Research on Desensitization Algorithm of Power Consumption Data Based on Anonymity and Differential Privacy Technology

Bin Xu, Xiaobing Liang, Feng Zhai
Electric Power Research Institute, Beijing, China

Abstract. With the development of information technology, the problem of privacy disclosure has become increasingly prominent. State Grid Corporation has a large number of sensitive data such as power marketing data, power customer data, personal electricity information, which has disclosure dangers during the production, transmission, storage, processing and sharing stage. This paper mainly studies the desensitization algorithm model of power consumption data based on anonymity and differential privacy technology, builds a general algorithm model and then realizes data desensitization according to each application scenario. The evaluation and analysis results show that the scheme can effectively protect the massive power marketing data, power customer data and personal power information from invasion and disclosure.

1. Introduction
With the rapid development of mobile Internet, cloud computing and Internet of things, the digitization of personal information is deepening. People can collect and use data through massive sensors and intelligent processing terminals, which promoted the arrival of the era of big data.

Big data is like a double-edged sword. It not only accelerates the application of various industries, but also makes personal privacy invisible. Sensitive data leakage is common. If we want to better meet the data sharing between users and systems, and better protect the privacy information and sensitive information between customers in enterprises, we must study data desensitization methods and technologies to ensure data security while using data. State Grid Corporation of China attaches great importance to the work of network security. The marketing department has formulated the code for desensitization of customer sensitive information of marketing specialty of State Grid Corporation of China, which guides all provincial companies and relevant directly affiliated units to carry out the work of desensitization of customer sensitive information. For the current large amount of sensitive data such as power marketing data, power customer data, personal electricity information, etc., a perfect data desensitization and privacy protection scheme is designed based on each application scenario. Carry out the desensitization research of electricity sensitive data, build the desensitization algorithm model of electricity data based on anonymity and differential privacy technology. The model can more accurately meet the needs of different users, better implement all kinds of sensitive data desensitization specifications formulated by the company, and effectively guarantee the information security within the system of State Grid Corporation.

Data anonymity is a privacy protection technology based on restriction of publication. Its purpose is to cut off the one-to-one correspondence between the data owner and the sensitive information in the original data set, and produce the anonymous data set that can meet the privacy protection needs and
ensure the availability of data, so as to achieve better results. Differential privacy is a new privacy protection model, which does not need to consider the background knowledge of the attacker. Based on the powerful mathematics, the model defines the privacy protection strictly and provides a quantitative evaluation method, which makes the level of privacy protection provided on the dataset comparable. Based on anonymity and differential privacy technology, the model of electricity data desensitization algorithm has better overall performance than other models, so it is selected as the final model of State Grid Corporation for data desensitization.

2. Related technologies

2.1. Anonymization Model

Data anonymity technology is a kind of privacy protection technology based on restriction of publication. The application of data anonymity technology can effectively cut off the corresponding relationship between data owners and sensitive information. Among them, K-anonymity, L-diversity and T-closeness are the most widely used data anonymity technology models, which can effectively guarantee the confidentiality and integrity of data transmission and storage process. Next, each anonymization model will be introduced in detail.

2.1.1. K-Anonymity Model: The concept of k-anonymity model was proposed by Samarati and Sweeney. The model requires that the records in the data set are at least indistinguishable from a certain number of records. Because it is usually assumed that the attacker uses the semi identity attribute group to identify someone, the idea of k-anonymity technology is that the published data set must ensure that each equivalent class (a set of the same semi identity attribute values) has at least the same K group records, reducing the probability of being re identified. The details of k-anonymity model are as follows:

Firstly, user data is divided into explicit identifier, quasi identifier, sensitive attribute and non-sensitive attribute according to different data attributes.

Next, k-anonymity is realized by generalizing and suppressing identification attributes. The combination of generalization and suppression can minimize the number of generalizations needed to satisfy k-anonymity. K-anonymity model has been widely studied since it was proposed. However, in k-anonymity model, only the constraints of quasi identifier attributes are considered, while the constraints of sensitive attributes are not considered, as well as the privacy disclosure caused by attacking anonymous data sets through privacy reasoning. Moreover, k-anonymity is realized by using generalization and suppression operation. Due to the reduction of data details, data its utility will inevitably decrease. To maximize the utility of data and calculate the optimal k-anonymity protection data set is a very difficult problem. Therefore, the data after k-anonymization may still be attacked, such as homogeneity attack and background knowledge attack, which leads to the disclosure of privacy information.

