Journal of Multimedia Information System VOL. 4, NO. 4, December 2017(pp. 263-270): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/JMIS.2017.4.4.263

A Study on De-Identification of Metering Data for Smart Grid Personal Security in Cloud Environment

Donghyeok Lee[1], Namje Park[2,*]

Abstract

Various security threats exist in the smart grid environment due to the fact that information and communication technology are grafted onto an existing power grid. In particular, smart metering data exposes a variety of information such as users' life patterns and devices in use, and thereby serious infringement on personal information may occur. Therefore, we are in a situation where a de-identification algorithm suitable for metering data is required. Hence, this paper proposes a new de-identification method for metering data. The proposed method processes time information and numerical information as de-identification data, respectively, so that pattern information cannot be analyzed by the data. In addition, such a method has an advantage that a query such as a direct range search and aggregation processing in a database can be performed even in a de-identified state for statistical processing and availability.

Key Words: de-identification, smart grid, cloud environment, privacy, personal security

I. INTRODUCTION

The smart grid, which is an output of grafting information and communication technology onto the existing power grid, is a technology that improves energy efficiency by using real-time information exchanges between power consumers and suppliers. This smart grid technology is expected to come into effect in the future and will have a positive impact on efficient energy consumption. In the future, the smart grid and information and communication technology will be integrated, and thereby various technologies will be introduced to ensure the high capacity, high availability and high efficiency required for the smart grid. A preliminary estimate made by a power company predicts that the smart grid will generate data at the level of 22 gigabytes per day on the basis of a population of two million people [1]. In addition, this amount of information is expected to increase over time.

According to this view, the smart grid-related cloud market is expected to attract attention in the future. However, the smart grid, which will be realized in the future, has various uncertainties compared to current power grid system. In other words, besides the security threats that existing power grids have, there are a number of factors that can cause more security threats. In addition, a new attack pattern may occur due to the grafting of information and communication technology onto it. In particular, it is possible to analogize various lifestyles such as when to go out and what kind of electronic appliances to use based on the power consumption obtained from a smart meter, which can lead to a serious infringement on personal information. Therefore, metering data in the smart grid environment needs to be recognized as important data which should be protected carefully, and security technology suitable for this purpose should be developed to safely manage the data.

Fig.1 shows a smart grid data cloud model [2]. This model grafts the smart grid environment onto the cloud to provide a number of advantages in data management. However, in order for a cloud-based smart grid environment to be settled, it is essential to take security into consideration, in particular, to fully prepare personal information protection. This is because the introduction of a cloud environment to the smart grid may leave a number of security threats intact depending on cloud environment characteristics.

Manuscript received December 21, 2017; Revised December 24, 2017; Accepted December 25, 2017. (ID No. JMIS-2017-0055)
Corresponding Author (*): Namje Park, 61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, 690-781, Rep. of Korea, +82-64-754-4914, namjepark@jejunu.ac.kr.
[1]Elementary Education Research Institute, Jeju National University, bonfard@jejunu.ac.kr
[2]Department of Computer Education, Teachers College, Jeju National University, namjepar@jejunu.ac.kr

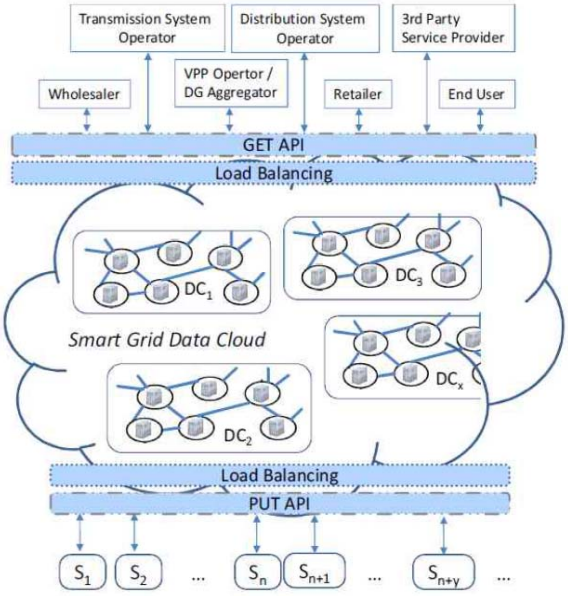

Fig. 1. Smart Grid Data Cloud Model [2].

