This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2017.2701817, IEEE Transactions on Big Data

Game Theory Based Correlated Privacy Preserving Analysis in Big Data

Xiaotong Wu, *Student Member, IEEE,* Taotao Wu, *Student Member, IEEE,*
Maqbool Khan, *Student Member, IEEE,* Qiang Ni, *Senior Member, IEEE,*
and Wanchun Dou, *Member, IEEE*

Abstract—Privacy preservation is one of the greatest concerns in big data. As one of extensive applications in big data, privacy preserving data publication (PPDP) has been an important research field. One of the fundamental challenges in PPDP is the trade-off problem between privacy and utility of the single and independent data set. However, recent research has shown that the advanced privacy mechanism, i.e., differential privacy, is vulnerable when multiple data sets are correlated. In this case, the trade-off problem between privacy and utility is evolved into a game problem, in which payoff of each player is dependent on his and his neighbors' privacy parameters. In this paper, we firstly present the definition of correlated differential privacy to evaluate the real privacy level of a single data set influenced by the other data sets. Then, we construct a game model of multiple players, in which each publishes data set sanitized by differential privacy. Next, we analyze the existence and uniqueness of the pure Nash Equilibrium. We refer to a notion, i.e., the price of anarchy, to evaluate efficiency of the pure Nash Equilibrium. Finally, we show the correctness of our game analysis via simulation experiments.

Index Terms—Differential privacy, privacy preservation, game theory, big data.

1 Introduction

With rapid development of information and communication technologies, people have stepped into the age of big data. Every day, a tremendous amount of raw data from various sources, such as social websites, online shopping and transportations, is generated rapidly. The occurence of big data brings us a great opportunity to significantly improve our insights in all aspects of human society. Even though, privacy preservation has been one of the most serious problems, which hampers further growth of big data [1], [2], [3]. As one of extensive applications in the big data processing, privacy-preserving data publication (PPDP) has attracted a lot of attention from academia and industry [4], [5]. That is, data needs to be sanitized via privacy preservation mechanisms so as to provide the privacy guarantee against the leakage of the sensitive information before it is published to the public. Among the existing privacy mechanisms (e.g., k-anonymity [6], l-diversity [7], t-closeness [8]), differential privacy [9], [10] has emerged as a rigorous mathematical definition of privacy requirement, which ensures that adding or deleting a single record doesn't affect the outcome of any analysis.

While those privacy-preserving mechanisms prevent users' privacy leakage, it inevitably causes the utility loss of data [11]. Naturally, one of the most important challenges is to research the trade-off problem between privacy and utility, which aims to maximize the data utility compromised by privacy constraints. There have been a large number of studies about the trade-off problem, which is not limited to big data. In general, it is divided into two categories. The first one is to compare the utility of different privacy-preserving mechanisms [12], [13], [14], while the second one is to choose the optimal privacy parameter to maximize the utility [15], [16]. For these issues, it always assumes that the data sets are independent with each other in terms of privacy. That is, it just needs to consider the privacy-utility trade-off of a single data set without the privacy influence of the other data sets.

However, in the age of big data, someone's information may be stored by multiple data sets with the same or similar types. For example, someone uploads his true trajectory to the social website (e.g., Twitter and/or Facebook). Meanwhile, this trajectory is also submitted to a third party for location-aware applications in location based services. In recent years, there have been a series of research works [11], [17], [18], [19], [20] showing that differential privacy is vulnerable when multiple data sets are correlated, though it provides a strong privacy guarantee to the independent data sets. In other words, the privacy level of some data set is influenced not only by this data set's privacy parameter, but also by its neighbor data sets' privacy parameters. Here, we take a simple example similar to the literature [11], [20] to illustrate this correlated privacy relationship, which is presented as follows.

Example 1. Consider a relatively extreme example, which has two different groups of data sets about health information. (1) The first group has ten same data sets, which are strongly correlated with each other. (2) The second group has ten fully disjoint data sets, all of which

- *X. Wu, T. Wu, M. Khan and W. Dou are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, P. R. China. W. Dou is the corresponding author.*
  E-mail: {wxt199003, wutaotaoxpy}@gmail.com, maqbool@163.com, douwc@nju.edu.cn
- *Q. Ni is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K.*
  E-mail: q.ni@lancaster.ac.uk

Manuscript received X, X; revised X, X.

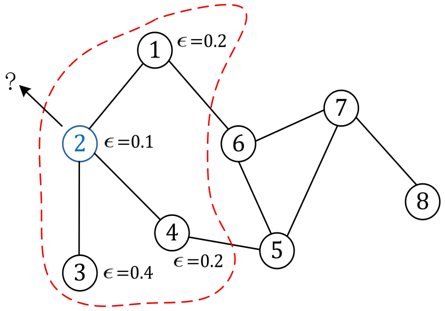


Fig. 1: Correlated data publication via ϵ-differential privacy

have no correlation with each other. Suppose that an "adversary" knows all the information of individuals in data sets except Jack. This adversary attempts to ask "whether or not Jack is in the data sets?". In order to protect Jack's health information, we use Laplace mechanism giving ϵ-differential privacy, where the final answer to the adversary's query in every data set is the true query plus the noise query $Laplace(1/\epsilon)$. In the first group, the sum of the query result is 0 or 10. In the second group, the sum of the query result is 0 or 1, since only one data set includes Jack's health information. As a result, to guarantee privacy of the same level in two groups, the noise added to each data set in the first group is $Laplace(1/0.1\epsilon)$, while the noise in the second group is $Laplace(1/\epsilon)$. $0.1\epsilon < \epsilon$ implies that the privacy requirement of each data set in the first group is higher than that in the second group.

Under the case of the correlated data sets in terms of privacy, it should not be just limited to analyzing the trade-off problem between privacy and utility in a single data set. More precisely, the trade-off problem in the single data set can be evolved into a game problem among different individuals, who publish their data sets via the perturbed process i.e., the privacy-preserving mechanism and the privacy parameter, respectively. As shown in Fig. 1, the privacy level of data set 2 with ϵ-differential privacy is dependent not only on his own parameter $\epsilon = 0.1$, but also on the privacy parameters of his neighbors (e.g., data set 1 with $\epsilon = 0.2$). In this case, what is the true privacy level of data set 2? More importantly, the fundamental question is: "*how many privacy parameters does each data set adopt to maximize his utility in correlated data sets?*".

1.1 Objective and Challenges

The objective of this paper is to analyze the differential privacy parameter choices in correlated network data sets and maximize the utility of each data set. However, studying this problem still suffers from several main challenges listed as follows.

- The first challenge is how to describe the correlated relationship between different data sets in terms of privacy. Especially, when the correlated data sets change the privacy parameters, what is the influence on the objective data set?
- The second challenge is how to design the reasonable measure about the utility of a sanitized data set. In general, the higher the privacy parameter, the worse the utility of data is. According to the utility measure, the following is to compute the utility of data.
- The third challenge is how to evaluate data owners' value of privacy. Although someone has taken some privacy preserving actions, it is still possible to cause the privacy leakage. Therefore, it is a key to measure the value of privacy.

1.2 Contributions

The contributions in this paper can be summarized as follows.

- In order to construct a game model of multiple players to release their own data sets which each is sanitized by anonymization mechanisms, we measure the differential privacy relationship of correlated data sets, the utility of sanitized data, and the value of privacy.
- We make game analysis based on the game model. We demonstrate the sufficient conditions of the existence and uniqueness of the pure Nash Equilibrium in the game. Especially, we show that the existence of the pure Nash Equilibrium fully depends on some player and his neighbors, when considering a single player.
- We utilize the price of anarchy to evaluate the efficiency of the pure Nash equilibrium. We demonstrate the lower bound of the price of anarchy.

The remainder of this paper is organized as follows. In Section 2, we briefly survey the related works. In Section 3, we introduce the preliminaries about differential privacy and its composition properties. In Section 4, we present a definition of differential privacy in correlated datasets. In Section 5, we construct a game model of correlated privacy. In Section 6, we make game analysis. In Section 7, we evaluate the correctness of our game analysis via simulation experiments. Finally, we conclude our work in Section 8.

2 Related Work

2.1 Tradeoff Between Privacy and Utility

The tradeoff between privacy and utility of data is one of the most challenging problems in privacy-preserving data publication. In detail, it aims to maximize the utility, compromised by a certain number of privacy requirements. There have been a large number of works to study this problem. The works can be split into two categories, i.e., the privacy mechanism design and the privacy parameter choice.

For privacy mechanism design, researchers generally discussed this from the perspective of theory and application, respectively. The first and key step in theoretical analysis is how to evaluate utility of anonymous data, regardless of adopting which privacy-preserving mechanism. To this end, Lin et al. [13], [14], [21] continuously studied the information-preserving properties of utility measures and then proposed a series of utility axioms. Based on these utility axioms, they [22] demonstrated that the solution to maximize the utility is a mechanism, whose matrix form consists of linearly independent rows. However, the proposed utility measure is not necessarily practical, since it ignores the actual application of data (e.g., prediction in data mining). Li et al. [23] proposed a unifying framework, i.e., membership privacy, to compare the utility of the existing privacy notions. Even though, to the best of our knowledge, the most of the previous research still focuses on designing the optimal mechanism based on differential privacy from the perspective of application.

As a rigorous standard of privacy definition, differential privacy has attracted a large number of researchers to design the optimal mechanism for privacy preservation in different scenarios. Ghosh et al. [12] pointed out that for any count query and differential privacy, the geometric mechanism is simultaneously expected loss-minimizing of the data set, subject to the differential privacy constraint. However, some literature showed that no universally optimal mechanism exists in some scenarios, such as histograms and two or more count queries [24] and count-range queries [25]. Brenner et al. [24] found that there is a universally optimal mechanism when it is a single count query. Zeng et al. [25] showed that the optimal differentially private mechanisms exist in “threshold” queries. Yuan et al. [26] showed that it is very difficult to design an optimal strategy maximizing the result accuracy, since this is a complex constrained and nonconvex optimization program. Meanwhile, they utilized (ϵ, δ)-differential privacy to design a suitable mechanism.

