International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-13

Article

Study on the Mechanism for Guaranteeing the Right to Automated Processing of Personal Information in Environmental Governance

Bona Song[1,*]

[1] School of Marxism, Jining Normal University, Ulanqab, Inner Mongolia, 012000, China
* Correspondence author: 302405@jnnu.edu.cn

Abstract: Aiming at the privacy risk brought by the automated processing of personal information in environmental governance, this paper proposes a set of systematic rights protection mechanisms. Firstly, the kernel principal component analysis method is used to reduce the dimensionality of information attributes by reducing the dimensionality and noise reduction of personal information. Then based on the privacy protection technique of stream cipher, the peak density clustering (CFSFDP) algorithm is introduced to improve the k -TBM anonymization algorithm, and a privacy protection method based on the k-TBM anonymization model of CFSFDP (CFSFDP-k-TBM) is designed. Finally, this paper conducts experimental tests on the CFSFDP-k-TBM algorithm, and compares and analyzes it. The test results show that CFSFDP-k-TBM algorithm has more effective privacy protection effect while ensuring users' personalized privacy needs. Therefore CFSFDP-k-TBM based anonymization model is an applicable privacy protection method for personal information in environmental governance.

Keywords: k -anonymization; peak density clustering; personal information privacy protection; environmental governance

1. Introduction

Environmental governance, as an important means of coping with global climate change and ecological degradation, has been the focus of attention of policymakers and academics in various countries [1-2]. The early environmental governance model was based on “command-and-control”, relying on government regulations and mandatory measures, but due to the high implementation cost, information asymmetry, and low public participation, it was difficult to fully mobilize social resources to achieve environmental protection goals [3-5]. With the complexity of environmental problems, the governance model has gradually transformed from single government-led to “polycentric governance”. In recent years, with the rapid development of technology, digital transformation has gradually become one of the key strategies to improve environmental governance [6-7]. Digital technologies, such as the Internet of Things, big data, artificial intelligence, etc., provide new tools and methods for environmental governance, which can more accurately monitor the environmental conditions, optimize resource allocation, enhance governance efficiency, and improve the scientific and effective environmental governance [8-11].

However, digital governance also brings serious risks of infringement of personal information rights. Technological development has enabled algorithms to cover many areas of social life with the help of ever-expanding data and ever-strengthening arithmetic power, processing personal information data at all times [12-14]. Under the impact of the rise of algorithmic power, the information subject gradually loses the ability to control and dominate personal information, at which time, informed consent can no

longer meet the needs of personal information protection [15-16]. In this context, the importance of systematic research on the rights protection mechanism of automated processing of personal information in environmental governance has become increasingly prominent.

As digital technology is widely used in environmental monitoring, pollution source tracking and environmental protection enforcement and other governance scenarios, the automated processing of personal information has become increasingly common, and there are some risks of privacy leakage. In this paper, we focus on the protection of personal information rights, and carry out dimensionality reduction processing of personal information to eliminate the noise data in the information. Combined with the theoretical technique of stream cipher encryption, the k-TBM anonymization model based on CFSFDP is proposed. The CFSFDP-k-TBM algorithm is tested, and the privacy protection method of k-anonymization is used as a reference comparison, and the algorithm is evaluated in five indexes: the success rate, the data availability rate after encryption, the loss metric, the identification of sensitive values, and the execution time.

2. Automated Processing of Personal InformationPrivate Information Data Preprocessing

2.1. Downgrading of personal information

In order to avoid a series of problems caused by the high dimensionality of the data, the idea of kernel local holding projection is introduced into the kernel principal component analysis method to do the dimensionality reduction of personal information, and the specific steps are as follows:

1) Denote the personal information dataset with $X=[x_1,x_2,\cdots,x_n]^T \in R^{n\times m}$, and establish the local structure preserving function $K_l(\beta)$:

$$\begin{aligned} K_l(\beta) &= \min_{\beta}\left\{\beta^T\Gamma(x_i)\Gamma^T(x_i)(F-D)\Gamma(x_i)\Gamma^T(x_i)\beta\right\} \\ &= \min_{\beta}\left\{\beta^T\Gamma(x_i)\Gamma^T(x_i)Z\Gamma(x_i)\Gamma^T(x_i)\beta\right\} \end{aligned} \tag{1}$$

where β represents the linear representation of the personal information samples in the feature space; $\Gamma(x_i)$ and $\Gamma^T(x_i)$ represent the kernel function; and $Z=F-D$ represents the Laplacian matrix, where F stands for the diagonal matrix, and D represents the weight matrix.

2) The global variance-maximizing objective function $K_g(\beta)$ is built on the basis of $K_l(\beta)$:

$$K_g(\beta)=\max_{\beta}\sum_{i=1}^{n}\left[\Gamma^T(x_i)\sum_{j=1}^{n}\beta_l\Gamma(x_i)\right]^2 \tag{2}$$

3) Combine $K_l(\beta)$ and $K_g(\beta)$ to establish the overall objective function $K(\beta)$:

$$K(\beta)=\max_{\beta}\left[\chi_C K_g(\beta)+\chi_L K_l(\beta)\right] \tag{3}$$

where χ_C represents the weight parameter of the global variance maximization objective function $K_g(\beta)$; χ_L represents the weight parameter of the local structure preservation function $K_l(\beta)$.

4) Calculate the first p eigenvalues $\mu_1,\mu_2,\cdots,\mu_p$ of the personal information by using the function $K_g(\beta)$, and construct the eigenvectors $S=[\beta_1,\beta_2,\cdots,\beta_n]$.

