Jiang Hancheng

Privacy preserving Data Collection for smart grid using Self-organizing Map

Master's thesis in Information Technology

October 22, 2016

University of Jyväskylä

Department of Mathematical Information Technology

Author: Jiang Hancheng

Contact information: jhc19930318@gmail.com

Supervisor: Chang Zheng & Wang Shuaiqiang

Title: Privacy preserving Data Collection for smart grid using Self-organizing Map

Project: Master's thesis

Study line: Web Intelligence and Service Engineering

Page count: 58

Abstract: Homomorphic encryption is widely researched in the smart grid area to publish and transfer electricity consumption data between electricity companies. This method makes it feasible to calculate total electricity consumption of neighborhoods without sharing any raw electricity consumption data. In the area of demand response(DR), calculating the total consumption of electricity is important in order to create DR reports which are published by third party to reduce the peak period of electricity usage such as 7 am or 6pm. Nevertheless, the possibility of data exposing or data decryption may lead to individual households private information revealing, for example, the timing of leaving home, timing of arriving home, appliances usage, detailed information of electricity devices. To avoid privacy disclosure, this thesis proposes a new framework based on self-organization map(SOM) which is an unsupervised learning method. The framework can share and publish electricity power consumption data between electricity providers securely and accurately and fulfill DR called SOM with the k-means framework. SOM with the k-means framework enables electricity providers sharing data without raw data published. Meanwhile, nearly 2.5% to 3% error and lower entropy can be achieved, which is a satisfactory result. SOM and k-means framework is a robust and effective approach for DR in the smart grid.

Keywords: SOM, K-Means, Privacy-preserving, Smart Grid, DR

Glossary

AM	Analyzing modules
AIS	Artificial immune system
AMI	Advanced metering infrastructure
AO	Asset/System Optimization
CS	Customer Side Systems
CAC	Central access controller
CM	Controlling module
DR	Demand Response
DMS	Distribution Management System/Distribution Automation
DSM	Data segmentation module
DER	Distributed Energy Resources
E&SC	Energy and service corporations
EDS	Energy distribution system
HAN IDS	Home area network IDS
IDS	Intrusion detection system
IAM	Information acquisition module
ICT	Information and Communications Integration
MM	Metering module
NAN IDS	Neighbor area network IDS
NILMN	On-intrusive load monitoring
OSGP	Open smart grid protocol

OM	Output module
PM	Preprocessing module
SM	Service module
SVM	Support vector machine
SOM	Self-organization map
SMD	Smart meter data collector
SCADA	Supervisory Control And Data Acquisition Controller
TA	Transmission Enhancement Applications
WAN IDS	Wide area network IDS

List of Figures

List of Tables

Contents

1 Introduction

Increasing number of customers' demands and scientific-technical progress have motivated the development and research of smart grid. Smart grid improves the normal functionality and capabilities of electric grids in generation, transmission and distribution parts in order to supply needs for customer-side self control and management, different energy sources combination, distributed system management. Smart grid will provide a stable, secure and effective infrastructure for users.

Smart meters are advanced electricity meters which are installed in every household. It has the ability to measure the real-time electricity power consumption data and transmits this to utilities and electricity power providers that have the contract with each household to calculate electricity data consumption and decide the tariff in an area. Based on the fluctuation of tariffs, customers can change their normal habits to save the cost of electricity power accordingly. Meanwhile, DR which is a demand-side management program reduces and controls peak period electricity consumption through varying electricity price.

Electricity consumption data is the basic information for DR forecast. While, electricity power consumption data may reveal household privacy information, for instance, the waking up time, leaving home time, arriving home time and electric appliances details[1]. Electricity power consumption data can be acquired from Internet. With advanced technology such as data mining, data analysis, or data decryption may get customer private information causing several crimes.

Non-intrusive load monitoring(NILM)[2][3] is one of the methods used to figure out households daily activities by analyzing current as well as voltage variation going into individual household to infer electricity appliances' usage condition and electricity power consumption. Utilities get detailed information of every household to analysis their activities through NILM technology with smart meter. Fig. 1 is one example which displays NILM technology through one household electricity consumption power data. If criminals acquire this kind of information will lead to serious consequences.

