Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220096

Face to Electricity Data Transmission of Composite Differential Private Recommend Method Research

Zhimin ZHAN[a,1], Jie XING[a], Ke ZHANG[a], Xiao LI[b] and Bin LUO[c]
[a] *State Grid Hubei Electric Power Co. Ltd., China*
[b] *State Grid Wuhan Power Supply Company, China*
[c] *Hubei Central China Technology Developmentd Of Electric Power Co.Ltd, China*

Abstract. With the construction and development of China's smart power grid, it has realized the power information interconnection, but also realized the collection of electricity user information fine, to bring convenience to users, but also led to the risk of privacy leakage. The traditional method of privacy protection has certain limitations to the protection of users' privacy data: only to a certain extent to protect the privacy of users. With the advent of advanced technologies such as machine learning, the ability of attackers to speculate on privacy has improved significantly, and traditional methods of privacy protection have been difficult to work with. This paper summarizes and analyzes the centralized differential privacy method and the localized differential privacy method in smart grid data transmission. The characteristics and advantages of The Laplace mechanism, Gaussian mechanism, and index mechanism are analyzed and compared on the addition of noise disturbance mechanism in differential privacy method. In addition, this paper introduces the current researchers on the local differential privacy methods and noise-making mechanism improvement methods. Finally, a K-Means user clustering method based on K-Means is proposed, the main method is to do K-Means clustering analysis based on the sensitivity of different user groups, and then use different differential privacy methods according to different group categories.

Keywords. Smart Power Grid, Differential Privacy, K-Means.

1. Introduction

1.1. Research Background

With the gradual improvement of the position of data application in the global economic operation, China is also a country with a huge population and economy, which produces a huge amount of data all the time, and big data, as a new round of disruptive technological innovation, is also a very important strategic highland in the "Made in China 2025" plan. China has formed a technologically advanced, application-prosperous, and powerful big data industry system, which provides support for data-driven industrial development and innovation.

[1]Corresponding Author, Zhiming ZHAN, State Grid Hubei Electric Power Co. Ltd; E-mail:524846102@qq.com.

Big data applications have been widely used in recent years, such as recommendation systems, data mining, etc. One commonality of these technologies is that big data is based on mining specific patterns from large amounts of data and learning specific tasks to accomplish complex tasks. Because of the large amount of data, some methods can be used to extract and calculate private data from a large amount of data, resulting in data privacy security being threatened. And with the improvement of people's awareness of privacy protection, the relevant privacy laws are also developing, how to protect data security relationship between individuals and even the security and interests of the country. How to transfer data under the premise of protecting privacy security and transmission availability in data transmission has become one of the important issues of contemporary privacy security. Grid data records the user's electricity data all the time, that is to say, records the user's various actions, status, lifestyle, etc., so grid privacy data protection is one of the main directions of contemporary privacy protection.

1.2. Data Privacy Issues

Data privacy is sensitive data that needs to be considered when analyzing data junctions such as financial transactions, web search logs, power consumption logs, medical records, and other data breaches that may result in associated data breaches.

In 2006 AOL released web search logs containing more than 650,000 users within three months, AOL replaced user id with random numbers, but Arnold's search log contained her last name information, nearby address information, life information, and so on, and the attacker combined the search log information with the Phone Book information to find out Arnold's true identity. In 2009, Netflix released a movie scoring dataset for data mining competitions, and although the data is anonymized, attackers use IMDB users' scoring records to locate the dataset and get the entire user's record. In summary, the data security issues described above are a major challenge to privacy issues.

1.3. Smart Grid Power Data Privacy Protection Status

Smart grid is to intelligent power grid, whether it is power generation, distribution, or transmission process has produced a large amount of data. One of the basic services in the smart grid is data acquisition aggregation, where data privacy security is one of the primary concerns.

In the smart grid, the privacy and security protection of industrial, commercial, and civil power data is a problem that must be solved in the further development of the smart grid. In order to better manage the use of energy, it involves requiring users to share information about the use of electricity, which in turn can lead to violations of user privacy. For example, through massive data mining can be inferred that a user has a few people in the home, a certain period of time several people at home, which electrical appliances used, and other private information.

Smart grid data acquisition scenario has several characteristics: first, the most basic intelligent meter data processing capacity is limited, if through the localization of differential privacy method its computational overhead may exceed the upper limit of the meter, followed by the high frequency of data acquisition, the number of nodes per acquisition is very large. Considering the characteristics of smart grid data acquisition, this requires improving data transmission efficiency, reducing computing and

communication transmission overhead, and improving the availability of the scheme while protecting the privacy of users.