5) Do the projection of the sample set of personal information in the low-dimensional orthogonal feature subspace by the following formula:

$$Y=L^T S \tag{4}$$

where K represents the kernel function; Y denotes the personal information after the dimensionality reduction process.

2.2. Personal Information Denoising

The personal information after the dimensionality reduction process is input into the recurrent consistency generative adversarial network, and the denoising process is carried out on it.

1) Generative Adversarial Network

Generative adversarial network consists of discriminator and generator, the basic idea of the network is "zero-sum game". The discriminator and generator in the network training process is the main body of the "zero-sum game", in the process of model optimization and training to achieve Nash equilibrium, to obtain the optimal parameters of the generative adversarial network model, the generative adversarial network objective function is as follows:

$$\min_{G}\max_{D} B(D,G) = R_{y\sim P_d(y)}\left[\lg D(y)\right] + R_{x\sim P_x(x)}\left\{\lg\left[1-D(G(x))\right]\right\} \tag{5}$$

where G represents the generator; D represents the discriminator; $B(D,G)$ represents the value function, which is related to the generator and the discriminator; $P_d(y)$ represents the probability distribution of the personal information y; $D(y)$ represents the probability of deciding that the personal information y belongs to the real data; $P_x(x)$ is the probability distribution corresponding to x.

2) Cyclic consistency generative adversarial network

Compared with generative adversarial network, cyclic consistency generative adversarial network has the following advantages:

(1) The generative adversarial network does not need a one-to-one mapping relationship between personal information, and can accomplish the free conversion between target data and input data.

(2) The training process is strengthened by the cyclic consistency loss function.

The network consists of two generative adversarial networks, which belong to the loop network, sharing generators $G_{X,Y}$, $G_{Y,X}$ and having their own discriminators D_X, D_Y during the training process.

The cyclic consistency loss constraint is introduced into the data transformation process to prevent the generators $G_{X,Y}$, $G_{Y,X}$ from transforming all the personal information into the same region.

The cyclic consistency loss and the adversarial loss form the objective function of the cyclic consistency generating network:

1) The objective function of generating adversarial network 1 is denoted as $l_1(G_{X,Y},D_Y,x,y)$, and the objective function of generating adversarial network 2 is denoted as $l_2(G_{X,Y},D_x,y,x)$:

$$\begin{cases} l_1(G_{X,Y},D_Y,x,y) = R_{y\sim P_d(y)}\left[\lg D_y(y)\right] \\ \qquad\qquad + R_{x\sim P_x(x)}\left\{\lg\left[1-D_y(G_{X,Y}(x))\right]\right\} \\ l_2(G_{X,Y},D_x,y,x) = R_{x\sim P_d(x)}\left[\lg D_x(x)\right] \\ \qquad\qquad + R_{y\sim P_y(y)}\left\{\lg\left[1-D_x(G_{Y,X}(y))\right]\right\} \end{cases} \tag{6}$$

2) Only a single adversarial loss during the network training process cannot guarantee that the input personal information x generates the desired output y through the mapping process. In order to improve the training effect, the mapping functions $G_{X,Y}(\cdot)$ and $G_{Y,X}(\cdot)$ need to fulfill the cyclic consistency condition. Noise exists in the X-domain and no noise exists in the Y-domain, the personal information present in the X-domain is converted to the Y-domain, at which point the information x' is obtained, and the process can be expressed as $x \to G_{X,Y}(x) \to G_{Y,X}\left[G_{X,Y}(x)\right] \to x'$, and in the process of converting the data, it is avoided to convert all personal information existing in X domain to Y domain, and similarly convert personal information existing in Y domain to X domain to obtain data y', i.e., $y \to G_{Y,X}(y) \to G_{X,Y}\left[G_{Y,X}(y)\right] \to y'$, and according to the above process to obtain the cyclic consistency loss :

$$\begin{aligned} l_c\left(G_{X,Y},G_{Y,X}\right) = {} & R_{x\sim P_d(x)}\left\{\left\| G_{Y,X}\left[G_{X,Y}(x)\right]-x\right\|_1\right\} \\ & + R_{y\sim P_d(y)}\left\{\left\| G_{X,Y}\left[G_{Y,X}(y)\right]-y\right\|_1\right\} \end{aligned} \tag{7}$$

where $\|\cdot\|_1$ describes the L1 paradigm.

The final loss function $l_a\left(G_{X,Y},G_{Y,X},D_x,D_y\right)$ is established from the above analysis:

$$l_a\left(G_{X,Y},G_{Y,X},D_x,D_y\right)=l_1\left(G_{X,Y},D_y,x,y\right) \\ +l_2\left(G_{Y,X},D_x,y,x\right)+\mu l_c\left(G_{X,Y},G_{Y,X}\right) \tag{8}$$

where μ represents the control coefficients.

The objective function of the recurrent consistency generating adversarial network is constructed based on the loss function $l_a\left(G_{X,Y},G_{Y,X},D_x,D_y\right)$:

$$\min_{G_{X,Y},G_{Y,X}}\max_{D_x,D_y} l_a\left(G_{X,Y},G_{Y,X},D_x,D_y\right)=\min l_1\left(G_{X,Y},D_y,x,y\right) \\ +\min l_2\left(G_{Y,X},D_x,y,x\right)+\mu l_c\left(G_{X,Y},G_{Y,X}\right) \tag{9}$$

The personal information after dimensionality reduction is input into the above objective function to complete the denoising of the data.