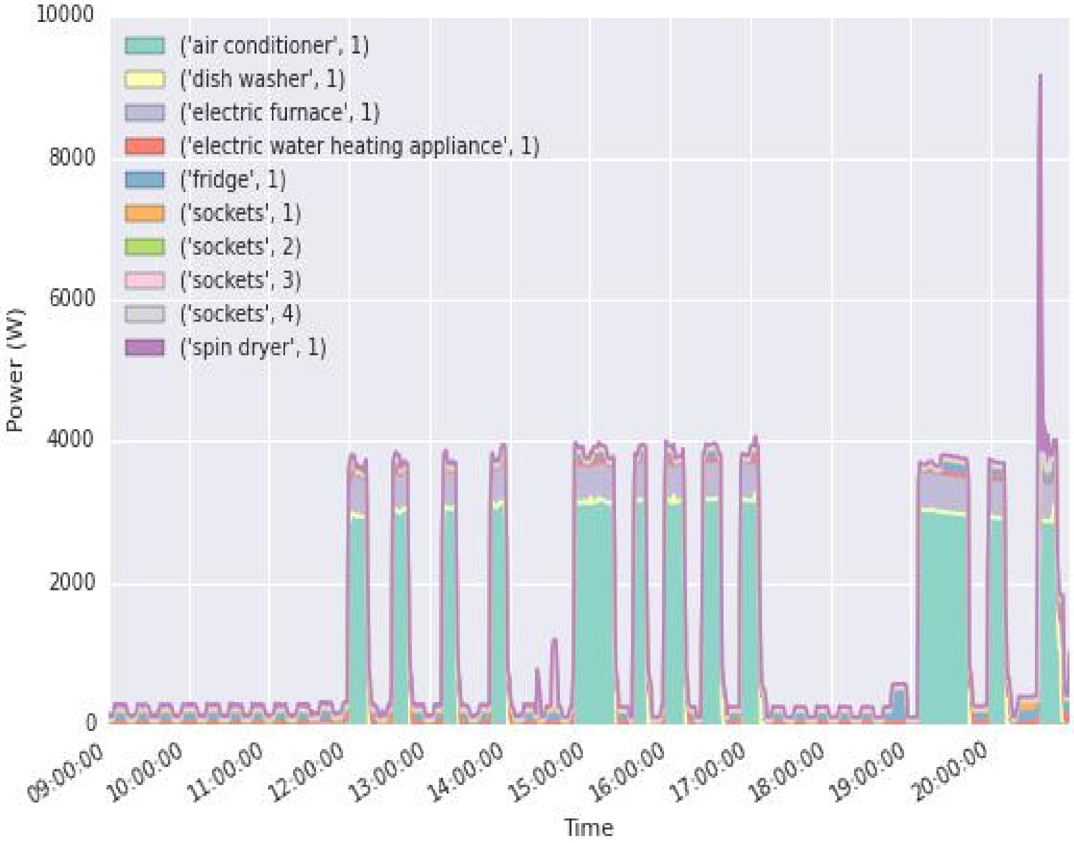


Fig 1. Example of NILM analysis from one household

Due to the implementation of electricity deregulation policy, every household may select different electricity providers because of their own daily activity pattern. In case of deregulation policy, electricity consumption power data which transmits to various electricity providers must be confidential and can not be shared between electricity providers. However,for DR, raw data needs to be shared by different electricity providers.

The motivation for this thesis is to find an effective method or platform to share raw electricity consumption power data securely and accurately for the sake of customers' privacy-preserving and DR needs. The results shows that SOM with K-means is a valid function than others with 97% or above accuracy as well as high privacy-preserving ability. The remainder of this paper is organized as follows.

Section 2 introduces the background of the smart grid. Section 3 contains the introduction of several privacy-preserving algorithms. Section 4 explains and compares several algorithms or frameworks for the DR purpose. Section 5 is the conclusion part which is the last chapter.

2 Background of Smart Grid

2.1 Electric grid evolution

As seen in Fig 2, the electric grid has two sub grids which are the transmission power grid and distribution power grid. The power plants generate three-phase alternating current voltage through a synchronized alternating current system. In order to transmit via transmission lines, the three-phase alternating current voltage needs to be increased by a generator set-up transformer. For the sake of reducing power losses during long distance transmission procedure, transmission lines have less surface area for lower electricity power capacity and resistance of conductors is lower in order to prevent power transforming into useless heats. The high-voltage alternating current will approach every substation step-down transformer decreasing alternating current voltage from high voltage to low voltage. The electricity distribution system which encompasses smaller,as well as lower,voltage distribution lines transmits lower alternating current voltage to companies, schools, stores or households.

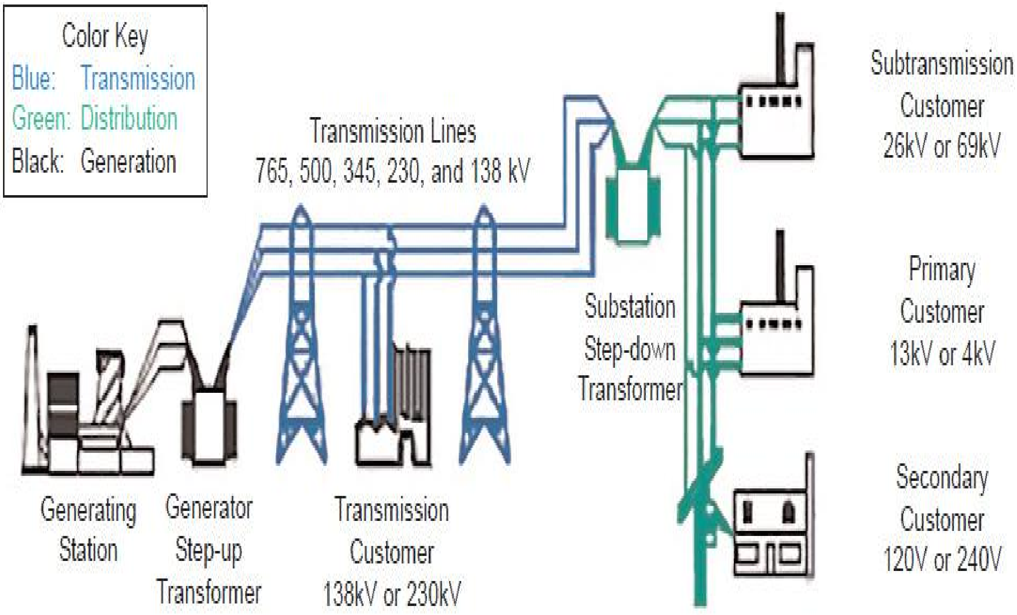

Color Key
Blue: Transmission
Green: Distribution
Black: Generation
Transmission Lines
765, 500, 345, 230, and 138 kV
Generating Station
Generator Step-up Transformer
Transmission Customer 138kV or 230kV
Substation Step-down Transformer
Subtransmission Customer 26kV or 69kV
Primary Customer 13kV or 4kV
Secondary Customer 120V or 240V

Fig 2. Electricity system

Fig 3 displays the present electricity system evolution process which only had the simplest function in the past as I explained in paragraph 2.1. Presently, electricity system has more control centers. For example, the transmission control center and distribution control center which can monitor, control, adjust the transmission and distribution procedure by means of communications with substations or generation station. However, in the future, smart grid provides more advanced technology to satisfy customers' increasing demands. Compared with present electricity systems, smart grid includes an energy storage part, using high-temperature superconductors, energy service providers, electric vehicles, combing different power source. As for the high-temperature superconductor, it may decrease power loss during transmission and distribution process by reducing resistance of power line. Combing different kinds of power sources such as wind energy source, nuclear energy source, solar energy, biomass energy, hydroelectric energy source will release high demand electricity consumption pressure. At the same time, electric vehicles are popular, which decrease carbon dioxide emission. Electric car will be introduced later in chapter 2.5.