II. RELATED RESEARCH

In this chapter, we first review the need for personal information protection in the smart grid environment, and then, the de-identification technology and order preservation encryption technology.

2.1. Need for Personal Information Protection in Smart Grid Environment

In the smart grid environment, power suppliers and consumers exchange real-time information to optimize energy efficiency. In order to perform this function, it is necessary to efficiently store, process, and analyze various related information. For smooth smart grid service, the minimum personal information necessary for the service should inevitably be collected. This is because suppliers should know the consumers' personal information in advance to smoothly serve consumers.

"Personal information" is defined as all the information about a person by which a specific individual can be identified, or by which he/she is likely to be identified if it is combined with other information. A variety of personal information can be exposed in the smart grid environment. Table 1 shows examples of typical personal information that can be exposed. Here, smart metering data exists as an item of personal information. This smart metering data may not be deemed to be personal information by itself, but it may be possible to identify a specific user based on an analysis of the user's power consumption pattern, etc. On the other hand, not only individual identification, but also various kinds of information such as individual power consumption patterns, lifestyle, and preferable electronic home appliances are likely to be exposed, which can directly lead to infringement on personal information, and even abuse by crime. Therefore, smart metering data also needs to be dealt with as personal information in a broad sense.

Table 1. Examples of Personal Information

Item	Examples
Name	The name of user
Address	Where the service provided
Smart Meter	Daily, monthly energy consumption
Finance	Arrears, unpaid charge
Life Cycle	Hour of rising, bedtime, used appliances
Identifier	IP address, network identifier

Figure 2 shows various kinds of information obtained by the analysis of smart metering data for one day only. This exposes the lifestyle of a specific family, such as when they have breakfast and what kind of electronic appliances are used (including whether they go out or not). The fact that the exposure of this information can lead to serious infringement on privacy suggests that metering data should be stored in a database in the safest manner.

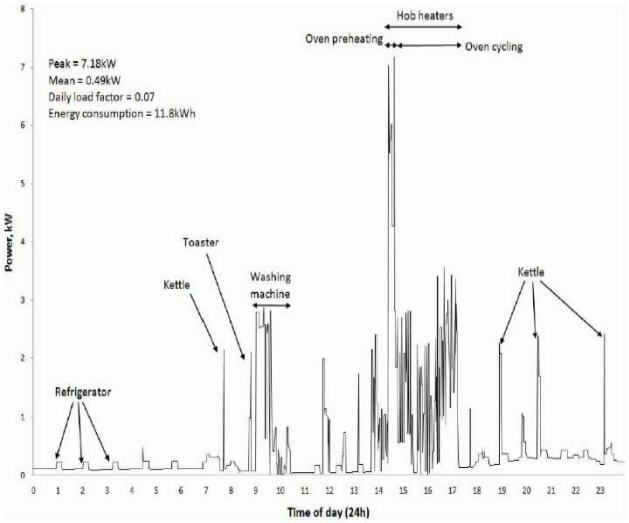

Fig. 2. Security Threats of Metering Data [2].

2.2. Technology for De-Identification and Order-Preserving Encryption

2.2.1. Personal Information De-Identification Technology

De-identification technology means the technology that makes it impossible to identify a specific individual by changing or deleting part of data for the purpose of minimizing the risk of infringement on personal

information. With the recent introduction of a big data environment, the risk of infringement on personal information such as the real-time analysis and provision of personal information for use in business is increasing. Accordingly, the Ministry of Security and Public Administration announced standards for the de-identification of personal information for the purpose of efficient protection of personal information in Sep. 2013.