3. Rights guarantee mechanism based on improved k-TBM anonymization model

3.1. Theory and Techniques of Privacy Protection Based on Stream Ciphers

1) RC4 stream cipher algorithm

The core implementation of RC4 algorithm is actually only two steps, which are key generator initialization algorithm (KSA), and pseudo-random generation algorithm (PRGA). The key generator initialization is mainly to set the initial state of the key generator, i.e., initialize the S box. The pseudo-random generation algorithm is mainly used to generate the key for each bit of the key stream, i.e., using the initialized S -box and the random pointers i and j to perform the dissimilarity operation.

The KSA process is mainly used to initialize the keystream generator using the seed key key to get the initial state S_0 .There are three main steps in KSA, which are:

(1) Initialize the S -box and initial original seed key key , which is stored in the S -box as $0 \sim 2^n$, usually $n=8$. The original seed key key is set internally by the system.

(2) Initialize a pointer variable $j=0$ as well as a pointer variable i traversing the S -box, using the state of the previous S and the current variable i to generate the value of a new pointer variable j , which is generated as shown in Eq. (10):

$$j=j+S[i]+key[i], i=0,1,\cdots,2^n \tag{10}$$

(3) Exchange the data in the S -box pointed by the pointer variables i and j . As the loop continues to iterate, the data in the S -box is continuously updated with transformations to obtain the keystream generator.

The principle of PRGA is the keystream generator generated according to the KSA algorithm, which generates a key for each transformation, i.e., the generated keystream is a sequence of pseudo-random numbers.The three steps in the work of PRGA keystream generator are as follows:

(1) Initialize the state table of the S -box according to the KSA, initialize two pointers i and j , pointer i cyclically traverses each element of the S -box, and pointer j as shown in equation (11):

$$j=j+S[i], i=0,1,\cdots,2^n \tag{11}$$

(2) Exchange the corresponding data in the S -boxes pointed to by the selected i and j positions;

(3) Transform the positions of the elements in the S -boxes in each round, and output the values of the positions pointed to by $S[i]+S[j]$ after each transformation, as shown in Equation (12). After many iterations of transformation, the key of n bit bytes is output until the key stream and the length of the plaintext are the same:

$$keystream=S\left[S[i]+S[j]\right] \tag{12}$$

2) Logistic mapping

Logistic mapping is one of the main algorithms for designing chaotic stream ciphers, which is

highly sensitive to the setting of the initial value, and the chaotic process of Logistic mapping is pseudo-random and non-periodic. In this paper, some properties of Logistic mapping are used to improve the stream cipher so that the key stream generated by the stream cipher is longer periodic and better randomized.Logistic mapping is a one-dimensional one-dimensional mapping as shown in Equation (13):

$$x_{i+1} = \mu x_i(1-x_i), i = 0,1,2,\cdots \tag{13}$$

3) Product Algebraic Systems

An algebraic system, referred to as an algebra, consists of a nonempty set S and k unitary or binary operations $f_1, f_2, \cdots, f_k$ on the nonempty set S, denoted $< S, f_1, f_2, \cdots, f_k >$. For example, an algebraic system denoted as $< M, \cdot >$ means that the nonempty set inside the algebraic system is M and the operations on M are ordinary multiplication operations.

Two algebraic systems are $V_1 = \langle A, \cdot \rangle$, $V_2 = \langle B, * \rangle$, where the operations $\cdot$, $*$ defined by V_1, V_2 are binary operations, and if $V_3 = \langle A \times B, \odot \rangle$, then we say that V_3 is the product algebra of V_1 and V_2. If $\langle a_1, b_1 \rangle \in V_1$, $\langle a_2, b_2 \rangle \in V_2$, then the product algebra $V_3 = \langle a_1 \cdot a_2, b_1 * b_2 \rangle$, where $\langle a_1, b_1 \rangle, \langle a_2, b_2 \rangle \in A \times B$.

3.2. CFSFDP-based data priming module

The CFSFDP algorithm attracted considerable attention as soon as it was proposed, and researchers have turned their attention to the algorithm to apply it to their own fields of research.The CFSFDP algorithm is based on the following two assumptions: 1) each cluster center is surrounded by neighboring data points with a local density lower than its own, and 2) each cluster center is very far away from other cluster centers with a higher density than its own .

For any data point i, CFSFDP needs to compute the values of 2 attributes: the local density attribute ρ_i of i and the distance attribute δ_i of data points with higher density than itself. The algorithm sets the dataset to be processed as $S = \{x_1, x_2, x_3, \cdots, x_n\}$, so that $IS = \{1, 2, \cdots, n\}$ is the corresponding indicator set and $d_{ij} = dist(x_i, x_j)$ is used to denote the distance between data points x_i and x_j, and the local density attribute ρ_i of data point i is defined as:

$$\rho_i = \sum\nolimits_j \chi\left(d_{ij} - d_c\right) \tag{14}$$

where: j belongs to IS when it is different from i, and $i \neq j$. The function $\chi(x)$ is:

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{15}$$

where: the parameter $d_c > 0$ is the truncation distance that needs to be specified manually; and ρ_i is the local density of the data point x_i, which is the number of data points in S that have a distance less than d_c from the data point x_i. Then the distance property δ_i for i is:

$$\delta_i = \begin{cases} \min\limits_{\substack{j \\ \rho_j > \rho_i}}(d_{ij}), & i \geq 1 \\ \max\limits_j(d_{ij}), & \rho_i \text{ is the highest in the entire system} \end{cases} \tag{16}$$

For each point x_i in the dataset S, its density attribute and distance attribute (ρ_i, δ_i) can be calculated. As in the figure clusters, the data point division is done in one step and does not have to be calculated repeatedly and iteratively like other clustering algorithms. The setting of threshold d_c in the peak density algorithm is the most important step, which is directly related to the merits of the algorithm. The recommended practice is to choose d_c so that the average number of neighbors for each data point is 1%-2% of the total number of data points.