For privacy parameter choice, the intuitive objective is to obtain more utility, constrained by a certain amount of privacy requirements. Although differential privacy generally has no assumptions about data, Kifer et al. [11] utilized no free lunch theorem to state that it is not possible to provide privacy and utility under no assumptions about the generation of data. They found that differential privacy still possibly causes privacy breaches when the notion of participation varies. In order to maximize the utility subject to the privacy constraints, He et al. [15] proposed a new definition of Blowfish Privacy to tune privacy-utility tradeoffs. Xu et al. [16] utilized a contract theoretic approach and designed optimal contracts to direct data owners how to decide their privacy parameters, when there is a collector to collect data from owners. In addition, some researchers attempt to design utility-aware privacy-preserving mechanisms. Makhdoumi et al. [27] propose utility-aware privacy mechanisms to defend inference attacks in the case where statistical knowledge is uncertain. Prasser et al. [28] proposed a utility-driven heuristic search strategy to anonymize high-dimensional datasets.

2.2 Correlated Privacy Definitions

Though differential privacy provides a strong privacy guarantee, it still suffers from some new challenges. One of the severe challenges is the increase of privacy breach risk in correlated data, which was firstly found in [11]. In order to remedy this defect, Kifer et al. [17] utilized differential privacy and defined a new privacy framework named Pufferfish, which considers the correlated data. Correspondingly, the existing differential-privacy-based mechanisms are not suitable for this new privacy framework. To this end, Wang et al. [29] proposed the Wasserstein Mechanism, which can be applied into any Pufferfish framework. By adding an extra parameter to evaluate the extent of correlation, Chen et al. [18] demonstrated that differential privacy still provides privacy guarantee in the correlated data and needs some adjustment. Considering the correlated level between records, Zhu et al. [19] defined a correlated sensitivity and designed a correlated data releasing mechanism in non-IID data set. Yang et al. [20] proposed a new definition called Bayesian differential privacy, by which a probabilistic perturbation algorithm is designed to evaluate the privacy level.

2.3 Game Theory

As a useful analysis tool, game theory [30] has been widely applied into data privacy game to analyze users’ behavior. For example, Freudiger et al. [31] constructed and analyzed a non-cooperative game model, in which each player, i.e., mobile node, attempts to maximize its location privacy with a minimum cost. Wang et al. [32] constructed a zero-sum stochastic game, in which users who utilize their contexts to obtain personalized services have the interactive privacy competition with each other. Shokri [33], [34] proposed user-centric obfuscation strategies to maximize utility with satisfying users’ privacy requirements, when users send their data to a potentially untrusted service provider. Based on game theory, Wang et al. [35] design a payment mechanism to control the quality of data from different users, who each considers his own privacy requirement. In addition, some auction mechanisms are proposed, in which privacy is viewed as a commodity to be sold [36], [37].

On the other hand, some researchers studied the game of multiple different roles. Xiao et al. [38] studied the interactions between a subjective cloud storage defender and a subjective Advanced Persistent Threats (APT) attacker. By game analysis, they also proposed a Q-learning based APT defense strategy for cloud storage. Chessa et al. [39] formulated a game-theoretic model, in which individuals take control over participation and data analysts set requirements for data precision. Gao et al. [40] formulated a repeated public-goods game between cloud users and cloud service providers about cloud data security and privacy protection. Bernhard et al. [41] focused on the privacy of voting schemes and proposed a game-based definition of privacy. Chen et al. [42] analyzed the effect of privacy concerns on the behavior of selfish agents and gave the Nash equilibrium analysis. Here, privacy concerns are usually explained as bringing loss for utility functions.

Although there have been a number of game studies about data privacy, these studies are just limited to some special scenarios, e.g., location privacy. Unfortunately, this cannot be applied into other general scenarios. Besides, correlated data makes self-interest users attempt to maximize their own utility, while their privacy is influenced with each other. To the best of our knowledge, there is little work to utilize game theory to analyze users’ behavior in correlated data sets. Therefore, we construct a game model and make game analysis.

3 Preliminaries

In this section, we mainly introduce the preliminaries about differential privacy and its composition properties.

3.1 Differential Privacy

Differential privacy is not only a gold standard of privacy definition, but also a rigorous mathematic definition. We present a formal definition of ϵ-differential privacy as follows.

Definition 3.1 (ϵ-Differential Privacy [9]). Suppose that $\mathbb{D}$ is a set of data sets, which are differing on at most one record. A privacy mechanism $\mathcal{M}$ gives ϵ-differential privacy if for any pair of data sets $D_1, D_2 \in \mathbb{D}$, and for any output $O \subseteq Range(\mathcal{M})$,

$$\Pr[\mathcal{M}(D_1) \in O] \leq \exp(\epsilon) \times \Pr[\mathcal{M}(D_2) \in O], \quad (1)$$

where the privacy parameter ϵ is the privacy budget. The higher value of ϵ corresponds to the lower privacy protection.

In general, a mechanism giving ϵ-Differential Privacy is related to the *sensitivity* of a query function, which is defined as follows.

Definition 3.2 (Global Sensitivity [9]). For any query $Q : \mathbb{D} \rightarrow \mathbb{R}^d$, the global sensitivity of Q is

$$\Delta Q = \max_{D_1, D_2} ||Q(D_1) - Q(D_2)||_1 \quad (2)$$

where data set D_1 and D_2 differ in at most one record.

In order to achieve ϵ-differential privacy, there are two standard mechanisms proposed by the previous literature. They are the *Laplace mechanism* [9] and the *exponential mechanism* [43]. Here, we mainly focus on the Laplace mechanism.

Definition 3.3 (Laplace Mechanism [9]). For any query $Q : \mathbb{D} \rightarrow \mathbb{R}^d$, the following mechanism

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

provides ϵ-differential privacy.

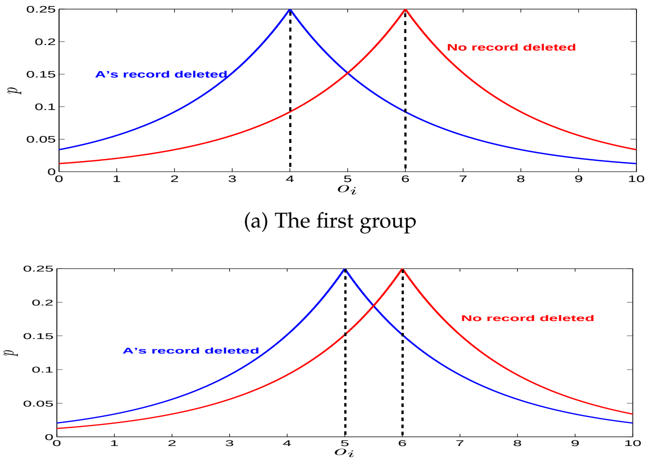
3.2 Composition Properties

Furthermore, differential privacy also considers the issues of privacy composition, in which multiple queries may be taken together so as to degrade the privacy guarantee. The literature [44] has shown that any group of the privacy mechanisms that each gives differential privacy in isolation also give differential privacy. This is called *sequential composition* as follows.

Lemma 3.4 (Sequential composition [44]). Suppose that every privacy mechanism $\mathcal{M}_i$ gives ϵ_i-differential privacy. A group of $\mathcal{M}_i(D)$ applied to the same data set D provides $(\sum_i \epsilon_i)$-differential privacy.

Here, from the point of view of the composition of data sets, it is obvious that if there are more and more data sets, the privacy guarantee under differential privacy is becoming worse and worse. Meanwhile, the number of data sets decides the upper bound of the privacy guarantee. For sequential composition, it assumes that the data sets are fully same. In contrast, when a group of queries are applied to a set of the disjoint data sets, the privacy guarantee is improving, compared to the same data sets. This is called *parallel composition*, which is given as follows.

Lemma 3.5 (Parallel composition [44]). Suppose that every privacy mechanism $\mathcal{M}_i$ gives ϵ_i-differential privacy. A group of $\mathcal{M}_i(D_i)$ over a set of disjoint data sets D_i provides $\max(\epsilon_i)$-differential privacy.



(a) The first group

(b) The second group

Fig. 2: Illustration for Example 2 with $\epsilon = 1/2$

The relationship between data sets in sequential or parallel composition of differential privacy is relatively simple. In fact, the relationship between data sets is rather complex. No matter whether the relationship between data sets is complex or simple in terms of privacy, we can get the following theorem.

Theorem 3.6. Suppose that the privacy mechanism M_i gives ϵ_i-differential privacy. A group of $\mathcal{M}_i(D_i)$ over a set of data sets D_i provides ϵ^*-differential privacy, where $\max(\epsilon_i) \leq \epsilon^* \leq \sum_i \epsilon_i$.

Proof: Obviously, Theorem 3.6 can be extended from Lemma 3.4 and Lemma 3.5. □

In order to illustrate this theorem, we give a simple example as follows.

Example 2. Similar to Example 1, suppose that there are two different data groups, which each has two data sets $D_i^j (i, j = 1, 2)$. Each data set has 3 records. In the first group, both of datasets have record A, while in the second group, only one has record A. Here, a query Q is to release the number of records in each group, perturbed by the Laplace mechanism $Laplace(\Delta Q/\epsilon)$, where $\Delta Q = 1$. Since the output of a query is $o_i = Q(D_i^1, D_i^2) + Lap(1/\epsilon)$, the density function of the output is $p(o_i) = \frac{\epsilon}{2}\exp(-\epsilon|o_i - Q(D_i^1, D_i^2)|)$. The concrete result is shown in Fig. 2, in which $\epsilon = 1/2$. Obviously, the results imply that the first group cannot give 1/2-differential privacy, while the second group still provides 1/2-differential privacy.

4 Differential Privacy in Correlated Data Sets

4.1 Characterizations of Correlated Relationship in Data Sets

Through the discussion in Section 3, a privacy mechanism $\mathcal{M}$ giving ϵ-differential privacy over a data set cannot hold the same privacy guarantee when there are multiple correlated data sets. It is a need to measure the relationship between correlated data sets in terms of privacy so as to compute the real privacy level of a data set.

In general, the relationship of records about some user in different data sets is split into two types as follows.

- **Direct Relationship**: This relationship is strictly defined as two fully same records. For example, a user simultaneously submits his tourist information to Facebook and Twitter. As a result, two different data sets have one same record about some user.
- **Indirect Relationship**: Different from the direct relationship, this relationship is more complex and defined as two different records about some user or his correlated users. For instance, information streams of some user's activity, e.g., GPS records and social networks records, are correlated with each other. Besides, [11] has shown that the privacy of correlated individuals may be compromised when their records are correlated.

