

Correlated data in differential privacy: Definition and analysis

Tao Zhang¹ | Tianqing Zhu¹ | Renping Liu² | Wanlei Zhou¹

¹Centre for Cyber Security and Privacy; School of Computer Science, University of Technology, Sydney, New South Wales, Australia

²School of Electrical and Data Engineering, University of Technology, Sydney New South Wales, Australia

Correspondence

Tianqing Zhu, Centre for Cyber Security and Privacy; School of Computer Science, University of Technology, 15 Broadway, Ultimo, Sydney, NSW 2007, Australia.
Email: tianqing.zhu@uts.edu.au

Funding information

ARC discovery project, Grant/Award Number: DP190100981, DP200100946

Summary

Differential privacy is a rigorous mathematical framework for evaluating and protecting data privacy. In most existing studies, there is a vulnerable assumption that records in a dataset are independent when differential privacy is applied. However, in real-world datasets, records are likely to be correlated, which may lead to unexpected data leakage. In this survey, we investigate the issue of privacy loss due to data correlation under differential privacy models. Roughly, we classify existing literature into three lines: (1) using parameters to describe data correlation in differential privacy, (2) using models to describe data correlation in differential privacy, and (3) describing data correlation based on the framework of Pufferfish. First, a detailed example is given to illustrate the issue of privacy leakage on correlated data in real scenes. Then our main work is to analyze and compare these methods, and evaluate situations that these diverse studies are applied. Finally, we propose some future challenges on correlated differential privacy.

KEYWORDS

correlated data, differential privacy, privacy leakage, Pufferfish

1 | INTRODUCTION

Over the last decade, the relationship between human and data has never been so inseparable. Meanwhile, the era of big data poses new challenges to human data management, especially in data privacy.¹ Privacy preserving data releasing has been adopted by academia and industry to protect individual privacy when datasets are published to the public.^{2,3} In order to guarantee data security, data need to be sanitized via privacy mechanisms, such as k-anonymity,⁴ l-diversity,⁵ t-closeness.⁶ Among these privacy mechanisms, differential privacy is one of the most promising privacy models to protect data privacy. The notion of differential privacy was firstly proposed by Dwork et al,⁷ which provides a rigorous mathematical framework of defining and protecting privacy. A common method to achieve differential privacy is to add random noise to the output of a query. It ensures that the adversary cannot distinguish the participation of the individual even if the adversary knows the entire background information.

In traditional differential privacy, a weak assumption is that records in a dataset are independent from each other. In practice, however, data in a dataset are usually correlated resulting from the process of data generation. As such, deleting one record will have impacts on other records. Such impacts may reveal more information for the adversary. Kifer et al confirms that the correlation between data may disclose more information than expected.⁸ This finding starts a new topic on how to preserve the privacy of correlated datasets when they are released to the public. Adding noise to correlated datasets has been proposed as one of methods to guarantee differential privacy. One of the challenges is how to add appropriate noise to preserve data privacy in correlated datasets since adding too much noise to the correlated dataset will degrade data utility, and adding insufficient noise will disclose data privacy.

Generally, the amount of noise added to correlated datasets depends the extent of data correlation, which is an inherent feature from data generation. In order to add appropriate noise, data generation or data correlation should be known by the curator or the adversary as

background information. Hence, many works modify traditional definition of differential privacy, and add background knowledge of data correlation by correlation parameters and correlation models, to cope with the issue of privacy loss for the correlated dataset. Pufferfish⁹ is a flexible privacy model without the assumption that data are independent in a dataset, which can be used to quantify the privacy loss due to data correlation.

In terms of how to describe data correlation, we classify existing research into three streams. The first stream uses parameters to describe simple data correlation in differential privacy. Chen et al used the method of multiplying original sensitivity with the number of correlated records, yet it may lead to too much noise.¹⁰ Other correlation parameters are proposed to describe data correlation in differential privacy, including the correlated degree matrix¹¹ and the dependence coefficient.¹² The second stream exploits correlation models to describe complex data correlation in differential privacy, such as Gaussian correlation model^{13,14} and Markov chain model.^{15,16} The last stream is built on the privacy framework, called Pufferfish,⁹ which is a flexible privacy model to guarantee the data sharing needs and is able to describe simple and complex data correlations. Inspired by Pufferfish, He et al proposed another privacy model, Blowfish to tune privacy-utility trade-off.¹⁷

1.1 | Outline and survey overview

Several surveys have been working on differential privacy. The first survey by Dwork⁷ summarized notions of differential privacy, mechanisms, and some differentially private algorithms for data publishing. Later, Dwork et al gave an overview on motivated applications and future directions for data publishing and data analysis.^{18,19} A book by Dwork presented comprehensive coverage of algorithms maintaining differential privacy against adversaries and differentially private methods for mechanism design and machine learning.²⁰ Sarwate et al studied differentially private algorithms for continuous data in signal processing.²¹ Recently, Zhu et al gave a summary on the data publishing and data analysis underlying differential privacy.²² Damien et al gave a systematic taxonomy of these variants and extensions of differential privacy.²³ Previous surveys mainly focus on the concepts, theories, and development of differential privacy. Different from existing works, this survey focuses on the issue of privacy leakage on correlated data, which is a vital issue in differential privacy. The contributions of this article are listed as below.

- First, we give a summary of existing research on correlated differential privacy, and roughly classify existing research into three research streams: correlation parameters, correlation models, and Pufferfish. This helps to understand the characteristics of existing methods on correlated differential privacy.
- Second, we compare the advantages and disadvantages, similarities and differences and the application scenarios of methods. This provides a guideline to use correlated differential methods in different scenarios.
- Finally, we propose a number of future topics on correlated differential privacy. This gives some sights on new issues and potential methods in correlated differential privacy.

The rest of this article is organized as follows. We describe the preliminaries in Section 2, and gives examples to illustrate the problem in Section 3. Sections 4 and 5 summarize the studies in differential privacy models by correlation parameters and correlation models, respectively. In Section 6, we introduce the framework of Pufferfish. Section 7 is the future direction, and finally, Section 8 is the conclusion.

2 | PRELIMINARIES

2.1 | Differential privacy

Differential privacy is a rigorous privacy model which is widely studied in the last decade. In brief, D is a dataset that contains a set of records. Two datasets D and D' are referred to as neighboring datasets when they differ in one record. A query f is a function that maps records $r \in \Omega$ to abstract outputs $f(D) \in \Omega$, where Ω is the whole set of outputs. Hence, the dataset is the input and the released information from the mechanism is the output. The relationship can be described as $f: D \rightarrow \Omega$.

Definition 1 (ϵ -Differential privacy⁷). Given neighboring datasets D and D' , a randomized algorithm \mathcal{M} satisfies ϵ -differential privacy for any possible outcome $f(D) \in \Omega$,

$$\Pr[\mathcal{M}(D) \in \Omega] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(D') \in \Omega], \quad (1)$$

where ϵ is privacy budget which determines privacy level. The lower ϵ represents the higher privacy level.

Definition 2 (Sensitivity⁷). For a query $f : D \rightarrow \Omega$, and neighboring datasets, the sensitivity of f is defined as

$$\Delta f = \max_{D, D'} \|f(D) - f(D')\|_1. \quad (2)$$

Sensitivity measures the maximal difference between neighboring datasets. When a dataset is given, sensitivity depends on the type of query f .