The initial partition module process is described as follows:

1) Data preprocessing to define the distance method.

2) Set the threshold d_c so that the neighbors of each data point are about 1%-2% of the total number of data points.

3) Calculate the density property $\rho_i = \sum_i \chi\left(d_{ij} - d_c\right)$ of the data points.

4) Sort the data points according to the density.

5) Calculate the clustering property δ for each data point.

6) Discover the clustering center: the data point with the largest $\rho \times \delta$ is selected as the clustering center; the remaining data points are automatically assigned to clusters where their respective nearest neighbors with densities greater than their own are located.

3.3. Improved k-TBM anonymization models

In the k-TBM algorithm although it can be better to cluster the closer tuples into an equivalence class, but if the amount of data is very large there will also be the problem of consuming too much time, and the NFPN partitioning with the possibility of grouping the more distant tuples into the same equivalence class, so before the k-TBM algorithm firstly use the CFSFDP to partition the tuples once a wide range of tuples, and then perform the k- TBM algorithm this can avoid the kind of situation above. In addition, separately performing k-TBM on each cluster after CFSDP can also greatly reduce the time consumed by the computation, to achieve both reduce the data loss and save the computational time overhead. The specific process of k-TBM anonymization algorithm based on CFSFDP clustering is described as follows:

1) Take the anonymized dataset D, the degree of anonymity K, and the truncation distance d_c.

2) Calculate the relative distances of all tuples stored in R.

3) Perform CFSDP clustering of all tuples based on relative distances, and present the noise temporarily.

4) Perform TBM for each cluster separately (first perform MDS dimensionality reduction and project it to a 2D plane. After that the tuples of Tn are concatenated with a partitioning algorithm and partitioned in k groups. Finally the partitioning result is used to generalize the data table. (Replace all data with the center of mass of each partition to form D').

5) Add the noise to the nearest equivalence class.

6) Output D_n.

4. Experiments and analysis of results

4.1. Experimental data and pre-processing

The experimental data were preprocessed before the experiment, and the user spatio-temporal data needed for the experiment were chosen to be retained, and finally the spatio-temporal data of 20,000 of them were preferred as the experimental dataset. In order to test the performance of the k-TBM anonymity model based on CFSFDP proposed in this paper, the programming language of compilation is chosen to be Python, and the corresponding pre-processing of the experimental data needed before the experiment is made, and the PS-TPA algorithm is chosen to judge the spatio-temporal privacy point-of-interest dataset P_n, and the personalized needs of the user's personalized needs are used with the parameter of privacy protection level $R = PR_{uk}$.

4.2. Experimental setup

In order to evaluate the method, the controlled experimental algorithm is a k-anonymity based privacy preserving method. The validated experiment has three main metrics: success rate, encrypted data availability rate and data processing time.

Where encryption success rate and encrypted data availability rate are defined as follows:

1) Success Rate

The metric of success rate ASR is defined to evaluate the effectiveness of this spatio-temporal data protection method, which is calculated as in Equation (17):

$$ASR = 1 - \frac{|P_n| - |\{P_e \mid P_e = S - TGES(p), p \in P_n\}|}{|P_n|} \tag{17}$$

2) Data Availability Rate

Define Data Availability Rate DAR this metric is used to evaluate the encrypted spatial and temporal privacy data in the associated users to detect the availability of spatial and temporal data services, calculated as in Equation (18):

$$DAR = \frac{|\{P_e \mid P_e = F(g_e^1, g_e^2), g_e^1 \in P_n, g_e^2 \in P_n\}|}{|P_n|} \tag{18}$$

In the experiment of availability of encrypted spatio-temporal data, the primary validation of spatio-temporal data availability in the Linked User Detection service after cryptographic protection by CFSFDP-k-TBM. In order to better quantify this evaluation metric of Linked User Detection, it is formally defined as Linked User Detection function as in Equation (19):

$$\begin{aligned} F(g_e^1, g_e^2) &= (g_e^1 \times h^- = g_e^2 \times h^-) \vee (g_e^1 \times h^+ = g_e^2 \times h^+) \\ &\vee (g_e^1 \times h^+ = g_e^2 \times h^-) \vee (g_e^1 \times v^- = g_e^2 \times v^-) \\ &\vee (g_e^1 \times v^- = g_e^2 \times v^+) \vee (g_e^1 \times v^+ = g_e^2 \times v^+) \end{aligned} \tag{19}$$

where: g_e^1, g_e^2 are each arrays of encrypted spatio-temporal data identifiers of the two users within the grid. Define $F(g_e^1, g_e^2)$ to be 1 if there is any overlapping intersection between the two lattice's horizontal identifier coordinates h^+ and h^-, or any overlapping intersection between the two vertical identifier coordinates v^+ and v^-, and 0 otherwise.

4.3. Experimental results and analysis

4.3.1. Comparison of success rates

According to the definition of ASR performance index in the actual verification experiments, as long as any item in the user's initial spatio-temporal data is re-identified or matched, it represents that this encrypted spatio-temporal data is successfully attacked. In the experiment, different numbers of spatio-temporal data p_n are given respectively, and the real spatio-temporal data of the user is re-recognized and judged by the relevant techniques. Validation experiments are conducted on CFSFDP-k-TBM service respectively from different sizes of privacy protection parameters chosen by users themselves to verify and analyze CFSFDP-k-TBM and k-anonymity based protection methods, and the results are shown in Figure 1.