It is easy to find that sequential composition and parallel composition are just suited for the direct relationship rather than the indirect relationship. In order to measure the privacy of the records with indirect relationship, a lot of work has been carried. Zhu et al. [19] utilized a correlated degree matrix to present the relationships between correlated data sets. In this case, the sensitivity of a query is changed into *correlated sensitivity*, which is the maximum among *record sensitivities*. However, this correlated privacy definition has a serious disadvantage. That is, *the correlated relationship between data sets is static but not dynamic*. Once some data set adjusts its privacy level, the static matrix doesn't change the privacy relationship. On the other hand, Yang et al. [20] proposed a Bayesian differential privacy (BDP) on correlated data sets. The idea is to utilize a Bayesian way to analyze an uncertain query, accompanied with some given and unknown tuples. In this paper, we follow [19], [20] and give our own definition of correlated differential privacy.

4.2 Definitions

Here, we draw lessons from these two approaches and define our own correlated privacy preserving of multiple data sets, which is given as follows.

Definition 4.1 (Correlated Differential Privacy). Suppose that $\mathbb{D} = \{D_1, D_2, \cdots, D_n\}$ is the vector of the data sets, all of which include one or more than records of some user or his correlated users. Data set D_i^1 and D_i^2 $(1 \le i \le n)$ generated from D_i are neighbor if they are differing on at most one record about some user, while the other records are fully same. A correlated privacy mechanism $\mathcal{M}$ is a randomized function whose domain is $\mathbb{D}$ and range is O. The correlated differential privacy leakage of $\mathcal{M}$ is

$$CDPL(\mathcal{M}) = \sup_{D_i^1, D_i^2, O, \mathbb{D}_{-i}} \log \frac{\Pr[\mathcal{M}(D_i^1) \in O \mid \mathbb{D}_{-i}]}{\Pr[\mathcal{M}(D_i^2) \in O \mid \mathbb{D}_{-i}]}, \tag{4}$$

in which $\mathbb{D}_{-i} = \mathbb{D} \setminus \{D_i\}$. As a result, we say a privacy mechanism $\mathcal{M}$ gives ϵ-correlated differential privacy if

$$-\epsilon \le CDPL(\mathcal{M}) \le \epsilon. \tag{5}$$

4.3 CDP vs. DP and CDP vs. BDP

The correlated differential privacy is an extension of differential privacy. Compromised by the correlated data sets, CDP also ensures that removing or adding a record in an objective data set doesn't (substantially) leak the privacy of some user. The only difference between CDP and DP is that CDP considers the influence of the correlated data sets, i.e., $\mathbb{D}_{-i}$. Obviously, if the data sets are independent, CDP is changed into DP.

CDP is actually an extreme case of BDP. In other words, CDP sets up $\mathcal{U} = [n] \setminus \{i\} \setminus \mathcal{K}$ and $\mathcal{K} = \emptyset$ in the definition of BDP, where an adversary denoted by $\mathcal{A}(i, \mathcal{K})$ is a person who knows all of the records in $\mathcal{K}$ and tries to attack the user i, and the set of unknown records is $\mathcal{U}$. In this paper, we assume that the correlated records are stored in different data sets so that the adversary cannot know these records and $\mathcal{U} = [n] \setminus \{i\}$.

4.4 Dynamic Privacy Choices

Here, we take an example with the truncated geometric mechanism (TGM) to clearly illustrate the dynamic privacy choices with CDP. The truncated geometric mechanism [12] is a discretized version and approximation guarantee for the Laplace Mechanism.

Definition 4.2 (Truncated Geometric Mechanism [12]). For any count query $Q : \mathbb{D} \to N$, the truncated α-geometric mechanism is defined as follows. For $\alpha \in (0, 1)$, the mechanism gives the output $Q(D) + \Delta$, where Δ is derived from the following geometric distribution:

$$\Pr[\Delta = \delta] = \begin{cases} \dfrac{1}{1+\alpha}\alpha^{Q(D)}, & \text{for } \delta = -Q(D) \\ \dfrac{1-\alpha}{1+\alpha}\alpha^{|\delta|}, & \text{for } -Q(D) < \delta < n - Q(D) \\ \dfrac{1}{1+\alpha}\alpha^{n-Q(D)}, & \text{for } \delta = n - Q(D) \\ 0, & \text{for otherwise} \end{cases} \tag{6}$$

Besides, for $\alpha = 0$, the mechanism always outputs the true query result $Q(D)$, while for $\alpha = 1$, the mechanism always outputs 0. Therefore, the mechanism gives ϵ-differential privacy, in which $\epsilon = \ln(1/\alpha)$.

Obviously, the higher value of α corresponds to the higher privacy protection, since ϵ is decreasing. The truncated α-geometric mechanism has a distinct characterization, which is given by the following theorem.

Theorem 4.3 (Main Characterization of TGM [12]). For $n \ge 1$, $\alpha \in [0, 1]$, and a count query Q, a privacy mechanism $\mathcal{M}$ is universally maximizing if and only if the mechanism $\mathcal{M}$ can be remapped into the truncated α-geometric mechanism via a mapping $\mathcal{Y}$.

Meanwhile, [12] has demonstrated that the Laplace mechanism cannot be universally utility maximizing. Theorem 4.3 implies that if someone uses the truncated α-geometric mechanism, he needn't consider the decision problem between different privacy preserving mechanisms and just focuses on the privacy parameter α. Therefore,

TABLE 1: Truncated $\frac{1}{2}$-geometric mechanism with $n = 2$

Input/Output	0	1	2
0	2/3	1/6	1/6
1	1/3	1/3	1/3
2	1/6	1/6	2/3

in the following sections to construct the game model, we always utilize the truncated α-geometric mechanism. A simple example of the α-truncated geometric mechanism is shown in Table 1, in which $\alpha = 1/2$ and $n = 2$. In the following, we focus on the influence on a data set when its correlated data set gives different ϵ-differential privacy.

Example 3. Assume that there are two data sets D_1 and D_2. The data set D_1 has a single record, whose value 0 and 1 have probability 50%, respectively. The data set D_2 adopts the α-truncated geometric mechanism with $n = 1$ and $Input = 1$. When α of D_2 is 1/2, the correlated relationship between D_1 and D_2 is shown in Table 2(a). In this case, when D_1 also adopts the 1/2-truncated geometric mechanism shown in Table 1 (i.e., $\epsilon = \ln 1/0.5$), the differential privacy leakage of D_1 is

$$\begin{aligned} & CDPL(\mathcal{M}) \\ = \;& \sup_o \log \frac{\sum_{D_2} \Pr[o|D_1 = 0, D_2] \Pr[D_2|D_1 = 0]}{\sum_{D_2} \Pr[o|D_1 = 1, D_2] \Pr[D_2|D_1 = 1]} \\ = \;& \log \frac{2/3 \cdot 2/3 + 1/3 \cdot 1/3}{1/3 \cdot 0 + 1/6 \cdot 1} \\ = \;& \log \frac{10}{3} \approx 1.20 > \log \frac{1}{0.5} \approx 0.69 = \epsilon, \end{aligned}$$

in which $o = 0$.

Here, the output of $Q(D_1)$ is 0 and $CDPL(\mathcal{M}) > \epsilon$ shows that it doesn't satisfy ϵ-differential privacy.

Example 4. Similar to Example 3, when parameter value α of D_1 and D_2 is 1/2 and 2/3 respectively, the correlated relationship between D_1 and D_2 is given in Table 2(b). So, the correlated differential privacy leakage of D_1 is

$$\begin{aligned} & CDPL(\mathcal{M}) \\ = \;& \sup_o \log \frac{\sum_{D_2} \Pr[o|D_1 = 0, D_2] \Pr[D_2|D_1 = 0]}{\sum_{D_2} \Pr[o|D_1 = 1, D_2] \Pr[D_2|D_1 = 1]} \\ = \;& \log \frac{2/3 \cdot 3/5 + 1/3 \cdot 2/5}{1/3 \cdot 1/5 + 1/6 \cdot 4/5} \\ = \;& \log \frac{8}{3} \approx 0.98 > \log \frac{1}{0.5} \approx 0.69 = \epsilon, \end{aligned}$$

in which $o = 0$.

The result of Example 4 is that $CDPL(\mathcal{M})$ of data set D_1 is greater than ϵ. This shows that it doesn't satisfy ϵ-differential privacy, either. More importantly, comparing Example 3 with Example 4, we can find that as the parameter α of D_2 increases from 1/2 to 2/3, i.e., ϵ decreasing from $\ln 2$ to $\ln 3/2$, the $CDPL(\mathcal{M})$ of D_1 decreases from 1.20 to 0.98. It means that as the correlated data set improves privacy preserving, the privacy of risk of the objective data set is degrading. Therefore, we can derive that the data sets affect the privacy level with each other.

TABLE 2: Correlation relationship between D_1 and D_2

(a) D_2 with $n = 1$, $\alpha = \frac{1}{2}$ and Input $= 1$

	$D_2 = 0$	$D_2 = 1$	Total
$D_1 = 0$	1/3	1/6	1/2
$D_1 = 1$	0	1/2	1/2
Total	1/3	2/3	1

(b) D_2 with $n = 1$, $\alpha = \frac{2}{3}$ and Input $= 1$

	$D_2 = 0$	$D_2 = 1$	Total
$D_1 = 0$	3/10	1/5	1/2
$D_1 = 1$	1/10	2/5	1/2
Total	2/5	3/5	1

5 Game Model of Correlated Differential Privacy

When some data publisher considers publishing his data, one of the most important problems is to adopt which privacy preserving mechanism is appropriate and decide what privacy parameter is optimal. Under the case of independent data sets, this is the well-known trade-off problem between privacy and utility [11], [16], [45]. However, since someone's private guarantee of data set depends not only on his own privacy parameter, but also on the privacy parameter of his neighbors shown in Section 4.4, his privacy choice is changed from the trade-off problem to a game problem. About privacy preserving mechanism, Section 4.4 has illustrated that the truncated α-geometric mechanism is a great option, because it is universally utility maximizing. Therefore, in the following, we assume all of data publishers adopt the truncated α-geometric mechanism and just focus on the choice of the privacy parameter.