2.2 | Differential mechanisms

Two common mechanisms are widely used to achieve ϵ -differential privacy: Laplace mechanism⁷ and exponential mechanism.²⁴

Definition 3 (Laplace mechanism). Given a query $f: D \rightarrow \Omega$ over the dataset D , Laplace mechanism satisfies ϵ -differential privacy if,

$$\mathcal{M}(D) = Q(D) + \text{Laplace}(\Delta/\epsilon), \quad (3)$$

where $\text{Laplace}(\cdot)$ denotes Laplace noise which is drawn from a Laplace distribution with the probability density function $p(x|\lambda) = \frac{1}{2\lambda} e^{-|x|/\lambda}$, where λ depends on privacy budget and sensitivity.

Definition 4 (Exponential mechanism). Given a score function $S(D, \phi)$ of a dataset D , exponential mechanism \mathcal{M} satisfies ϵ -differential privacy if

$$\mathcal{M}(D) = \left(\text{return } \phi \propto \exp\left(\frac{\epsilon S(D, \phi)}{2\Delta f}\right) \right). \quad (4)$$

Differential privacy exploits the exponential mechanism to randomize the outputs, and a score function $S(D, \phi)$ is used to evaluate the quality of an output ϕ for a query f .

2.3 | An analysis of data correlation

In this section, we will introduce some most frequent studied correlations in the literature. Many types of correlations are in real-world datasets, and the correlation is assumed to be known by the curator and the strong adversary. Generally, data correlation can exist in one dataset or in multiple datasets. In the first case, data correlation can disclose more information when the dataset is published from one entity. In the second case, personal information may appear in different entities. For instance, people would like to share their information in different social applications (e.g., Twitter or Facebook), and these information can be shared to the third party via social applications at the same time.

2.3.1 | Direct correlation

Direct correlation occurs when the curator has access to all knowledge of data correlation. For example, A, B, C are records in a dataset, and direct correlation between these records can be expressed as $A + B = C$ or $A * B = C$, etc. Direct correlation is deterministic; hence, it is relatively easy for the curator to handle. Direct correlation is simple data correlation.

2.3.2 | Indirect correlation

Different from the direct correlation, indirect correlation is more complex. It is nondeterministic and thus can not defined the correlation as a formula. Indirect correlation is complex data correlation.

Temporal correlation

A dataset with temporal correlation is generated by the predefinition of a time interval, and continuous released records falling into this time interval are regarded as correlated by time. Continuous generated data in the real world tend to be temporally correlated, like the dataset of user locations described in the example above. One characteristic of temporal correlation is that all records are usually correlated, which means the first record in the dataset may have an impact on the last record. The extreme case described in Section 3 is an example of such a case. And because all records in the temporally-correlated dataset are related, some studies have solely focused on differential privacy given temporal correlation.^{15,16,25}

Attribute correlation

Attribute correlation refers to correlations that can be revealed through a particular attribute, that is, when the value of two or more records is the same or similar. In reality, there are many attributes that can create correlations in real-world datasets, and many real-world datasets contain those attributes. For example, addresses, which are common to social network and ancestry datasets, are an attribute that can be used to identify members of the same family.

3 | PROBLEM STATEMENT

In this section, we show the issue of privacy loss due to data correlation. Most previous works assume that all records in the dataset are independent. Based on this assumption, differential privacy claims that it can limit the probabilistic inference when the attacker knows the whole information but one record. However, records in a real-world dataset are often correlated with each other, and it is likely to breach the privacy in a dataset when the adversary knows all but one record and the knowledge of data correlation. Here, we give an example to illustrate the temporal correlation in a dataset, and how data correlation will degrade privacy level.

4 | AN EXAMPLE TO ILLUSTRATE DATA CORRELATION

Considering the scenario of a traffic monitoring application, data of user mobility are collected by a trusted server continuously. In this scenario, one typical correlated dataset is generated—temporal correlated dataset. This type of dataset is generated continuously in a time interval, and released records are correlated due to time correlation. Users in a monitoring area are likely to have social relationships - perhaps friends or couples. Due to social relationships, location information is the same for some users during a period of time, and thus users' location information have some correlation in a dataset.

As shown in Table 1, users' location are given in different time points. In Table 1, we can note that $user_1$ and $user_2$ have the same location from the time point $t = 1$ to $t = 4$. The reason could be $user_1$ and $user_2$ are family members, they are likely to have the same track in a time period. In this case, changing the location of $user_1$ will also change the location of $user_2$, hence the records of $user_1$ and $user_2$ are referred to as correlated records.

Table 2 shows the sum of true counts with regard to user locations. When Laplace mechanism is applied in this case, the amount of $Lap(1/\epsilon)$ noise is added to perturb each count in Table 2 so that released information can achieve ϵ -DP at each time point. However, if the attacker knows the relationship between $user_1$ and $user_2$, the attacker can infer the location of $user_1$ and $user_2$. As a result, when the count of users location is released, the privacy of users location is unlikely to satisfy ϵ -differential privacy because the same count is considered to be released two times. $Lap(2/\epsilon)$ noise should be added to the query result in order to hold differential privacy in the dataset.

Based on the example above, we can find that data correlation can exist in a dataset. The change of one record can have an impact on other records, and it also leads to changes on the query response. This example proves that correlated data in a dataset will leak more information to the adversary when using differential privacy, and hence degrade the privacy level. Intuitively, one method is to inject more noise to the correlated

user \ t	1	2	3	4
u_1	loc_2	loc_2	loc_3	loc_4
u_2	loc_2	loc_2	loc_3	loc_4
u_3	loc_1	loc_4	loc_5	loc_2
u_4	loc_4	loc_5	loc_2	loc_5

TABLE 1 Users' locations at different time points

loc \ t	1	2	3	4
loc_1	1	0	0	0
loc_2	2	2	1	1
loc_3	0	0	2	0
loc_4	1	1	0	2
loc_5	0	1	1	1

TABLE 2 Sum counts of users' locations

dataset. The amount of noise added to the correlated dataset depends on the degree of correlated information in a dataset. This situation reveals that the level of challenge faced in dealing with the trade-off between data utility and data privacy.

5 | CORRELATION PARAMETERS IN DIFFERENTIAL PRIVACY

In this section, we will introduce some works that use correlation parameters to describe data correlation on correlated datasets in differential privacy. After data correlation is measured by correlation parameters, an appropriate amount of noise can be quantified to add in differential privacy, and thus privacy can be guaranteed with a desirable trade-off between data privacy and data utility.

5.1 | The number of correlated records

One of the simple methods to describe data correlation was proposed by Chen et al.²⁶ The correlation parameter k is used to measure the extent of correlated data. A dataset D with a correlation parameter k means that the maximum number of correlated records in the dataset is k . The correlation parameter k is assumed to be known by the curator or the strong adversary. After multiplying the original sensitivity with the number of k correlated records, any $\frac{\epsilon}{k}$ differentially private mechanism also satisfies ϵ differential privacy when the number of k correlated records are in the dataset.

Analysis: The advantage of this method is the simplicity and easy to implement. However, when a batch of correlated records are in a dataset, a large amount of noise will be added in the output since the correlated parameter k cannot describe data correlation accurately. Hence, this will lead to a severe degradation in dataset utility.

