Hindawi
Security and Communication Networks
Volume 2018, Article ID 3486529, 13 pages
https://doi.org/10.1155/2018/3486529

Research Article

A Privacy Protection Model of Data Publication Based on Game Theory

Li Kuang[1], **Yujia Zhu**[1], **Shuqi Li**[1], **Xuejin Yan**[1], **Han Yan**[1] and **Shuiguang Deng**[2]

[1]*School of Software, Central South University, Changsha 410075, China*
[2]*College of Computer Science, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Shuiguang Deng; dengsg@zju.edu.cn

Received 17 August 2018; Accepted 23 September 2018; Published 14 October 2018

Guest Editor: Xuyun Zhang

With the rapid development of sensor acquisition technology, more and more data are collected, analyzed, and encapsulated into application services. However, most of applications are developed by untrusted third parties. Therefore, it has become an urgent problem to protect users' privacy in data publication. Since the attacker may identify the user based on the combination of user's quasi-identifiers and the fewer quasi-identifier fields result in a lower probability of privacy leaks, therefore, in this paper, we aim to investigate an optimal number of quasi-identifier fields under the constraint of trade-offs between service quality and privacy protection. We first propose modelling the service development process as a cooperative game between the data owner and consumers and employing the Stackelberg game model to determine the number of quasi-identifiers that are published to the data development organization. We then propose a way to identify when the new data should be learned, as well, a way to update the parameters involved in the model, so that the new strategy on quasi-identifier fields can be delivered. The experiment first analyses the validity of our proposed model and then compares it with the traditional privacy protection approach, and the experiment shows that the data loss of our model is less than that of the traditional k-anonymity especially when strong privacy protection is applied.

1. Introduction

The rapid development of sensor networks and cloud computing has pushed the emergence of a great deal of data innovation applications for the IoT and other intelligent network systems in the fields of urban transportation, education, medical treatment, and living [1–3]. Most service development processes involve users, data collection organizations, and data development organizations. Data collection organizations are usually trusted by users, but data development organizations are likely to be untrusted by users. With the frequent leakage of sensitive data, individuals and society are paying more attention to the protection of privacy information. It has become a meaningful and challenging problem to resolve the contradiction between the privacy protection required by individuals and the data availability required by data development organizations [4–6].

In a dataset that contains privacy information, the attributes can be summarized into three categories: identification attributes, quasi-identification attributes, and sensitive attributes [7]. Identification attributes are those that can directly distinguish an individual's identity. Quasi-identification attributes (also known as nonsensitive attributes) are multiple attributes that together may infer the identity of an individual. Sensitive attributes are those that contain privacy data. Identification attributes are removed before the dataset is published, but simply hiding the identification attributes does not guarantee privacy because attackers may infer identification attributes based on quasi-identification attributes when attackers adopt link attacks or have users' background knowledge. If the quasi-identification attribute and the identification attribute are hidden together, although it is not possible to deduce a one-to-one association between privacy information and

personal identification, the dataset with sensitive attributes becomes useless.

The approaches related to the issue can be roughly divided into three kinds: data distortion, secure multiparty computing, and data anonymity. The algorithm of data distortion randomly modifies the sensitive data, and it has a high degree of data distortion and strong data dependence. The algorithm of secure multiparty computing forces untrusted third parties to complete data mining work without the direct communication of detailed data through data nodes, and it is very expensive in terms of calculation and communication overhead. At present, most scholars are investigating variations of k-anonymity algorithm, which guarantee that only groups of a minimum size k can be identified, rather than individuals [8]. However, it is proved to be a NP-hard problem to find the optimal information loss and time complexity based on k -anonymity.

For the development of big data-driven services in smart environments, the existing privacy protection algorithms about data publishing have the following deficiencies: (1) the time complexity of most algorithms are relatively high, so they are hard to apply in practice and (2) service development involves many factors, such as the profit of services, quality of services, and privacy and security; however, existing algorithms have not taken the trade-offs between the factors into consideration comprehensively.

Since the attacker identifies the user based on the combination of user's quasi-identifiers and the fewer quasi-identifier fields result in a lower probability of privacy leaks, therefore, in this paper, we aim to investigate an optimal number of quasi-identifier fields under the constraint of trade-offs between service quality and privacy protection. In the initial phase, we first construct a loss function for privacy leakage and a rating function for service quality and then model the service development process as a cooperative game between the data owner and consumers. The Stackelberg game model is employed to get the global optimal solution based on historical data, which comprehensively considers the service quality, privacy leakage loss, and service revenue, so that the number of quasi-identifier fields can be determined. Since the functions of the game model may be biased from reality, we then design a way to determine whether the new data need to be learned and a way to adjust weights of historical data and new data in the update phase, and game model is then used repeatedly to get the new strategy on quasi-identifier fields delivery.

The rest of the paper is organized as follows: Section 2 discusses the work related to data publication with privacy protection. Section 3 introduces the preliminary knowledge about game theory. Section 4 defines the problem and explains our proposed privacy protection model. Section 5 presents the experiments and analyses the results. Section 6 gives the conclusions and future work.

2. Related Work

Many scholars have conducted a series of research work on privacy protection in data publication. Their approaches can be mainly divided into three kinds: data distortion, data encryption, and data anonymization.

Approaches based on data distortion protect the privacy information by disturbing the original data. The attacker cannot reconstruct the original data through the published distorted data, while some information obtained from the distorted data is approximately equivalent to the information obtained from the original data. This kind of approaches [9–13] mainly studies how to perform data perturbation and how to mine the perturbed data, so that people cannot get the value of original data but can get high-quality mining results from the perturbed data. Since the distribution of the perturbed data is almost the same as the original data, the perturbed data can be used to train the learning models well [9]. The method focuses on the goal of preserving privacy by suppressing and perturbing the quasi-identifiers in the data without causing any loss to the information in the process [14]. However, such kind of approaches has a high degree of data loss and strong data dependence.

Approaches based on data encryption define privacy-based data mining applications as a secure multiparty computing problem involving untrusted third parties in many distributed environments. Each part only knows its own input data and the final results of all calculations among two or more sites are communicated through some kind of protocol. In secure multiparty computing [15–18], data are distributed and stored on multiple nodes. Each data node wants to perform data mining on the global data but does not want to disclose its own data. Therefore, information exchange protocols based on secure multiparty computing should be designed. Each data node does not exchange the detailed data samples directly, but it uses the protocols to exchange the information needed by the data mining algorithms in the absence of details of other nodes. However, such kind of approaches has a high computational and communication overhead.