When the number of spatio-temporal privacy interest points to be protected is small, the success rate of both protection algorithms is high because of the lack of background information that can be obtained. However, as the number of privacy interest points in the spatio-temporal data becomes larger, CFSFDP-k-TBM gradually outperforms the data protection method based on k-anonymization. If the LBS follows the grid-encrypted projection of the user's spatial data, its coordinates still enable services such as Linked User Detection, but in this way, there is no real personal information of the user in the user's spatio-temporal data that is finally saved to the server of the social networking site.

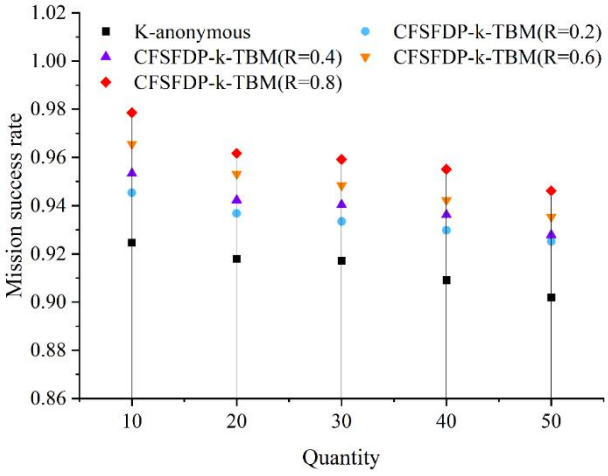

Figure 1. Success rate comparison

4.3.2. Comparison of availability rates

Validating the data availability of the CFSFDP-k-TBM method for connected user detection, the results are shown in Fig. 2. In the case where the number of privacy interest points in the user's

spatio-temporal data is relatively small, the use of the spatio-temporal privacy data is not efficient because the distribution of the spatio-temporal privacy data selected from the dataset is not very tight. On the whole, however, the CFSFDP-k-TBM-based privacy protection scheme outperforms the k -anonymization-based protection system scheme in terms of data availability for connected user detection.

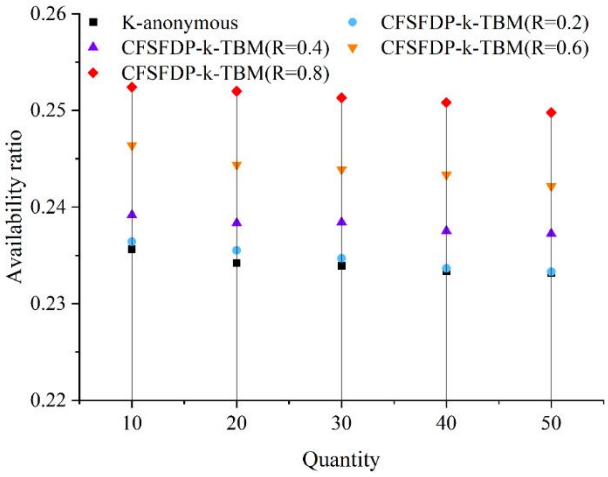

Figure 2. Availability comparison

Considering the user's own needs, according to the personalized privacy protection parameters that the user chooses to use, so the protection strength of the CFSFDP-k-TBM method is set accordingly to the user's own choice of privacy needs R. The results of the experimental validation are shown in Fig. 3. Overall, the data processing time of the method based on the k -anonymization idea varies too much as the number of privacy interest points in the user's spatio-temporal data becomes larger, while the data processing time of the CFSFDP-k-TBM method varies more steadily. When the number of privacy interest points in the spatio-temporal data is small, the CFSFDP-k-TBM data processing time is smaller than that of the data protection method based on k -anonymization. When the amount of spatio-temporal information of the user is large, CFSFDP-k-TBM is able to directly use the encrypted lattice that has already been established, which reduces the time for forming the encrypted lattice and also makes the data processing speed gradually better than the method that adopts the k -anonymization idea.

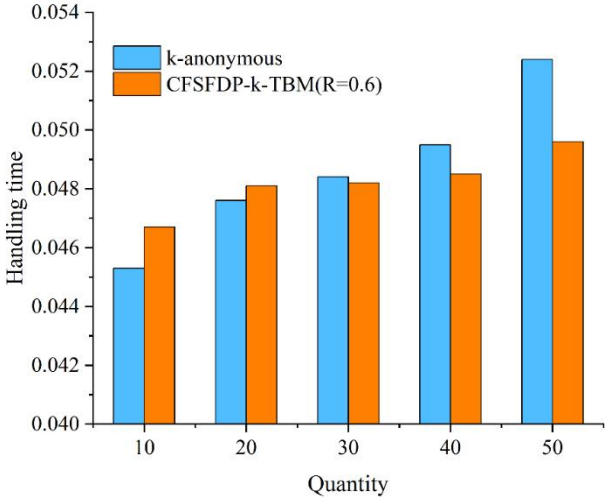

Figure 3. Processing time comparison

4.3.3. Information loss metrics

The spatio-temporal dataset will be compared and experimented based on k -anonymization algorithm, l -diversity algorithm and CFSFDP-k-TBM algorithm and analyzed in terms of the amount

of information loss of the algorithms, the recognition rate of the sensitive values, and the execution time to validate the feasibility of CFSFDP-k-TBMP algorithm.

