were proposed for all the above three stages using multiple variants of hierarchical grouping (HG) and hierarchical clustering (HC) techniques and they were compared with the algorithms proposed in [20] and the same metrics were used to evaluate the performance in both the works and algorithms of [23] provided better results for anomaly detection. When the two proposed methods were compared to each other, then it was observed that the CBA - HG framework obtained better performance in terms of all the used metrics.

Longbing Cao [24] 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 in. This work initially analyzed the IIDness of several classical algorithms and showed how they may not work properly in case of non-IID data. To measure the non- IIDness, authors in this paper have used a measure termed as Coupled Item Similarity (CIS). CIS between categorical items X and Y is defined as follows:

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

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.

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

The values of δ_j^{Ia} and δ_j^{Ie} can be calculated using [[1], Equation(4.2), and Equation (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 which subsequently result in more robust algorithms such as learning algorithms, pattern matching algorithms, etc.

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 was successful in achieving required privacy in the correlated dataset but as per the privacy and data utility trade-off, it set the major drawback of high data utility loss. Hence, an optimum balance of data privacy and data utility in case of correlated data was yet to be achieved. As used in this work,

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

$$GS(f) = \max_{\{D1, D2\}} \|f(D1) - f(D2)\|_1 \quad (7)$$

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

$$\begin{aligned} \text{ProposedSensitivitycalculation} &= \text{no.of correlated records} * \text{Globalsensitivity} \\ &= k * GS(f) \quad (8) \end{aligned}$$

Then in the year 2014, the Pufferfish mechanism was proposed in [17] 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 also, there were large number of proposed mechanisms to address the same threat. Zhu et al. [9] in 2015, intricately described about 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, it is also explained in detail about 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]. They clearly showed in this work that how the solution proposed in [8] induced a very large magnitude of noise which greatly 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 as compared to global sensitivity.

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

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

where q is a recordset 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

$$CS_i = \sum_{j=0}^n |\delta_{ij}| (\|Q(D_1) - Q(D_2)\|_1) \quad (10)$$

where $\delta_{ij} \in \Delta$. The record sensitivity measures the effect on all records in D when deleting a record r_i . δ is the correlated degree analysis matrix which holds the degree of correlation between all sets of records. Along with it, authors have also proposed a mechanism that satisfies Differential Privacy in case of correlated datasets and also preserves utility for further application. But during the practical implementation of the proposed mechanism, it was observed that it is dependent upon a no. of parameters. One of 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 cannot give privacy guarantee in case of correlated data and later in their work have proposed a perturbation mechanism for dependent data called 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 to demonstrate how an adversary can infer sensitive information such as location information of a user by using the noise added query outputs and by exploiting user's social relationships and thus violates the Differential Privacy guarantees. The performance of the Inference attacks have been measured using the following two metrics -

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

$$2.) LeakedInformation = \frac{1}{n} \sum_{i=1}^n H(D_i) - H(D_i \parallel \mu D_{-i}) \quad (12)$$

Dist(.) is defined in [[10], equation 5] and H(.) denotes the entropy information of

a random variable [25]. $H(D_i)$ measures the entropy of prior probability and hence measures the adversary's prior information for D_i without taking into account the social relationships. $(H(D_i) - H(D_i \parallel \mu D_{-i}))$ 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 prior to posterior beliefs. The above metrics applied over real-world datasets clearly showed that conventional Differential Privacy applied to a dependent dataset will disclose more sensitive information than expected and hence is a serious privacy violation issue. In the proposed perturbation mechanism called DDP, they have added an extra parameter called dependence coefficient to minutely measure the dependence relationship between tuples. The sensitivity under dependence relationship between two tuples D_i and D_j is proposed as -

$$\text{Sensitivity under dependence relationship} = \Delta D_i + \rho_{ij} \Delta D_j \quad (13)$$

where $\rho_{ij} \in [0, 1]$ and is a metric to measure dependence between two tuples D_i and D_j . Rest the utility and privacy superiority of the proposed DDP mechanism is been theoretically and experimentally proved in the current work. Furthermore, it is 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. It is 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 can 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,

Lemma1 - 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

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

The mathematical proof of Lemma 1 can be found in detail in [12]. In information theory, the relative entropy is used to measure the distance between two probability distributions [26]. 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

$$D(X_i||X_i|Y = r) \leq \epsilon \quad (15)$$

Using Lemma 1, we have

$$\begin{aligned} D(X_i||X_i|Y = r) \\ = \sum Pr[X_i = t] \log \frac{(Pr[X_i = t])}{(Pr[X_i = t|Y = r])} \leq \sum Pr[X_i = t] \log e^\epsilon = \epsilon \end{aligned} \quad (16)$$