At present, data anonymization is investigated widely and becomes the mainstream way to privacy protection. The k-anonymity model originally proposed by Sweeney [19] is used to defend against background knowledge attacks and link attacks, and generalization and compression techniques are widely used to achieve k-anonymity [20, 21]. Since Aggrawal [22] proved that the clustering method can achieve anonymity in a more efficient way, researchers began to investigate on clustering anonymity algorithm in privacy data publication. Li et al. [23] proposed anonymity scheme that applied the clustering idea, and the anonymity process merges the equivalence classes repeatedly and selects a certain equivalence class according to the principle of the minimum amount of loss of the consolidated generalization until all equivalence classes contain more than k tuples. Zhihui Wang and et al. [24] proposed an L-clustering method that could classify quasi-identifier attributes, they measure the degree of uncertainty of attribute values before and after generalization, and give us a measure of information loss that transformed the data anonymity problem into a clustering problem with specific constraints.

To enhance the performance on time efficiency and information loss in k-anonymity, Zhang et al. [25] proposed

a k-anonymity clustering algorithm based on information entropy, and the first step is to divide the table data into a number of record subsets according to the principle of the minimum average distance on quasi-identification attribute values, while the second step is to merge and split the subsets as appropriate so that the number of records for each subset is between k and $2k$. Huowen Jiang et al. [7] proposed a greedy clustering anonymization method based on the idea of the greedy method and clustering and they separately measured the information loss of the quasi-identifier, and the distances between tuples and the distances between tuples and equivalence classes. The methods mentioned above try to optimize the performance of k-anonymity algorithm, but the optimal information loss and time complexity for k-anonymity algorithm have been proved to be a NP-hard problem, and there is still space for improving the performance of existing algorithms.

3. Introduction to Game Theory

Game theory is originated from "Game Theory and Economic Behaviours," which was coauthored by von Neumann and Morgen stern in 1944 [26]. For the first time, this book presents a complete and clear description of the research framework of game theory and expounds the basic axioms. For a long time, the study of game theory focused only on the double zero-sum game. In the early 1950s, Nash [27] proposed the most important theory in game theory called the Nash Equilibrium, which determined the form and theoretical foundation of the noncooperative game and extended the research field of game theory to noncooperative games and nonzero-sum games. Game theory is suitable for solving conflicts and seeking a Nash equilibrium solution for the problem. In the problem of data publication, we need to publish the user's data to develop smart services, and untrusted data developers may leverage user's private information. Therefore, it is a conflict issue to publish user data for service development, and game theory can be used for modeling.

Game theory is about how smart and self-interested people act in the strategic layout and interact with their opponents. It has three parts: (1) a group of participants; (2) the actions that participants can take; (3) the benefits that participants may get. Each participant chooses the best action for their maximum benefit, and each participant will always think that other participants are also trying to get the best result. If game theory can provide a unique solution to the game problem, the solution must be a Nash equilibrium. The strategy chosen by each participant must be the optimal response to the strategy chosen by the other participants, and no participant is willing to abandon his selected strategy alone.

According to whether there is a binding agreement between the two parties, game theory can be divided into cooperative game and non-cooperative game; according to whether the sum of the revenue of both players is zero, game theory can be divided into zero sum game and nonzero sum game; according to the decision order of the players in the game, game theory can be divided into static games and dynamic games; according to whether the two parties understand the each other's strategy and revenue function, game theory can be divided into complete information game and incomplete information game. We model the problem of privacy data publishing as a cooperative game problem between data collector and data developer and establish the strategy space and revenue function of both parties. The game sequence is that the data collector first publishes privacy data, and then the data developer performs service development based on the data.

4. Privacy Protection Model Based on Game Theory

4.1. Problem Definition. There are a set of datasets with sensitive information for publication, and each dataset can be expressed as $T = \{t_1, t_2 \dots t_n\}$, where t_i is the ith record, and each record consists of q quasi-identifying attributes and one sensitive attribute, i.e., $t_i = (A^q, A^s)$, where $A^q = \{A^{q1}, A^{q2} \dots A^{ql}\}$ denotes all the quasi-identifying attributes in the data table and A^s is the sensitive attribute in the table. Table 1 shows an illustrating example of the dataset containing privacy information, where *age*, *sex*, and *zip code* are quasi-identifier attributes and *disease* is the sensitive attribute. The set of datasets can be classified by the sensitive attribute, such as disease, property, and religious beliefs.

Assume we have historical records of service development $R = \{r_1, r_2, r_3, \dots\dots, r_m\}$, where r_i is the loss, investment, revenue, and rating score of the service when delivering '*quasi_number*' pieces of quasi-identifiers on datasets with sensitive attribute '*privacy_type,*' and r_i can be expressed as a 7-tuple r_i=<*privacy_type, quasi_number, privacy_loss, technique_investment, A_revenue, B_revenue, score*>. '*privacy_loss*' is the loss of privacy, which consists of direct losses and indirect losses incurred by data collectors. The direct losses include users' privacy disclosure by competitors, and indirect losses include complaints and claims from users. '*Technique_investment*' is the technique costs contributed by data developer. '*A_revenue*' and '*B_revenue*' are the revenues of the data application services achieved by data collectors and data developers, respectively, and '*score*' is the quality score of the service. The samples are shown in Table 2.

The identifying attributes have been removed from Table 1, so that a specific field in the table cannot be mapped to a specific individual. However, simply hiding the identifying attribute does not guarantee privacy security when the attacker knows some background knowledge of the user, for example, the combination of age, sex, and zip code. Under this circumstance, the attacker may also infer and identify a person, which leads to personal privacy disclosure. In the development model of existing services, data collection organizations are trusted by users; however, data development organizations are likely to be untrusted by users.

Given R, we need to perform privacy protection on T assuming that the data developer is not credible; that is, we need to determine an optimal number of quasi-identifier

Table 1: Example of a dataset with sensitive information.

age	sex	zip code	disease
20	female	100018	bronchitis
28	male	300017	flu
32	male	100018	pneumonia
33	male	400015	indigestion
36	female	200017	rhinitis

Table 2: History usage records for the privacy field.

privacy_type	quasi_number	privacy_loss	technique_investment	A_revenue	B_revenue	score
disease	20	1000	600	1700	1000	3.9
disease	16	800	500	1500	7000	3.3
.....	.....	.....	.....	.....	.....	.....
disease	22	300	100	600	300	2.7

fields under the constraint of trade-offs between service quality and privacy protection.