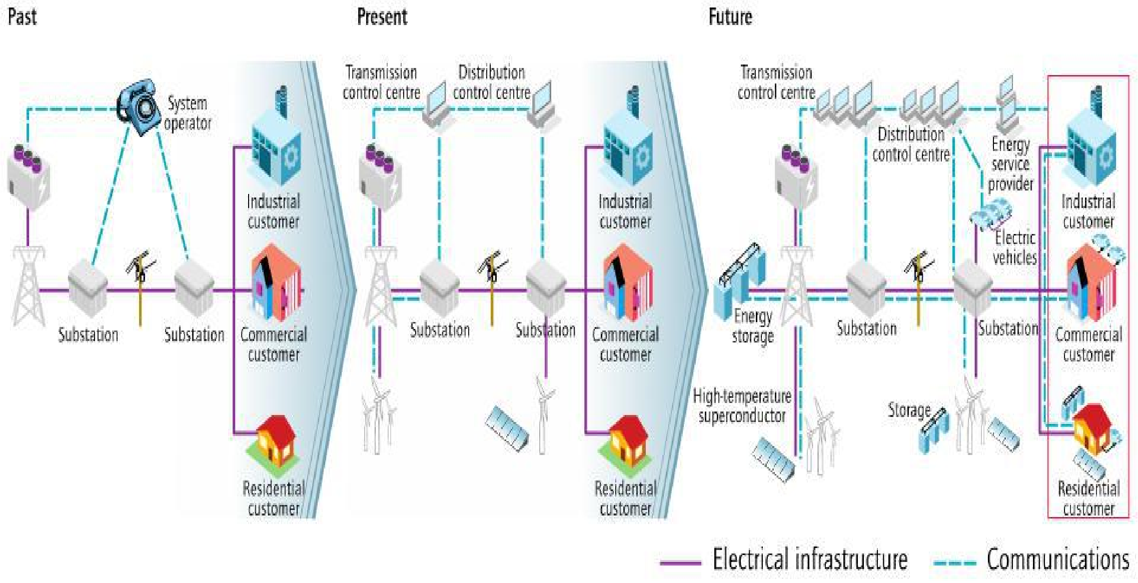


Fig 3. Electric system evolution

2.2 Structure of the smart grid

The structure of a smart grid is more complex than present electric grid system. Through Fig 4, an overview of smart grid will be introduced with a variety of service providers.

- Generation: Power plants provide electricity power energy through various power sources.
- Transmission: Transfer high-voltage from generation stations to substations through power line.
- Distribution: Substations step down high-voltage and deliver electricity power to every customer.
- Consumption: Customers use electricity power for various purposes, for example, watching TV, charging phones, washing clothes, cooking.

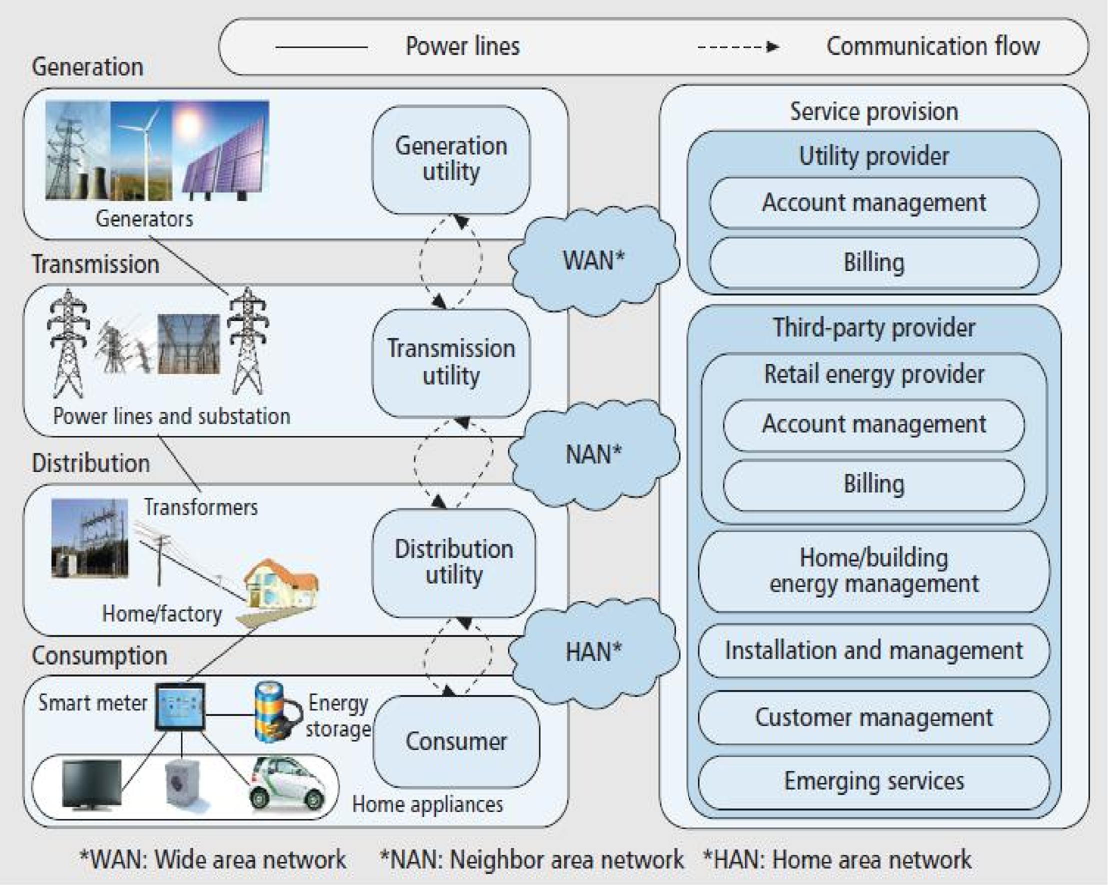


Fig 4. An overview of smart grid with service providers

- Service provision: Service providers support service for both power system generators and customers.

The service provision contains two parts, one is the utility provider and another is the third-party provider. The utility provider manages customer accounts and sends billing information about electricity power consumption to users and handles payment of each customer. Every month, customers pay cost via billing data. The third-party provider serves as a separate company. [8] discusses the concrete functions of third-party provider:

- Account management administrates the customer and retail energy provider accounts.
- Billing means third-party provider administrates customer electricity power consumption data and sends billing information with payments conducting.
- Building/home energy management supervises and manages electricity consumption and transmits controlling signals to smart grid.
- Installation and management indicates helping customers to install and maintain user equipment.
- Customer management means providing services and solving customer's problems and issues.
- Emerging services involve all kinds of existing services and innovations currently which will promote the smart grid development.