5.2 | Dependence coefficient

Liu et al studied data correlation with a new definition of dependent differential privacy, which considers more background knowledge of data correlation described by a correlation parameter ρ_{ij} , called the dependence coefficient.¹² First, the definition of dependent differential privacy (DDP) is given below.

Definition 5 (ϵ -Dependent differential privacy). A mechanism \mathcal{M} gives ϵ -dependent differential privacy for any pairs of dependent neighboring datasets $D(L, R)$ and $D'(L, R)$ and any possible outcomes Ω , if the mechanism \mathcal{M} satisfies

$$\max_{D(L,R), D'(L,R)} \frac{P(\mathcal{M}(D(L, R) = \Omega))}{P(\mathcal{M}(D'(L, R) = \Omega))} \leq \exp(\epsilon), \quad (5)$$

where L is the number of correlated records and R is the probabilistic dependence relationship between the records.

In the definition of ϵ -dependent differential privacy, we note two differences from the traditional differential privacy. One is the probabilistic dependence relationship is specified in the dataset and the other is the size of correlated records is specified in the dataset. From the definition of DDP, we see that the DDP can guarantee the data privacy and defend against the attacker who even has the background information of probabilistic dependence R between records. More specific, dependent sensitivity includes two parts: the sensitivity caused by the modification of the record itself ΔD_j and the sensitivity induced in other records $\rho_{ij} \Delta D_j$. The dependence coefficient $\rho_{ij} \in [0, 1]$ serves as a metric to evaluate the extent of the dependent relationship between tuples.

Definition 6 (Dependent sensitivity). For a query Q , dependent sensitivity is defined over a dependent dataset D as,

$$DS^Q = \max_j \sum_{i=C_{i1}}^{C_{iL}} \rho_{ij} \Delta Q_j, \quad (6)$$

where C_{i1}, \dots, C_{iL} denotes L records that are dependent with i th record and $\rho_{ii} = 1$. DS^Q denotes dependent sensitivity of a query Q over all records in the dataset D caused by the modification of one individual record D_i .

Analysis: The advantage of dependence coefficient or DDP is that this correlation parameter is able to measure the degree of data correlation. While the effectiveness of dependence coefficient depends on how well the correlation between records can be described and computed. For example, when the correlation in a dataset is exactly known by the curator, this method can model it well. When data correlation is unknown or partial known, the accuracy of dependence coefficient may be overestimated or underestimated.

5.3 | Zhao-dependent differential privacy

Another kind of dependent differential privacy studied in Reference 27, which we refer to as Zhao-DDP in this article. The goal of Zhao-DDP is to prevent the adversary from inferring the user's information with the combination of correlated records and query responses. The definition of Zhao-DDP is given below.

Definition 7 (Zhao- ϵ -dependent differential privacy). A mechanism \mathcal{M} provides ϵ -DDP, if for any neighbouring datasets and any possible outputs Ω , we have

$$\frac{\mathbb{P}[\mathcal{M}(x_i, x_K, X_{\bar{K}}) \in \Omega]}{\mathbb{P}[\mathcal{M}(x'_i, x_K, X_{\bar{K}}) \in \Omega]} \leq e^{\epsilon}, \forall i, K, x_i, x'_i, x_K, \Omega \quad (7)$$

where $i \in \{1, \dots, n\}$, $K \subseteq \bar{[i]} = \{1, \dots, n\} \setminus \{i\}$, $\bar{K} = \bar{[i]} \setminus K$; ϵ' is a segmented linear function of traditional ϵ -differential privacy.

When calculating the conditional probabilities, the correlation knowledge is needed from the curator. Comparing with differential privacy, we can find that K iterates through all subsets of $\bar{[i]} = \{1, \dots, n\} \setminus \{i\}$ to bound $\frac{\mathbb{P}[\mathcal{M}(x_i, x_K, X_{\bar{K}}) \in \Omega]}{\mathbb{P}[\mathcal{M}(x'_i, x_K, X_{\bar{K}}) \in \Omega]}$, while ϵ -DP bounds $\frac{\mathbb{P}[\mathcal{M}(x_i, x_{[i]}) \in \Omega]}{\mathbb{P}[\mathcal{M}(x'_i, x_{[i]}) \in \Omega]}$.

Analysis: Comparing with the DDP in Reference 12, Zhao-DDP considers more correlation information $\mathbb{P}[X_{\{1, \dots, n\} \setminus \{X_K\}} | X_K]$, while DDP considers the correlation information $\mathbb{P}[X_{\{1, \dots, n\} \setminus \{i\}} | X_i]$.

5.4 | Correlated degree matrix

Zhu et al used the correlation parameter, correlated degree matrix to describe data correlation.¹¹ In real-world datasets, the extent of correlation between records is different. For example, some records are fully correlated, which means that these records are same records. Some records are partially correlated, which means that changing one record has a probability to change other related records. When the generation of data is not known by the curator or the data correlation is not easy to specify, it is efficient to denote the relation between records with the method of Pearson correlation. With this method, the extent of the impact of a record on another record can be quantified and it is defined as the correlated degree in Reference 11. With the notion of correlated degree, correlated sensitivity is proposed and defined in the correlated dataset. The definition of correlated sensitivity is given below.

Definition 8 (Correlated sensitivity). Correlated sensitivity for a query Q is defined as,

$$CS_q = \max_{i \in q} \sum_{j=0}^n |\delta_{ij}| \{ \| (Q(D^j) - Q(D^{-j})) \|_1 \}, \quad (8)$$

where D_j and D_{-j} are neighboring datasets that differ in record j ; q is a set of records; θ_{ij} is correlated degree between record i and record j . Correlated sensitivity describes the maximal impact on all records in the dataset due to the deletion of one record. Then, the correlation between records can be expressed with the correlated degree and formed into a correlated degree matrix to show all relationships between records.

Analysis: The advantage is that this method can be applied in many cases when there is no special data correlation known by the curator. This is because Pearson correlation can indicate the extent to which records are linear correlated without any prior knowledge of data generation. However, Pearson correlation is a method to evaluate linear relationship between records, hence it may not model the correlation accurately in some cases.

5.5 | Discussion of correlation parameters

The above methods show how to describe data correlation with correlation parameters in different settings for correlated datasets, and we make a comparison of these method in Table 3. We can note that most these methods need more background knowledge of data generation. The background knowledge is the number of correlated records, and it is not enough to calculate the exact correlations, leading to a higher noise level.²⁶ In Reference 12, the background knowledge is the number of correlated records L and the probabilistic dependence relationship R between the records. However, the effectiveness of this method relies on how well the correlation between records can be modeled and computed by the probabilistic dependence relationship. It is not easy to compute dependent coefficient accurately unless the probabilistic models of the data is known.