4.2. The Framework of Privacy Protection Model. In order to ensure the data availability for high service quality and to protect users' privacy at the same time, we design a privacy protection model for data publication as shown in Figure 1. The model consists of two phases, namely, service development and service update. In the initial phase of service development, we define the relevant function of the service development to simulate the problem firstly, which includes the revenue functions of the game participants and the variables on which the revenue functions depend. Then we employ the game theory to model the problem and obtain a balanced solution that both parties are willing to accept, and publish the data according to the strategy. In the following phase of service update, we adopt statistical method to detect the constantly updated data, determining whether the service needs to learn the new data. Then, we design an adjustment way of the parameters in the functions based on the actual results in the previous development phase and analyze the data publishing strategy for service update.

In the following sections, we will illustrate the game process of the data collectors and data developer by four parts: (1) the establishment of complete information; (2) the strategy generation in service development based on Stackelberg game model; (3) the detection of the need for service update; (4) the renewal strategy generation in service update.

4.3. The Establishment of Complete Information on Both Sides of the Game. Strategic space is a collection of actions that are available to the parties of the game, and each strategy corresponds to a result. Since the data collector and the data developer cooperate to complete the service development, in this section we need to determine the strategic space and the revenue function of both parties.

We use the new strategic cooperation model which widely used in game theory economics to establish the cooperative relationship between data collector and data developer. Because the dominant party in the cooperation model will subsidize the other party to achieve better cooperation result, we set (Q, k) to represent the strategy combination of the data collector, where 'Q' is the loss of privacy information leakage, and 'k' is the percentage of the economic subsidy to the technology investment. The strategy of the data developer is the investment cost of data mining technology 'a'.

Furthermore, we also need to define the expression of privacy leakage loss 'Q'. The basic principle is that the loss is proportional to the probability of privacy leakage P. In order to obtain the relationship between the probability of privacy leakage and the number of quasi-identifiers, we use Taylor's third-order function to model it. The expression of Q is given in formula (1):

$$\begin{aligned} Q &= u_q P \\ P &= b_3 x^3 + b_2 x^2 + b_1 x + b_0, \end{aligned} \tag{1}$$

where: x represents the number of quasi-identifiers; b_0, b_1, b_2, b_3 represent the pending parameters in the third-order Taylor formula; P is the probability of privacy leakage; $u_q$represents a positive coefficient between P and Q.

We assume that service quality score for data developer is affected by the amount of data information and the investment cost of mining technology. Because the privacy leakage loss Q is defined by the number of quasi-identifiers, and the more the quasi-identifiers, and the amount of information in the data set can be expressed by the number of quasi-identifiers, we need to define the quality of service score as an increasing function of privacy leakage loss 'Q' and mining technology investment cost 'a', and the maximum value of the function is the highest score of service quality. Then, we need to set the parameters to indicate the impact of 'Q' and 'a' on the function and consider the interaction between the two variables which show a complementary relationship to quality of service score function. We consider quality of service score function as shown in

$$\mathrm{S}(\mathrm{Q}, a) = \omega - \beta a^{-\gamma} Q^{-\varsigma}, \tag{2}$$

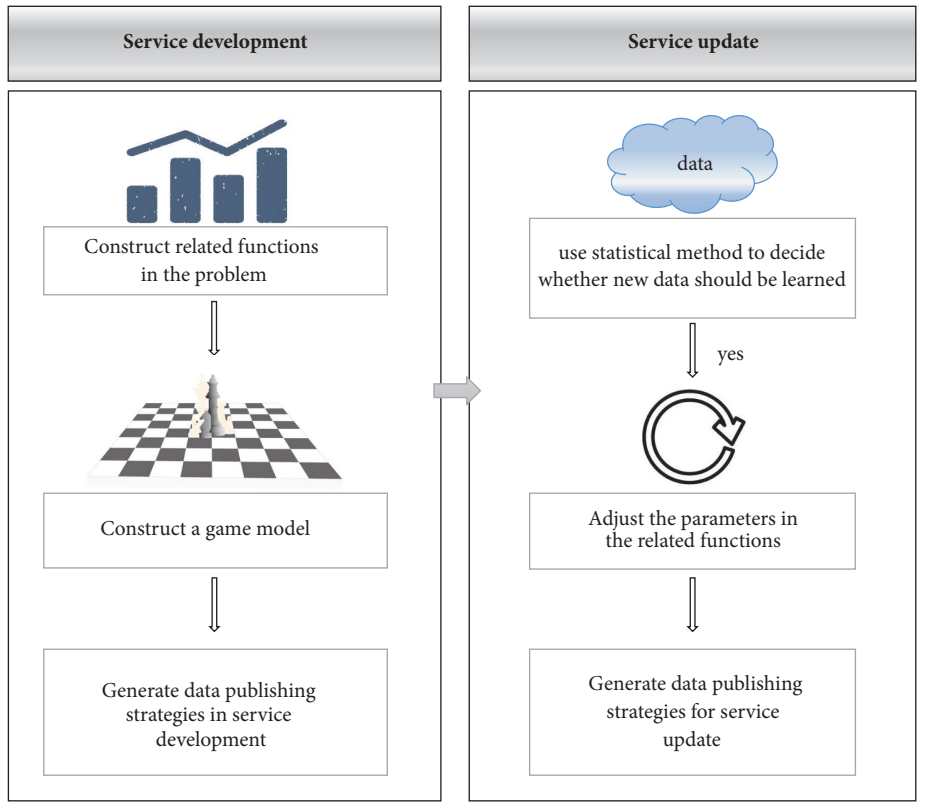

Figure 1: The framework of privacy protection model.

where ω represents the saturation value of the service level, β represents a normal number, γ represents the impact factor of a, and ς represents the impact factor of Q.

The return of both sides is proportional to the service quality score, and the form of revenue function is income minus cost. We define the revenue function of data collector r_m as (3), where ka and Q are the costs of the data collector. The revenue function of the data developer r_n is as (4), where $(1\text{-}k)a$ is the cost of service development.

$$r_m = \lambda S(Q, a) - ka - Q \tag{3}$$

$$r_n = \eta S(Q, a) - (1 - k)a, \tag{4}$$

where $S(Q,a)$ represents the score of service quality, Q represents the loss of privacy leakage, a represents the mining technology investment, k represents the proportion of subsidies to the data development organization, and λ and η represent direct proportionality coefficient.