When some data publisher decides what the privacy parameter is, he also needs to consider what the privacy parameter of his neighbors is. To this end, game theory is a well-suited analytical tool to discuss the case, in which players affect the payoff with each other. A game is a field to describe the strategic interaction, including the players' strategies and payoffs [30]. A desired solution is to systematically present the outcomes that may emerge in a specified game. For this, game theory suggests reasonable solutions for this game and examines their properties.

5.1 Formulation of Correlated Privacy Game

Suppose that there is a finite game $G = (N, S, U)$, which consists of:

- a finite set $N = \{1, \cdots, n\}$ of players, i.e., data publishers;
- a finite strategy space S_i for each player $i \in N$;
- a payoff function $u_i(\mathbf{s}) : S \to \mathbf{R}^+$ for each outcome $\mathbf{s} \in S = S_1 \times \cdots \times S_n$ and $U = \{u_1, \cdots, u_n\}$.

In this game, each player i utilizes the truncated α-geometric mechanism to anonymize his data set D_i. Let D be the set of data set $D_i, i = 1, \cdots, n$. The strategy space S_i of player i is the set of privacy parameter $\epsilon_i \in \mathbf{R}^+$. Correspondingly, some strategy $s_i \in S_i$ of player i is equal to ϵ_i.

For each player i, his payoff consists of two components, i.e., the utility of the anonymized data set $\mathcal{F}_i(\mathbf{s})$ (positive externalities) and the loss due to privacy leakage $\mathcal{L}_i(\mathbf{s})$ (negative externalities), which is given as follows:

$$u_i(\mathbf{s}) = \mathcal{F}_i(\mathbf{s}) - \mathcal{L}_i(\mathbf{s}). \quad (7)$$

In detail, for $\mathcal{F}_i(\mathbf{s})$ of player i with the truncated α-geometric mechanism, it is just dependent on the privacy parameter $\epsilon_i = \ln(1/\alpha)$ rather than the privacy parameter of its neighbors. Therefore, $\mathcal{F}_i(\mathbf{s}) = \mathcal{F}_i(s_i)$ and Equation (7) is rewritten as

$$u_i(\mathbf{s}) = \mathcal{F}_i(s_i) - \mathcal{L}_i(\mathbf{s}). \quad (8)$$

In order to construct a game model, the key step is to make clear the following factors: (i) the correlated relationship between different data sets under the case, in which every data set owner takes different privacy parameter, (ii) utility measures of anonymized data, i.e., $\mathcal{F}_i(\cdot)$, and (iii) loss measures if there is privacy leakage, i.e., $\mathcal{L}_i(\cdot)$.

5.2 Dynamic Correlation Relationship Model

Here, we use an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ to present the privacy relationship between players, in which $\mathcal{V} = \{1, \cdots, n\}$ is the set of players and $\mathcal{E}$ is the set of the edges $\{i, j\}$. If $\{i, j\} \in \mathcal{E}$, there is an undirected edge from player i to player j, which implies that there is the correlated relationship between player i's and j's data sets. Otherwise, there is no edge between player i and player j. A simple example of an undirected graph is shown in Fig. 3, in which the neighbors of player 2 are $ne(2) = \{1, 3, 4\}$ and the players who have the most neighbors are $2, 5, 6, 7$.

For player i, he chooses the truncated α-geometric mechanism to perturb the output result. Once player i decides the privacy parameter value $\epsilon = \ln 1/\alpha$, the output is derived from the distribution, i.e., Equation (6), which is indexed by s_i. Then, we define $\mathbf{s} = \{s_1, \cdots, s_n\} \in \mathbf{R}^n$ as a random vector. Next, the key is to measure the privacy relationship between multiple data sets. In the following, we take a sum count query as an example to illustrate how to derive the privacy level of player i.

5.3 Sum Count Query

When an adversary aims at some data set D_i of player i, he also utilizes the data sets of player i's neighbors to enhance the attack strength. In order to prevent the privacy leakage, player i adopts the truncated α-geometric mechanism to output the perturbed result. Suppose that there is a sum count query $Q(D)$ and the output of $sum(D)$ is t. Hence, the perturbed result is given by

$$o = t + \sigma, \quad (9)$$

in which $t = sum(D) = \sum_{i \in n} D_i$ and σ is the added noisy via the truncated α-geometric mechanism.

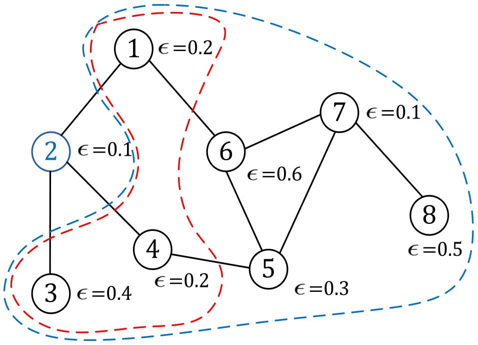


Fig. 3: An example of the privacy relationship in correlated data sets

As we discussed in Section 4.3, CDP is an extreme case of BDP. That is, CDP sets up $\mathcal{U} = [n] \backslash \{i\} \backslash \mathcal{K}$ and $\mathcal{K} = \emptyset$. We refer to [20] to compute the perturbed output of the sum count query. Meanwhile, [20] has demonstrated that the weakest adversary has the largest privacy leakage. Therefore, in our paper, the adversary has the strongest attack due to $\mathcal{K} = \emptyset$. According to the definition of correlated differential privacy, we should first compute $p(\sigma|D_i)$. From the marginalization rule, we get

$$p(o|D_i) \propto \int p(o|\sigma)p(\sigma|D_i)\mathrm{d}\sigma, \quad (10)$$

where $p(\sigma|D_i)$ is the correlation between the tuples and the true query result, and $p(o|\sigma)$ is the truncated α-geometric mechanism. Next, $p(o|D_i)$ is defined as follows.

$$p(o|D_i) = p(s_T = \sigma - D_i|D_i), \quad (11)$$

in which $s_T = \sum_{j:j \neq i} D_j$.

5.4 Correlation Analysis

According to Equation (11), it is a key to compute $p(s_T = \sigma - D_i|D_i)$, which is the privacy relationship of data sets between player i and the others. Therefore, we set up $D_0 = \sum_{j:j \neq i} D_j/(n-1)$ and $s_T = (n-1)D_0$. In the following, we refer to [20] and define the following matrix L to represent the relationship between D_0 and D_i, i.e.,

$$L^i = \begin{pmatrix} L_0^i & L_{0i}^i \\ L_{i0}^i & L_i^i \end{pmatrix}. \quad (12)$$

However, the difference from [20] is that for elements of matrix L^i, we think they are dynamic values rather than static values. The reason for this is that in our paper s_i is the privacy parameter of player i and can be also used to represent the probability distribution of the data and define the relationship between different data sets. Therefore, we re-define the elements of matrix L^i.

L^i represents the relationship between D_0 and D_i. More precisely, L_0^i is the relationship between D_0 and D_0 and thereby can be composed by D_{-i}, which is defined as $f_i(\mathbf{s}_{-i}) \geq 0$. $L_{0i} = L_{i0}$ is the relationship between D_0 and D_i and thereby can be defined by $g_i(ne(s_i) \cup s_i) \geq 0$. Suppose that $\partial g_i/\partial s_i \geq 0$ for s_i and $\partial g_i/\partial s_j \geq 0$ for $s_j \in ne(s_i)$ since the relationship is degrading as any variable decreases. Meanwhile, the function f_i has the same assumption, i.e., $\partial f_i/\partial s_j \geq 0$ for $s_j \in \mathbf{s}_{-i}$.

In [20], the authors have stated the result of the Bayesian differential privacy leakage on the sum count query[1]. Since in this paper the data publisher uses the truncated α-geometric mechanism rather than the Laplace mechanism, we refer to this and re-compute the result as follows.

Theorem 5.1. Suppose that the data publisher uses the truncated α-geometric mechanism to output the perturbed result on the sum count query, the correlated differential privacy leakage of L^i is

$$CDPL = M \ln \alpha^{-1} \left(1 + \frac{(n-1)g_i \ln \alpha^{-1}}{f_i}\right), \tag{13}$$

in which M is the maximum difference for different values of s_i, g_i represents the relationship between D_0 and D_i, and f_i represents the relationship between D_0 and D_0.

Proof: According to the literature [20], we replace $Lap(\lambda) = e^{-|z|/\lambda}$ with $\alpha^{|z|} = e^{|z| \ln \alpha}$. That is, $\lambda = -\frac{1}{\ln \alpha}$. In this case, it is easy to get Equation (13). □

5.5 Utility Function

On one hand, the data publisher needs to consider the privacy parameter. On the other hand, he computes the utility function of the anonymized data set, i.e., $\mathcal{F}_i(\cdot)$. Here, for the utility function $\mathcal{F}_i(\cdot)$ of player i, we make the following assumptions.

Assumption 1. The utility function $\mathcal{F}_i(\cdot)$ for each player i is a twice continuously differentiable function.

Assumption 2. In terms of privacy, the utility of each player is degrading as the privacy preserving level is decreasing. Therefore, the utility function $\mathcal{F}_i(\cdot)$ of player i is increasing with the privacy parameter s_i increasing.

Assumption 3. The utility function $\mathcal{F}_i(\cdot)$ of player i is a concave function. Similarly, the authors who considered the information measure about the statistical privacy theoretically make these same assumption [13]. Besides, some application papers [46] [47] and this paper (Section 7) make experiments to demonstrate this assumption.

According to the above assumptions, the utility function $\mathcal{F}_i(\cdot)$ of each player i has the following characterization, i.e.,

$$\frac{\partial \mathcal{F}_i(\cdot)}{\partial s_i} \geq 0, \text{ and } \frac{\partial^2 \mathcal{F}_i(\cdot)}{\partial s_i^2} \leq 0. \tag{14}$$

5.6 The Value of Privacy

As the idea of differential privacy [9], it is possible to cause the privacy leakage, no matter what privacy parameter the data publishers take. For example, when someone uses differential privacy to output the perturbation result, the adversary still has a certain probability to guess the true input. In this case, the data publisher has to consider the loss due to the privacy leakage. More precisely, Pai *et al.* [48] surveyed the previous literature related to the value of privacy and concluded that the value of privacy is negative! Hence, in our model, the value of privacy is a dis-utility.