2.1.2. I-Diversity Model: Due to the defects of k-anonymity model, Gehrke proposed an l-diversity model which can resist homogeneous attack and background knowledge attack. Based on k-anonymity, l-diversity model requires at least l distinguishable values in the same equivalence class. When the equivalence class meets the above premise, it is said that the equivalence class has l-diversity.

Homogeneity attack is an equivalent class with the same sensitive attribute value in the data table processed by k-anonymity. The attacker can infer the specific sensitive attribute value information according to the equivalent class with the same attribute value. The attack of background knowledge occurs when the attacker uses some additional information to reduce the uncertainty of sensitive attribute value to the target user. Thus, the sensitive attribute value information can be inferred.

I-diversity method requires each equivalence class of anonymous data to contain at least different sensitive attribute values, that is, each equivalence class must contain at least different sensitive attribute values. Data regulators generally use higher values to avoid privacy disclosure. Similar to k-anonymity, - diversity protects privacy by reducing the utility of data.
2.1.3. T-Closeness Model: Skew attack is an attack based on the frequency distribution of sensitive attribute values in an equivalent class. According to the known information of the target user and the probability of each situation, the sensitive attribute value of the user is inferred. Similarity attack occurs in anonymous tables after I-diversity processing. For example, if the value of "release" sensitive attribute on an equivalent class is infectious, it can be inferred that the target is a dangerous group.

In order to prevent the above two attacks, Li proposed the T-closeness model. The model overcomes the shortcomings of K-anonymity and I-diversity models, and requires that the distribution of sensitive attribute values in all equivalence classes be consistent with the global distribution of the attribute. In this way, the probability of skew attack to a certain target user's sensitive attribute value will be the same as the probability distribution of the whole data set. In other words, the value of the sensitive attribute of the target user will not change, and the attacker cannot get more privacy information from it. Similarly-closeness can reduce the efficiency of similar attack. In the T-closeness method, we need to calculate the distance between the frequency distribution of the sensitive attribute values in the published table and each equivalent class. There are different criteria for calculating these distances, such as the earth mover distance.

2.2. Anonymization Method

2.2.1. Generalization. Generalization is the use of more abstract, generalized values or intervals instead of precise values. The generalization technology based on big data can be divided into numerical type and subtype according to the attributes of quasi identifier. The way of numerical type and subtype generalization is different. After numerical type generalization, an interval covering accurate values will be formed, while the classification data will be generalized to a wider range of values.

There are two schemes for generalization: global generalization and local generalization. Global generalization includes three patterns: global generalization, subtree generalization and sibling node generalization. Local generalization includes two global generalization schemes: cell generalization and multi-dimensional generalization. All the same attribute values must be generalized. Global generalization means that all values of an attribute must be generalized at the same level, and the generalization granularity of all paths in the generalization tree must be the same. Subtree generalization refers to the generalization of all child nodes of the same parent node in the tree, either all or none. In the early stage of generalization, the whole generalization scheme was adopted, which resulted in a great loss of information. Therefore, researchers proposed a local generalization scheme. In local generalization, the same attribute values are not all generalized, and some of them remain unchanged. Unit generalization is a typical local generalization, which only generalizes a part of the value of an attribute, and keeps the other values unchanged. Multi dimension generalization is also a local generalization, but it is different from the first four generalization modes in that the first four generalize the value of a single attribute, and multi dimension generalization can generalize the value of individual attributes at the same time. Multidimensional generalization only needs to generalize the equivalence class that violates the specified threshold, and requires all records in an equivalence class to generalize to the same value.

2.2.2. Inhibition. Suppression operation is to delete or hide the data in the data table directly. In general, it uses "*" instead of the value to be suppressed, which is the most coarse-grained generalization. Inhibition is generally not used alone, as an auxiliary means in combination with generalization.

3. Desensitization algorithm model of power consumption data based on anonymity and differential privacy technology

3.1. General Algorithm Model

According to the actual situation of the State Grid, a desensitization algorithm model of power consumption data is constructed. The model framework consists of five stages: data preprocessing,
sensitive data identification and grading, desensitization strategy customization, desensitization task configuration and data distribution.