Through an overview of the smart grid, the generation station generates electricity power and transmits high voltage by transmission line. The substation steps down high voltage and distributes electricity energy to every households. In individual household, smart meter which will be introduced later measures the customer's real-time electricity power consumption and sends back to utility provider and power plants by home area network, neighbor area network and wide area network. The utility provider manages customer billing information and power plants will use precised raw data for DR in order to reduce electricity provision stress in peak time.

2.3 Smart electricity meter

[9]Smart electricity meters which were invented in the Great Britain is a new generation of electricity meter. A new generation of smart electricity meters have been widely used in the Great Britain. It will show you real-time precious electricity power consumption data. Meanwhile, the smart electricity meter will send accurate electricity consumption data to utility provider via advanced metering infrastructure(AMI) for billing purpose. At the same time, utility provider may transmit some responses to smart electricity meter. The communications between smart electricity meter are bi-directional.

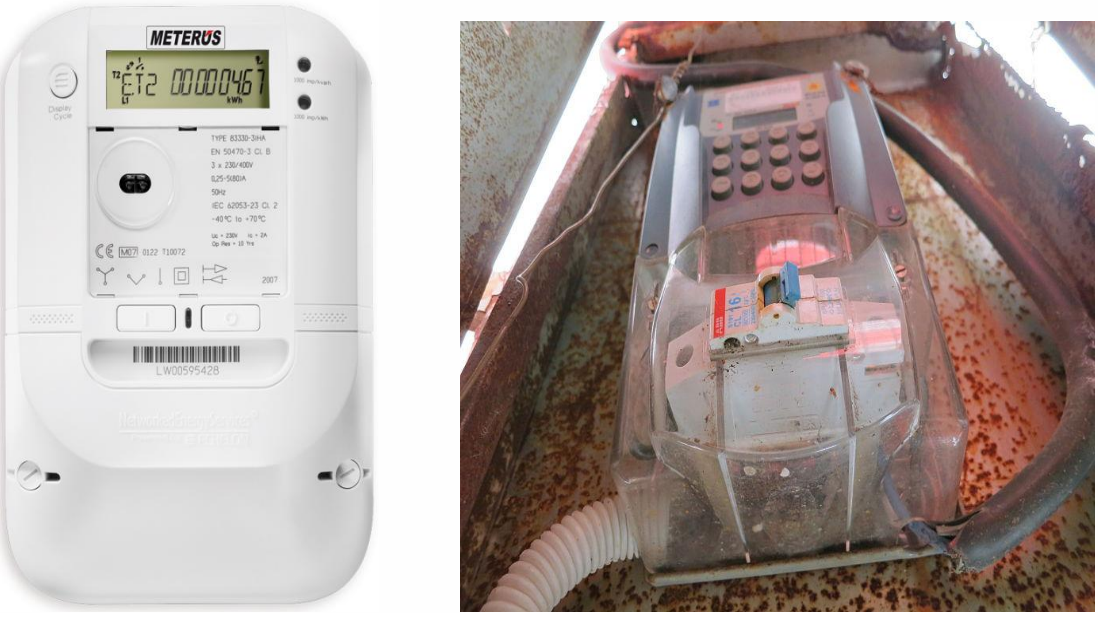


Fig 5. Example of smart electricity meter

Fig 5 shows two kinds of smart electricity meters. The left one which has the advantages of decreasing electricity load by disconnecting-reconnecting remotely is used in the European Union based upon open smart grid protocol(OSGP). The right side is a smart electricity meter that is wrapped with a transparent plastic box and is found near a supermarket in South Bali.

Once the smart electricity meter is installed in a customer's home, an in-home display equipment(Fig 6) will give to them. Through the in-home display equipment, customers are capable of checking[10]:

- The real-time electricity power consumption you are using
- How much electricity power was consumed in the past in the form of hour, day, week month, and year
- It can display if your electricity power consumption in one period is normal or abnormal(higher or lower than normal)
- Smart electricity meter updates data in high frequency(almost real time)

Moreover, if customer installs a prepay meter in home which interacts with smart electricity meter, prepay meter is able to show how much balance do you have.

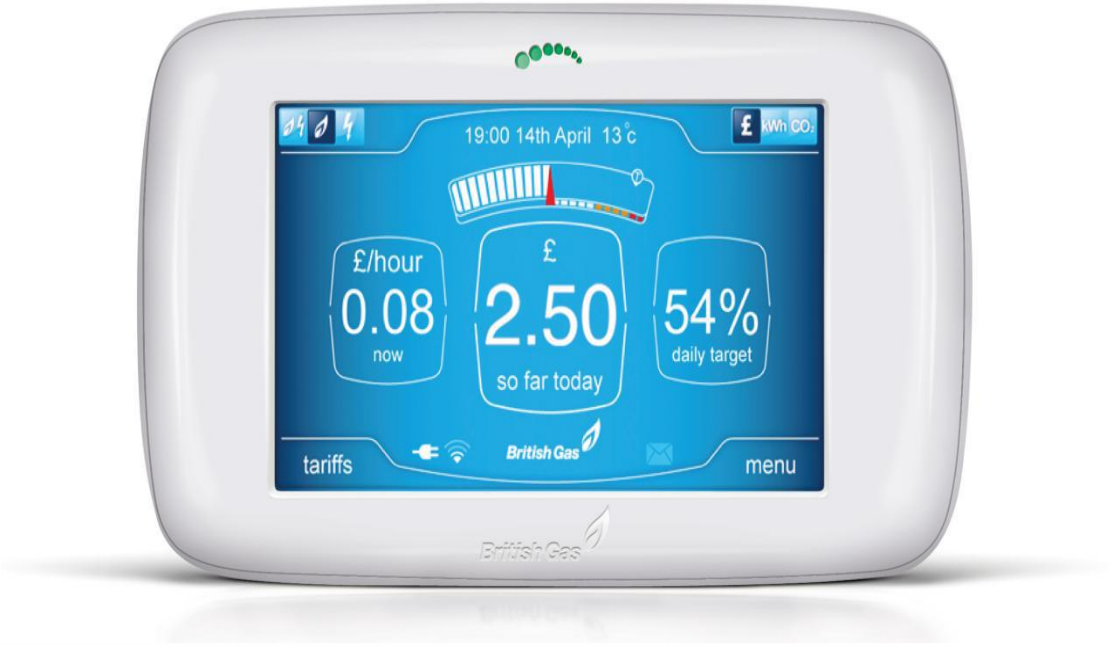


Fig 6. Example of in-home display equipment

Unfortunately, for the Victorian region, meter charge increases by about $60 for one smart electricity meter in order to make up AMI cost from users in 2010. As the table below shows, we can see that meter charge increased rapidly from 2010.