In addition, the NIST recently published a report titled "De-Identification of Personal Information." In Mar. 2012, it also established guidelines for the de-identification of personal information titled "Protecting Consumer Privacy in an Era of Rapid Change," and provided that any personal information that can be associated with a device with an identification function should be protected in any case. In particular, it specifies that de-identification measures should surely be taken in an appropriate manner such as deletion, modification, adding "noise," and sampling of data by which the information on an individual and computers and equipment can be identified [3]. In general, the method of de-identifying personal information includes pseudonymization, aggregation, data reduction, data suppression, and data masking. However, it is not easy to safely protect personal information in metering data in the smart grid with these methods alone. Such de-identification methods known as above are not compatible with the characteristics of smart metering data. This is because, if stored with an unclear value, the metering data, which is a user's power consumption data, may lead to interference with the smooth provision of service, and an occurrence of unclear statistical processing. For example, if you want to know at which particular time power is mostly consumed, you need to have such metering data be as detailed and specific as possible. However, it is very dangerous to have metering data, which is a kind of personal information, in the type of plain text. The analysis of metering data can estimate a user's lifestyle such as his/her habits, so the subject of the data can eventually be identified. Accordingly, security technology suitable for metering data is required.

2.2.2. Order-Preserving Encryption Technology

When applying a conventional encryption algorithm to a database, there is the problem that the order of the encrypted data becomes different from that of plain text, so that it cannot construct a database index. In this circumstance, order-preserving encryption was proposed.

This method allows indexing in an encrypted state of data, and also allows a range search, right-truncation search, statistical query, etc. without undergoing a separate procedure. As a typical study of order-preserving encryption, Agrawal proposed OPES, which has an advantage of minimizing the exposure of information on source data by changing the distribution of numerical data while preserving it. However, there is a fatal disadvantage in that the order-preserving encryption algorithm has the same sort order as that of plain text. In an extreme case where you know the set and order of plain text, you can get the same data from the encrypted database if you list them in the order of the plain text. On the other hand, due to order-preserving characteristics, there can be various kinds of attacks, and there is a lot of vulnerability in security because a large amount of information can be exposed by simply comparing the size of two specific values.

2.2.3. Order-Preserving Encryption Technology

With the spread of smart grid technology in the future, it is expected that the smart grid will be grafted onto a cloud environment due to various reasons such as a data storage space and availability. Figure 3 shows a smart grid service provision model in a cloud environment. The service provider here can be a cloud service provider that is a trusted third party. This is within a trusted zone. However, the actual data is stored in a cloud database outside the trusted zone. The service provider performs API communication between cloud servers to diversely process data within the cloud server.

In such a cloud environment, smart grid personal information such as metering data exists outside the trusted zone. The cloud service is thereby deemed to be an untrusted zone. This is because there is likely to be a variety of risks such as security threats to data stored in a cloud and data leakage by insiders.

This works as a direct threat to personal information. Therefore, when data is stored in a database of a cloud server, it is necessary to protect the data in an appropriate manner. In other words, when the service provider stores data in a cloud, it is safe to store it in a de-identified state, and when the data is converted into plain text, it is safe to process it within a trusted zone after retrieving the data from the cloud server.

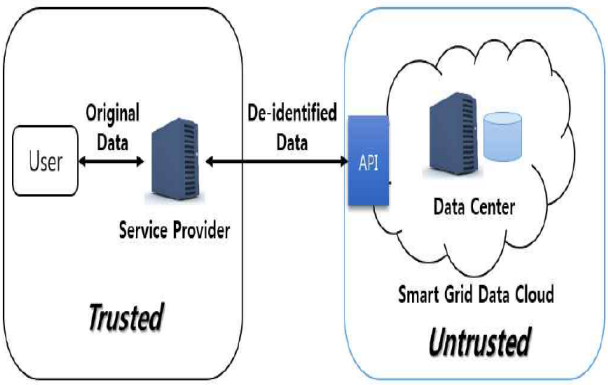

Fig. 3. Smart Grid Cloud Model

III. PROPOSED METHOD

In this chapter, we propose a method of de-identification of metering data from two viewpoints, i.e., time data and power consumption data, to protect personal information.