With the development of China's traditional power grid to digital smart grid construction, many researchers began to apply differential privacy algorithms to the privacy protection issues in smart grid, Cao Hui[1] proposed a localized differential privacy random response perturbation mechanism based on the multi-output hidden Markov model for highly sensitive users of smart grid, to ensure data feasibility and ensure user behavior privacy, and to propose a localized differential privacy random disturbance mechanism based on sparse coding (SCRAPPOR) for ordinary users. Na Gai[2] For the smart grid smart meter needs to collect data frequently and limited calculation, a lightweight privacy protection data aggregation scheme to meet local differential privacy is proposed. Yin designed a differential privacy aggregation mechanism combining homography encryption and differential privacy mechanism, and regarded user privacy data as a commodity, and put forward the pricing strategy of user privacy data.

2. Mixed Differential Privacy Algorithm

2.1. Differential Privacy

Because the smart grid produces huge amounts of data all the time, this poses great challenges to the transmission, management, storage, and processing of data in the power grid system. The hybrid differential privacy approach is used to protect and encrypt grid databases. Because differential privacy does not depend on the attacker's background, it is widely used in machine learning, data mining, smart grid, and other fields.

2.1.1. Differential Privacy Introduction

Differential privacy technology was first proposed by WORK in 2006 as a method of privacy protection against privacy disclosure, which is mainly used to address two aspects of the privacy protection process: a pair of data sharing a strict definition of privacy, and the premise of privacy protection under the premise of ensuring privacy availability. Differential privacy methods have gradually developed into the current mainstream of two major categories: centralized differential privacy and localized differential privacy, centralized differential privacy is divided into transaction-level differential privacy, approximate differential privacy, user-level differential privacy.

As the main method of protecting privacy, differential privacy is mainly used to protect user privacy by disturbing user data while collecting information and is based on the premise of data managers' reliable and secure protection of privacy. These technologies have a lot of room for improvement in the quality, efficiency and practicality of privacy protection, for example, transaction-level differential privacy has great advantages in injecting noise when uploading data, but because user data has a strong correlation between them, resulting in poor privacy protection. Therefore, the researchers put forward user-level differential privacy based on this method, which gives the user data sensitivity threshold by the algorithm to protect the user data security and then improves the privacy by injecting a relatively large amount of noise. Both of the first two add relatively large amounts of noise to improve privacy, but adding noise will lead to a decrease in availability, approximate differential privacy put

forward new privacy parameters δ control differential privacy failure rate, improve availability, but there is still the problem of data accuracy loss. The first few differential privacy is called centralized differential privacy, provided that the server is trusted, the server before the data is shared out of the differential privacy conversion, but there is a risk that the server will disclose the privacy data. The basic idea of localized differential privacy is that users handle privacy information differential privacy locally, adding appropriate noise to disturb, and avoiding the disclosure of privacy by uploading raw data to third-party servers.

As figure 1(a), figure 1(b) represents the centralized differential privacy framework and localized differential privacy framework, the main difference is whether the user is trustworthy to the server.

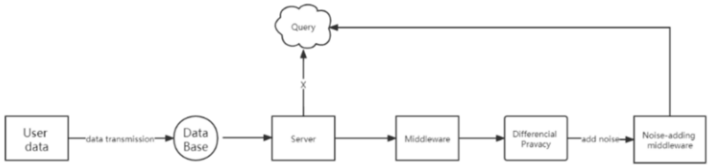

Figure 1(a). Centralized differential privacy process

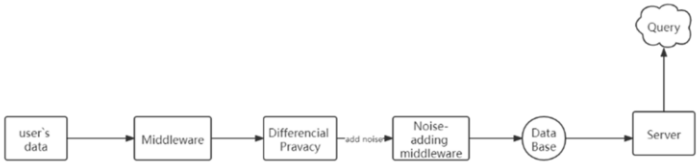

Figure 1(b). Centralized differential privacy process

2.1.2. The Life Cycle of Privacy

Differential privacy to achieve user data privacy security protection, first of all, the need for a strict definition of privacy information privacy, so the privacy life cycle, as shown in figure 2, that is, the generation of privacy information, perception, protection, publishing, storage, exchange, analysis, sales and recipient of 9 parts, differential privacy mainly solves the privacy release, storage and exchange of 3 aspects.