4.4. Strategy Generation in Service Development Based on Stackelberg Game Model. We then aim to describe the process how data collector and data developer play the Stackelberg master-slave game. The data collector first proposes a cooperation program, and then data developer makes decisions based on the behavior of data collector. In the first phase of the game, data collector determines the proportion of economic subsidy for technique inputs to encourage service development, as well as the loss of the privacy leakage by quasi-identifier fields. In the second phase of the game, the data developer decides to invest the technique based on the data collector's program. Due to the sequence of actions, data collector makes decisions first, and data developer makes corresponding decisions based on the data collector's decision. This is a two-stage game problem that can be solved by inverse induction.

In order to obtain the Nash equilibrium solution, we consider the revenue function of data developer r_n first and get the response function of the data developer about Q(the loss of privacy leakage) and k(the proportion of subsidies to the data developer) by deriving r_n and the process is as follows:

$$\begin{aligned} r_n &= \eta\left(\omega - \beta a^{-\gamma} Q^{-\varsigma}\right) - (1 - k)a \\ \frac{\partial_{r_n}}{\partial_a} &= 0 \end{aligned} \tag{5}$$

From (5), we can obtain the response function of the data developer:

$$a = \left[\frac{\gamma\eta\beta}{(1 - k)Q^{\varsigma}}\right]^{1/(\gamma+1)} \tag{6}$$

After obtaining the reaction function (6) of the data developer, we eliminate the parameter a in the revenue

function of the data collector r_m and then find the partial derivative of the variables in function r_m, the process is as follows:

$$\begin{aligned} r_m &= \lambda\left(\omega - \beta a^{-\gamma} Q^{-\varsigma}\right) - ka - Q \\ \frac{\partial_{r_m}}{\partial_Q} &= 0, \\ \frac{\partial_{rm}}{\partial_k} &= 0 \end{aligned} \tag{7}$$

From (6) and (7), we can obtain the value of Q and k of the data collectors Q_1 and k_1:

$$k_1 = \frac{\lambda - (1+\gamma)\eta}{\lambda - \gamma\eta} \tag{8}$$

$$Q_1 = \left[\varsigma^{\gamma+1}\beta\gamma^{-\gamma}(\lambda - \eta\gamma)\right]^{1/(\varsigma+\gamma+1)} \tag{9}$$

After getting the value of Q, we can get the number of quasi-identifiers through (1) in data publication.

4.5. Detection of the Need for Service Update. We then aim to build a dynamic mathematical model for the constantly update data and analyze, diagnose, predict, and optimize the system based on the model. We construct a statistical model to identify the extent of the data changes.

To achieve this goal, we construct a unilateral hypothesis test about mean μ when the variance σ^2 is unknown. First, we get the model error evaluation index E_i of the ith test set. Then, we calculate the average error of the model according to Definition 1 and finally construct the distribution function according to Definition 2.

Definition 1. The XinQin Law of Large Numbers state that if $\{x_n\}$ is an independent and identically distributed random variable sequence and $EX_n = \mu$ exists, then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu \tag{10}$$

Definition 2. The Central Limit Theorem states that regardless of $x_i \sim F(\mu, \sigma^2)$, in the case of large samples

$$\frac{1}{n}\sum_{i=1}^{n} x_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{11}$$

Definition 1 indicates that when the random variables are independent and identically distributed, the average of the random variables tends to the true average with a certain probability p, and the larger the value of n, the closer the value of p is to 1. We derive the mean of the error evaluation criteria on the original test data set from Definition 1 as shown in

$$\mu_0 = \frac{1}{i}\sum_{i=1}^{i} E_i \tag{12}$$

Definition 2 means that it is not necessary to consider the original distribution of the random variables. In the case of large samples, the average of n random variables obeys the normal distribution. Collecting the data sample of a certain time interval, we can get the error evaluation criterion E_{i+1} of the model. Then, we establish two assumptions $H_0 : \mu \le \mu_0$ and $H_1 : \mu > \mu_0$ and construct the statistics as shown in (13), where S represents the sample variance and n represents the sample size.

$$T = \frac{E_{i+1} - \mu_0}{S/\sqrt{n}} \tag{13}$$

When the level of significance $\alpha(0 < \alpha < 1)$ is given, we can get the rejection domain of $H_0 : \mu \le \mu_0$ as $w = \{T > t_\alpha(n-1)\}$. If T falls into the denial domain, it explains that because of the big data dynamic characteristic, the performance of the model with the original knowledge is not within our expectation and the new knowledge needs to be learned. The specific process of update algorithm is shown in Algorithm 1.

4.6. The Renewal Strategy Generation in Service Update. Since the initial model is trained by historical development records of similar services, it may not be consistent with the actual service development; therefore, we need to adjust the weights of new samples according to the accuracy of the learned model. The real values of service score S and the income r_m, r_n of both parties will be generated in the previous service lifecycle, we can compare the predicted values $S^*, r_m{}^*, r_n{}^*$ with the real ones and adjust the weight of new sample data for training involved in the game model.

When the difference between the real value and the predicted value is within a specified range, we believe that the historical data of similar services can train the modeling functions well, so we can add the new data sample using the same weight with the historical data. When the difference exceeds the specified range, the historical data cannot reasonably train the modeling functions, so we adjust the weight of the new sample, making it appear more times so that the new sample will have more important impact on the training process. In formulas (14)-(16), w_1, w_2, and w_3 represent the weights of the new sample in the training process of the correlative functions (2)-(4), respectively, l_1, l_2, and l_3 are the corresponding prediction error threshold values, and N is the number of historical training samples.

$$w_1 = \begin{cases} 1, & |S* - S| < l_1 \\ \frac{1}{2}N, & |S* - S| \ge l_1 \end{cases} \tag{14}$$

$$w_2 = \begin{cases} 1, & |r_m* - r_m| < l_2 \\ \frac{1}{2}N, & |r_m* - r_m| \ge l_2 \end{cases} \tag{15}$$

$$w_3 = \begin{cases} 1, & |r_n* - r_n| < l_3 \\ \frac{1}{2}N, & |r_n* - r_n| \ge l_3 \end{cases} \tag{16}$$

The Nash equilibrium of Stackelberg model is unique. If there is a unique Nash equilibrium in the staged game,

input : $n, s, E_{i+1}, \mu_0, oldservice$
output : *service*
1: $T = (E_{i+1} - \mu_0)/(S/\sqrt{n})$
2: *service* = 0
3: *IF*($T > t_\alpha(n-1)$)
4: *service* = *re.studyservice*
5: *Else*
6: *service* = *oldservice*
7: *return service*

Algorithm 1: Update Algorithm.

the Nash equilibrium solution of each stage in a game with repeated times is the same solution as the one-time game; therefore, in the next life cycle of the data application service, the decision expression is the same as that in the previous stage.