The model framework of electrical data desensitization algorithm is shown in Figure 1. The desensitization system extracts data from the relevant databases and data files of the power consumption information collection system and the marketing application system according to the user's needs, preprocesses the extracted source data, identifies and classifies the sensitive data, selects desensitization algorithm and sets parameters, and completes the desensitization strategy formulation. After the desensitization strategy is customized, users can choose the desensitization execution mode, in which static desensitization can be used for development, testing, data migration and storage; Dynamic desensitization can provide desensitization services for data analysis systems such as the whole business unified data center through agent mode. If there is no new data or configuration requirements, desensitization policies and tasks can be saved in the desensitization system for subsequent calls and execution.

The desensitization algorithm model flow is shown in Figure 2. After extracting data from the source system, the desensitization system should make appropriate strategies for these data.
3.1.1. Data Preprocessing
Due to the different methods of sensitive information identification of different types of data, the system needs to classify the source data first. In addition, the identification of multiple professional sensitive information will also bring a lot of interference to the identification process, seriously affecting the accuracy of sensitive information identification. According to different source business systems, source data can be classified into human property, planning, construction, operation, maintenance and marketing data. Then, the classification of the source data for feature extraction, through data feature matching to achieve sensitive information recognition. In order to reduce the amount of calculation in the process of sensitive information identification and classification, we adopt the form of automatic data dictionary to collect structured data, and collect a certain number of sample data from the business data table through the method of data table traversal, and then filter and generalize it to eliminate the data "impurities".

3.1.2. Classification of Sensitive Data Identification. Sensitive data recognition is the key premise of data desensitization. Through the training set, the characteristic database of text and structured data is obtained. The security department and business personnel identify and classify the corpus and the characteristic database together, and select the representative ones to form the sensitive information
database. Combining the sensitive information pattern matching and the importance of the source business system, the sensitive level value is set manually to be used for the classification of sensitive information. After preprocessing the target data for feature extraction, the extracted feature value is matched with the feature value of the sensitive information database. When the match hits, the system automatically records the sensitive level value of the current sensitive information. At last, the error classification is corrected by identifying quality evaluation, and the other sensitive information is supplemented.

3.1.3. **Desensitization Strategy Customization**

(a) **Data Encryption**

Data encryption technology can be divided into two types: irreversible encryption and reversible encryption. Among them, there are many ways to deal with the data, including using camouflage data to replace the sensitive data in the source data, using camouflage symbols to mask part of the content of the source data, directly deleting the sensitive data, and using random functions to control the transformation of the source data.

Reversible encryption has low operability. It can clone and mask the data in the production environment, generate the data with the same format and association with the original data, output it to the test environment, perform function test, performance test, simulation test and other tasks, or encrypt the plaintext in a specific format into the cipher text with the same format for storage and transmission.

(b) **Data Anonymization**

The purpose of data anonymity technology is to cut off the one-to-one correspondence between the data owner and sensitive information in the original data set, and to produce anonymous data sets that meet the needs of privacy protection and ensure the availability of data. See chapter 2 for details.

(c) **Differential Privacy**

Laplace mechanism is often used in local models, often used to protect the privacy of numerical results, by adding random noise obeying Laplace distribution to the exact query results to achieve \((\varepsilon, \alpha)\)-differential privacy protection; Gauss mechanism meets \((\varepsilon, \alpha)\)-differential privacy. In the process of data publishing, the Gauss mechanism is used to add the Gauss noise to the original data, which makes it realize differential privacy.

3.1.4. **Desensitization Task Configuration.** After completing the desensitization strategy formulation, in order to make the desensitization task work for a long time, first register the address and port number of the business system where the source data is located in the desensitization system. Then, the desensitization strategy is obtained, and the desensitization system generates desensitization code according to the selected desensitization algorithm and relevant parameters.

The user selects the desensitization implementation method according to the application scenario. For the static desensitization, the system first performs the desensitization operation, and caches the desensitization results in the local storage. When the target system needs to obtain the desensitization data, the user registers the address and port of the target system in the desensitization system, and finally transmits the local desensitization data to the target system. For dynamic desensitization, the user must first register the address, port and account number of the target system in the desensitization system, then send the desensitization code to the proxy server, and the proxy server will desensitize the online data, return the desensitization result to the desensitization system, and finally transmit it to the target system, which will be used by the account number of the data demander in the target system.