Distributor	2005	2006	2007	2008	2009	2010	2011	2012	2013	2015	2016 2017
SP AusNet	17.49	17.49	17.49	17.49	17.49	86.1	93.83	101.02	108.75	117.08	126.04
United Energy Distribution	6.60	6.60	6.60	6.60	6.60	69.21	89.18	99.57	107.62	116.33	125.73
Jemena Electricity Networks	12.87	12.87	12.87	12.87	12.87	134.63	136.7	155.84	159.86	162.34	164.88
Citipower	15.20	15.20	15.20	15.20	15.20	104.79	108.4	93.38	95.26	97.17	99.13
Powercor	17.20	17.20	17.20	17.20	17.20	96.67	105.35	92.72	93.91	95.12	96.34

Table 1. Meter charge increase from 2010 and forecast for 2017[9]

2.4 Three-layer network of smart grid

2.4.1 Components of three-layer network

Smart grid communication infrastructure consists of three layers which are home area network, neighbor area network and wide are network. The communication between smart electricity meter and utility provider relies on three layer network. [11] introduces concrete components of three-layer network.

Home are network which is the first layer of the smart grid three-layer network is composed of the metering module(MM) that contains the smart meter part, service module(SM) as well as intrusion detection system(IDS) module. Each household's real-time electricity power consumption data is supplied by SM. At the same time, the MM records each household's real-time electricity power consumption data. As for the home area network IDS that will monitor and track the ingoing and outgoing transmission information for the sake of checking problems or threats taking place accidentally.

Neighbor area network is the second layer of the smart grid three-layer network that will gather nearby home area network's metering and service data and transmit the data to upper layers. Neighbor area network is composed of the central access controller(CAC), the smart meter data collector(SMDC) as well as the neighbor area network IDS. The CAC is regarded as a communication connector between home area networks and utility provider or energy provider. As for the SMDC,a wireless node, it will be responsible for geographically nearby home area network's metering logs. Neighbor IDS has the same but advanced functions than the home area network IDS, which monitors the whole incoming and outgoing data stream with the purpose of probing security issues.

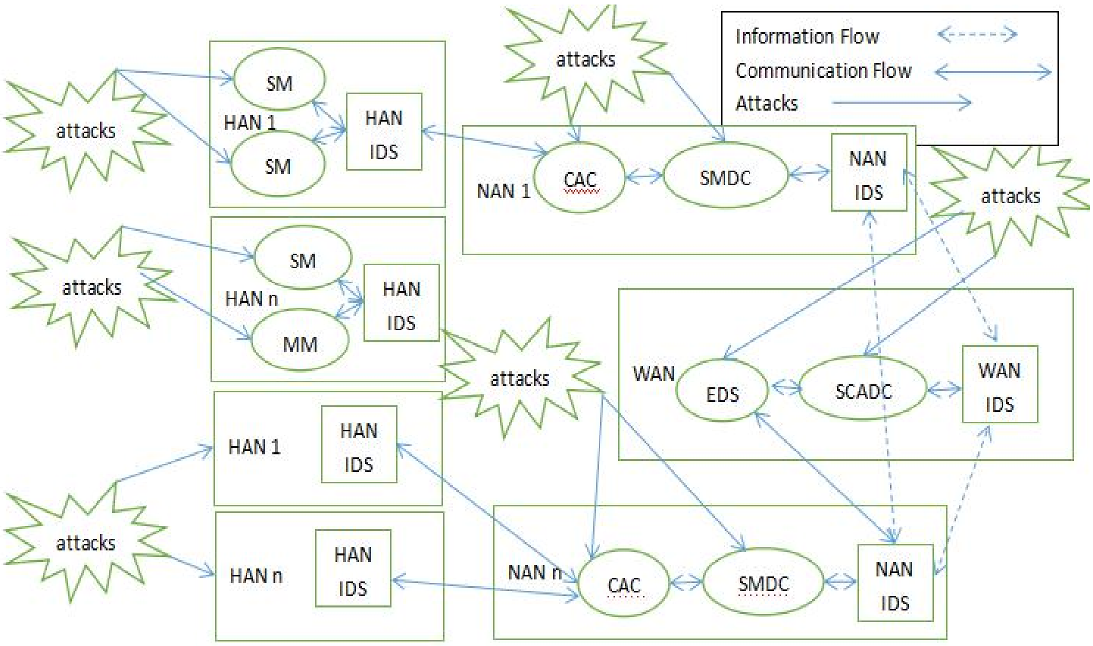


Fig 7. Three-layer network architecture

The third layer of the three-layer network is named as the wide area network. Wide area network provides not only the wireless transmission but also the wired network communication between neighbor area networks, substations, utility providers, power providers, remote smart grid devices. Wide area network encompasses three components which are the energy distribution system(EDS), the Supervisory Control And Data Acquisition Controller(SCADA) controller and the wide area network IDS. With regard to EDS, it will be responsible for the distribution of the metering data. With the purpose of administrating distribution smart grid devices, the SCADA controller supports distributed process control for the utility provider. Meanwhile, wide area network IDS is needed in charge of security problem between the SCADA controller and energy and service corporations (E&SC) because of the importance of metering data and control data. Data leakage will bring crucial consequences. Fig 7 describes the three-layer network architecture visually.