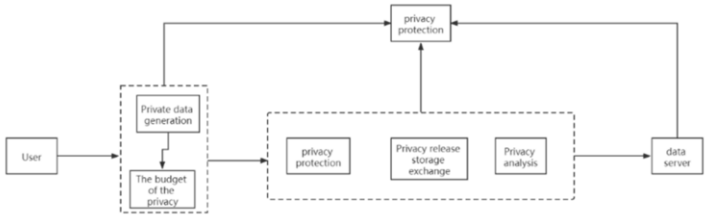

Figure 2. Privacy Life Safety Cycle

2.1.3. Real-World Application of Differential Privacy

Differential privacy algorithms can be combined with federal learning to solve the data silos that have been created by the current increase in privacy data regulation and the inability to share data. It can also be used as a primary means of privacy protection to protect everyone's privacy while collecting data. There are already many companies in the actual reference to use differential privacy technology, such as Xiaomi in user

positioning using an approximate differential privacy method to blur the user's specific location information, 2016 Apple in ios10 using differential privacy technology to collect data from Safari browser, thereby protecting user privacy, and now using LDP (local differential privacy) technology, before the data leaves the user noise processing. Internet companies such as Google and Samsung are also applying local differential privacy technology to areas such as mobile crowdsourced data collection.

2.2. Centralized Differential Privacy

Adjacent (brother) database definition: If two databases can get d' if and only if one piece of data in D is changed, that is, D and D' are adjacent (brother) databases.

The definition of centralized differential privacy, if algorithm M meets the $\varepsilon - difference$ privacy, when and only if any two adjacent databases x, y, then M's output is S, and satisfied

$$Pr[M(x) \in S] \leq e^{\varepsilon} Pr\,[M(y) \in S] \quad (1)$$

ε is a measure of privacy, the smaller the ε, the better the privacy protection, Pr indicates the probability of an event occurring.

2.3. Localize Differential Privacy Algorithm

According to the examples of AOL and Netflix (2.1), there is also a risk of data breaches by storing data on third-party servers, and users often do not trust entities that have access to the original data. Therefore, localized differential privacy technology is proposed to provide privacy data protection methods in distributed data acquisition scenarios, in the case of data without the need for third parties to perturb the data after transmission to the data server, the server based on all uploaded data for holistic analysis, so as to ensure that the privacy of individual users does not disclose the overall statistics.

2.4. Differential Privacy Encryption Policy

Traditional smart grid behavior privacy protection methods are based on physical perturbation methods and cryptography methods. Cryptography-based technology can be divided into two categories: one is the use of public-key encryption technology, the other is the use of homologous encryption technology. Public key encryption technology is the use of the information sender to use the public key to encrypt the information sent out, the recipient uses the private key to decrypt the encrypted information, this process is called public-key encryption. Homeopathic encryption technology is a kind of encryption method based on semantic security, the user processes data information in a way in a redacted way, and the information recipient uses the same method information to decrypt it, and homography encryption is divided into two categories: partial homography and full homography.

2.4.1. Paillier Adds Homography Encryption

At present, the literature[2] puts forward a localized differential privacy policy based on Paillier addition homographic encryption technology, using Paillier addition homographic encryption technology to encrypt data homography, can realize the addition, multiplication and other operations of data without disclosing clear text

content.

Homeopathic encryption technology generally uses clear text data to be mapped to the corresponding redaction operation without disclosing clear text information. The main processes of this method are: key generation, key encryption, key decryption and homography calculation, the algorithm content is as follows:

Generating a public key: Randomly select prime numbers p ,q:

$n = pq, \lambda = [p-1, q-1], and\big(pq, (p-1)\times(q-1)\big) = 1$,Random select $g \in Z_{n^2}^{※}, \mu = [L\big(g^{\lambda} mod n^2\big)]^{-1} mod\ n$, thereinto $L(x) = (x-1)\cdot n^{-1}$, (n, g) is the public key，(p, q, λ) is the private key。

1. Encryption: will be clear text m mapping to $[0, n)$ interval，select $r \in Z_n^{※}$, Calculate the redaction $c = g^m \times r^n (mod n^2)$
2. Decryption：$m = L(c^{\lambda} mod n^2)\mu(mod\ n)$
3. Addition homeopathic calculation: Two clear text messages will correspond m_1 and m_2 and the corresponding redactions c_1 and c_2, The addition homeopathy manifests itself as：

$$\boldsymbol{E}(m_1)\cdot\boldsymbol{E}(m_2) = c_1 \cdot c_2 = g^{m_1+m_2}\cdot(\boldsymbol{r_1 r_2})^{\boldsymbol{N}} = \boldsymbol{E}(m_1+m_2) \tag{2}$$