It is natural to propose a fundamental question: how to model the value for the privacy? In order to answer this question, we refer to the idea of the literature [48], [49] and think that there are two main factors to influence the value of privacy, i.e., the privacy parameter and the privacy value of player i. Suppose that for each player i, the value of privacy is proportional to the privacy parameter ϵ, i.e.,

$$\mathcal{L}_i(s_i) = s_i \cdot v_i, \tag{15}$$

where $v_i \geq 0$ is player i's privacy valuation.

It is worth emphasizing that for ϵ-differential privacy as well as the other related privacy definition (e.g., CDP in our paper), the privacy parameter ϵ just quantifies the worst case harm that befall each player from revealing his private data. This seems that $s_i \cdot v_i$ is not a good measure for the value of privacy. In order to illustrate this, we refer to the explanation in the literature [49]. That is, $s_i \cdot v_i$ is a good and approximate bound of different measurements about the value of privacy [50], [51]. Therefore, $s_i \cdot v_i$ is a suited measurement of the value of privacy and is applied into our model.

6 Game Analysis

One of the fundamental objectives in game theory is to study Nash equilibrium, i.e., the strategy profile $\mathbf{s}^* = (\mathbf{s}_1^*, \cdots, \mathbf{s}_n^*)$, in which a strategy decided by each player is the optimal response to the strategies of the other players. In this section, we mainly focus on discussing the pure Nash equilibrium in the game.

6.1 Nash Equilibrium

In order to describe the Nash equilibrium, we firstly introduce the notion of dominant strategies as follows.

Definition 6.1 (Dominant Strategies [30]). In a finite game $G = (N, S, U)$, for strategy $\mathbf{s}_i$ and $\mathbf{s}'_i$ of player i, strategy $\mathbf{s}'_i$ is dominated by strategy $\mathbf{s}_i$ if for each feasible combination of the other players' strategies, i's payoff with strategy $\mathbf{s}_i$ is greater than or equal to i's payoff with strategy $\mathbf{s}'_i$:

$$(\mathbf{s}_i, \mathbf{s}_{-i}) \succsim (\mathbf{s}'_i, \mathbf{s}_{-i}) \iff u_i(\mathbf{s}_i, \mathbf{s}_{-i}) > u_i(\mathbf{s}'_i, \mathbf{s}_{-i}). \tag{16}$$

Nash Equilibrium is one of the most common equilibriums. According to dominant strategies, the pure Nash Equilibrium is defined as follows.

Definition 6.2 (Pure Nash Equilibrium [52]). In a finite game $G = (N, S, U)$, the strategy $\mathbf{s}^* = (\mathbf{s}_1^*, \cdots, \mathbf{s}_n^*)$ is a pure Nash Equilibrium if and only if, for each player i and each feasible strategy $\mathbf{s}_i$ in S_i, $(\mathbf{s}_i^*, \mathbf{s}_{-i}^*) \succsim (\mathbf{s}_i, \mathbf{s}_{-i}^*)$.

Nash Equilibrium provides an important way of predicting what will happen if multiple players are interactive with each other. If there exists pure Nash Equilibrium, it means that no player will gain anything by changing only his own strategy at this special state. That is, each player makes his optimal choice, as long as the others' choices remain unchanged. In the following, we focus on analyzing the existence and uniqueness of pure Nash Equilibrium.

[1] For the literature [20], we find that the result given by Theorem 3 is not consistent with the corresponding proof. We check them and refer to the result in proof, i.e., $BDPL_{\mathcal{A}}(\mathcal{M}; M) = \frac{M}{\lambda}\left(1 + \frac{mw_{0i}}{\lambda w_0}\right)$.

6.2 Equilibrium Analysis

In the following, we focus on the game analysis. Applying Equation (13) and (15) into Equation (8), we have

$$\begin{aligned} u_i(\mathbf{s}) =& \mathcal{F}_i(s_i) - \mathcal{L}_i(\mathbf{s}) \\ =& \mathcal{F}_i(s_i) - v_i s_i M\bigg(1 + \frac{(n-1)s_i g_i}{f_i}\bigg), \end{aligned} \tag{17}$$

in which $s_i = \ln \alpha^{-1}$ and $\alpha \in (0, 1]$.

According to Definition 6.2, in a finite game, the strategy profile $\mathbf{s}^* = (\mathbf{s}_1^*, \cdots, \mathbf{s}_n^*)$ is Nash Equilibrium if, for each player i and every feasible action $\mathbf{s}_i$ in S_i, $(\mathbf{s}_i^*, \mathbf{s}_{-i}^*) \succsim (\mathbf{s}_i, \mathbf{s}_{-i}^*)$. In the following, we discuss the existence of Nash equilibrium points. Based on Kakutani's Fixed Point Theorem, Nash [52] has demonstrated that there always exists at most one mixed Nash Equilibrium point for a game with a finite number of players and a finite number of strategies. However, for a pure Nash Equilibrium, it is not necessary. Here, we give Theorem 6.3 to present the sufficient conditions of the existence of the pure Nash Equilibrium in the correlated privacy model.

Theorem 6.3. In the n-player game of correlated differential privacy, in which the strategy profile is s and the payoff function is u_i, there are at least one Nash equilibriums if it satisfies two sufficient conditions as follows. The first one is given by

$$\frac{\partial^2 (s_i^2 g_i)}{\partial s_i^2} \geq 0. \tag{18}$$

And the second one is that when player j is the neighbor of player i,

$$\frac{\partial^2 (s_i^2 g_i / f_i)}{\partial s_i \partial s_j} \leq 0, \tag{19}$$

and when player j is not the neighbor of player i,

$$\frac{\partial (s_i^2 g_i)}{\partial s_i} \geq 0. \tag{20}$$

Proof: In the game, it is easy to find that the strategy set of player i, i.e., $\alpha \in [0, 1]$ and $\epsilon = \ln 1/\alpha \in [0, +\infty)$. Suppose that the set of the strategy of each player is compact and convex.

When we discuss the existence of the pure Nash Equilibrium in the game, it is a key to evaluate the payoff function of each player, i.e., u_i. In general, most of the previous literature is to demonstrate the existence of Nash Equilibrium via verifying concavity of the payoff function [53]. Meanwhile, another alternative approach is to demonstrate the existence of Nash Equilibrium through verifying whether the game is supermodular game [54]. That is, the payoff function of a game is supermodular if and only if $\partial^2 u_i / \partial s_i \partial s_j \geq 0$ for $i = 1, \cdots, n$ and $j \neq i$. Combined with these two cases, if there exists at least one pure Nash Equilibrium strategy in the game, it needs to satisfy the following conditions, i.e., $\partial^2 u_i / \partial s_i^2 \leq 0$ or $\partial^2 u_i / \partial s_i \partial s_j \geq 0$ for all s_j and $j \neq i$.

At first, we derive the first-order partial derivative of u_i with respect to s_i, which is given as

$$\frac{\partial u_i}{\partial s_i} = \frac{\partial \mathcal{F}_i}{\partial s_i} - v_i M - \frac{(n-1)v_i M}{f_i}(2 s_i g_i + s_i^2 \frac{\partial g_i}{\partial s_i}). \tag{21}$$

The second-order partial derivative with respect to s_i is

$$\frac{\partial^2 u_i}{\partial s_i^2} = \frac{\partial^2 \mathcal{F}_i}{\partial s_i^2} - \frac{(n-1)v_i M}{f_i}\bigg(2 g_i + 4 s_i \frac{\partial g_i}{\partial s_i} + s_i^2 \frac{\partial^2 g_i}{\partial s_i^2}\bigg). \tag{22}$$

Next, the second-order cross-partial derivative with respect to s_i and s_j is that (i) when j is the neighbor of i, we have

$$\begin{aligned} \frac{\partial^2 u_i}{\partial s_i \partial s_j} =& -\frac{(n-1)M}{f_i^2}\bigg[\bigg(2 s_i \frac{\partial g_i}{\partial s_j} + s_i^2 \frac{\partial^2 g_i}{\partial s_i \partial s_j}\bigg) f_i \\ & - \bigg(2 s_i g_i + s_i^2 \frac{\partial g_i}{\partial s_i}\bigg)\frac{\partial f_i}{\partial s_j}\bigg], \end{aligned} \tag{23}$$

and (ii) when j is not the neighbor of i, we have

$$\frac{\partial^2 u_i}{\partial s_i \partial s_j} = \frac{(n-1)M}{f_i^2}\frac{\partial f_i}{\partial s_j}\bigg(2 s_i g_i + s_i^2 \frac{\partial g_i}{\partial s_i}\bigg). \tag{24}$$

In Equation (22), since $\partial^2 \mathcal{F}_i / \partial s_i^2 \leq 0$, $v_i \geq 0$, $M > 0$ and $f_i > 0$, the sufficient condition of $\partial^2 u_i / \partial s_i^2 \leq 0$ is $2g_i + 4s_i \partial g_i \partial s_i + s_i^2 \partial^2 g_i \partial s_i^2 \geq 0$. That is, the second-order partial derivative of $s_i^2 g_i$ with respect to s_i is greater than or equal to 0. For Equations (23) and (24), both of them need to be greater than or equal to 0. In order to satisfy this, we derive the following requirements. In Equation (23), the second-order cross-partial derivate of $s_i^2 \dfrac{g_i}{f_i}$ with respect to s_i and s_j is less than or equal to 0. In Equation (24) with $\partial f_i / \partial s_j \geq 0$, the first-order partial derivate of $s_i^2 g_i$ with respect to s_i is greater than or equal to 0. □

In the following, we consider a special case, in which there exists only one pure Nash Equilibrium in the game, and give the following sufficient condition. It is worth emphasizing that it is hard to verify the uniqueness of pure Nash Equilibrium in the game, since at present no known works dominate [55].

Theorem 6.4. There exists a unique pure Nash Equilibrium in the game if the corresponding Hessian Matrix, i.e., Equation (25), is negative quasi-definite.