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

$$I(X_i; Y) \leq \epsilon \quad (17)$$

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

Authors Kumar et al. [27] in 2018 have tried to do correlation analysis over some selected datasets to enhance the accuracies when some classification problem arises. For the same, the datasets used are - Depressive disorder symptom dataset for evaluating depression severity, Local weather dataset for classifying depression severity, and 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:

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

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, $\{a_1, a_2, \dots, a_n\}$ and $\{b_1, b_2, \dots, b_n\}$, the correlation coefficient can be calculated as:

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

where n is the sample size, a_i and b_i are the i th 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 strong negative correlation, and those closest to 0 show no relation. Then for feature selection, 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 [20]. And then an optimum number of features was selected as per the assigned ranks. But better correlation analysis technique must be used 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.

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. But 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 time complexity of the proposed mechanism was calculated 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],[9], and [10] (b) As per the curves drawn using experimental results, the influence of different privacy budget parameter ϵ is not stable for mechanisms provided in [8], [9] but its influence on the proposed method in the current work is quite 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 further work may be done are - (a) Examination of external correlation of datasets as in this only internal correlations have been considered. (b) The Traditional MIC algorithm used in this is unable to 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. [14] in 2019, proposed a slight modification to Differential Privacy to give privacy guarantee for correlated data. To do so, they had considered all viable combinations that could be done by combining tuple correlations and the query responses. This work was concluded by stating that the modification done to Differential Privacy, termed as Dependent Differential Privacy (DDP) facilitated privacy

guarantee to any data correlation model and adversary's knowledge and authors also showed how quantitatively DDP can be deduced from DP by using a stronger privacy parameter.

Proposed Solution using the concept of Privacy Leakage with time

Cao et al. [28] midway in the year 2019, tried to address the issue of privacy loss of temporal data release in their work. If data is recorded with respect to time then such a correlated dataset is said to be a temporally correlated dataset or simply a temporal dataset. 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 Markov Chain. The parameter of the Markov Chain is a transition matrix. They have accordingly used transition matrices P_{Bi} and P_{Fi} which are called backward temporal correlation and forward temporal correlation respectively in this work. In this, it was also considered rational that the

matrices may be known by the adversary.

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

$$BPL(A_i^T, M_t) = \sup_{l_i^t, l_i^{t1}, r^1, \dots, r^t} \log \frac{Pr(r^1, \dots, r^t \mid l_i^t, D_k^t)}{Pr(r^1, \dots, r^t \mid l_i^{t'}, D_k^t)} \quad (20)$$

$$BPL(M_t) = \max_{\forall A_i^T, i \in U} BPL(A_i^T, M_t) \quad (21)$$

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

$$FPL(A_i^T, M_t) = \sup_{l_i^t, l_i^{t1}, r^1, \dots, r^t} \log \frac{Pr(r^1, \dots, r^t \mid l_i^t, D_k^t)}{Pr(r^1, \dots, r^t \mid l_i^{t'}, D_k^t)} \quad (22)$$

$$FPL(M_t) = \max_{\forall A_i^T, i \in U} FPL(A_i^T, M_t) \quad (23)$$

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

$$TPL = BPL + FPL - PL_0(A_i^T, M_t) \quad (24)$$

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

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

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

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

After quantifying TPL, α -DPT i.e. α -DP under temporal correlations was proposed. Using equation 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,

$$\alpha - DPT = (\alpha_t^B + \alpha_t^F - \epsilon^t) - DP \quad (27)$$

The current work has consequently evinced the sequential composition theorem of α - DPT and showed how temporal correlations only affect event-level privacy (i.e. privacy at a particular time t) and does not affect user-level privacy (i.e. privacy of each individual during the whole timeline). Then further, to calculate the mathematical values of BPL and FPL to mathematically find TPL, Cao et al. [28] have formalized the objective function and constraints and have shown that their optimum values can be found 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 in this, the number of different methods was discussed 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 in this. 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. Future scope of work lies in studying privacy leak under temporal correlations combined with other types of correlation models.

Proposed solution using Weighted Hierarchical Graph Technique

Li et al. [15] 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 they can be used by an adversary with little or full prior knowledge. Mechanisms proposed in [29],[30],[31], and [18] 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) whose node and edge values are calculated using the formulae specified in the current work. Few key points became evident in the process of finding the edge values and node values of WHG- (a) The weakest adversary can cause privacy leakage when there exists a positive correlation. (b) When WHG has both positive and negative edge, then to compute the correct privacy leakage, the whole WHG will have to be traversed. To traverse the full WHG, they have initially used a full space searching algorithm with time complexity $O(n^4 2^{n-1})$ where n is the number of tuples in a database. To reduce this computational time complexity, a fast searching algorithm on a subspace of original full space is proposed with little sacrifice of accuracy and time complexity of $O(n^4)$. To analyze the impact of the above-mentioned factors 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 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 and the accuracy of the fast searching algorithm is quite close to the full space searching algorithm hence in practical applications, fast searching algorithm is better suited. Figure 4 and 5 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. But it should be applied to nonlinear queries as well to better evaluate its performance.