3.1. Outline of Proposed Method

3.1.1. Outline

Figure 4 shows an approximate concept of the method proposed in this paper. A cloud server stores the smart grid metering data. There are details of the hourly power consumption of customers in it. When de-identification is not made, the details of a customer's power consumption will be stored intact in the cloud server. However, if an appropriate method of de-identification is applied based on the method proposed in this paper, it is difficult to know the details of a consumer's power consumption with the information stored in the server. This source metering data can only be imported through a trusted server. The trusted server retrieves the de-identified data with a query to a database of the cloud server, and reconstructs the source data. Based on this trusted server, users can acquire the source data, and the trusted server can construct an aggregation of search queries using query transformation so that statistical processing can also be possible. In addition, in this process, if exposed by a man in the middle attack, etc. the query and result value of the query between a trusted server and a cloud server are safe because they cannot be reconstructed.

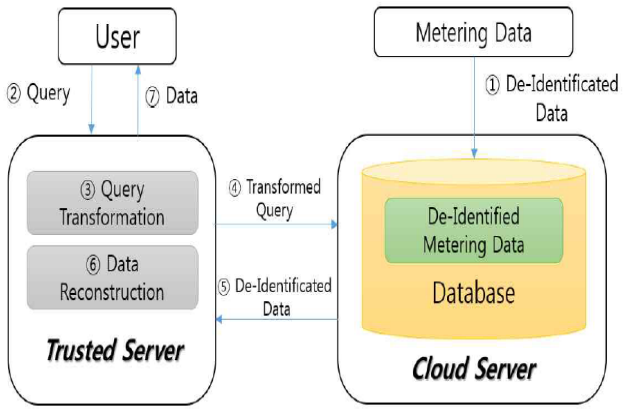

Fig. 4. Overview of Proposed Scheme

On the other hand, the proposed method assumes that the trusted server is a safe domain. Therefore, the cloud server and the trusted server share a symmetric key and a pseudo-random number, and the query about a specific subscriber and reconstruction are possible within a trusted server. Accordingly, the trusted server has an obligation to securely manage confidential information.

The overall working procedure is as follows:

① The metering data is de-identified on a real-time basis and stored in a cloud server.

② The user's client sends a data query to the trusted service provider (trusted server).

③ The trusted service transforms and reconstructs the query.

④ The trusted server makes a transformed query to a database of the cloud service to retrieve transformed numerical data.

⑤ The cloud service returns de-identified data intact to the trusted server.

⑥ The trusted server performs re-identification of the data.

⑦ The re-identified data is transmitted to a user.

3.2. Detailed Procedure

3.2.1. Abbreviation

Abbreviations are given in Table 2 for a better explanation.

Table 2. Notations

Abbreviation	Content
S	Initial seed of pseudo-random number
Pns	The nth pseudo-random number value
GIDn	The n-th group ID
Gn	n-th group
H(•)	Hashed result value
K	Pre-shared encryption key
E(•)K	Encrypted result value using K as a key
DTn	n-th time value

3.2.2. De-Identification Processing Phase

(1) Time-Information Encryption and Grouping

The meaningful information in terms of personal information exposure in metering data is time information and power consumption per hour. In addition, an analysis of power consumption requires analysis based on time information. In other words, power data alone without time information can make a meaningful combination of information difficult. Accordingly, in this paper, we encrypt and store such time information.

However, if time information is encrypted, a range search cannot be performed based on the time information. When encryption is performed, the sort order becomes completely different from that of plain text, so the return value in a range search is meaningless. In other words, it is very difficult to construct a query for the power value generated during a specific time. If a query is performed for all records in a database, the overhead will be very large.

The per-hour grouping method is used to solve this problem. In other words, data on adjacent time are grouped into a unit. However, when the number of data in a group is uniform, it is possible to roughly estimate power consumption for a specific time by analyzing power consumption for a specific period. To avoid this risk, the number of data in a group is set at random. This number of data can be determined based on a pseudo-random number, and at this stage, the size value for a specific group (the number of data in the group) is determined as follows: The sum of the initial seed and n, which is s+n, is set as the seed, i.e., s'. Then, the group size can be determined by the result value of Pns'. GIDn, which is the ID of the nth group and is deduced as follows:

$$\mathrm{GID_n} = \mathrm{H}(\sum_{i=1}^{n} P_i^s)$$

To get the GIDn, we need to know the seed specifically. An attacker who does not know pre-information on the seed and K, which are pre-shared by the trusted domains, cannot generate GIDn correctly.