3.1.5. **Release and Application of Desensitization Data.** Static desensitization data is generally used in non-production environment or offline desensitization processing for online data. After desensitization, it is used in non-production environment, mainly used for data batch sharing, system development and testing and other typical business requirements. Dynamic data desensitization is generally used in the production environment. It does not change the original data in the production database. It only desensitizes the "input request" and "output data" in real time to prevent the leakage of sensitive data.
This form of desensitization is suitable for dynamic access and retrieval of production data, usually combined with access rights.

3.2. Application Scenario Validation

The external flow of power consumption data is one of the important application forms of state grid. Whether it is the access request of power consumption customers or the data transaction of external units, it involves that the user's personal information should not have the risk of privacy disclosure in the process of outgoing, so it is necessary to analyze and desensitize the user's privacy information in the large data environment. Before the desensitization algorithm model is developed, data demand analysis shall be carried out according to the application scenarios of each power consumption data first. According to the desensitization specification for marketing professional customer sensitive information of State Grid Corporation of China, the desensitization data to be desensitized and the data not to be desensitized shall be distinguished, and the desensitization strategy adapted to this scenario shall be formulated in combination with the demand analysis results and technical perspective.

3.2.1. Public Release Scenario

(a) Government Agencies

According to the application scenario of government agencies, an algorithm model based on the maximum information coefficient and data anonymity is proposed. Through the combination of information theory, we use data anonymity and differential privacy to solve the problem of big data publishing and sharing. Firstly, a small number of quasi identifier attributes and sensitive attributes are selected from the original dataset as the feature set, then the maximum information coefficient is used to analyse the correlation between the remaining attributes and the feature set attributes in the original dataset, and the data with high correlation is selected as the privacy dataset. Finally, the collaborative privacy protection algorithm proposed for the sensitive attribute dataset is applied, and the publication meets the difference Privacy protected big data sets.

(1) Data preprocessing and sensitive data identification and grading

According to the privacy level, we classify the electricity data set, select a few attributes from the attributes that are not desensitized as the feature set, and use the maximum information coefficient to select the data with high correlation as the output data set, only protect the privacy attributes, so as to improve the efficiency of data processing.

(2) Desensitization strategy customization

Firstly, the data set is divided into mutually exclusive sub data sets by using hybrid micro aggregation, and then the size of each sub data set is fixed to, finally, each sub data set is aggregated and Laplace noise is added to it to realize differential privacy protection. Data anonymity based on Hybrid Micro aggregation can differentiate query function sensitivity and improve data availability.

(3) Algorithm implementation

Firstly, some sensitive attributes and quasi identifier attributes are selected as the feature set from the original data set, and the remaining attributes of the original data set are selected as the candidate set. Using the maximum information coefficient and heuristic search algorithm, the correlation between the candidate set and the feature set is calculated, and the feature data with high correlation is selected as the output data set. Through the maximum information coefficient, we can accurately find the dependence between features, classify the original data set, and select the attributes that need privacy protection. Then, based on the improved hybrid micro aggregation algorithm and differential privacy, HM-DP algorithm is proposed to protect the privacy of the data in the dataset.

(4) Desensitization data

After the desensitization algorithm processing, the privacy protection data set is output and released to the National Energy Administration for data analysis.

(b) Ministry of Housing and Urban Rural Development

For the scenario of the Ministry of housing and urban rural development, a desensitization model based on clustering anonymity and differential privacy protection is proposed. Firstly, the data sets are
classified according to the quasi identifier, so that the data can meet the requirements of k-anonymity model. In the process of anonymization, all data attributes are regarded as quasi identifiers. Then, Laplacian noise is added to the anonymous grouped data to disturb the real value of the data record, so as to achieve the privacy requirements of the differential privacy protection model. Compared with the conventional Laplacian mechanism, clustering operation is mainly used to reduce the sensitivity of query function, and the reduction of sensitivity makes the added noise reduce, so as to greatly enhance the data availability.