Study of Correlation Constraints

Hyma et al. [32] by the end of 2015, studied various correlation constraints that need to be considered 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, the privacy level is scrutinized. The value decides the impact of the correlation on the privacy level and mechanism.

c.) Attribute-based Correlation Constraints- In this, the associations between multiple attributes are examined. If there is any association then chances are high that knowing the values of all attributes of one record will help adversary to 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 prominent impact over 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 it is handed over by the owner to the publisher. A proper balance between the two is required to achieve the optimistic privacy of data.

These constraints were important to study before understanding correlation and its effect on data privacy. In this work, it was strongly concluded that the privacy mechanism applied to correlated data and non-correlated data are very different. And also an accurate correlation constraint mechanism is required to achieve the required level of privacy. But this work has given theoretical study of the above-stated facts over real-time examples and also it lacks practical implementations of the same.

Conclusion and Future Work

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 them, proposed algorithms to reduce the amount of noise to be added so that the data utility could be maintained at the required level along with ensured data privacy. We have also presented a few graphs to describe results in a better way. Few researches have utilized Data Correlation Analysis, which is a machine learning concept, to study the relationship between data in a better way. This has given some promising results as well. One more observation is, most of the works have considered normal data to study the above mentioned issue, whereas real world datasets are often large in size, technologically termed as big data. 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. For future work, more methods of calculating the correlation between data may be studied and applied to datasets, their application to big datasets must be analysed, more efficient mechanisms for providing correlated data privacy protection may be explored and applied, more machine learning concepts must be studied and considered for ensuring data privacy for correlated datasets and correlated big datasets.

Appendix

Acknowledgements

Not applicable

Ethics Approval

The content provided is as per the ethics of the research fraternity.

Availability of Data and Material

All the data and matter is readily available.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

All authors have given consent for publication of the matter.

Funding Details

Not Applicable

Authors' contributions

All authors have equally contributed for the building of this survey paper.

Figure 6 Categorisation of Proposed Solutions

Authors' information

- 1 Sreemoyee Biswas, is currently pursuing 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 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.
- 2 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. Dr. Nilay Khare, publications are in 54 International and National Conferences, and International Journal. He is a Life member of ISTE.
- 3 Dr. Pragati Agrawal, is working as Assistant Professor in MANIT Bhopal. She has more than 5 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.
- 4 Dr. Priyank Jain, is working as a 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 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 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.

Author details

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

References

1. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkatasubramanian, M.: L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**(1), 3 (2007). doi: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, pp. 106–115 (2007). doi: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, ??? (2006). <https://www.microsoft.com/en-us/research/publication/differential-privacy/>
4. Yang, X., Wang, T., Ren, X., Yu, W.: Survey on improving data utility in differentially private sequential data publishing. *IEEE Transactions on Big Data*, 1–1 (2017). doi:10.1109/TBDDATA.2017.2715334
5. Jain, P., Gyanchandani, M., Khare, N.: Big data privacy: a technological perspective and review. *Journal of Big Data* **3** (2016). doi:10.1186/s40537-016-0059-y
6. Wang, Y., Song, S., Chaudhuri, K.: Privacy-preserving analysis of correlated data. *ArXiv abs/1603.03977* (2016)
7. Chen, J., Ma, H., Zhao, D., Liu, L.: Correlated differential privacy protection for mobile crowdsensing. *IEEE Transactions on Big Data* **PP**, 1–1 (2017). doi:10.1109/TBDDATA.2017.2777862
8. Chen, R., Fung, B., Yu, P., Desai, B.: Correlated network data publication via differential privacy. *The VLDB Journal* **23**, 653–676 (2014). doi: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 Transactions on Information Forensics and Security* **10**(2), 229–242 (2015). doi:10.1109/TIFS.2014.2368363
10. Zhao, J., Zhang, J., Poor, H.V.: Dependent differential privacy for correlated data, pp. 1–7 (2017). doi: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). doi:10.1145/1989323.1989345. <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 and Security* **82** (2019). doi:10.1016/j.cose.2018.12.017
14. Zhao, J., Zhang, J., Poor, H.V.: Dependent differential privacy for correlated data, pp. 1–7 (2017). doi:10.1109/GLOCOMW.2017.8269219
15. Li, Y., Ren, X., Yang, S., Yang, X.: Impact of prior knowledge and data correlation on privacy leakage: A unified analysis. *IEEE Transactions on Information Forensics and Security* **14**(9), 2342–2357 (2019). doi:10.1109/TIFS.2019.2895970
16. 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*, pp. 432–449 (2011)
17. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* **39** (2014). doi:10.1145/2514689
18. Yang, B., Sato, I., Nakagawa, H.: Bayesian differential privacy on correlated data. (2015)
19. 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). doi:10.1145/2508859.2516686. <https://doi.org/10.1145/2508859.2516686>