You can see it, $E(m_1)\cdot E(m_2)$can get ciphertext $E(m_1+m_2)$。

2.5. Differential Privacy Disturbance Implementation Mechanism

DWORK and his team proposed the Laplace Mechanism and Gaussian Mechanism Add noise to the data set, and these data can still maintain the statistical significance of the data set. McSherry and many more Proposed Exponential Mechanism[3] in 2007, This machine is mainly for adding noise to non-numerical data. In addition to these mechanisms, there are other noise-increasing mechanisms, for example Geometric Mechanism[4],Functional Mechanism[5], Matrix Mechanism[6] and so on. Chen Q and Liu Y[7] Proposed a Laplacian noise algorithm based on mostly decomposition, The algorithm needs to add the amplitude of the Laplacian noise according to the adaptive decision of the privacy sensitivity of each dimension, which improves the usability and robustness of the privacy algorithm. Shangqian Li[8] put forward Wavelet-Gaussian Differential Privacy, WGDP, The algorithm compresses user data by means of the wavelet transform, which improves the efficiency of data transmission.

The differential privacy implementation mechanism is to randomly add noise interference when the algorithm is output. The current three mainstream mechanisms are the Laplacian mechanism, the Gaussian mechanism and the exponential mechanism. Laplace mechanism and Gaussian mechanism deal with numerical data, and exponential mechanism deal with non-numerical data.

Sensitivity definition:

Sensitivity is related to the query function f, which represents the distance between the two databases, If the two databases are adjacent databases D and D', Function Δf The maximum range of change is 1, sensitive Δf Function definition:

$$\Delta f = max_{D\Delta D'}\ ||f(D) - f(D')||_1 \tag{3}$$

Laplace Mechanism[9]:

The Laplacian mechanism means that when noise disturbance is added, this noise satisfies a specific Laplacian distribution. Its Laplace function is

$$M(D) = f(D) + Y\ （\xi） \tag{4}$$

$f(D)$ represents the query function, Y represents the Laplacian random noise and $M(D)$ represents the returned result, then the algorithm M satisfies the ε-differential privacy. Laplace noise is

$$Pr[Lap(\beta) = x] = (2\beta)^{-1}e^{-|x|/\beta} \quad (5)$$

Laplace mechanism like figure 3:

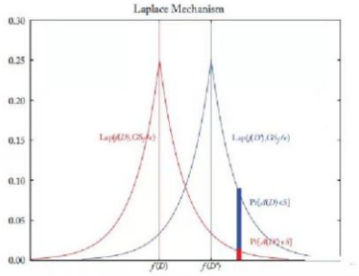

Figure 3. Laplace mechanism

Among them, the red curve is the output distribution of $A(D)$, and the blue is the output distribution of $A(D')$. The central axes are $f(D)$ and $f(D')$, respectively.

$$Pr\,[Lap(\beta) = x + d]^{-1}Pr\,[Lap(\beta) = x] \le e^{(d\beta^{-1})} \le e^{(\Delta f\beta^{-1})} = e^{\varepsilon} \quad (6)$$

It means that differential privacy is satisfied. The amount of noise is proportional to Δf and inversely proportional to ε. When Δf is constant, the smaller the ε, the greater the noise, and the greater the privacy protection, but the usability will decrease, as shown in Figure 4.

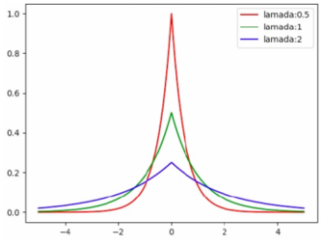

Figure 4. The relationship between ε and the amount of noise

Gaussian Mechanism[9]:

The Gaussian Mechanism means that when noise disturbance is added, this noise satisfies a specific Gaussian distribution. Where the Gaussian distribution function is：

$$P[M(D) \in S] \le e^{\varepsilon}P[M(D') \in S] + \delta \quad (7)$$

Among them, any

$$\delta \in (0,1), \delta > [2\ln(1.25 \cdot \delta^{-1})]^{0.5}\Delta f/\varepsilon \quad (8)$$

Noisy Y~$N(0,\sigma^2)$ satisfies $(\varepsilon,\delta) - DP$, where $M(D) = f(D) + Y$,σ is the standard deviation of the Gaussian distribution, which determines the scale of the noise, and ε represents privacy budget, δ represents slackness.

Index mechanism[9]:

The exponential mechanism mainly deals with non-numerical data, and its output is an element in a set of discrete data. The sensitivity of non-numerical data,

$$\Delta q = max_{D\Delta D'}||q(D,R_i) - q(D',R_i)||1 \quad (9)$$

Among them, $q(D',R_i)$ is the scoring function, D is the data set, and $q(D,R_i)$

represents the score of a certain output result R_i Theoretically, the index mechanism is based on

$$M(D, q, R_i) \sim e^{\varepsilon q(D,R_i)/2\Delta q} \tag{10}$$

The probability output result R_i.

