Analysis of Privacy Preserving for Big Data in the Environment of Power Systems

Yun Ye
China Electric Power Research Institute, Nanjing Branch
Nanjing, China, 211106, PRC
veyun@epri.sgcc.com.cn

Wei Yang
Depart.of Comp. Sci.& Tech., USTC
NHPCC 416, East Campus, USTC
Hefei, China, 230026, PRC
86-551-3602445
smartyw@mail.ustc.edu.cn

Wei-min Lin
China Electric Power Research Institute, Nanjing Branch
Nanjing, China, 211106, PRC
linweimin@epri.sgcc.com.cn

Tao Zhang
China Electric Power Research Institute, Nanjing Branch
Nanjing, China, 211106, PRC
zhangtao@epri.sgcc.com.cn

Abstract: As the development of Smart Grid deeply progressing, the data in power systems grows extremely fast. Information security has become a very important support for smart grid’s safety and stably operation. To build unified, reasonable, safe and efficient information facilities is necessary for safeguarding the development of Smart Grid. The privacy problems become more and more outstanding as the big data technologies progressing. Electric power informatization branch of Chinese Society of Electrical Engineering will publish a white book called the big data development in Chinese power systems, and it will be the first white book on big data problems in electric systems. Big data problems in power systems is the inevitable trend for Smart Grid, and the corresponding privacy preserving problems also become the important guarantee for various business in power systems. In this paper, we will deeply investigate and survey the privacy preserving problems and technologies, and will provide some positive suggestions, such as a framework of privacy preserving approaches for the apps of big data in power systems.

Keywords: Power Systems, Big Data, Privacy Preserving, Smart Grid

I. INTRODUCTION

Smart grid is a concrete application based on the integration of two-way communication, high-speed network, through advanced sensing and measuring technology and equipment, to realize a reliable, safe, economic, efficient and environmental friendly purpose. It’s main characteristics include self-healing, resist attack, and provide the high quality power supplies for the whole user. As marched into the 21st century, the data in power systems expanding in rapid speed as the smart grid developing greatly itself. In our Country, the information security technology has been the main foundation for the safe and stable operation of the smart grid. To construct a unified, reasonable, safe and efficient information infrastructure is an important guarantee of smart grid. After entering the era of big data, especially the problem of data security and privacy in the smart grid is increasingly prominent. If there is not a good solution to the existing safety problems, the potential safety hazards will affect our process of intelligent power grid. As a national fundamental energy facility, Smart Grid is related to social development and people’s life. In addition, it is an important condition for national economy society sustainable progressing healthily and stably. To make active use of Big data technology to promote Chinese power big data research has great practical significance for realizing the development of Chinese power industry technologies. As the big data in power systems progressing, there is a new challenge for privacy preserving technologies. To act as a huge system, the construction of smart grid should firstly consider the defense capability, namely the ability to against the attack. On the one hand, the power net will suffer from attacks outside, such as virus, network attack, research achievements been stolen and so on. On the other hand, suffers from the attacks from inside, such as unauthorized access or modify some important data at random and so on. Both the two aspects will bring great damage in economy and the bad fame to company. The former can be protected by keys, firewalls, network isolation devices and so on, while the latter cannot achieve this goal by the same methods. We should employ a technology namely privacy protection technology to protect the employees malicious operation on data sets. That is to say to make the data accession for everyone transparently. To realize the purpose, we can do the computations among data in the encrypted form, instead of the plain text. Thus, no one will get access to secrets. Smart grid is characterized as wide openness and complexity. It becomes more and more frequent to interact with the outside information systems. As the power systems march into the era of big data, the enormous strength of the big data technology brings to the electric power system is very significant. For example, through analyzing electric power marketing side of big data, we can make more optimal planning of distribution power system in all aspects of the optimal strategy. Also, we can implement dynamic improvement optimization for power systems. To analyze the big data of Power Generation systems, we can get an estimation of equipment fault probability, thus we can take some necessary measures to avoid some possible major issues.
accidents. To make use of the big data in power systems for data mining, we can get the long-term potential threat lurking within the power system, such as advanced persistent threat (APT) attack. Some hackers will research the target deeply, and APT will be tailored for that. After a very patiently invasion, latent, then then use Trojan steal high-value information continuously, as well as the sensitive information. In the very period, the possibility of outbreak of damage increases, and the big data environment is more serious. Thus, we should use big data technologies to defense similar attacks. At present, the flux audit scheme has strong ability of real-time detection and ex post facto. It is a great solution for us to combined with big data analytics technology to discover the hidden attacks and security threats. Assistant to the head of the EU data protection supervision organization Giovanni Buttarelli said that these data is very useful for our analysis of energy usage, but at the same time, there may be some people with ulterior motives for marketing, advertising, differential pricing, etc will use the data. The big data in power systems not only has the 3-v characters, but also has the 3-e characters. That is to say, 3-v indicates that Volume, Variety, Velocity, and 3-e indicates Energy, Exchange and Empathy (figure 1).