20. Cao, L., Ou, Y., Yu, P.: Coupled behavior analysis with applications. *Knowledge and Data Engineering, IEEE Transactions on* **24**, 1–1 (2012). doi:10.1109/TKDE.2011.129
21. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. *2012 IEEE Conference on Computer Vision and Pattern Recognition* **0**, 994 (1997). doi:10.1109/CVPR.1997.609450
22. Ghosh, A., Kleinberg, R.: Inferential privacy guarantees for differentially private mechanisms. *CoRR abs/1603.01508* (2016). 1603.01508
23. Song, Y., Cao, L., Wu, X., Wei, G., Ye, W., Ding, W.: Coupled behavior analysis for capturing coupling relationships in group-based market manipulations. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012). doi:10.1145/2339530.2339683
24. Cao, L.: Non-iidness learning in behavioral and social data. *The Computer Journal* **57**, 1358–1370 (2013). doi:10.1093/comjnl/bxt084
25. Cover, T.M., Thomas, J.A.: *Elements of Information Theory* 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, ??? (2006)
26. 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). doi:10.1145/2063576.2063715. <https://doi.org/10.1145/2063576.2063715>
27. 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* **15**(12) (2018). doi:10.3390/ijerph15122907
28. Cao, Y., Yoshikawa, M., Xiao, Y., Xiong, L.: Quantifying differential privacy under temporal correlations. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 821–832 (2017). doi:10.1109/ICDE.2017.132
29. 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). doi:10.1145/2213556.2213571. <https://doi.org/10.1145/2213556.2213571>
30. 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). doi:10.1145/2339530.2339695. <https://doi.org/10.1145/2339530.2339695>
31. Cover, T.M., Thomas, J.A.: *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA (2006)
32. Janapana, H., Prasad, P.V.G.D., Damodaram, A.: A study of correlation impact on privacy preserving data mining. *International Journal of Computer Applications* **129**, 22–25 (2015). doi:10.5120/ijca2015907152

Table 2 Proposed Solutions

S.No	Privacy Measure	Definition	Limitations
1	Pufferfish Mechanism [17]	Formed strong foundation for other similar research	Could not satisfy Differential Privacy Guarantee.
2	Coupled Behaviour Analysis (CBA) [20]	Good experimental results were obtained on real datasets.	Other challenges of CBA are yet to be explored.
3	CBA-HG and CBA-HC [23]	Experimental comparison showed that CBA-HG outperformed the mechanisms of [20]	Applicability on other datasets with different couplings is uncertain.
4	Coupled Item Similarity (CIS) [24]	Proposed an effective mechanism to measure the non IIDness.	No solution to deal with the non IIDness was proposed.
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 a large amount and greater data utility as compared to [8].	Few parameters held trade off with utility.
7	Bayesian Differential Privacy [18]	Mechanism provided privacy for correlated data and against 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 that correlation of distant nodes need not be considered.	When compared to the results of [19], 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	Did not study about other privacy leakages without which its practical implementation is not suggested.
11	Bayesian Network Perturbation Mechanism [7]	Proposed perturbation mechanism provided a decreased privacy budget and increased data utility.	The requirement of modelling the Bayesian Network in advance may not be practically feasible.
12	Statistical Correlation Analysis [27]	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 [14]	Proposed DDP and proved mathematically how it can be derived from DP	Lacks practical implementation.
15	Temporal Privacy Leakage [28]	Temporal correlation was studied along with the study of relationship between data privacy and data utility.	Other correlation models were not studied for temporal leakages.
16	Weighted Hierarchical Graph Mechanism [15]	Mechanism offers privacy guarantee in case of negative correlation as well.	Not applicable to nonlinear queries.

Figures

Figure 1

Steps of ACHMM

Figure 2

Steps of the proposed framework

Figure 3

Flowchart of r-CBDP [16]

Figure 4

Increasing Privacy Leakage with increasing Correlation

Figure 5

Decreasing Privacy Leakage with increasing Prior Knowledge

Figure 6

Categorisation of Proposed Solutions