5. Experiment

In the experiment, we use Adult dataset as the dataset T with sensitive information, which is the protection target. The original dataset has a total of 15 fields, 32561 records. We remove the nonquasi-identifier attributes and add a field of user's ID. A sample data set is shown in Table 3. We then simulate and generate the historical records of service development R, in which parameters are set according to the relevant literature [28, 29]. Table 4 shows the six sets of simulation parameters in the functions.

In the following experiment, we first aim to analyze the rationality of modeling functions and Nash equilibrium solution under the constraints of service quality and privacy protection in Section 5.1, and we then visualize the process that when we have a biased estimate of the functions, the estimation curve will approximate the actual curve by parameters adjustment in Section 5.2. At last, we compare the proposed model with the traditional k-anonymity method in Section 5.3.

We adopt GCP (Global Certainty Penalty) [30] to compare the proposed method with the traditional k-anonymity method and to measure the data availability under the same level of privacy protection. The range of GCP is [0, 1], where 0 means no information loss and 1 means total information loss.

5.1. Analysis of Modelling Functions. We randomly extract 80 sets of samples from Table 3, and each set has 100 records. The extracted sets of samples are used to calculate the probability of identifying a specific user based on the combination of quasi-identifier fields. The detailed process is to (1) calculate the probability that identifies each user in the samples according to the quasi-identifier fields and (2) calculate the average probability of all users. For example, when calculating the average probability of identifying a user in the samples under one quasi-identifier, we randomly select a quasi-identifier in the dataset and calculate the probability that value of the selected quasi-identifier for each user can identify the user, and the average probability of identifying a specific user under different numbers of quasi-identifier fields is shown in Table 5.

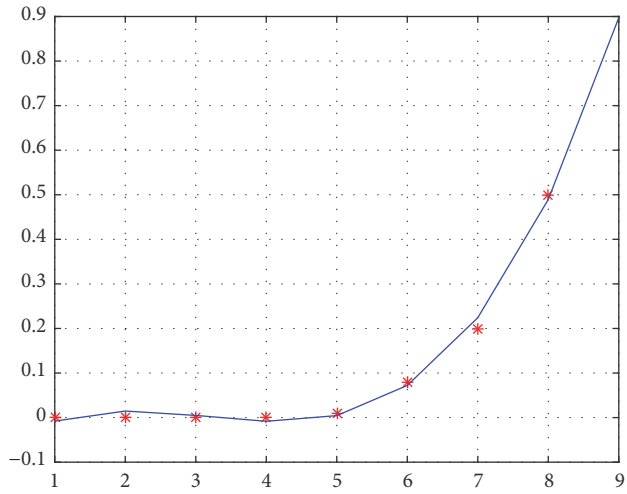

Figure 2: Simulation of formula (1).

We aim to prove that the Taylor third-order function mentioned in formula (1) can well simulate the relationship between the probability of privacy leakage and the number of quasi-identifiers first. In Figure 2, the red dot identifies the real sampling points, which are calculated from Adult dataset, and the blue curve is third-order Taylor function (the value of parameters can be determined by red dots: b_3 = 0.0049, b_2 = -0.0454, b_1 = 0.1248, and b_0 = -0.0921). We can see that the function curve fits the data very well, and it is reasonable to use the Taylor function to model the target variables.

We then aim to prove that the quality of service function and the Nash equilibrium solution are consistent with the reality, which is mentioned in formula (2), Section 4.3. As shown in Figure 3, we perform visual analysis of the quality of service function, and we can see that the curves under six sets of parameters in Table 4 basically follow the similar shape. The quality of service increases with the increase of a (mining technology investment) and Q (the loss of privacy leakage), and the function curve increases rapidly first, then the growth tends to be gentle. During the growth of the curve, the value tends to reach the highest service quality score, i.e., the full score ω in the function. Considering the user's privacy and the

Table 3: The sample of adult data set.

Id	workclass	education	marital_status	occupation	relationship	race	sex	native_country	income	agerank	numrank
A1	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	United-States	<=50K	30~40	13~16
A2	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	<=50K	50~60	13~16
A3	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	United-States	<=50K	30~40	9~12
A4	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	United-States	<=50K	50~60	5~8
A5	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	Cuba	<=50K	20~30	13~16
A6	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States	<=50K	30~40	13~16
A7	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	Jamaica	<=50K	40~50	5~8
A8	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K	50~60	9~12
A9	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	United-States	>50K	30~40	13~16
A10	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K	40~50	13~16

Table 4: The simulation of parameters in the functions.

simulation	λ	η	ω	β	γ	ς
1	246	82	50	10	0.2	0.4
2	459	153	67	13.4	0.3	0.6
3	30	10	78	15.6	0.1	0.2
4	87	29	73	14.6	0.1	0.2
5	318	106	58	11.6	0.3	0.6
6	291	97	45	9	0.2	0.4

Table 5: The average probability of identifying specific users.

Test	The number of quasi-identifier fields							
	1	2	3	4	5	6	7	8
test 1	0.00005	0.00016	0.00100	0.00261	0.01639	0.08333	0.16667	1.00000
test 2	0.00021	0.00060	0.00179	0.00307	0.01003	0.08001	0.14285	0.20000
test 3	0.00009	0.00013	0.00201	0.00211	0.01173	0.07990	0.14285	0.25000
test 4	0.00028	0.00078	0.00101	0.00324	0.01032	0.07720	0.25000	0.25000
test 5	0.00010	0.00014	0.00181	0.00201	0.01709	0.08070	0.25000	1.00000
test 6	0.00005	0.00092	0.00209	0.00200	0.02009	0.08003	0.16667	1.00000
test 7	0.00006	0.00040	0.00130	0.00209	0.02013	0.07693	0.16667	0.16667
test 8	0.00004	0.00072	0.00271	0.00350	0.01631	0.08000	0.16667	0.16667
test 9	0.00005	0.00055	0.00160	0.00311	0.02079	0.07932	0.14285	1.00000
test 10	0.00017	0.00031	0.00101	0.00301	0.01089	0.08702	0.14285	0.25000
average	0.00011	0.00048	0.00167	0.00269	0.01501	0.08077	0.17099	0.57121

response strategy of the data developer, the Nash equilibrium strategy is not close to the strategy combination with the highest service quality score, and the strategy takes a high point rather than the optimal point. In the untrustworthy cooperation process of reality, our intuition is that the target variable increases with the increase of both sides, the target variable will gradually approach the maximum value of the target variable, and because the two sides do not trust each other, they cannot achieve the best solution and the quality of service function can reasonably simulate the cooperation between data developer and data collector.