![Big Data Framework in Power Systems](image)

**Figure 1. Big Data Framework in Power Systems**

II. RESEARCH ON PRIVACY PRESERVING MODEL AND SECURITY ALGORITHM IN POWER SYSTEMS

Privacy preserving technology is mainly for protecting the malicious users from identifies some sensitive information to leak messages. In the 1970s, a group of scholars started to research the Statistical Database (Statistical Database, SDB), including the data in the form of a relational table and in the form of individual records sets. This information often needs to release for other organization Statistical analysis, while the release will cause some of the sensitive information leaked. The publisher wants to publish as more information as he can but the privacy should be preserved. That is to say, try your best to mediate the contradiction between privacy and availability of data. Until now, researchers in many different areas have put forward the corresponding solution in order to prevent leakage of privacy in data publishing problems, but a lot of leakage situation still exists. Data sets released by combining different time-periods, for example, or link to published data sets and additional information [2], a malicious person can with high accuracy deduce some individual identity information. The privacy preserving methods already exist can be analyzed as follows:

1) Interference method [3]: By modifying the data or add noise data, make data centralized data cannot identify in order to protect the real raw data, and the modified data can still be used in statistical analysis. Use data after interference method, raw data will has certain interference data, so even if a certain data items are linked to the designated individuals will not completely exposed to the data of real value, so as not to reveal private information. However, this method will destroy data primitiveness, and lead to the low data availability, so that data released no actual value.

2) The data transformation method [4]: The main idea is to reduce the support or confidence of private information in the database to a threshold value, by deleting or adding data items. This method can obviously prevent leaking of private information, but destroy the integrity of the data. Thus, it is not conducive to the actual statistical analysis.

3) The data block method [5]: the principle is to add placeholder "" to the raw data table, in order to make the support degree and confidence interval uncertain. In this way, the privacy preserving purpose can be achieved.

4) data reconstruction method [6]: the principle is to use data mining method to dig the original data set of frequent patterns, based on the frequent pattern to build a new data set, and the original data set will not be as public data sets. As a relatively new method of constrained association rules hiding, it is different from the data transformation and data blocking method that need to transform to modify the original data set to hide the private information.