Proof: According to the literature [55], if the following matrix, i.e., the Hessian matrix,

$$H = \begin{bmatrix} \dfrac{\partial^2 u_1}{\partial s_1^2} & \dfrac{\partial^2 u_1}{\partial s_1 \partial s_2} & \cdots & \dfrac{\partial^2 u_1}{\partial s_1 \partial s_n} \\ \dfrac{\partial^2 u_2}{\partial s_2 \partial s_1} & \dfrac{\partial^2 u_2}{\partial s_2^2} & \cdots & \dfrac{\partial^2 u_2}{\partial s_2 \partial s_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 u_n}{\partial s_n \partial s_1} & \dfrac{\partial^2 u_n}{\partial s_n \partial s_2} & \cdots & \dfrac{\partial^2 u_n}{\partial s_n^2} \end{bmatrix}, \tag{25}$$

is negative quasi-definite, there is a unique pure Nash Equilibrium. □

6.3 Price of Anarchy

After we have demonstrated the existence and uniqueness of pure Nash Equilibrium in the game, it is interesting and important to evaluate the utility efficiency in the case of Nash Equilibrium. Therefore, we refer to price of anarchy (PoA) to measure the utility efficiency in the pure Nash Equilibrium, which is the lowest ratio between the total utility at a pure Nash Equilibrium and social optimum. That is,

$$PoA = \frac{\sum_i^n u_i(\mathbf{s}^*)}{\sum_i^n u_i(\mathbf{s}')}, \tag{26}$$

in which $\mathbf{s}^*$ is the strategy profile at pure Nash Equilibrium and $\mathbf{s}'$ is the strategy profile to maximize $\sum_i u_i(\cdot)$. Note that the value of PoA must be between 0 and 1.

Theorem 6.5. If $\partial^2(s_i^2 g_i)/\partial s_i^2 \leq 0$ and $\partial(g_i/f_i)/\partial s_j \geq 0$ for all i and j, the lower bound of PoA in the game is

$$\min\left\{1, \min_k \left\{\left(\sum_i \frac{\partial U_i(\mathbf{s}^*)}{\partial s_k}\right)/(v_k M)\right\}\right\}, \tag{27}$$

in which $U_i = \mathcal{F}_i - \dfrac{(n-1)v_i M s_i^2 g_i}{f_i}$.

Proof: In order to demonstrate the theorem, we refer to the method in [56], in which the authors demonstrated the upper bound of PoA in a cost-minimization game. Instead, we aim to derive the lower bound of PoA in a utility-maximization game.

Firstly, we have

$$\begin{aligned} u_i(\mathbf{a}, G) =& \mathcal{F}_i - v_i s_i M\left(1 + \frac{(n-1)x_i g_i}{f_i}\right) \\ =& \mathcal{F}_i - \frac{(n-1)v_i M s_i^2 g_i}{f_i} - v_i s_i M. \end{aligned}$$

Then, we set up $U_i = \mathcal{F}_i - \dfrac{(n-1)v_i M s_i^2 g_i}{f_i}$ and get $u_i = U_i - v_i s_i M$. At Nash Equilibrium, there may be two possible cases listed as follows,

$$\begin{cases} \partial U_i(\mathbf{s}^*)/\partial s_i = v_i M & \text{if } s_i^* > 0 \\ \partial U_i(\mathbf{s}^*)/\partial s_i < v_i M & \text{if } s_i^* = 0. \end{cases} \tag{28}$$

In Equation (28), the first equation means that there exists the value $s_i^* > 0$ so that $\partial U_i/\partial s_i = 0$, while the second equation implies u_i is decreasing via s_i. Theorem 6.3 has shown that $\partial^2(s_i^2 g_i)/\partial s_i^2 \geq 0$ is the sufficient condition of $\partial^2 U_i/\partial s_i^2 \leq 0$, implying that U_i is concave. So, $U_i(\mathbf{s}^*) \geq U_i(\mathbf{s}') + (\mathbf{s}^* - \mathbf{s}')^T \nabla U_i(\mathbf{s}^*)$. Next, we have

$$\begin{aligned} & PoA \\ =& \frac{\sum_i u_i(\mathbf{s}^*)}{\sum_i u_i(\mathbf{s}')} \\ =& \frac{\sum_i U_i(\mathbf{s}^*) - \sum_i s_i^* v_i M}{\sum_i U_i(\mathbf{s}') - \sum_i s_i' v_i M} \\ \geq& \frac{\sum_i U_i(\mathbf{s}') + (\mathbf{s}^* - \mathbf{s}')^T \sum_i \nabla U_i(\mathbf{s}^*) - \sum_i s_i^* v_i M}{\sum_i U_i(\mathbf{s}') - \sum_i s_i' v_i M} \\ =& \frac{\sum_i U_i(\mathbf{s}') - \mathbf{s}'^T \sum_i \nabla U_i(\mathbf{s}^*) + \mathbf{s}^{*T}(\sum_i \nabla U_i(\mathbf{s}^*) - \mathbf{v}M)}{\sum_i U_i(\mathbf{s}') - \sum_i s_i' v_i} \\ =& \frac{\sum_i U_i(\mathbf{s}') - \mathbf{s}'^T \sum_i \nabla U_i(\mathbf{s}^*) + \sum_i s_i^* \left(\sum_k \frac{\partial U_k(\mathbf{s}^*)}{\partial s_i} - v_i M\right)}{\sum_i U_i(\mathbf{s}') - \sum_i s_i' v_i M} \end{aligned} \tag{29a}$$

$$\geq \frac{\sum_i U_i(\mathbf{s}') - \mathbf{s}'^T \sum_i \nabla U_i(\mathbf{s}^*)}{\sum_i U_i(\mathbf{s}') - \sum_i s_i' v_i M} \tag{29b}$$

in which $\mathbf{v}$ is the vector of v_i. The reason from Equation $(29a)$ to $(29b)$ is that if $s_i^* = 0$, then it holds and if $s_i^* > 0$, $\partial U_k(\mathbf{x}^*)/\partial s_i > 0$ and thereby $\sum_k \partial U_k(\mathbf{s}^*)/\partial s_i - v_i M > 0$.

Since PoA must be less than or equal to 1, it means that $\mathbf{s}'^T \sum_i \nabla U_i(\mathbf{s}^*) < \sum_i s_i' v_i M$. Otherwise, PoA is equal to 1. Besides, it is easy to find that if $A > B$, $A > C$, $B < C$ and $A, B, C > 0, \dfrac{A-B}{A-C} > \dfrac{B}{C}$. Therefore, we have

$$PoA \geq \frac{\mathbf{s}'^T \sum_i \nabla U_i(\mathbf{s}^*)}{\sum_i s_i' v_i M} \geq \min_k \left\{\left(\sum_i \frac{\partial U_i(\mathbf{s}^*)}{\partial s_k}\right)/(v_k M)\right\}.$$

Besides, when $\mathbf{s}' = \mathbf{0}$, the right-hand side of Equation $29(b)$ is equal to 1. Therefore, we complete this proof. □

6.4 Remarks

Here, we mainly analyze the above theoretical results and present some useful remarks.

If there exists the pure Nash Equilibrium in the game, we have demonstrated two sufficient conditions. For the first sufficient condition, i.e., $\partial^2(s_i^2 g_i)/\partial s_i^2 \geq 0$, it shows that the existence of pure Nash Equilibrium is related to $s_i^2 g_i$. More precisely, the function g_i only includes those variables, which are the neighbors of i. Therefore, whether or not the pure Nash Equilibrium is existing in the first sufficient condition fully depends on player i and his neighbors. Meanwhile, since $\partial(s_i^2 g_i)/\partial s_i = 2s_i g_i + s_i^2 \partial g_i/\partial s_i \geq 0$, $s_i^2 g_i$ is an increasing and convex function.

For the second condition, it has two different cases. When j is not the neighbor of i, since $\partial(s_i^2 g_i)/\partial s_i$ is always greater than or equal to 0, this condition always holds. When j is the neighbor of i, the condition is $\partial(s_i^2 g_i/f_i)/\partial s_i \partial s_j \leq 0$, i.e., $\left(2s_i \dfrac{\partial g_i}{\partial s_j} + s_i^2 \dfrac{\partial^2 g_i}{\partial s_i \partial s_j}\right) f_i - \left(2s_i g_i + s_i^2 \dfrac{\partial g_i}{\partial s_i}\right)\dfrac{\partial f_i}{\partial s_j} \leq 0$. Especially, since $\left(2s_i g_i + s_i^2 \dfrac{\partial g_i}{\partial s_i}\right)\dfrac{\partial f_i}{\partial s_j} \geq 0$ holds, $\dfrac{\partial(s_i^2 g_i/f_i)}{\partial s_i \partial s_j} \leq 0$ holds if $\dfrac{\partial^2(s_i^2 g_i)}{\partial s_i \partial s_j} \leq 0$. Similar to the analysis of the first sufficient condition, whether or not the pure Nash Equilibrium is existing in the second sufficient condition only depends on i and his neighbors. Combined with these two sufficient conditions, the existence of the pure Nash Equilibrium fully depends on each player and his neighbors.

On the other hand, we use PoA to evaluate efficiency of the pure Nash Equilibrium and present the lower pound, i.e., $\min\left\{1, \min_k \left\{\left(\sum_i \dfrac{\partial U_i(\mathbf{s}^*)}{\partial s_k}\right)\Big/(v_k M)\right\}\right\}$, in which $U_i = \mathcal{F}_i - \dfrac{(n-1)v_i M s_i^2 g_i}{f_i}$. For $\sum_i \dfrac{\partial U_i(\mathbf{s}^*)}{\partial s_k}$, when $i = k$ as well as Equation (27), the value of $\dfrac{\partial U_k(\mathbf{s}^*)}{\partial s_k}$ is equal to $v_k M$ or $\dfrac{\partial \mathcal{F}_k}{\partial s_k} \leq v_i M$. If $\dfrac{\partial U_k(\mathbf{s}^*)}{\partial s_k} = v_k M$, then $\left(\sum_i \dfrac{\partial U_i(\mathbf{s}^*)}{\partial s_k}\right)\Big/(v_k M) \geq 1$ and thereby PoA is equal to 1. Therefore, considering the lower bound of PoA, we can first consider player k, whose $\dfrac{\partial U_k}{\partial s_k}$ is less than $\leq v_i M$.

7 Numerical Experiments

7.1 Experimental Setup

7.1.1 Dataset

In order to describe the relationship between utility and privacy, we attempt to conduct the experiment and derive

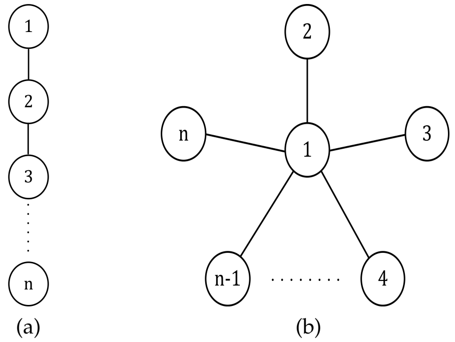


Fig. 4: Two simple models for correlated privacy relationship. Two models are (a) Linear Model and (b) Start Model, respectively.

the function $\mathcal{F}_i(\cdot)$, based on the real data set. We adopt the Adult data set from UCI machine learning repository [57], which was extracted from the Census database and has been widely applied to data analysis and anonymization. The data set contains 32561 instances, which each consists of 15 attributes. In general, the data set can be used to predict whether or not income of an individual exceeds $\$50,000$ a year.