Next, we aim to prove that the profit function is reasonable and the maximum point is the Nash equilibrium solution. Since the revenue function of the data collector is related to three variables as shown in formula (3), Section 4.3, it is necessary to eliminate one variable a according to formula (6), Section 4.4. We then plot equivalent profit curves under (Q, k) space in Figure 4. We can find the six curves are all convex, and the Nash equilibrium of all curves corresponds to the maximum values. The profit of data collector increases first with respect to Q and k (the proportion of subsidies to the data developer), and the rate of increase becomes smaller as Q and k increase, and when it increases to the highest point, the function curve begins to drop. The revenue function of the data collector takes into account the quality of service, the loss of privacy leakage, and the strategic variables of the data developer, the maximum value of the function corresponds to the Nash equilibrium solution. The function curve is consistent with our expectation as follows: (1) the function value increases with the service quality score and (2) it is a convex function about the input cost.

5.2. Analyze the Iterative Process. In this section, we aim to prove that the training after sample weight adjustment can approximate the real function when the initial modeling is biased from reality, and the actual parameters of the functions in the experiment are $\lambda = 246$, $\eta = 82$, $\omega = 50$, $\beta = 10$, $\gamma = 0.2$, and $\varsigma = 0.4$.

When the absolute difference between the real value and the estimated value is greater than the specified threshold, we will adjust the weights of the new samples. As shown in Figure 5(a), we mark the real function curve and set the threshold of error to 2. We use MATLAB randomly generate 20 reasonable sample scatter points larger than the threshold and draw the initial estimation curve by gradient descent algorithm. At the end of the service, a real sample scatter is generated, so we take the point on the real curve and use the gradient descent algorithm to estimate the curve by the sample weight adjustment method in Section 4.6. When the function simulation curve is corrected several times according to the real data, the error between the predicted value and the true value is less than the threshold and the sample weight is no longer changed. Figure 5(b) represents the revenue function of the data collector, and the threshold is set to 100 and the iterative simulation process is similar with Figure 5(a).

5.3. Effectiveness Analysis of Privacy Protection. We aim to compare the effect of privacy protection of our proposed model with the traditional method k-anonymity, and the result is shown as Table 6, where P is the probability of privacy leakage and K presents the size of anonymity group when k-anonymity achieves the same privacy protection. We compare

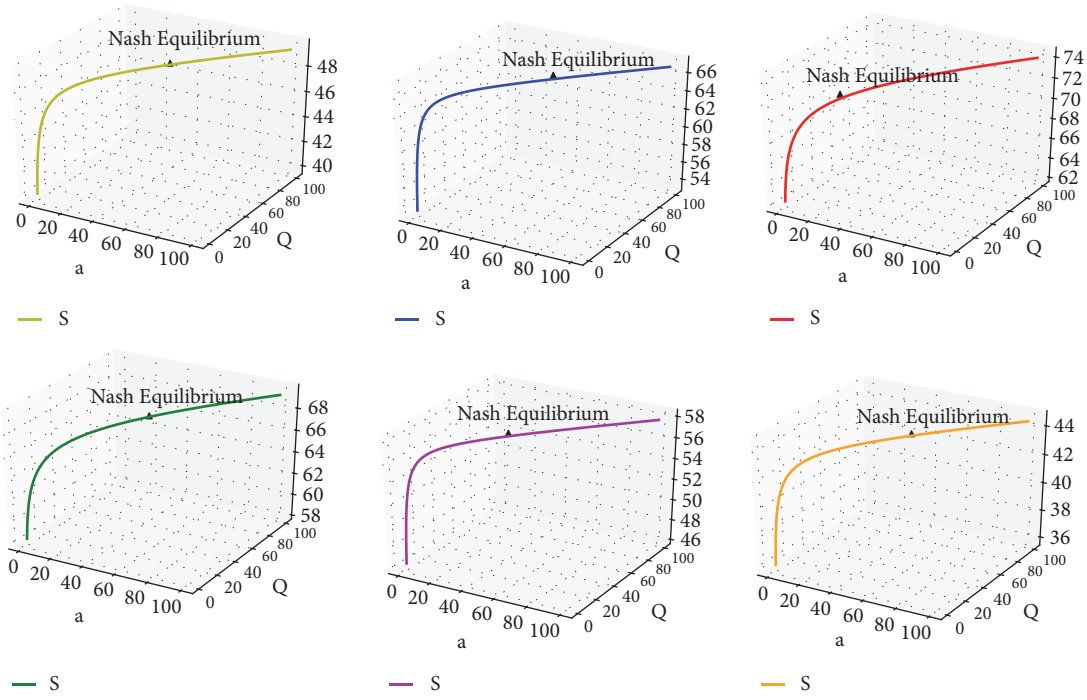

Figure 3: Simulation of formula (2).

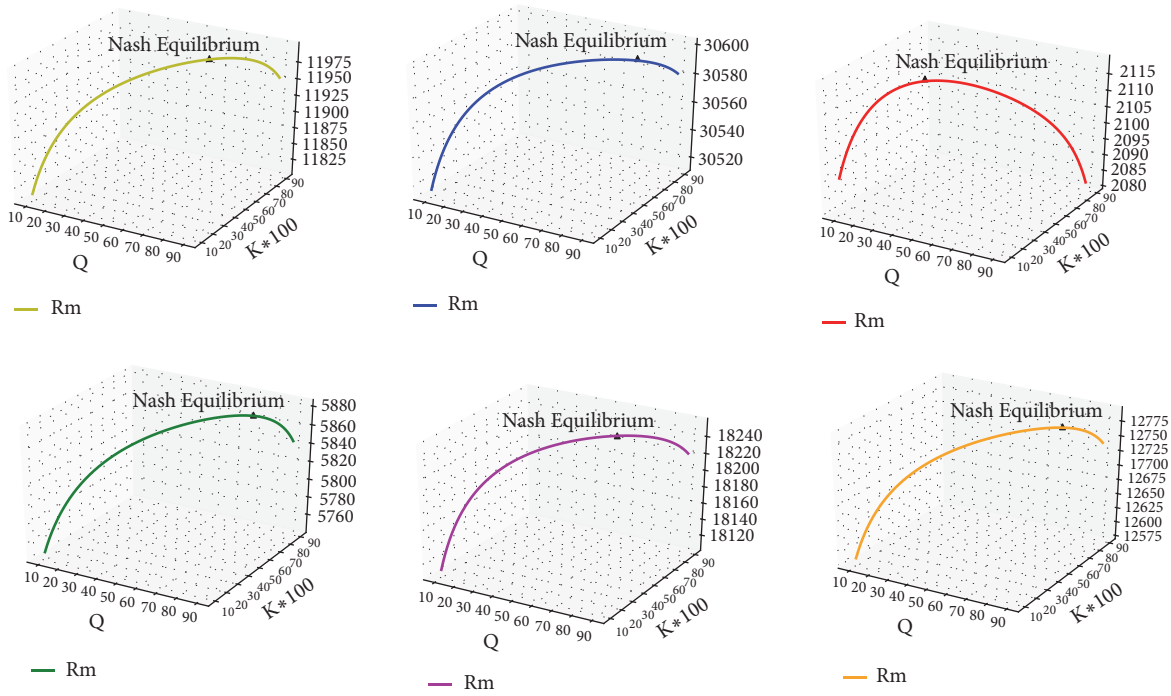

Figure 4: Simulation of formula (3).