1) Desensitization strategy customization
Firstly, each data record in the original data set is transformed and normalized according to the data attribute type, and the distance between any two data records is calculated respectively. The data of classified attributes are transformed into data of numerical attributes, and the distances of classified attributes in any two data records are calculated respectively. The data of numerical attributes are normalized, and the distances of numerical attributes in any two data records are calculated respectively. The transformed and normalized data sets are clustered to obtain multiple data clusters with independent attributes.

2) Algorithm implementation
Firstly, the transformed and normalized data sets are clustered to obtain multiple data clusters with independent attributes, and then the multiple data clusters are divided by micro aggregation anonymous method to obtain multiple equivalence groups with preset threshold number of size, and calculate the centroid value of each equivalent group, and replace the data record value of the corresponding equivalent group with the centroid value, To obtain multiple equivalent groups after anonymization. For any equivalence group in the multiple equivalence groups, after calculating and deleting any data record in the equivalence group respectively, the query function is obtained to query the sensitivity of the equivalence group. According to the parallel combination property of differential privacy protection, noise is added to the equivalence group of deleting the data record. For any equivalence group in the multiple equivalence groups, respectively delete any equivalence group After a piece of data is added with noise, the query set satisfying differential privacy of the equivalence group and the real query set of the equivalence group are calculated for similarity, and when the similarity is greater than 0, the data set satisfying differential privacy protection of the equivalence group is released.

3) Desensitization data
After the desensitization algorithm processing, the privacy protection data set is output and published to the Ministry of housing and urban rural development for data analysis.

3.2.2. Exchange and Share Scenarios
(a) Statistical Query
Power consumption data in the company is mainly used for statistical query, audit management and other exchange and sharing business scenarios. In the process of data exchange and sharing, there is a risk of privacy information disclosure. It is necessary to desensitize the relevant data to ensure the security and reliability of the user's privacy information in the internal use process.

Power grid companies usually analyze the payment data of users, and need to use the power consumption data in the process of carrying out various businesses and management. The data for desensitization should be selected according to the corresponding roles of operators. Therefore, it is necessary to extract the relevant fields in the relevant data table from the power consumption information collection system and marketing business application system as the algorithm input.

In this scenario, $(\alpha_{ij}, k, m)$-anonymous model based on multi sensitive attribute classification is proposed to classify the attribute values of multiple sensitive attributes. The concept of classification table is introduced. Each sensitive attribute is set with a classification table and a frequency constraint for each level $\alpha_{ij}$. 
(1) Model design
To protect personal privacy information is mainly to protect sensitive information about personal identity or identity. The electricity data set is composed of explicit identifier, quasi identifier, sensitive attribute and non-sensitive attribute.

\((\alpha_{ij}, k)\)-Anonymous model. According to the sensitivity of attribute \(SA\), the data set is divided into levels \(\alpha_i\) from high to low, and a corresponding constraint frequency is set for each level. The sensitive values in the equivalence class are required to meet the frequency constraint of the sensitive level to which they belong. When completing \((\alpha_{ij}, k)\)-anonymous grouping, the model is said to satisfy \((\alpha_{ij}, k)\)-anonymous if records with the same level sensitive value are prevented from being in the same group.

\((\alpha_{ij}, k, m)\)-Anonymous model. On the basis of satisfying \((\alpha_{ij}, k)\)-anonymous, the sensitivity of an equivalence class with sensitive attributes to the sensitive values of other sensitive attributes can be divided into two categories: set a corresponding constraint frequency according to the corresponding level of semantics, and requiring that all the sensitive attribute values of the equivalence class meet the frequency constraint of their own level. When completing \((\alpha_{ij}, k, m)\)-anonymity grouping, preventing if records with the same level of sensitive values exist in the same group, then the model is called \((\alpha_{ij}, k, m)\)-anonymous.

(2) Algorithm design
In the process of algorithm implementation, in order to form an equivalence class, the record insertion with the highest local priority is preferred, and then the record insertion with the lowest local priority is selected, which circulates in turn, and finally forms an equivalence class that meets the conditions. In the same equivalence class, the greater the sensitivity distance between the two levels, the higher the degree of privacy protection, so the highest and lowest priority is selected By inserting the two levels with the largest local sensitivity distance, the privacy protection degree of the improved \((\alpha_{ij}, k, m)\)-anonymous model is improved.