In addition to the above several methods, there are some special methods, such as probability of independent method [7]. However, to some extent, they achieved the goal of privacy preserving, without exception the destruction of the authenticity, and availability of the data, which lead to release of data set information too much distortion, and the available range is too narrow. In 1998, L.Seweney and the others put forward a K - anonymous technology, which is free from the links to attack in the PODS international conference. Its basic idea is to ensure that each record can be hidden in the record set of size K. In 2002, L.Seweney put forward K - anonymous privacy protection model. In that same year, he expressed in literature [9] how to realize the K -anonymous privacy protection model. In 2004, Williams et al. illustrated the most preferred K -unseen name problem is NP (Nondeterministic Polynomial) problem [10]. Therefore, the researchers have proposed many heuristic and approximation algorithms [11]. K -anonymous model, although to a certain extent can ensure the privacy, but there is still a lot not applicable fields. The researchers also targeted putting forward many improved models. In literature [12], 1 - diversity model is put forward, and it requires that the anonymous data records with the
highest frequency of the number of sensitive attribute values is not greater than 1 / l. It can prevent an equivalent in all or most of the records with the same set of attribute values that cause privacy. P - Sensitive k - Anonymity [13] requires the same equivalent of at least p different sensitive attribute values. \((a, k)-\text{Anonymity}[14]\) requires frequency of any sensitive attribute value should be no greater than \(a\). Both of them has the same similarity with l-Diversity method. \((k, e)-\text{Anonymity}[15]\) requires that sensitive attribute value in the equivalent set of range should be \(e \) at least. It tries to avoid similarity attack by minimum \(e\) value, but may cause information loss, also cannot resist the deflection of the sensitive attribute value attack, mainly for numeric sensitive attribute. \(t\)-Closeness[16] requires that the published data sets in the \(K\) – anonymous form, also it should ensure that the distribution of the equivalent sensitive attribute value in the group and the sensitive attribute value in the anonymization population distribution difference should be no more than \(t\) in the table. In [17], a privacy protection strategy is put forward, which supports incremental dynamic release. When the data set increases the new data, we not directly add it to the group, which has divided by an equivalent, but new data such as reaches a certain amount to be added to the next, in the release, eliminate the attacker by different release difference to infer information privacy. m-invariance [18] support at the same time increasing the release strategy, and deleting data dynamic in order that requires two records in the release version is equivalent with exactly the same set of privacy collection of attribute’s value. Thus, eliminate between multiple versions of inference channel to prevent privacy leaks.

In anonymous research, to study the model is the whole research framework. In addition, the corresponding algorithm is a model support, which is to say, we should adopt what kind of algorithm to build the mold. MinGen [9] is the earliest for K - anonymous model to put forward the implementation of the algorithm. The main idea is to search space completely and to select the optimal generalization operation, with the ending conditions that the data meet the principle of data with \(K\) – anonymous. In theory, MinGen algorithm can meet the requirements of data availability and privacy, but because of using the entire process is fully search result in algorithm time complexity is too high, it cannot be realized in the actual application. Incognito [19] is a widely used in actual K - anonymous algorithm, the algorithm about process is: first of all, according to the data features to build global generalization figure. Then targeted trim figure to narrow your search select optimal generalization scheme, the purpose of the whole process and generalization of the original data is a bottom-up. Finally ended up with data with K - anonymization principles for conditions, otherwise continue the above operation until the condition is satisfied. Due to the application in using the trim generalization strategy, the result maybe in data availability declined slightly, but the algorithm-time-complexity control in the range of the affordable, conform to the requirements of the practical application of the algorithm.

III. RESEARCH ON RELIABLE FAULT-TOLERANT MECHANISM FOR BIG DATA STORAGE AND ACCESS IN SMART GRID

Data storage and access control is a fundamental problem in data security, which generally includes the audit technology, disaster tolerance technology and data access control and authorization mechanisms. Audit is a kind of typical data security mechanisms, which can record the user to the operation of the data, according to every data operation records for later review any data record no matter who, when, where, what information of the operation. By using this mechanism, we can guarantee the integrity of the data. At present, the technology of the data audit is mainly for small data audit. Audit of big data technology research has not yet been reported. Due to the small data audit’s efficiency is not high, so predictably, big data era could not use the previous audit technology, and it requires a different approach.

Disaster tolerance is to ensure that after the disaster damage occurred, the system will still be able to maximize the normal service of computer information system by specific disaster mechanism. The disaster backup system according to security can be divided into disaster data and applications, and we mainly consider the data disaster. Disaster data is to set up a long-distance data system, which is a key application of real-time data replication on the local system. When a disaster occurs, it can quickly replaces local system by long-distance system and ensure the continuity of the business. RPO is an important index for data disaster backup, which mainly refers to the application system that can tolerate the amount of data loss. Design a disaster tolerance system, need to consider many factors, such as the quantity and disaster backup/restore data recovery rate, redundancy required by center of funds management and investment, etc.