When the curator has no knowledge of how the data generated, the method proposed in Reference 11 can help identify data correlation. Comparing with the method in Reference 12, the method in Reference 11 may not have a better performance. This is because in Reference 11, the

TABLE 3 Comparison of correlation parameters underlying differential privacy

Correlation parameter	Sensitivity	Advantage	Challenge
k^{26}	$k\epsilon$	It is easy to compute.	It may introduce a large amount of noise to the output.
ρ_{ij}^{12}	$DS^Q = \max_i \sum_{j=C_{i1}}^{C_{in}} \rho_{ij} \Delta Q_j$	The correlation between record i and record j can be presented clearly.	The utility of correlated dataset depends on how well the dependence coefficient is computed.
ϵ^{27}	Δf	It considers all possible cases of data correlations.	The data correlation is not presented clearly with the correlation parameter.
δ_{ij}^{11}	$CS_q = \max_{i \in q} \sum_{j=0}^n \delta_{ij} \{ \ Q(D^j) - Q(D^{-j})\ _1 \}$	Correlated degree is able to measure the degree of data correlation.	Calculating correlated matrix degree is computational comparing with other methods.

sensitivity measures the effect on all records in the dataset according to the Pearson correlation, which may not describe data correlation accurately as the method in Reference 12. The above analysis shows that the background knowledge of how data are generated or correlated is essential when addressing the issue of privacy leakage on correlated data. Usually, with more background information of data correlation, such as References 12 and 27, the correlation can be computed more precisely, leading to a better performance in terms of noise level or data utility. In summary, the effectiveness of each method depends on the background knowledge known by the curator for correlated datasets.

6 | CORRELATION MODELS IN DIFFERENTIAL PRIVACY

In this section, we introduce two widely used models to describe complex data correlations: Gaussian correlation model and Markov chain model. In the previous section, we introduce some correlation parameters to describe data correlation for simple correlated datasets. However, it may still be difficult to measure some complex data correlations, like social network datasets and temporal correlated datasets. In this article, simply correlation refers to the correlation that can be described by correlation parameters, and complex correlation refers to the correlation that is difficult to be measured by correlation parameters and measured by correlation model.

6.1 | Gaussian correlation model

Gaussian correlation model is proposed to describe the complex data correlation and quantity the privacy loss in a new privacy model, called Bayesian differential privacy (BDP).¹³ First, we give the definition of Gaussian correlation model as,

Definition 9 (Gaussian correlation model). Let $G(x, W)$ be a weighted undirected graph, where the vertex $x_i \in X$ denotes the record i in X and the weight w_{ij} denotes the correlation between records i and j . Let $\mathbf{W} = (w_{ij})$ be weighted adjacent matrix which contains all weights; $\mathbf{D} = \text{diag}(w_1, \dots, w_n)$ be the diagonal matrix of $G(x, W)$ where $w_i = \sum_{j \neq i} w_{ij}$; $\mathbf{L} = \mathbf{D} - \mathbf{W}$ be the Laplacian matrix of $G(x, W)$,

$$\mathbf{L} = \mathbf{D} - \mathbf{W} = \begin{pmatrix} w_1 & -w_{12} & \dots & -w_{1,n} \\ -w_{12} & w_2 & \dots & -w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{1,n} & -w_{2,n} & \dots & w_n \end{pmatrix}. \quad (9)$$

The pair (x, \mathbf{L}) is called Gaussian correlation model, denoted as $G(x, \mathbf{L})$. The conditional joint probability of $\mathbf{x}_{-i} = \mathbf{x}_{[n] \setminus \{i\}}$ denoted as,

$$p(\mathbf{x}_{-i} | x_i) \propto \exp\left(-\frac{\mathbf{x}_{-i}^T \mathbf{L} \mathbf{x}_{-i}}{2}\right). \quad (10)$$

With Gaussian correlation model, unknown correlation between records can be described, and maximum correlated data can be computed. Gaussian correlation model is often used with Bayesian differential privacy. Overall, the main idea of Bayesian differential privacy is to connect the uncertain query answer with given records in a Bayesian way. The definition of Bayesian differential privacy is given below,

Definition 10 (Bayesian differential privacy¹³). Given an adversary $\mathcal{A} = \mathcal{A}(i, \mathcal{K})$ and a randomized perturbation mechanism $\mathcal{M}(x) = Pr(r \in \Omega|x)$ on the dataset X , Bayesian differential privacy leakage of \mathcal{M} related to \mathcal{A} is

$$BDPL_{\mathcal{A}}(\mathcal{M}) = \sup_{a,b,X_{\mathcal{K}},\Omega} \log \frac{p(\mathcal{M}(X) \in \Omega | X_i = a, X_{\mathcal{K}})}{p(\mathcal{M}(X) \in \Omega | X_i = b, X_{\mathcal{K}})}, \quad (11)$$

where $\mathcal{K} \subset [n] \setminus \{i\}$ be a tuple set and $\mathcal{A}(i, \mathcal{K})$ denotes the adversary with knowledge \mathcal{K} to attack x_i . Then we say \mathcal{M} satisfies ϵ -Bayesian differential privacy, if

$$\sup_{\mathcal{A}} BDPL_{\mathcal{A}} \leq \epsilon. \quad (12)$$

Bayesian differential privacy leakage shows the largest difference between $Pr(r \in S | x_i, x_{\mathcal{K}})$ and $Pr(r \in S | x'_i, x_{\mathcal{K}})$ and the leakage is bounded by ϵ . In the definition of Bayesian differential privacy, the background knowledge is $x_{\mathcal{K}}$, rather than x_{-i} in the differential privacy, which means the adversary in Bayesian differential privacy is weaker than the adversary in differential privacy. However, a weaker adversary may have a greater risk in the Bayesian differential privacy, which depends on the prior and posterior of the data distribution.

BDP vs DP There are two cases when Bayesian differential privacy is equivalent to differential privacy: (1) the data are independent in the dataset and the adversary has full knowledge of the dataset except the object of its attack; and (2) the adversary has full knowledge of the dataset except the object of its attack and the correlation between records.

Analysis: To describe data correlation in the setting of Bayesian differential privacy, Gaussian correlation model is used to measure data correlation. Advantages of Gaussian correlation model include: (1) any arbitrary correlation between records can be described by a weighted network with an arbitrary topology structure; (2) Gaussian correlation model assumes that the joint distribution of all records are Gaussian distribution. Since Gaussian distribution is easy to compute, the conditional distribution of some records when given other records can be easy to obtain; and (3) Gaussian correlation model can describe both infinite continuous data and discrete data. Due to these pros, Gaussian correlation model is suitable to describe data correlation in Bayesian differential privacy, which fits the background knowledge that the adversary partially knows knowledge of individuals, and the unknown individuals can be estimated by the Bayesian theorem.

6.1.1 | Applications of Gaussian correlation model in BDP

Some works studied correlated data in real-world applications based on Bayesian differential privacy. For example, Gaussian correlation model is used to describe the correlated data under the method of Bayesian differential privacy in Mobile CrowdSensing (MCS).¹⁴ MCS is a new sensing paradigm that people can use personal mobile devices to collect data from the surrounding environment.²⁸ Data collected from some mobile applications, such as traffic monitoring and advertisement delivering are from personal devices, and these information are likely to be correlated and leads to information leakage. In Reference 14, the correlated records come from the correlated group that is divided according to the relationship among participants, and the probabilistic relationship among sensing data records is modeled by the Gaussian correlation model. Furthermore, Gaussian correlation model is used to describe the correlation structure among sensing data with different prior knowledge. Besides, Liu et al analyzed the issue of location privacy preserving caused by the effects of temporal and spatial correlations based on Bayesian Geo-indistinguishability.²⁹ Another application uses Gaussian correlation model to describe data correlation in the game theory.³⁰ Data correlation can exist in multiple datasets. To preserve privacy in multiple datasets, Wu et al constructed a game model of multiple players (or publishers) to preserve the data privacy by controlling the privacy parameters of publishers. In multiple correlated datasets, the privacy of a dataset not only depends on its privacy budget but also depends on the privacy budget of other datasets. Gaussian correlation model is used to describe the background knowledge of the correlation between multiple datasets.