(3) Desensitization data
After \((\alpha_{ij}, k, m)\)-anonymity processing, the power data set of privacy protection can be output, and the customized desensitization strategy can be realized.

(b) Audit Management
The internal restriction system of electric power marketing is a reliable guarantee for the smooth development of electric power marketing. With the continuous development of enterprise internal marketing information, all relevant business processes, information and control are closely integrated in the marketing system, and they penetrate each other. Business risk also exists in the information system. In order to conduct comprehensive audit management, it is necessary to continuously collect, sort out and calculate marketing related business data and data by means of information technology, verify and analyze the authenticity and integrity of marketing system data, find clues by using the cross check relationship between reports and marketing system and the logical relationship between internal data of information system, and finally realize audit management by finding problems.

For this scenario, a differential privacy data publishing algorithm model based on principal component analysis is proposed

The main purpose of differential privacy data publishing algorithm based on principal component analysis is to retain the information of the original data as much as possible, which has general generality. By analysing the characteristics of the original data set, it projects the original data into another space according to the different targets, and obtains the corresponding projection matrix. Then it adds the random noise obeying the Laplace distribution to the projection matrix, and finally returns the projection matrix to the original publication, so that the published data can be related queries and numbers in the framework of differential privacy According to the excavation and other work to provide support.

After the differential privacy data publishing algorithm based on principal component analysis is processed, the data set for privacy protection is output.
3.2.3. Development and test scenarios. In the process of software development, it is necessary to extract relevant fields from relevant data tables from the power consumption information collection system and marketing business application system for function test. Considering the impact of desensitization data on function test, when desensitizing the corresponding fields, the reserved lattice encryption algorithm model is used to process the power consumption data. The relevant fields in the relevant data table are extracted from the power consumption information collection system and the marketing business application system as the algorithm input. There are a large number of Chinese fields in the marketing business system, which are user privacy data and need to be desensitized. Therefore, an encryption model based on user name reservation format is proposed.

1) Model design
One of the most difficult points of the reserved format encryption of user name is to ensure that the encrypted text conforms to the naming habit of Chinese name. In order to solve the above problems, the Cycle-Prefix algorithm and the reserved format encryption model based on user name are proposed.

2) Algorithm flow
The algorithm first initializes the parameters according to the key, reads the user database, and generates the replacement table of last name and first name. Then the name is transferred into the name extraction algorithm, which will automatically try to distinguish the last name and the first name, and extract the last name and the first name. If the extraction fails, the name library extension algorithm will be triggered to maintain the name extension. Once the extraction of surname and first name is successful, we can build a mapping relationship to get the surname code and first name code, then calculate the adjustment factor and use Cycle-Prefix algorithm to encrypt the surname code and first name code.

3) Algorithm implementation
Combined with the idea of cyclic encryption, this paper proposes a Cycle-Prefix format preserving encryption algorithm. Firstly, the validity of the input data is detected, and the illegal data is filtered to reduce unnecessary operations and avoid interference to the algorithm. After that, initialization is needed to read the name database data and set the basic construction method of encryption. Taking the secondary key provided by the user as the random seed generation adjustment factor, the number of encryption rounds of the first name is generated according to the generated adjustment factor, and the number of encryption rounds of the last name is generated according to the adjustment factor and the length of the first name. Then we use the mode of "encryption after coding" to encrypt the last name and last name in specified rounds. Finally, according to the mapping relationship, we get the corresponding cipher text and splice it into the final result.

4) Desensitization data
After the desensitization algorithm processing, the privacy protection data set is output.

4. Evaluation and analysis of data desensitization Technology

4.1. Evaluation Standard
The difficulty of data desensitization is to keep the integrity of data. Based on shielding sensitive data in non-production environment data, the production data can be extracted and deformed to ensure that the data after deformation can maintain the dependence between the original production data attributes and data.

The evaluation of data desensitization technology can be mainly considered from the aspects of sensitive information removal degree, data defect, calculation cost, communication cost, etc.