2.4.2 Communication technology for three-layer network in smart grid

For the three-layer network in the smart grid, different layers require different data transmission rates and signal cover ranges. Fig 8 shows requirements. [12] describes the specific information as below:

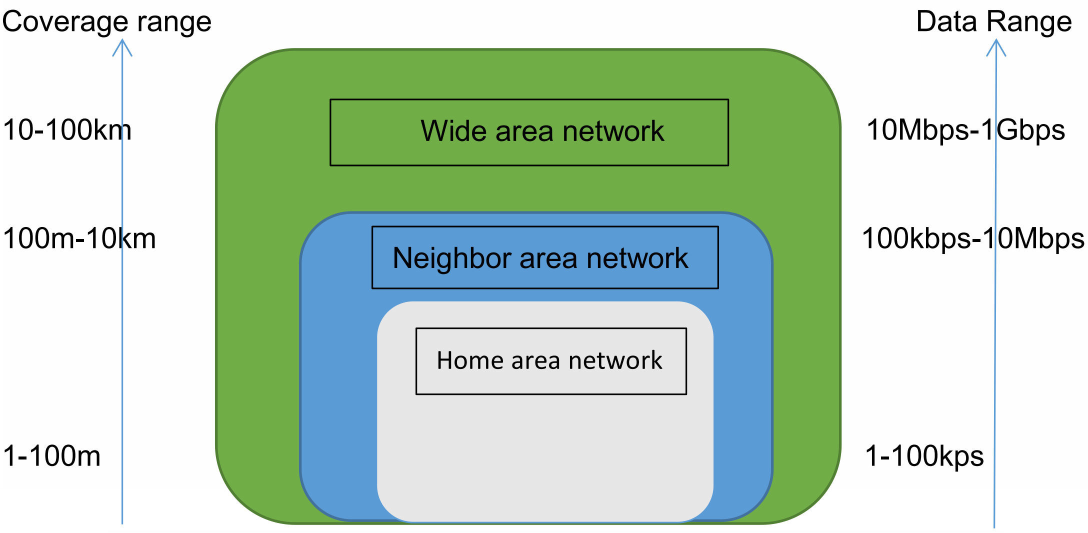


Fig 8. Data rate and coverage range for home area network, neighbor area network, wide area network

House automation and industry automation(home area network applications) need to transmit the electricity power consumption data to a controller based on wireless transmission technology. The data rate does not request high speed and frequency and coverage range is smaller because all applications are inside limited houses or industries. Accordingly, lower electricity power consumption, security, reliability are main characteristics for the home area network applications' data transmission. Therefore, the data rate achieves 100 kbps as well as coverage range up to 100 meters are adequate for the home area network. The communication technology such as the

ZigBee, ZWave, WiFi, Bluetooth, Ethernet and power line carrier is common employed in home area network automation applications.

As for the neighbor area network, it includes several applications, for example, DR, distribution automation and smart metering. Data transmission from the customer side to substation side to customer side is burdensome and frequent. In consequence, data rate needs to be higher than the home area network communication rate and the rate is from100 kbps up to 10 Mbps. Meanwhile, coverage range requires up to 10 Km because of the long distance between substations and customers' devices. Accordingly, mesh networks Fig 9(ZigBee mesh networks and WiFi mesh networks), power line carrier, WiMax, cellular, digital subscriber line and Coaxial cable can be used to satisfy requirements.

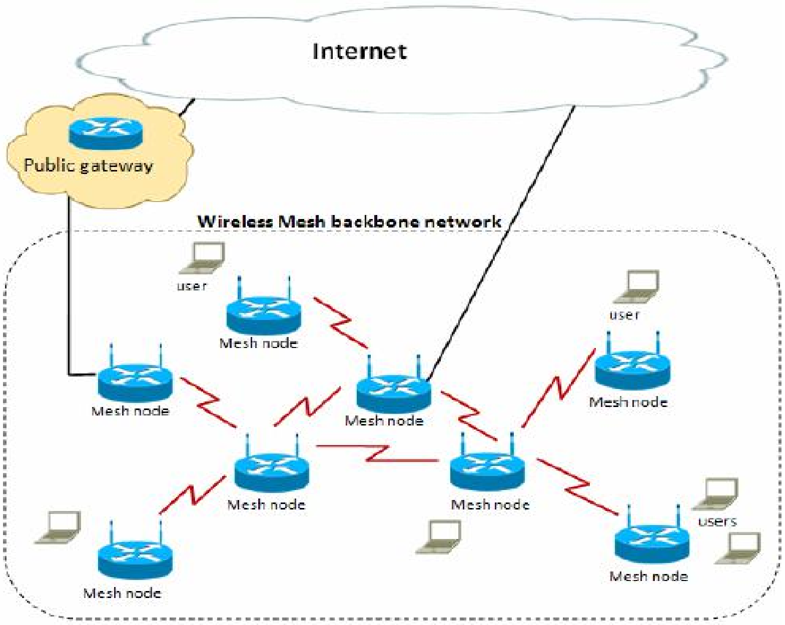
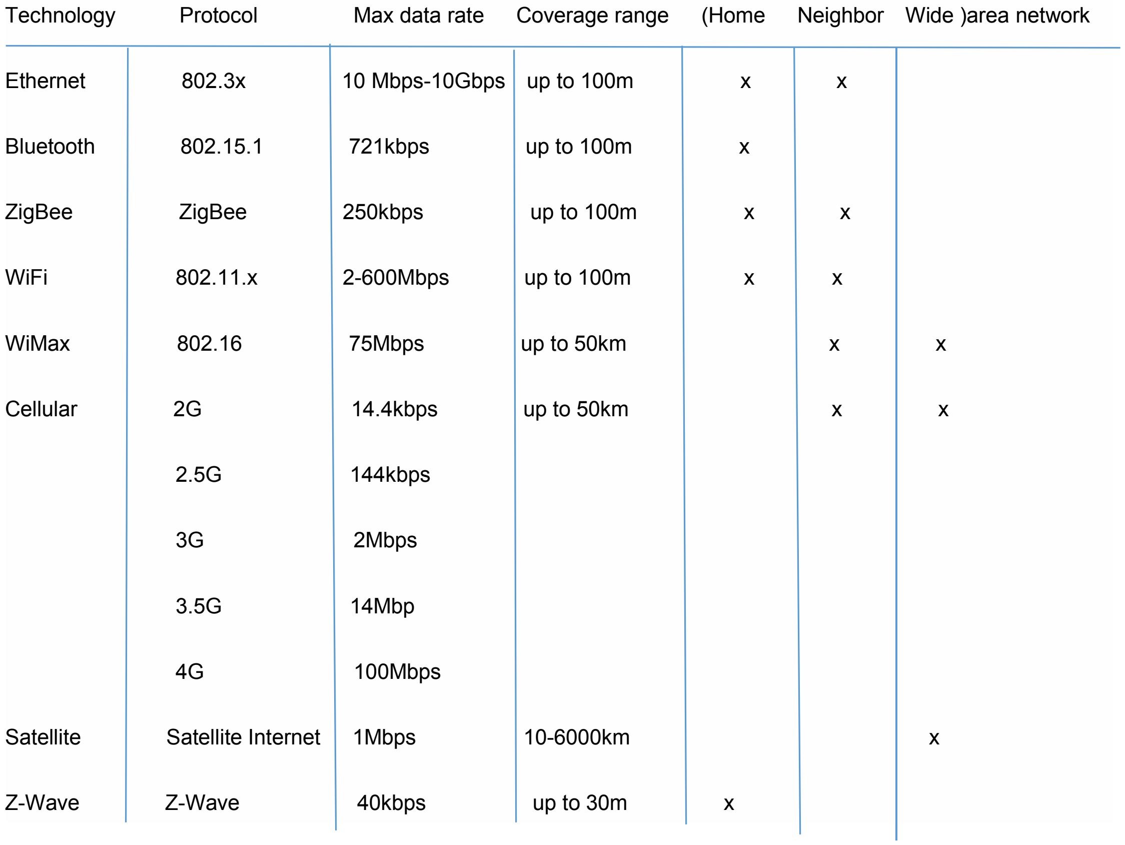