6.2 | Markov chain model

Markov chain model is a stochastic model used to describe a sequent of possible events. One feature of this model is that the probability of each event depends only on the state attained in the previous event. Due to this, Markov chain is widely used in modeling user mobility.^{31,32}

A Markov chain contains two components: states and transitions. More precisely, $P = p_1, \dots, p_n$ denote a set of states, in which each state corresponds to a value and a current value only depends on the previous one. A set of transitions, such as t_{ij} denotes the probability of moving from state p_i to state p_j . If an individual move from a state to an occasional position before returning to this state, then the transition from a state to itself may occur. The sum of the probabilities in each row of the transition matrix is 1. Here, we give an example of location data that is modeled by Markov chain.

TABLE 4 Transition matrix

	loc_1	loc_1	loc_1
loc_1	0.1	0.1	0.8
loc_1	0.2	0.3	0.5
loc_1	0.6	0.6	0.2

In Table 4, the first column denotes the time point t , and the first row denotes the time point $t + 1$. We can note that $Pr(l^t | l^{t+1}) = 0.6$, which means that one user is at loc_3 at time point t , then the probability of being loc_1 at time point $t + 1$ is 0.6.

6.2.1 | Application of Markov chain model

One popular application of Markov chain model is on temporal correlated datasets. When users' locations are continuously recorded, these records can be considered as a temporal correlated dataset. Cao et al studied the potential privacy loss under the temporal correlated dataset with a traditional mechanism.¹⁵ The background knowledge includes the individual information except the attacker's object and the temporal correlation. The parameters of Markov chain can be formed into a transition matrix to describe temporal correlation. Due to temporal correlation, temporal privacy leakage comes, and it is defined as temporal privacy leakage.

Definition 11 (Temporal privacy leakage). Let $D_{\mathcal{K}}^t$ be the tuple knowledge of the adversary A_i . Temporal privacy leakage of \mathcal{M}^t for the A_i is defined as follows.

$$TPL(A_i, \mathcal{M}^t) = \sup_{l_i^t, l_i^{t'}, r^1, \dots, r^T} \log \frac{Pr(r^1, \dots, r^T | l_i^t, D_{\mathcal{K}}^t)}{Pr(r^1, \dots, r^T | l_i^{t'}, D_{\mathcal{K}}^t)}, \quad (13)$$

where D_t and D'_t are neighboring dataset. l_i^t and $l_i^{t'}$ are two different values of user i 's data at time t and we have $D^t = D_{\mathcal{K}}^t \cup \{l_i^t\}$ and $D^{t'} = D_{\mathcal{K}}^t \cup \{l_i^{t'}\}$. Temporal privacy leakage includes backward privacy leakage and forward privacy leakage. Dividing the temporal privacy leakage in the t^t , then we can have backward privacy leakage and forward privacy leakage. The analysis shows that backward privacy leakage is likely to accumulate from previous privacy leakage and forward privacy leakage increases with future release.

Another work about temporal correlation is related to the location privacy. The background knowledge of data correlation is also modeled through the Markov chain and the neighboring dataset and sensitivity are redefined to fit the applicable problem.¹⁶ Let p_t^- be the prior probability of a user's location at time t . δ -location set is a set of minimum number of possible locations that the sum of prior probabilities is no more than $1 - \delta$, and the equation is given below.

$$\Delta X_t = \min \left\{ s_i \mid \sum_{s_i} p_t^-[i] \leq 1 - \delta \right\}. \quad (14)$$

The goal of δ -location set is used to form a dataset that reflects a set of probable locations the user might appear, which is equivalent to the dataset of outputs in differential privacy. We can note that the difference from traditional differential privacy lies in the neighboring dataset. The neighboring dataset in the new definition is any possible location x_1 and x_2 in the δ -location set. This definition states the output of location z_t is differentially private at time t for continual released locations under temporal correlations. Moreover, due to the modification of neighboring dataset, two dimensional space turns into multidimensional space. Based on the notion of convex hull, the sensitivity hull is proposed to capture the geometric meaning of sensitivity.

6.3 | Discussion of correlation models

The advantage of correlation model is that it can model complex data correlation and can be used to express in the form of the posterior probabilities with other background knowledge expressed in the form of prior probabilities. The comparison of correlation models is in Table 5. The experiments¹⁴ shows the proposed perturbation mechanism based on Bayesian differential privacy introduces less noise to the query results of correlated sensing data, comparing with Zhu's scheme¹¹ and Chen's scheme.¹⁰ For simple data correlation, adopting correlation parameters to model data correlation is easy to compute and to protect data privacy with a high utility of data. For example, the background knowledge of correlation is described by dependent coefficient¹² and by correlated degree matrix.¹¹ For complex data correlation, Gaussian correlation model and

Correlation model	Direction	Circulation	Applied situation
Gaussian correlation model	Directed graph	Acyclic	It can describe induced correlation.
Markov correlation model	Undirected graph	Cyclic	There is no direction for data correlation and can represent cyclic dependencies.

TABLE 5 Comparison of correlation models underlying differential privacy

Markov correlation model are considered as powerful methods to describe complex correlated dataset. For temporal correlated datasets, Markov chain model is a suitable method to present data correlation and guarantee the data privacy. Due to the advantage of correlation model, more and more works tend to adopt it to describe the correlated dataset. However, the best choice of correlation model depends on the specific type of correlations.

7 | THE FRAMEWORK OF PUFFERFISH

In this section, we first introduce Pufferfish and its variant, and then give mechanisms for Pufferfish. Pufferfish is a privacy framework that is proposed to cope with the issue of privacy leakage in correlated data. Kifer and Machanavajhala confirmed that correlated data is likely to leak more unexpected privacy⁸ under differential privacy. In order to break the limitations of correlated differential privacy, they proposed a new privacy model called Pufferfish which can provide different privacy definitions to the needs of customized applications. Pufferfish is quite different from differential privacy, and this privacy model includes the protected target, background knowledge of the adversary, and the neighboring dataset.

7.1 | Pufferfish

In Pufferfish, three components are used to specify the privacy requirements: S , a set of secrets that are needed to be protected; S_{pairs} , a set of secret pairs that need to be indistinguishable to the adversary; Θ , a class of distributions that represents how the data are generated. The definition of Pufferfish is described as follows.

Definition 12 (ϵ -Pufferfish⁹). A privacy mechanism \mathcal{M} satisfies ϵ -Pufferfish in a framework (S, \mathcal{Q}, Θ) if for datasets $X \sim \theta$ and for all secret pairs $(s_i, s_j) \in S_{pairs}$ and for all possible output $\omega \in \Omega$,

$$e^{-\epsilon} \cdot \frac{P(s_i|\theta)}{p(s_j|\theta)} \leq \frac{P(s_i|\mathcal{M}(X) = \omega, \theta)}{P(s_j|\mathcal{M}(X) = \omega, \theta)} \leq e^{-\epsilon} \cdot \frac{P(s_i|\theta)}{p(s_j|\theta)}, \quad (15)$$

where θ ($\theta \in \Theta$) is to represent a probability distribution which denotes the attacker's probabilistic belief and background knowledge. $P(s_i|\theta)$ and $P(s_j|\theta)$ are conditional probabilities and the attacker has uncertainty about s_i and s_j ($P(s_i|\theta) \neq 0$, $P(s_j|\theta) \neq 0$). When the ϵ is small, seeing the sanitized output w leaks nearly no information to the attacker who is trying to figure out whether s_i or s_j is true.