1) Removal degree of sensitive information
The degree of removal of sensitive information is relative to the original data. For example, if the last four digits of the customer ID number are randomly replaced, the region and birthday information with sensitive information will still be retained after desensitization. If all the digits are replaced, all the sensitive information will be removed.
(2) Data defect
Data defect is a measure of data quality after desensitization, which is reflected by data information loss after data desensitization. The more information is lost, the higher data defect is, and the lower data utilization rate is. For example, if * is used to replace the middle birthday segment of ID card number, the data defect is high.

(3) Calculation cost
The cost of computation is mainly evaluated by time and space complexity, which is related to hardware and software environment. Generally speaking, the smaller the computation cost is, the better. The encryption and decryption algorithm consumes a lot of computation cost, and the data distortion / interference has a little computation cost.

(4) Communication cost
The communication cost is mainly evaluated by the amount of interactive information and the number of rounds. Generally speaking, the smaller the communication cost is, the better the communication security is.

4.2. Comparison and Analysis
Compare and analyze the adaptability of desensitization technology based on data distortion / disruption technology (differential Privacy), data anonymity technology (k-anonymity, l-diversity, t-proximity), data encryption technology (encryption technology based on attribute and identity) in the data environment of electricity information collection, as shown in table 1.

| Technical comparison | Removal degree of sensitive information | Data defect degree | Computational overhead | Communication overhead | Comprehensive performance |
|----------------------|----------------------------------------|--------------------|------------------------|------------------------|--------------------------|
| Based on data distortion / scrambling technology | in                                    | high               | low                    | low                    | high                     |
| Based on data encryption technology                  | high                                  | low                | high                   | high                   | low                      |
| Data anonymization Technology                         | high                                  | in                 | in                     | low                    | in                       |

Traditional encryption technology is composed of symmetric, asymmetric and hash algorithms, which has high security strength and can ensure the confidentiality and integrity of data in the process of transmission. However, because the data must be completely decrypted when it is used, the sensitive data is still plaintext for the end-user, so it cannot meet the needs of sensitive data security and availability at the same time. Through the comparative analysis of the three data desensitization technologies, based on the data distortion technology, the performance efficiency is relatively high, but there is a certain degree of data defect and information loss; based on the encryption technology, the accuracy and security of data can be guaranteed, but the calculation cost is relatively large; based on the data anonymity technology, the authenticity of data can be guaranteed, but there will be information loss.

5. Conclusion
Based on the anonymity and differential privacy technology, a general algorithm framework is proposed for the power consumption information collection system in the big data environment, and a power consumption data desensitization algorithm model suitable for different business scenarios is
constructed. The experimental results show that the model has higher accuracy, better automation, better anti stealing ability and stronger expansion ability, which can better meet the sharing and exchange of information between users and systems, and better protect the privacy information and sensitive information between users. It provides a new idea for desensitization of power system data.

Acknowledgments
Sponsor Research on the application of artificial intelligence technology in the power consumption information collection system. JL71-18-022.

Author profile: Bin Xu, male, master, information security engineer

References
[1] Viktor M Schonberger, Kenneth Cukier. Big data: a revolution that will transform how we live, work and think [M]. Boston: Houghton Mifflin Harcourt, 2013.
[2] EMC Education Services. Data science and big data analytics: discovering, analyzing, visualizing and presenting data [M]. New Jersey: Wiley, 2015.
[3] David Belanger, Linda Pittenger. Big data [M]. California: APress, 2014.
[4] Sebastian L R, Babu S, Kizhakkethottam J J. Challenges with big data mining: a review[C]. 2015 International Conference on Soft-Computing and Networks Security, 2015, 1 - 4.
[5] B C M Fung. Privacy-preserving data publishing [D]. Simon Fraser University, 2007.
[6] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression [R]. SRI Computer Science Laboratory, 1998: 1 - 19.
[7] Sweeney L. K-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10 (5): 557 - 570.
[8] Machanavajjhala A, Gehrke J, Kifer D. l-Diversity: privacy beyond k-anonymity [C]. Proceedings of the 22nd IEEE International Conference on Data Engineering, 2006, 24 - 36.
[9] Sweeney L. Computational disclosure control: a primer on data privacy protection [D]. Massachusetts Institute of Technology, 2001.
[10] Lefevre K, Dewitt D J, Ramakrishnan R. Incognito: efficient full-domain k-anonymity [C]. Proceedings of the 24th ACM SIGMOD International Conference on Management of Data, 2005, 49 - 60.