Figure 5 shows a state where the group IDs are added and time values are encrypted.

Time period	Usage (KWH)
2/1/2016 1:00	0.385
2/1/2016 2:00	0.365
2/1/2016 3:00	0.425
2/1/2016 4:00	0.5

Group ID	Time period	Usage (KWH)
$H(\sum_{i=1}^{n} P_i^s)$	$E(DT_n)^K$	0.385
$H(\sum_{i=1}^{n+1} P_i^s)$	$E(DT_{n+1})^K$	0.365
$H(\sum_{i=1}^{n+2} P_i^s)$	$E(DT_{n+2})^K$	0.425
$H(\sum_{i=1}^{n+3} P_i^s)$	$E(DT_{n+3})^K$	0.5

Fig. 5. De-Identification of Time Field.

Figure 6 shows an example where a randomized group size is applied. The number of data that a group unit can have, i.e., the group size, is random. A cloud server can count the number of data related to a specific ID for de-identified data, and thereby know the number of data belonging to each group, but this does not have a big meaning in itself. This is because the group ID itself is not a sequential series of numbers or systems, so a specific group cannot be associated with another group in terms of order. In other words, if s, which is the seed of the pseudo-random number, is not known, it is impossible to know the group ID after a specific Gn is Gn+1. Therefore, data is characterized as that it is independent of each group unit.

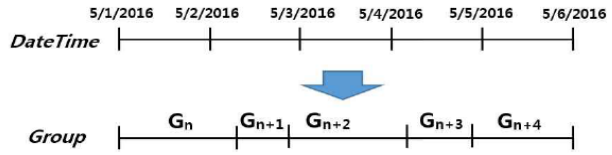

Fig. 6. Group Size Randomization

(2) Transforming Numerical Information

This is a phase where numerical data on actual power consumption is transformed based on a polynomial. For reference, the amount of power is measured using any real number greater than zero.

When there are randomly selected real numbers, x1 and x2, which are greater than zero belonging to X in a function f, which is f: X → Y, a strong monotonically increasing function where f(x1) < f(x2) is always true with any combination of x1 and x2, provided x1 < x2, is required. At this stage, transformed numerical data can be deduced by the expression based on the values of Pn-2s, Pn-1s, and Pns as follows:

$$f(x) = P_{n-1}^s x^2 + P_{n-2}^s + P_n^s$$

Numerical data is transformed based on a polynomial like this. This means that a specific group has the distribution of source data intact, while the actual value was transformed, making it difficult to perform various energy analysis attacks based on the data. Where the transformation expression is composed of a high-order function, which has higher-order terms, security may be improved; however, the size of the transformed data may greatly increase. This may cause a problem in securing a field size and inefficiency of calculation depending on the characteristics of a database, so an appropriate trade-off is necessary. Therefore, the transformation expression in this paper is composed of a quadratic function suitable for this purpose.

The values of Pns are identical for each group unit. Therefore, the distribution of orders is maintained within a single group. Accordingly, various statistical queries such as a range search are possible within a specific group, and a meaningful value can actually be obtained.

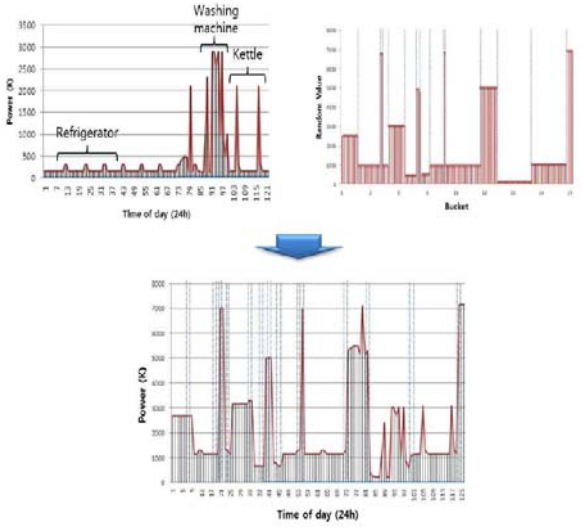

Fig. **7.** Transformation of Numeric Data

Figure 7 shows an example where numerical data was transformed. It was easy to analyze the usage status of the refrigerator, washing machine, and electric kettle using metering data alone, but such analysis can be hardly conducted with the transformed data. In particular, the data in the lower part lists transformed data by hour for the sake of better understanding; however, an actual attacker cannot list metering data in chronological order, as shown in the Figure, because he/she does not know the pre-shared Key, i.e., K. Accordingly, the attacker cannot make a power analysis attack.