There are two advantages of why Pufferfish privacy framework is able to deal with correlated data: (1) Pufferfish is able to hide private information against data correlation in the dataset since data correlation is assumed to be specified. (2) Pufferfish is capable of dealing with a large number of correlated records, and it can provide a high utility of data. This is because the sensitive information is specified, and various discriminative pairs can be used to protect the sensitive information.

Pufferfish vs DP Three main differences are between Pufferfish and differential privacy. (1) The information that we want to protect in Pufferfish is specified and can be various information, while the information we want to protect in differential privacy is whether one user (or record) is in the dataset in the ϵ -DP. (2) The discriminative pairs can be various in Pufferfish, while the discriminative pair can be regarded as "one record is in the dataset" and "one record is not in the dataset" in ϵ -DP. (3) Assumptions are made in data generation in Pufferfish, while data are assumed to be independent in differential privacy. When satisfying some conditions, differential privacy can be regarded as a special case of Pufferfish. Hence, Pufferfish is a kind of generalization of differential privacy, which provides rigorous statistical guarantee to prevent the information leakage.

BDP vs Pufferfish Bayesian differential privacy can be considered as a special case of Pufferfish. When the potential secrets to be the set of all possible values of records in the dataset and discriminative pairs to be the corresponding set of all pairs of secrets, and data are generated by the Bayesian network, and then Pufferfish transforms into Bayesian differential privacy.

7.2 | Blowfish

Based on the framework of Pufferfish, another privacy model Blowfish privacy is proposed to provide a rich interface for implementation.¹⁷ The key feature of Blowfish is a policy that the sensitive information is specified, and adversary knowledge is in the form of a set of deterministic constraints Q that are known by the public. With these policies, mechanisms can be expected to permit more utility since not all properties of an individual need to be kept secret and adversarial attacks on correlated records can be limited due to public known constraints. The definition of Blowfish is given below.

Definition 13 ((ϵ, P) Blowfish). Given a privacy budget ϵ and a policy $P(\mathcal{T}, \mathcal{G}, \mathcal{I}_Q)$, a randomized mechanism \mathcal{M} satisfies (ϵ, P) -Blowfish privacy if for any pairs of neighboring datasets D, D' and for all possible outputs $\omega \subset \Omega$, we have

$$\Pr[\mathcal{M}(D \in \Omega)] \leq e^\epsilon \Pr[\mathcal{M}(D' \in \Omega)]. \quad (16)$$

Here, the policy $P(\mathcal{T}, \mathcal{G}, \mathcal{I}_Q)$ is a new notion proposed in Blowfish. For a policy $P(\mathcal{T}, \mathcal{G}, \mathcal{I}_Q)$, \mathcal{T} is the domain of the dataset; $\mathcal{G} = (V, E)$ is a discriminative graph used to present the secret pairs, in which $V \subset \mathcal{T}$ and $E \subset \mathcal{T} * \mathcal{T}$ denotes values in the domain that an adversary must not distinguish between them; \mathcal{I}_Q denoting the set of datasets that are possible under the publicly known constraints Q .

Blowfish vs Pufferfish Blowfish borrows the notion of a set of specified secrets that need protection from Pufferfish. In Pufferfish, the adversary knowledge is defined as the assumption about how data are generated, and it tends to be described by the probabilistic correlation function. In Blowfish, the knowledge of correlation is defined as a set of publicly known constraints. This indicates that Blowfish without constraints is equivalent to Pufferfish in the case of adversaries who believe records in the dataset are independent. Both Pufferfish and Blowfish are helpful for the data publisher who can customize privacy definitions by carefully defining sensitive information and background knowledge.

Blowfish vs DP Differential privacy can be considered as a special case of Blowfish when two conditions are satisfied: (1) The second parameter of the policy \mathcal{G} is the complete graph on the domain, instead of a part of the domain. (2) There is no publicly known constraints on the dataset.

7.3 | Mechanisms for Pufferfish

7.3.1 | The Wasserstein mechanism

Every privacy model needs corresponding mechanisms to perform. Some mechanisms proposed are proposed to implement Pufferfish. Since there is no general mechanism for the framework of Pufferfish, Wang and Song first proposed general mechanisms that can apply in the Pufferfish.^{33,34} They used the Wasserstein distance as a metric to measure the maximum distance between distributions $P(\mathcal{M}(X)|s_i, \theta)$ and $P(\mathcal{M}(X)|s_j, \theta)$ for a secret pair (s_i, s_j) . Here, the maximum distance for a secret pair is similar to the maximal difference between the query result on neighboring datasets in differential privacy. Hence, the goal of the Wasserstein mechanism is to measure the maximum distance for a secret pair. The definition of maximum distance is given below.

Definition 14 (∞ -Wasserstein distance). Suppose for some (s_i, s_j) and θ , $P(f(X)|s_i, \theta)$ can be transformed into $P(f(X)|s_j, \theta)$. Then the maximum distance of two probability mass function is Wasserstein distance which is given as

$$W_{i,j,\theta} = W_\infty(P(f(X)|s_i, \theta), P(f(X)|s_j, \theta)), \quad (17)$$

where f is the query.

Adding Laplace noise with the scale of $W_{i,j,\theta}$ to the query answers will guarantee the odds ratio of s_i to s_j in the range of $[e_{-\epsilon}, e_\epsilon]$. The odds ratio of s_i to s_j is the probability of s_i being s_j after the attacker seeing the sanitized output. After iterating all pairs $(s_i, s_j) \in \mathcal{Q}$ and all $\theta \in \Theta$, the maximum Wasserstein distance can be obtained. In Wasserstein mechanism, the amount of noise added to the correlated dataset is similar to the form in Laplace mechanism,

$$Z = \text{Lap}\left(\frac{W}{\epsilon}\right), \quad (18)$$

where $W = \sup_{(s_i, s_j) \in \mathcal{Q}, \theta \in \Theta} W_\infty(u_{i,\theta}, u_{j,\theta})$.

7.3.2 | The Markov Quilt mechanism

As Wasserstein mechanism may have a complex computation, another mechanism called the Markov Quilt mechanism, based on Bayesian network is proposed for Pufferfish. As we mentioned in Section 6, Bayesian network is a popular method to describe data correlation. Hence, there is also a mechanism which uses Bayesian network to design mechanism in Pufferfish.

The mechanism will attempt to find a proper set X_A such that X_i has low max-influence on X_A under Θ . Here, X_A can be regarded as a set of nodes that have correlation with X_i . First, we need to quantify the extent of changing the value of a variable $X_i \in X$ can affect a set of nodes $X_A \in X$. The maximum influence of the variable X_i on a set of variables X_A is defined as

$$e(X_A|X_i) = \max_{a,b \in X} \sup_{\theta \in \Theta_{X_A \in X}} \log \frac{P(X_A = x_A | X_i = a, \theta)}{P(X_A = x_A | X_i = b, \theta)}. \quad (19)$$

Hence, the maximum influence is the maximum divergence between distributions $P((X_A = x_A | X_i = a, \theta))$ and $P((X_A = x_A | X_i = b, \theta))$. In order to find the set X_A efficiently, Markov Quilt is proposed to find the set and the definition is given below.