Data access control and authorization mechanism is a kind of a mechanism of ensuring data integrity and confidentiality, which originated in the 70 swenthe mainframe system is to meet the management need to grant access to Shared data. Now there are already several
important access control technologies, such as discretionary access control, mandatory access control, role-based access control model, tasks and workflow based access control model, roles and tasks based access control model, and attribute-based access control, etc. Their basic aim is to prevent the illegal users from entering the system and to legal illegal use of system resources. Because attribute-based access control can solve complex fine-grained access control in information system and the massive user dynamic scaling issues, it has become a complex computing system security field at home and abroad as a research hot spot. In attribute-based access control, the requestor, the requested resource and some restrictions are attributes to describe the use environment, namely the description of all entities are unified use attributes to describe. This makes the access control decision function in determining when to access control decision by consensus according to processing. At present, the study of attribute-based access control mainly includes the entity properties research, description and semantic interoperability research, synthesis and conflict resolution strategy research, formalization model and properties and strategic security interaction research.

In entity properties research, [20] put forward the ABAM model described in attribute values accessing tuple access matrix of subject-object relations. Because in attribute-based access control, access control decisions may need to be from different authority of attribute information, which allows the user only in the authority of different attributes have the same available to identify global user name that on the limit unfavorable to protect users' privacy. According to the theory of grid computing credentials management, the MyProxy system in [21] has carried on the thorough discussion. It provides users with an encrypted x.509 certificate agent services. Accession to the main body may be entrusted to MyProxy for short-term certificates of attribute information.

In description and semantic interoperability of strategy research, XACML has been the industry wide support extensible access control markup language, which is a kind of using attributes in access control policy description language, and provides the basic authorization framework. It is the most ideal choice for attributes based of the access control policy description language. In [22], it uses semantic Web technologies to propose a semantic perception function attribute based access control model (SABAC). Web ontology language (OWL) SABAC model is used to describe the metadata of resources and the user's attributes and the reasoning. XACML is used as control strategy description language for Web services provides extensible access control management, and semantic interoperability.

In synthesis and conflict resolution strategy research which is proposed in [23], Bonatti and so on combination with synthetic security policy, the form of an access control policy into triples authorized collection subject, object and behavior. In addition, [24] based on the synthesis of the FIA's proposed the XACML policy architecture and proves that the proposed algebra of the pram and completeness, when can deal with granular synthesis of XACML strategy. Synthesis formal framework structure was calculated by the attribute value extension strategy in [25], and put forward the strategies of the based on attribute synthetic ApoCA algebraic model. Then it discusses the policy expressions of algebraic properties. Due to the complexity of strategy description, differences and different organizational domain control strategy makes it easier for the synthetic strategy.

If synthetic strategy decision conflicts are detected and solved, the access control security authorization will not be able to guarantee related entities. The conflict rules of logic is used to analyze the strategies of the basic method. Polymorph system in [26] can detect the logical constraints, suggest to eliminate conflict. Dynamic assessment help system administrator access control policy with logic constraints. [27] in description logic (DL) to formalize the XACML strategy, with the analysis of existing DL validation tool detection strategy of redundant rules.