We take the prediction performance as the utility of data. Most of the previous studies used the area under the ROC curve (AUC) on the data set to measure the prediction performance [47]. The higher the AUC, the better the prediction performance is. In our experiments, we will firstly use the privacy-preserving mechanism to anonymize the data and then analyze the utility. Here, we use PrivBayes [58] to achieve the anonymization of data and predict the result, which is a differentially private method. In order to reduce the error, we will repeat the experiment with different private parameters in PrivBayes 50 times.

For the correlated privacy relationship, we construct two simple models, i.e., linear and star model. In linear model, only the first and last player have a neighbor, while the others have two neighbors. Different from the linear model, player 1 has $n-1$ neighbors and the others have only one in star model. It is obvious that these two models are different and used to make comparison.

7.1.2 Function Setup

Once we obtain the results between utility and privacy from the experiments based on the real data set, we need to choose the proper function to fit. Here, we use $\ln$ function to fit the relationship between utility and privacy. Therefore, we set

$$\mathcal{F}(s_i) = \alpha_1 \ln(s_i + \alpha_2) + \alpha_3, \qquad s_i > 0. \tag{30}$$

When we consider function $f_i(\cdot)$ and $g_i(\cdot)$, it requires two functions need to satisfy the conditions in Theorem 6.3. We set

$$f_i(\mathbf{s}_{-i}) = \sum_{j \neq i} \frac{w_{ij}}{e^{-s_j} + \beta_1}. \tag{31}$$

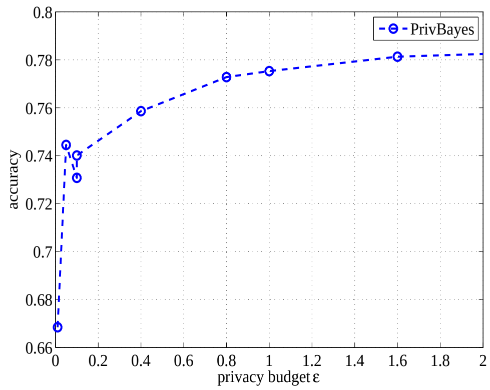


Fig. 5: Prediction accuracy by PrivBayes.

Then, we use λ_{ij} to represent whether or not player j is the neighbor of player i, which is given as follows

$$\lambda_{ij} = \begin{cases} 1, & \text{if } j \text{ is connected to } i \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

Hence, function g_i is given as

$$g_i(\mathbf{s}) = \sum_{j \neq i} \frac{\lambda_{ij} e^{x_i}}{\alpha_{ij} e^{-s_j} + e^{x_i}}. \tag{33}$$

7.2 Results

In our experiments, we mainly discuss the relationship between AUC and privacy budget in four different cases, including no breach model (NBM), breach model (BM), linear model (LM), and star model (SM). NBM is a case where we don't consider the loss due to privacy breach, while BM refers to the case of loss due to privacy breach in a single data set. Both LM and SM consider the loss of privacy breach in correlated data sets. The correlated privacy relationship between different players in LM and SM is shown in Fig. 4.

We firstly observe the relationship between privacy budget and the prediction accuracy. In these experiments, we repeat each privacy budget value 50 times to reduce the error. Fig. 5 shows the corresponding results between privacy budget and the prediction accuracy. It is easy to find that as the privacy budget increases, the prediction accuracy is increasing. We can also find that the growth rate from 0 to 0.2 of privacy budget is faster than that from 0.2 to 2. Meanwhile, this illustrates the importance of the choice of privacy budget.

Although Fig. 5 is able to show the relationship between privacy budget and prediction performance, we still use AUC as the standard performance metric, which is widely applied to most prediction systems. Meanwhile, we use curve fitting toolbox in MATLAB to fit the function $\mathcal{F}_i(\cdot)$. It is intuitive to see that AUC has the similar changes with prediction accuracy. In addition, Fig 6. shows the fit function is $\mathcal{F}_i(s_i) = 0.05811 * \ln(s_i + 0.001) + 0.7652$. According to the R-square index which is equal to 0.8368, it is proper to use the fit function.

After we obtain the fit function $\mathcal{F}_i(\cdot)$ and other important functions, it is able to observe the AUC in different

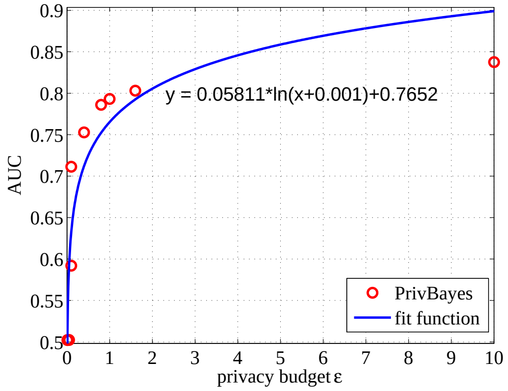


Fig. 6: Relationship between privacy budget and AUC. We use PrivBayes to obtain the AUC in different budget privacy. Then, we use curve fitting toolbox in MATLAB to fit the function between privacy budget and AUC. The related parameter to show the goodness of fit is $SSE = 0.02342$, $R - square = 0.8306$, and $RMSE = 0.05784$.

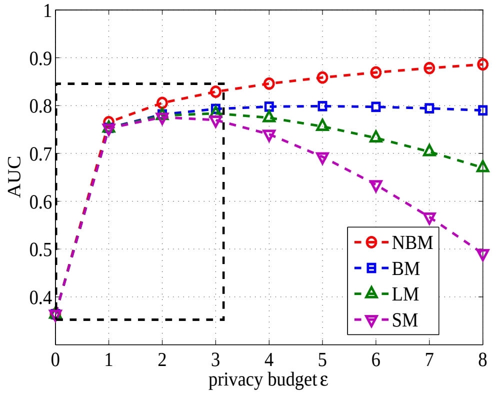


Fig. 7: AUC in different scenarios.

scenarios, including NBM, BM, LM, and SM. Here, we mainly consider player 1 in two models shown in Fig. 4 and assume that the other players choose the same and constant privacy budget. Fig. 7 shows results about the relationship between privacy budget and AUC in these four scenarios. Comparing NBM and BM, we can see that AUC of NBM is greater than that of BM due to considering the loss of privacy breach. It implies that when someone is sensitive to privacy, it inevitably decreases the utility of data due to data anonymization. On the other hand, AUC of BM is greater than that of LM and SM, which means that under the same privacy requirement, the cost to preserve the privacy in correlated data sets is higher than that in a single dataset. The cost of player 1 in SM is greater than that in LM, since player 1 in SM faces the more correlated data sets. Last but not least, the function of both LM and SM is a concave function, which shows that the game has the Nash Equilibrium. This was demonstrated by the literature [59]. This shows the requirements in Theorem 6.3 are correct.

Fig. 8 shows AUC of player 1 in LM and SM when the privacy budget of the other players changes. Since player 1 in SM is in a key position, the increasing privacy budget of the other players significantly influences his privacy compared with that in LM, which is shown in Fig. 8. On the other hand, if value v_1 of player 1 in SM is greater than others, the lower bound of PoA is possibly determined by player 1 when we don't consider the other factors. Meanwhile, when privacy budget of the neighbors of player 1 is less than 1, the privacy of player 1 has little change.

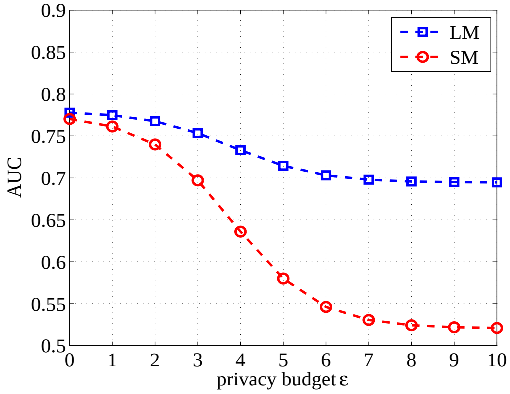


Fig. 8: AUC influenced by the others' privacy.

8 Conclusions and Future Work

In this paper, we studied the problem of the privacy preserving analysis in correlated data publication. As the privacy level of some data set is dependent on its and its neighbors' privacy parameters, it is changed from the trade-off problem to a game problem when data publisher considers his privacy parameter to maximizing his utility. We defined the correlated differential privacy to evaluate the real privacy level of a single data set influenced by the other data sets. Then, we constructed a game model of multiple players, who each publishes the data set sanitized by differential privacy. Next, we demonstrated the sufficient conditions of the existence and uniqueness of the pure Nash Equilibrium. We also referred to the price of anarchy to get the efficiency of the pure Nash Equilibrium. Finally, we presented the correctness of our game analysis via extensive experiments.

In the future, based on the game analysis in this paper, we attempt to find the most critical data publisher, who has the greatest impact on the privacy level of his neighbors. Thereby, we will make some useful measures to improve the overall utility in the game.

Acknowledgments

Our thanks to Mingxuan Yuan for his assistance in experiments. This research is partially supported by the EU FP7 CROWN project under grant number PIRSES-GA-2013-610524, Security Lancaster project SEF1015, the National Science Foundation 490 of China under Grant No. 91318301, No. 61273232 and No. 61321491, the Key Research and Development Project of Jiangsu Province under Grant No. BE2015154, BE2016120 and the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University.

References

[1] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.

[2] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Privacy-preserving big data publishing," in *Proceedings of International Conference on Scientific and Statistical Database Management, SSDBM*, 2015, pp. 26:1–26:11.

[3] X. Zhang, C. Leckie, W. Dou, J. Chen, K. Ramamohanarao, and Z. Salcic, "Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation," in *Proceedings of ACM International on Conference on Information and Knowledge Management, CIKM*, 2016, pp. 1793–1802.