Fig 9. Mesh network architecture

Wide area network applications consist of the wide-area control, monitoring and protection. Therefore, data transmission is huge and more frequent than neighbor area network applications in order to guarantee stability in the smart grid system. Data rate needs to be 10 Mbps until 1 Gbps, which is higher than mentioned above. At the same time, coverage range is longer up to 100 Km. Nowadays, the transmission between utility providers or power providers and substations uses the optical communication technology which has the advantages of higher data capacity and shorter delay. Meanwhile, Cellular and WiMax communication technologies are suitable for the data transmission because of the long cover range and rapid data rate.

Table 2 presents the different kinds of communication technologies with data rates and coverage ranges which satisfy various layer applications.



Technology	Protocol	Max data rate	Coverage range	(Home	Neighbor	Wide)area network
Ethernet	802.3x	10 Mbps-10Gbps	up to 100m	x	x	
Bluetooth	802.15.1	721kbps	up to 100m	x		
ZigBee	ZigBee	250kbps	up to 100m	x	x	
WiFi	802.11.x	2-600Mbps	up to 100m	x	x	
WiMax	802.16	75Mbps	up to 50km		x	x
Cellular	2G	14.4kbps	up to 50km		x	x
	2.5G	144kbps				
	3G	2Mbps				
	3.5G	14Mbp				
	4G	100Mbps				
Satellite	Satellite Internet	1Mbps	10-6000km			x
Z-Wave	Z-Wave	40kbps	up to 30m	x		

Table 2. Communication technologies comparison to apply in different applications

2.4.3 Distributed intrusion detection system (IDS) modules in three-layer network

Fig 10 displays the general structure of the home area network IDS which consists of several intelligent modules[11]. For The information acquisition module(IAM), it gathers electricity power consumption data packages and preserves all data packages into a matrix. Collected data packages will be divided into sizeable segments by the data segmentation module(DSM). Segmentation files are transmitted to the preprocessing module(PM) to preprocess. Afterwords, the analyzing modules(AM) which encompasses three sub components as displays in Fig 11 has the ability to probe dubious issues. PM sends preprocessing data to the intrusion data acquisition module that is the first sub-part of AM. Next, the trained support vector machine(SVM) or artificial immune system(AIS) models classifies suspicious intrusions. In the end, accuracy evaluations and result recordings that include the types of intrusions, location information and time of attacks are shown through output module(OM). Controlling module(CM) is regarded as the brain for humans, which controls all of probe procedures that occur in the home area network.

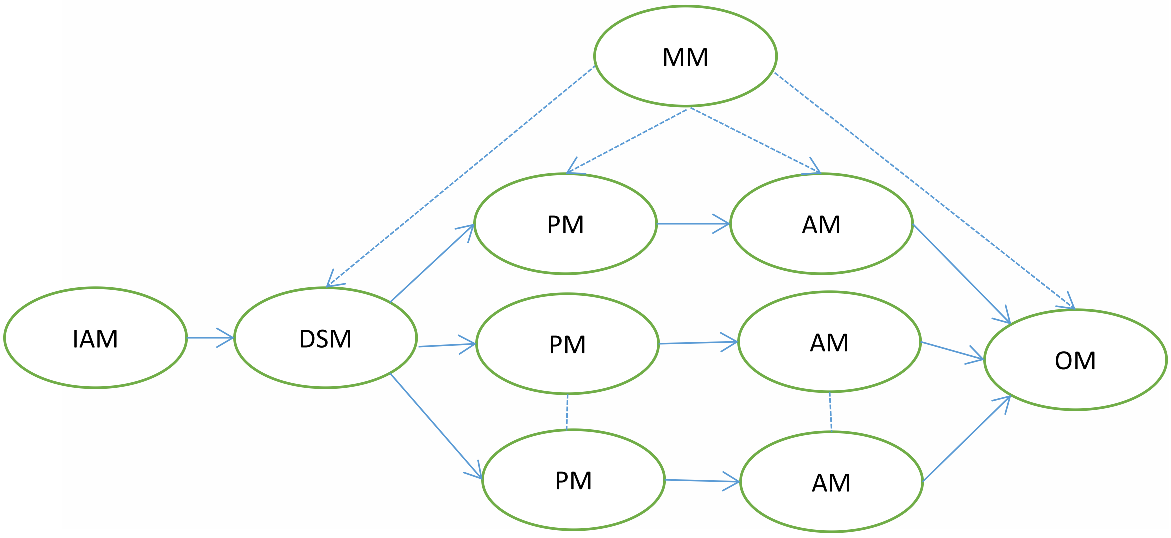


Fig 10. IDS in home area network

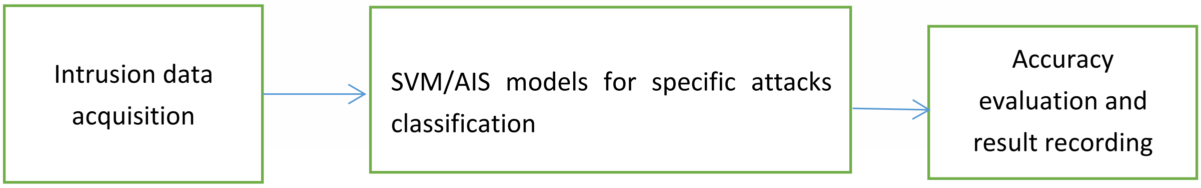

Intrusion data acquisition
SVM/AIS models for specific attacks classification
Accuracy evaluation and result recording