the two methods from the GCP and time complexity. In our approach, *reduced number of fields* represents the number of quasi-identifier fields that are reduced.

In data applications, deleting the appropriate quasi-identifier fields will not affect the data mining results, but deleting too many quasi-identifier fields will reduce the upper bound of the accuracy. When k-anonymity algorithm is used for privacy protection, no matter how much K is equal to, it will cause loss of data information which limits the application of mining algorithm and the upper bound of accuracy. Generally, P is thought to be weak when P is between $[0.1, 0.5]$, middle when P is $[0.01, 0.1)$, and strong

Table 6: Comparison of the two methods.

P	k-anonymity			Game theory		
	GCP(%)	time complexity	K	GCP(%)	time complexity	Reduced number of fields
0.5	8.60		2	27.27		3
0.2	21.71		5	36.36		4
0.08	34.59		13	45.45		5
0.01	54.93	$O(n^2)$	100	54.54	$O(n)$	6
0.002	71.27		500	63.63		7
0.001	73.81		1000	72.72		8
0.0005	79.95		2000	81.81		9

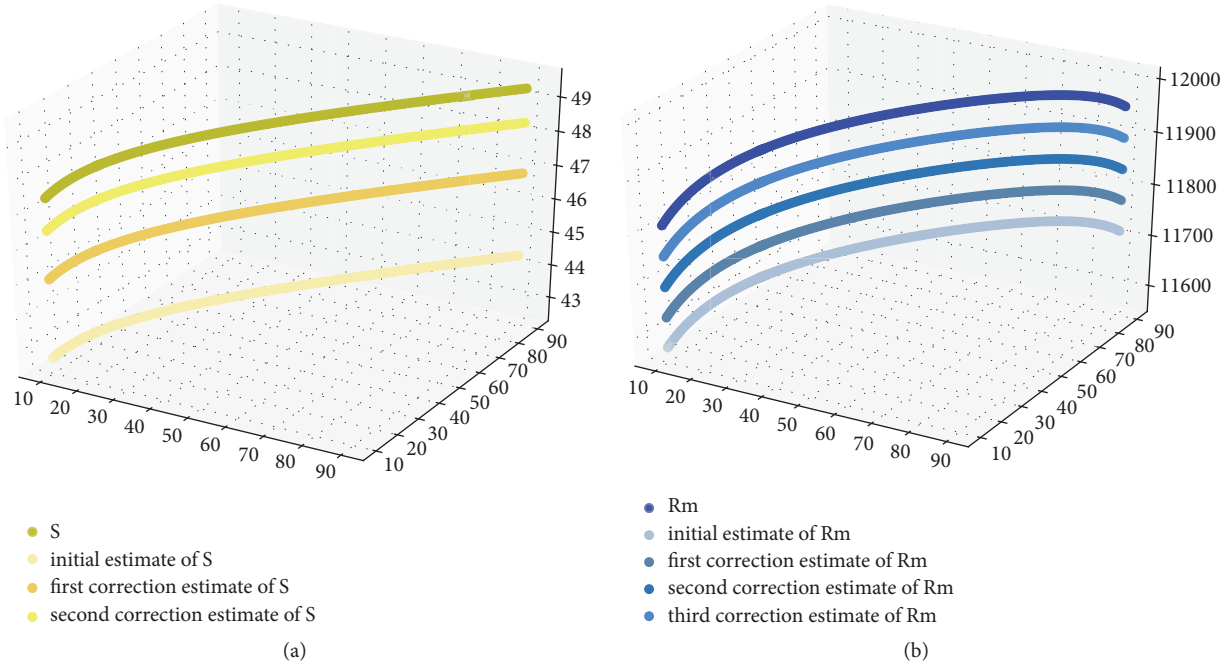

Figure 5: Iterative process of modeling functions.

when P is $[0, 0.01)$. GCP represents the information loss, 0 means no information loss and 1 means total information loss. The less the GCP, the better the data availability.

From Table 6 we can see that the GCP of k-anonymity increases stably when P changes from weak to strong, as well, and the size of k increases from 2 to 1000. The GCP of our proposed model also increases with P, and the reduced number of quasi-identifiers increases from 3 to 8. When users need week privacy protection, the GCP of k-anonymity is smaller than game theory model, and it is appropriate to select k-anonymity to process the data. But when users need middle privacy protection, our model performs better than k-anonymity. On the other hand, we calculate the number of published quasi-identifiers through the privacy protection model, and the data developer selects the corresponding number of quasi-identifier field from the source data fields, and when data needs to be applied with week privacy protection, fewer quasi-identifier fields need to be reduced and the truncated fields may have little impact on the target variable in practice.

Furthermore, we compare the time complexity of the two methods. Because our proposed privacy protection model relies on modeling functions, where the unknown parameters of the functions can be determined by gradient descent algorithm, we can use the stochastic gradient descent (SGD) [31] to solve the problem in practice and the time complexity of the SGD is $O(n)$. We use an improved k-anonymous algorithm [7] for comparison, and its time complexity is $O(n^2)$. Therefore, the game theory method is simple to implement in practical application and can adapt to the large amount of data.

6. Summary

This paper introduces the developmental characteristics of the data application service in the intelligent network system and analyzes the shortcomings of the privacy protection algorithm in solving such problems. On this basis, this paper proposes a privacy protection model based on game theory, which protects users' sensitive information by reducing the

number of quasi-identifier fields in the released data table, and the strategy calculated by the game model can simultaneously protect user privacy and service quality. This paper introduces the proposed architecture of the privacy protection model that is based on game theory and contains the realization of the process. Finally, we verify by experiments that the proposed privacy protection model can effectively protect the user privacy and service quality. However, in the development of the big data services of the intelligent network system, there is still a lack of algorithms and models that are effective and have the greatest degree of protection for user privacy. To achieve the maximum privacy and security of users, the relevant laws need to be further improved and a deeper study of the relevant issues in the industry is needed.