(3) De-Identified Data Query and Reconstruction

When a trusted server accesses a de-identified database for a query, it first needs to obtain a group ID for the data. Since this trusted server knows a seed, it can extract the group ID based on it. If the data to be queried belongs to a specific group, only one query is needed, and the number of queries is the same as when querying plain text. In other words, all data in the group can be retrieved by setting the group ID as a retrieving search condition, and the previously encrypted time information can be converted into the source data using decryption. Then, it is necessary to reconstruct the source numerical data based on the transformed numerical data. If f(x) = y, the inverse function, i.e., f-1(y) that can deduce a source numerical data, x can be deduced as follows:

Previously, the source numerical data was transformed as follows:

$$f(x) = P^s_{n-1}\, x^2 + P^s_{n-2} + P^s_n$$

If the expression above is transposed as y - Psn-2 -Psn = Psn-1 x2, then it can be changed as follows:

$$\frac{y - P^s_{n-2} - P^s_n}{P^s_{n-1}} = x^2$$

Therefore, the inverse function f-1(y) is as follows:

$$f^{-1}(y) = \pm\sqrt{\frac{y - P^s_{n-2} - P^s_n}{P^s_{n-1}}}$$

A specific example of a numerical transformation and reconstruction is as follows: If we want to transform the data, 0.385 first shown in Fig.5, assuming that the values of Pns, Pn-1s, and Pn-2s are 254, 691, and 759, we can calculate transformed numerical data according to the numerical transformation formula as follows:

$$(691)(0.385)^2 + 759 + 254 = 1115.42$$

The transformed numerical data cannot be converted into the source data if all values of Pns, Pn-1s, and Pn-2s are not known. In other words, the data is a de-identified value, and cannot be converted into the source data unless the seed of a pseudo-random number is known. This transformed numerical data can only be converted into the source data through the inverse function by a trusted server that can generate a pseudo-random number based on the seed.

$$\sqrt{\frac{1115.42 - 759 - 254}{691}} = 0.385$$

In addition, some meaningful results can only be obtained in statistical processing by querying a de-identified database itself, even without the reconstruction of value. Reconstruction is required if the source data is inevitable. However, where it is necessary to know when the maximum or minimum value was recorded, or the range of upper usage section, it is an advantage that the statistical processing can be performed only by querying the de-identified database without a separate decryption.

IV. CONCLUSION

The smart grid environment is likely to adopt a cloud environment in the future since the data to be processed is expected to increase going forward. However, security should be taken into consideration before the adoption of a cloud environment. Without security, an activated service is not expected, and infringement on personal information may cause various big issues and lead to a disaster. In particular, since smart metering data exposes a variety of information such as an individual's privacy patterns and devices in use, de-identification is absolutely required if an untrusted cloud environment is adopted. Hence, this paper proposes a new de-identification method for metering data, which is vulnerable in terms of personal information security.

To this end, in Chapter 2, we first review the need for personal information protection from the perspective of the smart grid, and also review the de-identification method and the order-preserving encryption. Then, in Chapter 3, we propose a new de-identification method for smart metering data. For a secure smart grid environment, the protection of various kinds of personal information used in the smart grid environment should be taken into consideration. In this paper, we propose a solution focusing on the protection of metering data, in particular. A variety of personal information in addition to metering data exists in the smart grid, and if appropriate protection measures are studied, it will be possible to realize a secure smart grid environment. Henceforth, we plan to conduct a quantitative evaluation of the proposed method with speed measurement and continuously study the method to protect smart grid personal information.

Acknowledgement

This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-00207, Development of Cloud-based Intelligent Video Security Incubating Platform). Corresponding author (Namje Park).