Definition 15 (Markov Quilt). A set of nodes X_Q is Markov Quilt set for a node X_i if the following conditions are satisfied in the Bayesian network $G = (X, E)$. (1) Deleting the X_Q can separate G into two sets X_N and X_R and thus $X = X_N \cup X_R \cup X_Q$ and $X_i \in X_N$. (2) X_R is independent of X_i conditioned on X_Q .

The main insight behind the Markov Quilt mechanism is that if X_i and X_j are distant from each other, then X_j is largely independent of X_i . Thus, adding noise to the local nodes can obscure the effect of X_i in the query result. Using Markov Quilt, it is efficient to find X_R which is a set of remote nodes far from X_i and X_N which is a set of local nodes near X_i .

7.4 | Discussion of Pufferfish

The drawback of differential privacy is that it is not enough to erase the participation of a single individual's private value when there are multiple records correlated with each other. Hence, another privacy model, Pufferfish is proposed to cope with the issue of privacy loss on correlated data. The key of Pufferfish is that it considers how the data are generated and the knowledge of potential attackers. Inspired by Pufferfish, more privacy models study the privacy leakage with the consideration of the background knowledge of data generation. Also, increasing mechanisms for Pufferfish are proposed for the application of this privacy model.

8 | FUTURE DIRECTIONS

In this section, we will introduce some promising future directions on correlated differential privacy. In Section 5 to 7, we summarize most works on correlated data under different privacy models. Three lines include: (1) describing data correlation with correlation parameters, and (2) using correlation models to describe complex data correlation under the setting of differential privacy, and (3) using the framework of Pufferfish to measure data correlation. However, there are still some issues on correlated data that have not been considered yet.

8.1 | Correlated differential privacy in machine learning

Differential privacy is privacy model also used in artificial intelligence to prevent data leakage,³⁵⁻³⁷ especially in machine learning.^{38,39} Chaudhuri et al provided an output perturbation⁴⁰ and objective perturbation mechanism.⁴¹ Abadi et al studied differentially private stochastic gradient descent mechanisms, where noise is added to gradients.⁴² However, data correlation has not been considered when adding noise during the learning. Data correlation in the training data is likely to lead to more changes on the training result, and consequently the adversary is able to obtain more information. So far, Zhang et al have proposed a feature selection method to reduce data correlation in the training dataset.⁴³ Privacy loss due to data correlation in machine learning still have some open issues, such as quantifying the privacy loss due to data correlation.

8.2 | Multiple correlated relationships

Real-world datasets are likely to be multiple relationships between records. As the example illustrated in the Section 4, two types of relationships can be in the location dataset, such as the user mobility pattern and the social relationship. In previous studies, researches assumed that only one correlation is in the dataset, which is not practical. One intuitive way to cope with multiple correlations in the dataset is to treat different kinds of correlations as the same correlation. And then methods illustrated in Sections 5 to 7 can be used to deal with the correlated records and protect privacy leakage. Obviously, it is not an optimal solution because the number of correlated records are enlarged and more noise will be added to the

dataset. Hence, the utility of datasets will not be desirable. The method to model multiple correlated relationships and the mechanism to guarantee differential privacy under multiple correlations are open issues that need to be explored in the future.

8.3 | Correlations in different datasets

Currently, most research focuses on the issue of correlated data in a dataset, while correlated data can be distributed in different datasets. The sensitive information may be leaked when multiple entities publish their data sequentially. If the adversary has enough background information of the dataset, the privacy level of these datasets will be degraded especially when records are correlated in different datasets. The privacy level of datasets not only depends on its privacy parameter, but also depends on the privacy parameter of its neighboring datasets. Most studies may be not applicable when correlated data are in different datasets, like the framework of Pufferfish. This is because there are multiple entities, and they need to negotiate with each other and then make the best choices according to each publisher's privacy request and the utility of whole datasets. Wu et al constructed a game model of multiple players and study the uniqueness of pure Nash Equilibrium.³⁰ However, there are still many issues that need to be considered, like the weight of each publisher and each publisher's own privacy requirement. One promising method of this issue can be modeled as a multiagent systems to achieve the optimal data utility for multiple data entities.

8.4 | Continuous query release

When a dataset deals with a large number of queries, data privacy is more vulnerable since the adversary may infer more information via multiple queries. This will leak more information, especially when the records in the dataset are not independent. Zhu et al studied the continuous query release for the correlated dataset.⁴⁴ An iterative-based mechanism⁴⁵ is adopted to answer a set of queries on the correlated datasets. During the process of continuous queries, when a query finds an obvious difference between the current dataset and true dataset, the mechanism will have an update on the current dataset in next query. Continual query release is a difficult topic in privacy preserving, especially for correlated datasets. There are still many unsolved problems, for example, how to deal with various types of statistical queries, and how to incorporate with multiple correlations for continuous query release.

8.5 | Inference attacks on correlated data

As we mentioned, data correlation in datasets can leak more information than expected when using differential privacy. This makes inference attacks more easier on correlated datasets. Even though strong protection provided by differential privacy obfuscates the original data using stochastic noise to avoid privacy leakage, privacy leakage is still breached by some inference attacks. Shao et al proposed a novel location inference attack framework, which is able to recover multiple trajectories from differentially private trajectory data using the structured sparsity model.⁴⁶ In the future, more and more attacks are aiming on correlated differential privacy, and how to defend these attacks is a challenging topic.

9 | CONCLUSION

This article presents a survey on correlated data under different privacy models. Since correlated dataset are expected to leak more privacy than expected, many works focus on how to address with this issue. Basically, these research are mainly classified into three streams: the first focuses on how to use parameters to describe the correlation in differential privacy; the second uses correlation models to describe data correlation in differential privacy; the last method is a new privacy model, called Pufferfish to protect data privacy while keeps a good utility of datasets. In the first two lines, we analyze different correlation parameters and correlation models of how to describe data correlation, and compare cons and pros of these methods. Simple data correlation can be described by correlation parameters, and complex data correlation can be described by correlation models. In the last research line, we analyze Pufferfish, compare the difference of this model with differential privacy, and present mechanisms for Pufferfish. Our goal is to provide an overview of existing work on the issue of correlated dataset. Finally, we propose some interesting issues that have not been studied or solved in correlated differential privacy.

ACKNOWLEDGEMENT

This work is supported by an ARC Discovery Project (DP190100981, DP200100946) from the Australian Research Council, Australia.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this article.