The information loss is calculated by calculating 35483 validation data. The values of k were taken in increasing order from 3, 5, 7 to 10, respectively. The amount of information loss of the privacy preserving model obtained after the calculation of the information loss generated by the average generalization height is shown in Figure 4. It can be seen that as the value of k increases, the amount of information loss in the dataset increases, because the size of k directly affects the size of the equivalence classes in the data, and the larger the value of k is, the more values of quasi-identifier data that need to be generalized for the same equivalence class aggregation, and so the greater the amount of information loss. When the value of k increases to a certain level, the difference in the amount of information loss between the l -diversity algorithm and the k -anonymization algorithm is not significant, this is because a fixed value of l does not constrain the dataset significantly when the value of k increases to a certain level. Overall, the CFSFDP-k-TBM algorithm has a greater amount of information loss relative to the other two algorithms, which is caused by its increased personalized protection of sensitive values.

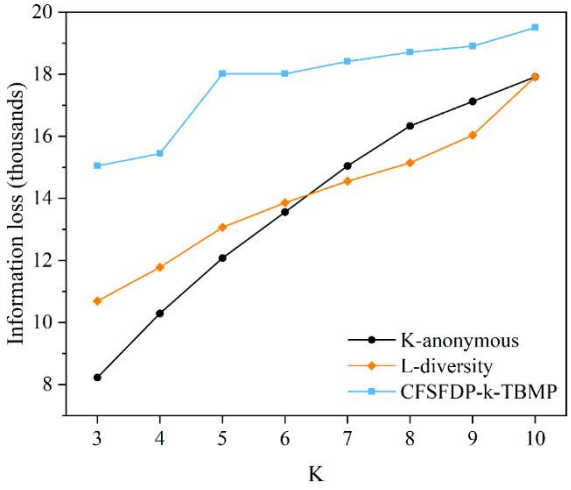

Figure 4. The amount of information loss varies with the k value

The data set size takes the value of respectively from $5k-30k$ sequentially increasing. A comparison of the amount of information loss of the algorithms under the condition that the dataset is constantly changing when $k=5$ is shown in Figure 5. The amount of information loss is all growing as the dataset increases, which is due to the fact that as more and more data needs to be continued to be anonymized, which leads to an increase in information loss. The reason for the largest amount of information loss in the CFSFDP-k-TBM algorithm is the increased level of privacy protection due to the addition of personalized anonymization operations for sensitive attribute values in the CFSFDP-k-TBM algorithm.

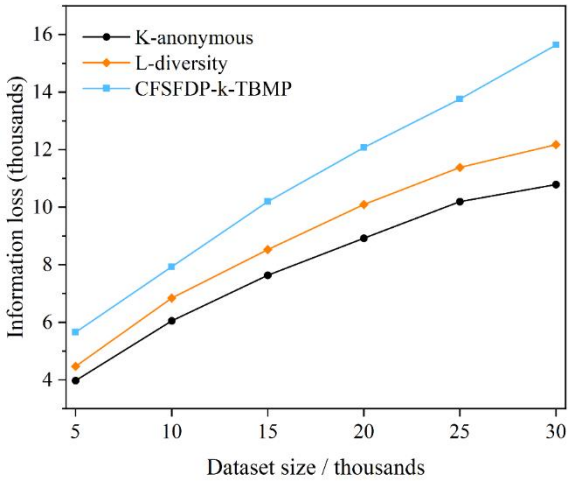

Figure 5. The amount of information loss varies with the dataset size

4.3.4. Analysis of the recognition rate of sensitive values

The privacy breach risk metric focuses on assessing the security of anonymized datasheets, and the metric principle is to assess the likelihood of inferring records in the original datasheet from records in the anonymized datasheet, which is generally measured using the degree of correlation between the original datasheet and the same record in the anonymized datasheet. For an anonymized dataset, the lower the probability that its sensitive attributes are identified by an attacker, the better the privacy protection of the dataset. In this paper, we use the average identification rate of sensitive attribute values to quantify the degree of privacy protection of a dataset.

Definition (Recognition rate of sensitive attribute values): given a dataset D and an equivalence class C, and s is the sensitive attribute value of a piece of data t in the equivalence class C, the recognition rate of the sensitive attribute value s of t in C is computed as:

$$RR_t(s,C) = \frac{|(s,C)|}{|C| \times |f(s)|} \tag{20}$$

where $|(s,C)|$ is the number of sensitive attribute values s in the equivalence class C, $|C|$ is the size of the equivalence class, and the value of $|f(s)|$ is equal to the number of leaf nodes of the subtree in which the parent is located after the generalization of the hierarchical tree of the generalization of the sensitive attribute, and the name of the original location in the tuple is retained after the anonymization if the sensitive value has not been generalized, when $|f(s)|=1$.

The recognition rate of sensitive values in the dataset is calculated using Equation (20), with the values of k taken in increasing order from 3, 5, 7 to 10, respectively. Figure 6 gives the sensitive value recognition rate of the privacy preserving algorithm when the value of k varies in the same data context as in Figure 4. As the value of k changes, the sensitive value recognizable rate of the algorithm decreases and eventually levels off. Among them, the CFSFDP-k-TBM algorithm has the lowest sensitive value recognizable rate and the best privacy preserving effect compared to the other two algorithms.