REFERENCES

[1] M. Shargal and D. Houseman, "The Big Picture of Your Coming Smart Grid", *Smart Grid News*, 2009.

[2] Rusitschka, Sebnem, Kolja Eger, and Christoph Gerdes. "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain," *Smart Grid Communications (SmartGrid Comm), 2010 First IEEE International Conference on. IEEE*, 2010.

[3] NIST, "Guidelines for Smart Grid Cybersecurity, Volume 2 - Privacy and the Smart Grid," U.S. Department of Commerce, pp. 291-473, Sep. 2014.

[4] Namje Park, Jin Kwak, Seungjoo Kim, Dongho Won, and Howon Kim, "WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment," LNCS, *Advanced Web and Network Technologies and Applications*, 3842, pp.741-48, Jan. 2006.

[5] E. L. Quinn, "Privacy and the New Energy Infrastructure," *Social Science Research Network (SSRN)*, Feb. 2009.

[6] Namje Park, Hongxin Hu, Qun Jin, "Security and Privacy Mechanisms for Sensor Middleware and Application in Internet of Things (IoT)", *International Journal of Distributed Sensor Networks*, vol.2016, 2016.

[7] Dongkook Kim and Hyeok Lee, "Personal Information De-Idenification Trends based on Big Data," *Review of Korean Society for Internet Information*, vol.16, no.2, pp.15-22, Dec. 2015.

[8] Namje Park and Hyo Chan Bang, "Mobile middleware platform for secure vessel traffic system in IoT service environment," *Security and Communication Networks*, vol.9, no.6, pp. 500-512, Apr. 2016.

[9] Donghyeok Lee and Namje Park, "A Study on Metering Data De-identification Method for Smart Grid Privacy Protection," *Journal of the Korea Institute of Information Security & Cryptology*, vol.26, no.6, Dec. 2016.

[10] Dan Keun Sung, Nah-Oak Song, Kab Seok Ko, Jiyoung Cha, Kuk Yeol Bae and Hanseung Jang, "Convergence of Power System Technology and Information Communication Technology in Smart Grid," *Communications of The Korea Information Science Society*, vol.31, no.3, pp.10-21, Mar. 2013.

[11] Namje Park, "UHF/HF Dual-Band Integrated Mobile RFID/NFC Linkage Method for Mobile Device-based Business Application", *Journal of KICS*, vol.38, no.10, pp. 841-851, Oct. 2013.

[12] Daeseon Choi and Younho Lee, "Privacy Protection Technology for Public Information Open & Sharing," *Journal of KIISE : Computer Systems and Theory* vol.41, no.3, pp.109-115, Jun. 2014.

[13] Namje Park and Marie Kim, "Implementation of load management application system using smart grid privacy policy in energy management service environment," *Cluster Computing*, vol.17, no.3, pp. 653-664, Sep. 2014.

Authors

Donghyeok Lee received the BSc degree in information industry from dongguk university, Korea, and received his M.S. degrees in E.C.T from dongguk university in 2007, respectively. He is a researcher, elementary education research institute at jeju national university since 2015. Prior to joining the researcher at jeju univ., he had worked as a researcher at KT co. ltd. for 7 years. And he had an appointment as the researcher of the information security research division of the Electronics and Telecommunication Research Institute for 1 year. He has many talks related in information security technologies, cloud security.

Namje Park received the BSc degree in information industry from Dongguk University, Korea in 2000, and received his M.E., and Ph.D. degrees in Information Engineering from Sungkyunkwan University in 2003, and 2008 respectively. He is a Professor of Department of Computer Education in Teachers College at Jeju National University since 2010. He has been serving as a Research Scientist of Arizona State University since 2010. Prior to joining the researcher at ASU, he had worked as a post-doc at University of California, Los Angeles for 1 year. And he had an appointment as the senior engineer of the information security research division of the Electronics and Telecommunication Research Institute for 6 years. He is concerned in the information security technology field for the mobile environments, IoT system, Smart Grid, Mobile XML Security, Web Services Security, Ubiquitous computing including RFID/WSN and a variety of cryptographic technologies. He has many talks related in mobile and information security technologies, computer education.