In formalization model research, [28] use the limit theory presents a logical framework LABC, and uses the CLP in the set to describe attributes and services. Introduced in [20] attribute to the access control matrix, the ABAM attribute-based access control matrix model is put forward, which supports the properties based authorization and dynamic access restrictions to enhance the access matrix expression ability. [29] authorized delegation model based on attribute ABDM, when given a role of reference, requirements must meet the role attribute expression constraint. In terms of attributes and strategic security interaction research, [30] employ the third party guarantee in the access control can provide reference for attribute and the strategy of security interaction. In [31] ACK strategy is put forward to strengthen the protection of sensitive attribute. In an ACK mechanism, consultation of one party should know each other whether meet the requirements of a property, which should first meet each other corresponding ACK strategy of sensitive attribute. In [32], ACK target figure (TTG) strategy and trust agreement are employed to enhance the security in the process of negotiations. In [33], access control model based on trust and attributes proposed to limit resource providers need to ask the requester properties and describe freedom to its attributes. Thus, trusted relationship is established between both sides. In [34], trust negotiation isin the form of hidden attribute credentials and access policies to minimum disclosed credentials and access policies, according to the need to implement different levels of privacy protection. In [35], attribute federal (attribute federation) concept is proposed so that the user does not perform trust negotiation when interacting in a property.

IV. RESEARCH ON PRIVACY PRESERVING DATA MINING FOR BIG DATA IN SMART GRID

Research on the privacy of the smart grid at home and abroad large analysis of data mining technology research is
still in its infancy. The existing solutions [36-40] can be only used for aggregation operation. In [36]and [37], before the implementation of data aggregation operation, optimal relationship tree is constructed in the first place. This tree can reach a certain relationship between the entire neighbor node, including any specified nodes of the shortest path to the neighbors. Tree on the basis of data gathered in the relationship, will greatly reduce the data traffic. In [38], super increasing sequence is adopted to construct multi-dimensional user data and USES Pailler homomorphic encryption scheme of the data is encrypted. Encryption is only for local gateway, and we will not leak any privacy information. Eventually all local gateway encryption gathered results are sent to the data center, which through the decryption can get the result. In [39], a data aggregation scheme is designed based on diffie-hellman key exchange protocol. This scheme not only can complete the data gathered, but also can be applied to fraud and leak detection, and the grid data of deeper statistical analysis. In [40], an addition secret sharing scheme is proposed, will add secret sharing and Pailler homomorphic encryption scheme organic unifies in together, significantly improve the efficiency of gathering. However, the work of customer information in the smart meter only can do simple addition operations. For large data in smart grid in more valuable information, there is no effective privacy preserving data mining solutions.

V. POSSIBLE FRAMEWORK OF PRIVACY PRESERVING FOR POWER SYSTEMS

In power systems, big data can be produced in many businesses, such as generation period, marketing side and so on. Take marketing side for example, the interact between the power system and users becomes more and more often, there is great amount of metering data. The users’ private habit of using electricity or some other users’ behavior will be revealed. Therefore, efficient measures should be taken to protect the users’ privacy data.

Take the smart metering for example, the data can be in various forms. Firstly, we should uniform the data form, either get some characteristic data or uniform the whole data in order to get the regular pattern. Secondly, take some measures to hide the private data, by adding some noise or transform the data into specific form. Usually, the noise data should be transparent to users. That is to say, through the statistics measures, the noise will not affect the real results. Thirdly, take the anonymity measures to hide the users’ identification. Then, privacy preserving protocols should be employed to protect the privacy of the data itself. In this period, the secure multiparty computation is the primary method. Finally, the final statics results can be got by removing the noise data. Usually, it is carried out by using specific protocols.

Figure 3. Frame Work of Privacy Preserving for Power System

In this way, we can get some precious message such as the users’ behavior which will give the department great amount suggestion through the privacy preserving methods. For some other businesses in power systems, we can implement the similar procedures to realize the specific privacy preserving method for big data environment.

ACKNOWLEDGMENT

The subject is sponsored by the Science and Technology Research of State Grid (SG11034), National 863 High Technology Research Program of P.R.China (No. 2012AA050802)

REFERENCES

[37] F Li, B Luo, P Liu. Secure information aggregation for smart grids using homomorphic encryption, Smart Grid Communications, 2010
[38] R Lu, X Liang, X Li, X Lin, X Shen, Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications, IEEE Transactions on Parallel and Distributed Systems, 2012
[40] A Rial, G Danezis, Privacy-preserving smart metering, annual ACM workshop on Privacy in the electronic, 2011