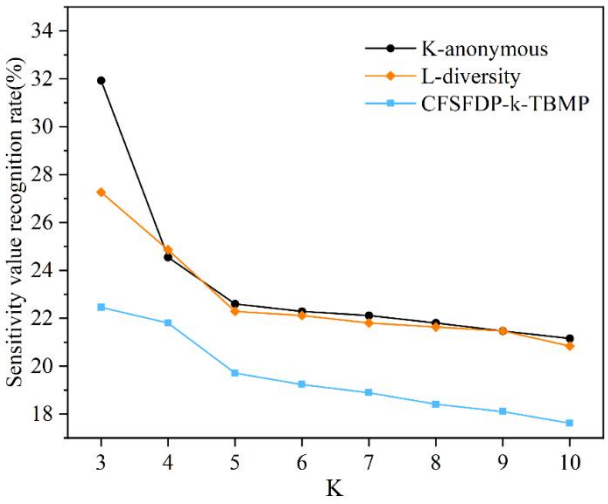

Figure 6. The recognition rate of sensitive values varies with the k value

The dataset sizes are taken in increasing order from 5 k -30 k, respectively. Figure 7 gives a comparison of sensitive value recognition under different datasets when k = 5. Overall, the dataset size has basically no effect on the data sensitive value recognition, indicating that the stability of the algorithm does not fluctuate significantly with the increasing dataset size. Under the same parameter conditions, the CFSFDP-k-TBM algorithm has the smallest sensitive value recognition rate and the best privacy protection.

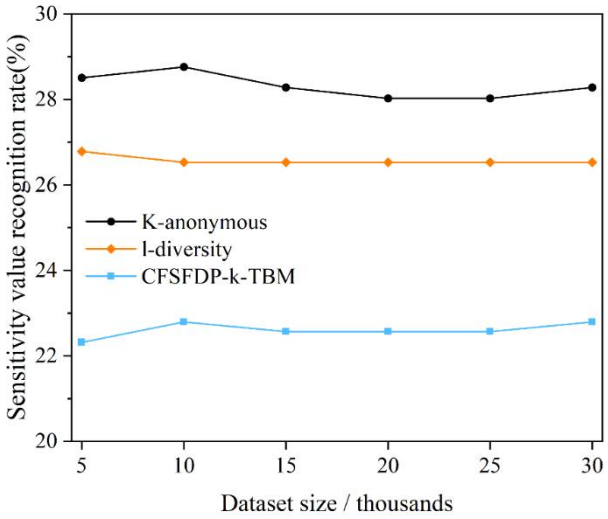

Figure 7. Identification rate of sensitive values varies with dataset size

4.3.5. Implementation time analysis

The values of k are taken in increasing order from 3, 5, 7 to 10, respectively. The execution time of the algorithm according to the change in the value of k is shown in Fig. 8. When the value of k changes, the execution time of the algorithms will have some ups and downs. When k = 4, the l -diversity algorithm and the CFSFDP-k-TBM algorithm will have a long execution time as the privacy parameter l and α in the privacy-preserving model are set to increase the determination time when dividing the equivalence classes. And as the value of k increases, there is little difference in the execution time of the algorithm, so the CFSFDP-k-TBM algorithm is better at preventing privacy leakage.

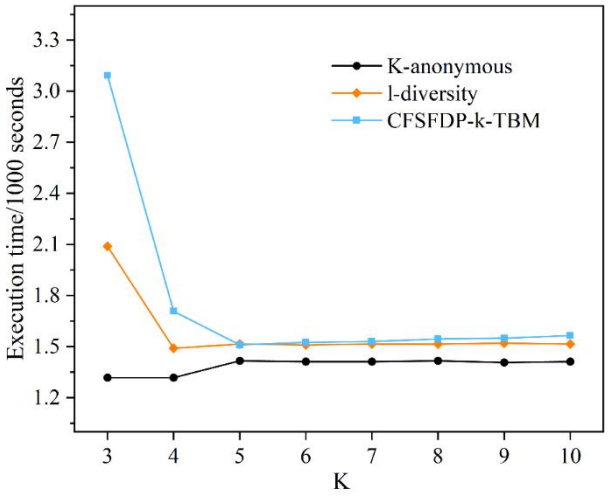

Figure 8. Execution time varies with the k value

The dataset size takes the value of respectively from $5k-30k$ in increasing order. The execution time of the algorithms varying with the size of the dataset is shown in Fig. 9. The execution time of the algorithms are growing as the size of the dataset changes, which is due to the fact that the amount of data to be processed becomes larger. However, it can still be seen from the comparison graph that the growth of the CFSFDP-k-TBM algorithm tends to level off as the dataset grows in $15k-20k$, and it is hypothesized that as the data continues to grow, the CFSFDP-k-TBM algorithm's execution time will eventually maintain a smaller gap between it and the l -diversity and k -anonymization algorithms.