ORCID

Tao Zhang  <https://orcid.org/0000-0003-4696-641X>

REFERENCES

1. Yu S. Big privacy: challenges and opportunities of privacy study in the age of big data. *IEEE Access*. 2016;4:2751-2763.
2. Wang K, Chen R, Fung B, Yu P. Privacy-preserving data publishing: a survey on recent developments. *ACM Comput Surv*. 2010;42(4).
3. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. *IEEE Access*. 2016;4:1821-1834.
4. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst*. 2002;10(05):557-570.
5. Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M. l-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data (TKDD)*. 2007;1(1):3-es.
6. Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. Paper presented at: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul; 2007:106-115; IEEE.
7. Dwork C. Differential privacy: a survey of results. Paper presented at: Proceedings of the International Conference on Theory and Applications of Models of Computation, Berlin, Heidelberg; 2008:1-19.
8. Kifer D, Machanavajjhala A. No free lunch in data privacy. Paper presented at: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, New York; 2011:193-204.
9. Kifer D, Machanavajjhala A. Pufferfish: a framework for mathematical privacy definitions. *ACM Trans Database Syst (TODS)*. 2014;39(1):3.
10. Chen R, Fung BC, Philip SY, Desai BC. Correlated network data publication via differential privacy. *VLDB J*. 2014;23(4):653-676.
11. Zhu T, Xiong P, Li G, Zhou W. Correlated differential privacy: hiding information in non-IID data set. *IEEE Trans Inf Forens Sec*. 2015;10(2):229-242.
12. Liu C, Chakraborty S, Mittal P. Dependence makes you vulnerable: differential privacy under dependent tuples. *NDSS*. 2016;16:21-24.
13. Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. Paper presented at: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, New York; 2015:747-762.
14. Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. *IEEE Trans Big Data*. 2017. https://ieeexplore.ieee.org/abstract/document/8126240?casa_token=uitv3mKEYnIAAAA: vde_cMUp5xSj7MhrcAoP695zj2DDtZrjzY87KPafllgBTPtfwQGv7V_sSkpXH1-qrUlkAoVK
15. Cao Y, Yoshikawa M, Xiao Y, Xiong L. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Trans Knowl Data Eng*. 2018;31(7):1281-1295.
16. Xiao Y, Xiong L. Protecting locations with differential privacy under temporal correlations. Paper presented at: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, New York; 2015:1298-1309.
17. He X, Machanavajjhala A, Ding B. Blowfish privacy: tuning privacy-utility trade-offs using policies. Paper presented at: Proceedings of the 2014 ACM SIGMOD International Conference on Management of data; 2014:1447-1458.
18. Dwork C. Differential privacy in new settings. Paper presented at: Proceedings of the 21 Annual ACM-SIAM Symposium on Discrete Algorithms, USA; 2010:174-183.
19. Dwork C. A firm foundation for private data analysis. *Commun ACM*. 2011;54(1):86-95.
20. Dwork C, Roth A. *The Algorithmic Foundations of Differential Privacy*. Vol 9. Hanover, MA, USA: Now Publishers Inc; 2014.
21. Sarwate AD, Chaudhuri K. Signal processing and machine learning with differential privacy: algorithms and challenges for continuous data. *IEEE Sig Process Mag*. 2013;30(5):86-94.
22. Zhu T, Li G, Zhou W, Philip SY. Differentially private data publishing and analysis: a survey. *IEEE Trans Knowl Data Eng*. 2017;29(8):1619-1638.
23. Desfontaines D, Pejó B. Sok: differential privacies. *Proc Priv Enhanc Technol*. 2020;2020(2):288-313.
24. McSherry F, Talwar K. Mechanism design via differential privacy. Paper presented at: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science; 2007.
25. Bozkir E, Günlü O, Fuhl W, Schaefer RF, Kasneci E. Differential privacy for eye tracking with temporal correlations; 2020. arXiv preprint arXiv:2002.08972.
26. Chen R, Fung BC, Yu PS, Desai BC. Correlated network data publication via differential privacy. *Int J Very Large Data Bases*. 2014;23(4):653-676.
27. Zhao J, Zhang J, Poor HV. Dependent differential privacy for correlated data. Paper presented at: Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps); 2017:1-7.
28. Ganti RK, Ye F, Lei H. Mobile crowdsensing: current state and future challenges. *IEEE Commun Mag*. 2011;49(11):32-39.
29. Liu B, Zhu T, Zhou W, Wang K, Zhou H, Ding M. Protecting privacy-sensitive locations in trajectories with correlated positions. Paper presented at: Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM); 2019:1-6.
30. Wu X, Wu T, Khan M, Ni Q, Dou W. Game theory based correlated privacy preserving analysis in big data. *IEEE Trans Big Data*. 2017. <https://ieeexplore.ieee.org/document/7920335>
31. Gamba S, Killijian MO, Prado-Cortez DMN. Next place prediction using mobility Markov chains. Paper presented at: Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility, New York; 2012:1-6.
32. Mathew W, Raposo R, Martins B. Predicting future locations with hidden Markov models. Paper presented at: Proceedings of the 2012 ACM Conference on Ubiquitous Computing; 2012:911-918.
33. Cao Y, Yoshikawa M, Xiao Y, Xiong L. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Trans Knowl Data Eng*. 2018;31(7):1281-1295.
34. Wang Y, Song S, Chaudhuri K. Privacy-preserving analysis of correlated data. *CoRR*. 2016;abs/1603.03977.
35. Song S, Wang Y, Chaudhuri K. Pufferfish privacy mechanisms for correlated data. Paper presented at: Proceedings of the 2017 ACM International Conference on Management of Data; 2017:1291-1306.

36. Ye D, Zhu T, Zhou W, Philip SY. Differentially private malicious agent avoidance in multiagent advising learning. *IEEE Trans Cybern.* 2019. <https://ieeexplore.ieee.org/document/8685696>
37. Zhu T, Philip SY. Applying differential privacy mechanism in artificial intelligence. Paper presented at: Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas; 2019:1601-1609.
38. Zhu T, Ye D, Wang W, Zhou W, Yu P. More than privacy: applying differential privacy in key areas of artificial intelligence. *IEEE Trans Knowl Data Eng.* 2020;1-1. <https://ieeexplore.ieee.org/document/9158374>
39. Zhu T, Xiong P, Li G, Zhou W, Philip SY. Differentially private model publishing in cyber physical systems. *Future Generat Comput Syst.* 2018;108:1297-1306.
40. Yang M, Zhu T, Liu B, Xiang Y, Zhou W. Machine learning differential privacy with multifunctional aggregation in a fog computing architecture. *IEEE Access.* 2018;6:17119-17129.
41. Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. *Adv Neural Inf Process Syst.* 2009;289-296. <https://papers.nips.cc/paper/3486-privacy-preserving-logistic-regression>
42. Chaudhuri K, Monteleoni C, Sarwate AD. Differentially private empirical risk minimization. *J Mach Learn Res.* 2011;12(Mar):1069-1109.
43. Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. Paper presented at: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016:308-318.
44. Zhang T, Zhu T, Xiong P, Huo H, Tari Z, Zhou W. Correlated differential privacy: feature selection in machine learning. *IEEE Trans Ind Inform.* 2019;16(3):2115-2124.
45. Zhu T, Li G, Xiong P, Zhou W. Answering differentially private queries for continual datasets release. *Future Generat Comput Syst.* 2018;87:816-827.
46. Hardt M, Rothblum GN. A multiplicative weights mechanism for privacy-preserving data analysis. Paper presented at: Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, Las Vegas; 2010: 61-70.
47. Shao M, Li J, Yan Q, Chen F, Huang H, Chen X. Structured sparsity model based trajectory tracking using private location data release. *IEEE Trans Dependable Sec Comput.* 2020.