3. Clustering of User Differential Privacy Scheme Based on K-Means

It is difficult to dig out the user's true privacy needs from the user's original power data. Therefore, this solution proposes sensitive factors from different user groups, performs cluster analysis based on the sensitive factors, and recommends different differential privacy solutions for different user groups.

The K-Means algorithm is an unsupervised learning strategy. The clustering algorithm can effectively classify different groups, such as the differential privacy policy division of residential power groups, commercial power groups, and industrial power groups. Factors such as electricity period, electricity consumption, etc. are proposed to calculate the sensitivity factor, and K-Means clustering is performed for each user based on the sensitivity factor, and finally K clusters $P = \{P\_1, P\_2, P\_3, \ldots, P\_K\}$ are obtained. The basic steps are as follows:

Step 1: data preprocessing, the sensitivity factor S is obtained according to the user's various electricity consumption data, and the data is normalized to the K-dimensional space according to the sensitivity factor $S = \{s\_1, s\_2, \ldots, s\_3\}$ to obtain S data point.

Step 2: At the initial point, select K samples from the data set according to the sensitivity factor s_i of each sample c_i as the initial cluster center $C = \{c\_1, c\_2, \ldots, c\_k\}$;

Step 3: For the sample x_i in the data set, calculate the distance from the sample to the K cluster centers, and assign it to the class corresponding to the closest cluster center, that is, calculate the distance

$$d(x_j, c_i) = ||x_j - c_i||_2 \tag{11}$$

Group x_j into clusters:

Step 4: For each category c_i, calculate the cluster center$c_i' = |c_i|^{-1}\sum_{x \in c_i} x$, The cluster center is the centroid of the sample of the category.

Step 5: Repeat steps 3 and 4 until the cluster center position does not change.

4. Summary and Outlook

With the completion of the basic construction of my country's smart grid, refined user data collection has long been realized. When the smart grid collects electricity consumption data, it will contain the user's personal privacy information. A large amount of collected electricity data may expose the user's behavior status and other personal privacy. Therefore, how to protect personal privacy while collecting data on the power grid is a problem that needs to be solved.

This article first analyzes the historical background of smart grids and differential privacy methods and introduces the implementation principles of centralized differential privacy methods and localized differential privacy methods. The

implementation principles and application scenarios of the Laplacian mechanism, Gaussian mechanism and exponential mechanism of noise addition technology are analyzed. Finally, a user clustering method based on K-Means is proposed, which clusters and analyzes different users, and then uses different differential privacy methods according to the privacy sensitivity of different users.

There are still some challenges in the application of the differential privacy method in the field of smart grid data protection, such as adding noise to the data, which sacrifices the accuracy of the data in exchange for privacy protection. Therefore, it is an important topic to develop a more accurate, robust and better privacy protection scheme.

References

[1] CAO Hui Research on local differential privacy in smart grid [D]. : School of Computer Wuhan University, 2020.
[2] Na Gai. Research on Privacy-preserving Data Aggregation Mechanism Based on Local Differential Privacy in Smart Grid [D]. : University of Science and Technology of China, 2021.
[3] McSherry, Frank, and Kunal Talwar. Mechanism Design via Differential Privacy. FOCS. Vol. 7.2007.
[4] TALWAR K, HARDT M A W. Geometric mechanism for privacy-preserving answers: U.S. Patent8,661,047[P]. 2014-2-25
[5] ZHAO J, CHEN Y, ZHANG W. Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions[J]. IEEE Access, 2019, 7:48901-48911.
[6] LI C, MIKLAU G, HAY M, et al. The matrix mechanism: optimizing linear counting queries under differential privacy∫J1. VLDB Joumal - the International Journal on Very Large Data Bases, 2015, 24(6):757-781.
[7] CHEN Qian, LIU Yun. Optimization of multi-dimensional decomposition and plus noise algorithm in intelligent grid privacy protection [J]. Journal of Chongqing University,2018, 41(9):86-93.
[8] Shangqian Li Research on Data Privacy Protection and Secure Transmission Technology Based on Smart Meter System [D] Southwest University, 2021.
[9] DWORK C, McSherry F, NISSIM K.et al.Calibrating Noise to Sensitivity in Private Data Analysis[C] // Proceedings of the Third Conference on Theory of Cryptography. New York: Springer-Verlag,2006:265-284.