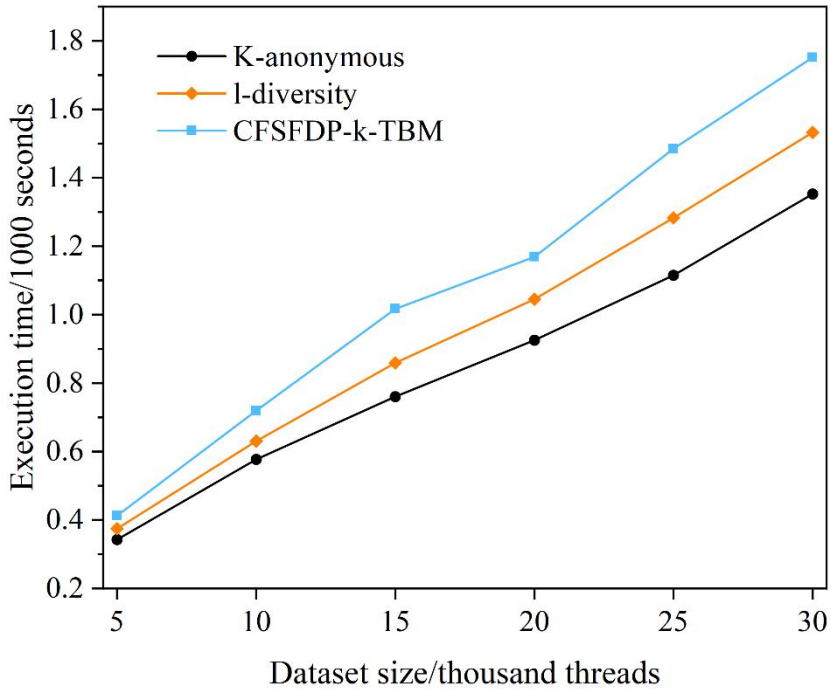

Figure 9. Execution time varies with dataset size

5. Conclusion

The study firstly launched the degradation and denoising process for personal information, and then studied the privacy protection mechanism based on stream cipher, RC4 stream cipher algorithm, chaotic mapping, and product algebra system to propose a micro-aggregation-based k-anonymization model (CFSFDP-k-TBM) oriented to privacy protection of personal information, and finally proved the applicability of CFSFDP-k-TBM model through experiments. The following conclusions are obtained:

1), the effectiveness of the service is evaluated in terms of three indicators: the success rate of CFSFDP-k-TBM algorithm, the data availability rate after encryption and the data processing time. The experimental results show that CFSFDP-k-TBM algorithm has a better privacy protection effect while ensuring users' personalized privacy needs.

2) Both k -anonymization algorithm and l -diversity algorithm cannot satisfy the privacy constraints of all tuples. Users can set their own sensitivity attribute values for the identifiers in the data release according to their privacy protection sensitivity level.The CFSFDP-k-TBM algorithm can satisfy users' personalized privacy protection needs, which verifies the effectiveness of the algorithm.

About the Author

Bona Song was Born in Ulanqab, Inner Mongolia, China in 1985.She received her Doctor of Laws (LL.D.) degree from China University of Political Science and Law. She is currently working at Jining Normal University. Her main research interests including environmental and resources protection law, constitutional law, jurisprudence and legal practice.

References

1. Bennett, N. J., & Satterfield, T. (2018). Environmental governance: A practical framework to guide design, evaluation, and analysis. Conservation Letters, 11(6), e12600.
2. Brondizio, E. S., & Tourneau, F. M. L. (2016). Environmental governance for all. Science, 352(6291), 1272-1273.
3. Agrawal, A., Brandhorst, S., Jain, M., Liao, C., Pradhan, N., & Solomon, D. (2022). From environmental governance to governance for sustainability. One Earth, 5(6), 615-621.
4. Moon, K., Blackman, D., Brewer, T. D., & Sarre, S. D. (2017). Environmental governance for urgent and uncertain problems. Biological Invasions, 19(3), 785-797.

5. Zhang, B., Cao, C., Gu, J., & Liu, T. (2016). A new environmental protection law, many old problems? Challenges to environmental governance in China. Journal of Environmental Law, 28(2), 325-335.
6. Ren, S., Hao, Y., & Wu, H. (2023). Digitalization and environment governance: does internet development reduce environmental pollution?. Journal of Environmental Planning and Management, 66(7), 1533-1562.
7. He, G., Jiang, H., & Zhu, Y. (2024). The effect of digital technology development on the improvement of environmental governance capacity: A case study of China. Ecological Indicators, 165, 112162.
8. Zhu, B., Li, S., & Chevallier, J. (2025). The impact of digitization on ecological environment governance capacity: Evidence from China. Environmental Impact Assessment Review, 115, 108014.
9. Kostka, G., Zhang, X., & Shin, K. (2020). Information, technology, and digitalization in China's environmental governance. Journal of Environmental Planning and Management, 63(1), 1-13.
10. Guo, J., & Shen, X. (2024). Does digitalization facilitate environmental governance performance? An empirical analysis based on the PLS-SEM model in China. Sustainability, 16(7), 3026.
11. Li, P. (2024). Digital technologies, environmental governance and environmental performance: Empirical evidence from China. Engineering Economics, 35(2), 236-248.
12. Gong, X., Wang, J., & Xiao, Z. (2025). Effects of digitization on environmental governance in public services: evidence from China. Public Money & Management, 1-13.
13. Kloppenburg, S., Gupta, A., Kruk, S. R., Makris, S., Bergsvik, R., Korenhof, P., ... & Toonen, H. M. (2022). Scrutinizing environmental governance in a digital age: New ways of seeing, participating, and intervening. One Earth, 5(3), 232-241.
14. Mendes, V., & Viola, E. (2023). Green digitalization? Agriculture 4.0 and the challenges of environmental governance in Brazil. In Sustainability challenges of Brazilian agriculture: Governance, inclusion, and innovation (pp. 207-226). Cham: Springer International Publishing.
15. Layode, O., Naiho, H. N. N., Adeleke, G. S., Udeh, E. O., & Labake, T. T. (2024). Data privacy and security challenges in environmental research: Approaches to safeguarding sensitive information. International Journal of Applied Research in Social Sciences, 6(6), 1193-1214.
16. Xu, J., She, S., & Liu, W. (2022). Role of digitalization in environment, social and governance, and sustainability: Review-based study for implications. Frontiers in psychology, 13, 961057.