Data Availability

Previously reported Adult data were used to support this study and are available at http://archive.ics.uci.edu/ml/datasets/Adult. These prior studies (and datasets) are cited at relevant places within the text as [7].

Conflicts of Interest

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

Acknowledgments

The research is supported by National Natural Science Foundation of China (no. 61772560), National Key R&D program of China (nos. 2018YFB1003800, 2017YFB1400601), and National Science Foundation of China (no. 61772461).

References

[1] X. Zhang, W. Dou, Q. He et al., "LSHiForest: A generic framework for fast tree isolation based ensemble anomaly analysis," in *Proceedings of the 33rd IEEE International Conference on Data Engineering, ICDE 2017*, pp. 983–994, USA, April 2017.

[2] J. Zhang, Z. Zhou, S. Li et al., "Hybrid computation offloading for smart home automation in mobile cloud computing," *Personal and Ubiquitous Computing*, vol. 22, no. 1, pp. 121–134, 2018.

[3] L. Kuang, L. Yu, L. Huang et al., "A Personalized QoS Prediction Approach for CPS Service Recommendation Based on Reputation and Location-Aware Collaborative Filtering," *Sensors*, vol. 18, no. 5, p. 1556, 2018.

[4] L. Kuang, Y. Wang, P. Ma et al., "An Improved Privacy-Preserving Framework for Location-Based Services Based on Double Cloaking Regions with Supplementary Information Constraints," *Security and Communication Networks*, vol. 2017, Article ID 7495974, 15 pages, 2017.

[5] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, and J. Chen, "A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment," *Future Generation Computer Systems*, vol. 88, pp. 636–643, 2018.

[6] L. Qi, X. Zhang, W. Dou, and Q. Ni, "A distributed locality-sensitive hashing-based approach for cloud service recommendation from multi-source data," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2616–2624, 2017.

[7] H. W. Jiang, G. S. Zeng, and H. Y. Ma, "Greedy clustering anonymous method for privacy preservation of table-data publishing," *Journal of Software. Ruanjian Xuebao*, vol. 28, no. 2, pp. 341–351, 2017.

[8] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD)*, pp. 439–450, Dallas,Texas, 2000.

[9] J. Thanveer, "A Multiplicative Data Perturbation Method to Prevent Attacks in Privacy Preserving Data Mining," *International Journal of Computer Science and Innovation*, vol. 2016, no. 1, pp. 45–51, 2016.

[10] Z. Ming, W. Zheng-Jiang, and H. Liu, "Random projection data perturbation based privacy protection in WSNs," in *Proceedings of the 2017 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2017*, pp. 493–498, USA, May 2017.

[11] T. Jahan, G. Narsimha, and C. V. Rao, "Multiplicative Data Perturbation Using Fuzzy Logic in Preserving Privacy," in *Proceedings of the International Conference on Information and Communication Technology for Competitive Strategies. ACM*, pp. 1–5, 2016.

[12] V. S. Reddy and B. T. Rao, "A combined clustering and geometric data perturbation approach for enriching privacy preservation of healthcare data in hybrid clouds," *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 1, pp. 201–210, 2018.

[13] Y. Shen, R. Chen, and H. Jin, "Differentially Private User Data Perturbation with Multi-level Privacy Controls," in *Machine Learning and Knowledge Discovery in Databases*, vol. 9852 of *Lecture Notes in Computer Science*, pp. 112–128, Springer International Publishing, 2016.

[14] A. Kaur, "A hybrid approach of privacy preserving data mining using suppression and perturbation techniques," in *Proceedings of the 2017 IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017*, pp. 306–311, India, February 2017.

[15] K. Gai, M. Qiu, H. Zhao, and J. Xiong, "Privacy-Aware Adaptive Data Encryption Strategy of Big Data in Cloud Computing," in *Proceedings of the 3rd IEEE International Conference on Cyber Security*, pp. 273–278, China, June 2016.

[16] K. Gai, M. Qiu, and H. Zhao, "Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing," *IEEE Transactions on Big Data*, 2017.

[17] A. H. Aljammal, A. Alsarhan, A. Qawasmeh, H. Bani Salameh, and A. F. Otoom, "A new technique for data encryption based on third party encryption server to maintain the privacy preserving in the cloud environment," *International Journal of Business Information Systems*, vol. 28, no. 4, p. 393, 2018.

[18] H. Zhou, "Classification of Large Data Privacy Encryption Simulation Research," *Computer Simulation*, 2016.

[19] L. Sweeney, "k-Anonymity privacy protection using generalization and suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.

[20] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain *k*-anonymity," in *Proceedings of the Proceeding of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, pp. 49–60, June 2005.

[21] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, 25 pages, April 2006.

[22] G. Aggarwal, T. Feder, K. Kenthapadi et al., "Achieving anonymity via clustering," in *Proceedings of the Proceeding of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '06)*, pp. 153–162, New York-NY-USA, June 2006.

[23] J. Li, R. C. Wong, A. W. Fu, and J. Pei, "Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures," in *Data Warehousing and Knowledge Discovery*, vol. 4081 of *Lecture Notes in Computer Science*, pp. 405–416, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[24] Z. Wang, J. Xu, W. Wang, and B. Shi, "A Clustering-Based Approach for Data Anonymization," *Journal of Software*, vol. 21, no. 4, pp. 680–693, 2010.

[25] J. Yang, B. Zhang, J. P. Zhang, and J. Xie, "A k-anonymity clustering algorithm based on the information entropy," in *Proceedings of the 2014 IEEE the 18th Int' l Conf.on Computer Supported cooperative work in Design*, pp. 319–324, 2014.

[26] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, USA, 1944.

[27] J. Nash, "Equilibrium points in N-person games," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 36, pp. 48-49, 1950.

[28] N. Amrouche, G. Martín-Herrán, and G. Zaccour, "Pricing and advertising of private and national brands in a dynamic marketing channel," *Journal of Optimization Theory and Applications*, vol. 137, no. 3, pp. 465–483, 2008.

[29] R. Frank H and B. S. Bernanke, "Principles of microeconomics," McGraw-Hill Irwin, New York, 2007.

[30] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007*, pp. 758–769, Austria, September 2007.

[31] S. Ruder, *An overview of gradient descent optimization algorithms*, 2016.