[4] B. C. M. Fung, "Privacy-preserving data publishing: a survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, pp. 14:1–14:53, 2010.

[5] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertianty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, pp. 3:1–3:52, 2007.

[8] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," in *Proceedings of IEEE International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of Theory of cryptography*. Springer, 2006, pp. 265–284.

[10] C. Dwork, "Differential privacy: A survey of results," in *Proceedings of International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[11] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of ACM SIGMOD*. ACM, 2011, pp. 193–204.

[12] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[13] B. Lin and D. Kifer, "Information measures in statistical privacy and data processing applications," *ACM Transactions on Knowledge Discovery from Data*, vol. 9, no. 4, pp. 28:1–28:29, 2015.

[14] D. Kifer and B.-R. Lin, "An axiomatic view of statistical privacy and utility," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 5–49, 2012.

[15] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proceedings of ACM SIGMOD*. ACM, 2014, pp. 1447–1458.

[16] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. Liu, "Privacy or utility in data collection? a contract theoretic approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1256–1269, 2015.

[17] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Transactions on Database Systems (TODS)*, vol. 39, no. 1, pp. 671–683, 2014.

[18] R. Chen, B. C. M. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," *The VLDB Journal*, vol. 23, pp. 653–676, 2014.

[19] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: hiding information in non-IID data set," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 2, pp. 229–242, 2015.

[20] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proceedings of ACM SIGMOD*. ACM, 2015, pp. 747–762.

[21] D. Kifer and B. Lin, "Towards an axiomatization of statistical privacy and utility," in *Proceedings of ACM Symposium on Principles of Database Systems, PODS*, 2010, pp. 147–158.

[22] B.-R. Lin and D. Kifer, "Geometry of privacy and utility." in *Proceedings of IEEE GlobalSIP*, 2013, pp. 281–284.

[23] N. Li, W. H. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: a unifying framework for privacy definitions," in *Proceedings of ACM Conference on Computer and Communications Security, CCS*, 2013, pp. 889–900.

[24] H. Brenner and K. Nissim, "Impossibility of differentially private universally optimal mechanisms," *SIAM J. Comput.*, vol. 43, no. 5, pp. 1513–1540, 2014.

[25] C. Zeng, J. Cai, P. Lu, and J. F. Naughton, "On optimal differentially private mechanisms for count-range queries," in *Proceedings of ICDT*, 2013, pp. 261–271.

[26] G. Yuan, Y. Yang, Z. Zhang, and Z. Hao, "Convex optimization for linear query processing under approximate differential privacy," in *Proceedings ofACM International Conference on Knowledge Discovery and Data Mining SIGKDD*, 2016, pp. 2005–2014.

[27] A. Makhdoumi and N. Fawaz, "Privacy-utility tradeoff under statistical uncertainty," in *Proceedings of Annual Allerton Conference on Communication, Control, and Computing*, 2013, pp. 1627–1634.

[28] F. Prasser, R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, and K. A. Kuhn, "Lightning: Utility-driven anonymization of high-dimensional data," *Transactions on Data Privacy*, vol. 9, no. 2, pp. 161–185, 2016.

[29] Y. Wang, S. Song, and K. Chaudhuri, "Privacy-preserving analysis of correlated data," *CoRR*, vol. abs/1603.03977, 2016.

[30] M. J. Osborne and A. Rubinstein, *A course in game theory*, 1994.

[31] J. Freudiger, M. H. Manshaei, J. Hubaux, and D. C. Parkes, "On non-cooperative location privacy: a game-theoretic analysis," in *Proceedings of ACM Conference on Computer and Communications Security, CCS*, 2009, pp. 324–337.

[32] W. Wang and Q. Zhang, "A stochastic game for privacy preserving context sensing on mobile phone," in *Proceedings of IEEE Conference on Computer Communications, INFOCOM*, 2014, pp. 2328–2336.

[33] R. Shokri, "Privacy games: Optimal user-centric data obfuscation," in *Proceedings of PETs*, 2015, pp. 299–315.

[34] R. Shokri, G. Theodorakopoulos, and C. Troncoso, "Privacy games along location traces: A game-theoretic framework for optimizing location privacy," *ACM Trans. Priv. Secur.*, vol. 19, no. 4, pp. 11:1–11:31, 2017.

[35] W. Wang, L. Ying, and J. Zhang, "A game-theoretic approach to quality control for collecting privacy-preserving data," in *Proceedings of Annual Allerton Conference on Communication, Control, and Computing*, 2015, pp. 474–479.

[36] A. Ghosh and A. Roth, "Selling privacy at auction," *Games and Economic Behavior*, vol. 91, pp. 334–346, 2015.

[37] L. Olejnik, M. Tran, and C. Castelluccia, "Selling off user privacy at auction," in *Proceedings of Network and Distributed System Security Symposium, NDSS*, 2014.

[38] L. Xiao, D. Xu, C. Xie, N. B. Mandayam, and H. V. Poor, "Cloud storage defense against advanced persistent threats: A prospect theoretic study," *Journal on Selected Areas in Communications*, in accept.

[39] M. Chessa, J. Grossklags, and P. Loiseau, "A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications," in *Proceedings of IEEE Computer Security Foundations Symposium, CSF*, 2015, pp. 90–104.

[40] L. Gao, Z. Yan, and L. Yang, "Game theoretic analysis on acceptance of a cloud data access control system based on reputation," *IEEE Transactions on Cloud Computing*, in press.

[41] D. Bernhard, V. Cortier, D. Galindo, O. Pereira, and B. Warinschi, "Sok: A comprehensive analysis of game-based ballot privacy definitions," in *Proceedings of IEEE Symposium on Security and Privacy, SP*, 2015, pp. 499–516.

[42] Y. Chen, O. Sheffet, and S. P. Vadhan, "Privacy games," in *International Conference Web and Internet Economics, WINE*, 2014, pp. 371–385.

[43] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proceedings of IEEE FOCS*. IEEE, 2007, pp. 94–103.

[44] F. D. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of ACM SIGMOD*. ACM, 2009, pp. 19–30.

[45] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of ACM KDD*. ACM, 2009, pp. 517–525.

[46] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of ACM KDD*. ACM, 2010, pp. 493–502.

[47] X. Hu, M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, H. Guan, and J. Zeng, "Differential privacy in telco big data platform," in *Proceedings of the VLDB Endowment*, vol. 8, no. 12. VLDB Endowment, 2015, pp. 1692–1703.

[48] M. M. Pai and A. Roth, "Privacy and mechanism design," *ACM SIGecom Exchanges*, vol. 12, no. 1, pp. 8–29, 2013.

[49] K. Nissim, C. Orlandi, and R. Smorodinsky, "Privacy-aware mechanism design," in *Proceedings of ACM Conference on Economic Commerce (EC)*. ACM, 2012, pp. 774–789.

[50] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proceedings of ACM Conference on Economic Commerce (EC)*. ACM, 2011, pp. 199–208.

[51] D. Xiao, "Is privacy compatible with truthfulness," in *Proceedings of ACM conference on Innovations in Theoretical Computer Science*. ACM, 2013, pp. 67–86.

[52] J. Nash, "Noncooperative games," *Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[53] G. Debreu, "A social equilibrium existence theorem," *Proceedings of the National Academy of Sciences*, vol. 38, no. 10, pp. 886–893, 1952.

[54] D. M. Topkis, *Supermodularity and complementarity*. Princeton university press, 2011.

[55] G. P. Cachon and S. Netessine, "Game theory in supply chain analysis," in *Handbook of Quantitative Supply Chain Analysis*. Springer, 2004, pp. 13–65.

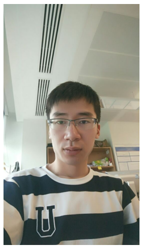
[56] L. Jiang, V. Anantharam, and J. Walrand, "How bad are selfish investments in network security?" *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 549–560, 2011.

[57] R. Kohavi and B. Becker, "UCI machine learning repository," [Online].Available:http://archive.ics.uci.edu/ml, 2016.

[58] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: private data release via bayesian networks," in *Proceedings of International Conference on Management of Data, SIGMOD*, 2014, pp. 1423–1434.

[59] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave N-person games," *Econometrica*, vol. 33, pp. 520–534, 1965.

Xiaotong Wu is currently working towards the PhD degree at the Department of Computer Science and Technology, Nanjing University, China. He has received his Bachelor's degree and Master's degree in Software Engineering from Central South University and Department of Computer Science and Technology from Nanjing University of China, respectively. His research interests include data privacy, network security, cloud computing and big data.



Taotao Wu received the B.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China. He is currently working toward the Ph.D. degree in the Department of Computer Science and Technology, Nanjing University, Nanjing, China. His research interests include cloud computing and applicationsmultimedia computing and communications social computing.

Maqbool Khan received B.Sc. and M.Sc. degrees in Computer Science from Gomal University D.I.Khan Pakistan, in 2002 and 2004 respectively. He worked as a Lecturer in Govt. College Abbottabad, Pakistan. He won the cultural scholarship award for abroad study from Ministry of Education Pakistan and received MS degree in Information Security from Huazhong University of Science and technology Wuhan, China in 2013. Currently, he is pursuing Ph.D. from department of Computer Science and Technology in Nanjing University, China. His research interests include Big Data, massive graphs and cloud computing. He is a student member of IEEE.

Qiang Ni received his Ph.D. degree in Engineering from Huazhong University of Science and Technology, Wuhan, in 1999. He is a Full Professor and the Head of Communication Systems Group, School of Computing and Communications, Lancaster University, UK. He is with Data Science Institute and Security Lancaster Centre. His research interests include future generation communications and networking systems, big data analytics, mobile and cloud networks, 5G, SDN, security and privacy, etc. Up to now, he had published more than 180 research papers in international journals and conferences.

Wanchun Dou received his PhD degree in Mechanical and Electronic Engineering from Nanjing University of Science and Technology, China, in 2001. From Apr. 2001 to Dec. 2002, he did his postdoctoral research in the Department of Computer Science and Technology, Nanjing University, China. Now, he is a full professor of the State Key Laboratory for Novel Software Technology, Nanjing University, China. From Apr. 2005 to Jun. 2005 and from Nov. 2008 to Feb. 2009, he respectively visited the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, as a visiting scholar. Up to now, he has chaired three NSFC projects and published more than 60 research papers in international journals and international conferences. His research interests include workflow, cloud computing and service computing.