Fig 11. Analyzing module structure

Fig 12 and Fig 13 display the neighbor area network IDS(NAN IDS) and wide area network IDS(WAN IDS) respectively[11]. Not only the neighbor area network but also the wide area network include the advanced home area network IDS(HAN IDS). HAN IDS has SVM/AIS classification algorithm models for concrete attacks when disposing a large amount of electricity data packages or other information in different communication layers. Finally, accuracy evaluations and result recordings will be displayed. In addition, a central controller exists in the WAN IDS that is acquired for managing the NAN IDS. In the case of hard to classify suspicious attacks in the present layer, malicious attacks will be transmitted to upper layers depending on the decision of current layer's evaluation results.For example, if malicious attacks can not be classified by the HAN IDS, attacks will be transmitted to the NAN IDS or HAN IDS and results can be done in the same manner.

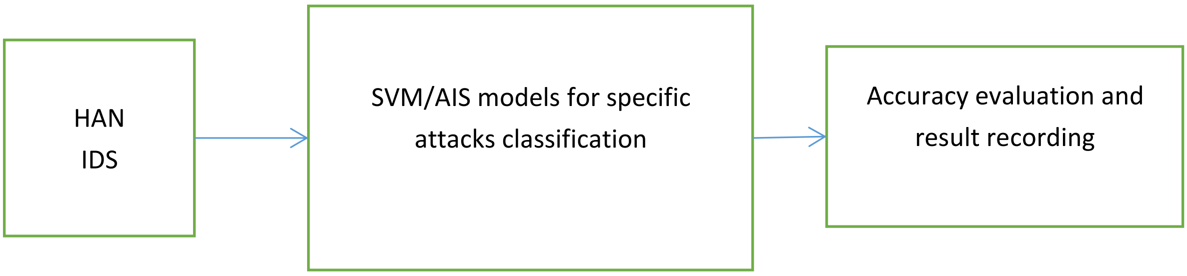

HAN
IDS
SVM/AIS models for specific attacks classification
Accuracy evaluation and result recording

Fig 12. Neighbor area network intrusion detection system

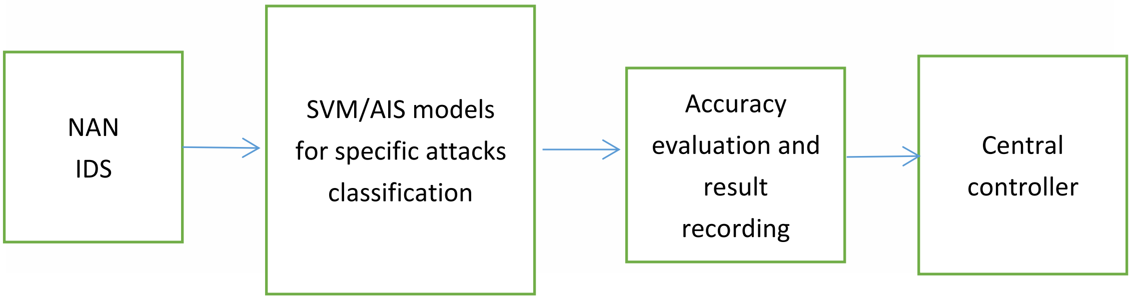

NAN
IDS
SVM/AIS models for specific attacks classification
Accuracy evaluation and result recording
Central controller

Fig 13. Wide area network intrusion detection system

2.5 Electric vehicle

2.5.1 Introduction of electric car

[14]Electric car is a kind of vehicle which includes several electric motors that can provide powerful and steady acceleration to drive. Meanwhile, compared to common internal combustion engines, electric motor is three times as efficient as them. Instead of gasoline or diesel, electric car uses electrical power that is saved in rechargeable batteries. In the 1880s, the first electric car was invented. [15][16] Until in the end of 19th century, electric car became fashionable. Nevertheless, with the development and advancement of internal combustion engines and gasoline cars' lower price, the sales volume of electric car reduced rapidly. However, according to the energy shortage problem between 1970s and 1980s, electric cars had a brief prosperity.

From 2008, due to the promotion of the new batteries technology and smart grid occurring, the former electric car manufacturing industry was reviving. At the same time, with the higher gasoline price and [17][18]encouragement of local governments for decreasing greenhouse effect, the market of electric car is boosting with high speed. Also, electric car can bring a large amount of advantages than gasoline cars nowadays. Firstly, the electric car is silenter compared to normal internal combustion engine cars. Secondly, as for electric cars' exhaust gas emission, such as, the nitrogen (N2), water vapor (H2O) (except with pure-carbon fuels), and carbon dioxide (CO2) and carbon monoxide (CO) from inadequacy combustion can be avoided. Electric car is benefit for environment protection[19] and reduces greenhouse effect[17][18]. In the meantime, people breath fresh air with lower PM 2.5, which may decrease lung cancer proportion.

As shown in Fig 14, an electric car is charged on Rome street in 2016.

Fig 14. Electric car is charged on street in Rome in 2016

2.5.2 Electric vehicles development history in China

[20]In 2009, China gained on the USA which had 10.43 million sales volume including electric cars and light trucks. However, 13.9 million electric vehicles sold in domestic(china) and was the largest electric vehicle market because of the high requirement for electric vehicles. Meanwhile, due to the income improvement in china, more and more young people have the ability to buy electric cars. Also, the government provides subsidy for those people who buy the electric cars and it is easier to get vehicle license plate to some extent especially in big cities, such as Beijing, Shanghai, Guangzhou, Shenzhen. In order to encourage the progress of electric vehicles, the government invests 15$ billion to electric vehicle factory[21]. Furthermore, electric vehicle industry’s creation will bring a large number of job opportunities and export revenue somehow. Also, with the development of electric vehicles, air pollution and reliance on gasoline will reduce accordingly[22]. By 2020, five million battery-electric and plug-in hybrid electric cars are an objective to

achieve for the government. Also, one million yearly output by 2020 is another target[23].

Electric vehicle industry milestones show as below[20]:

2001

In 2001, “863 Electricity Vehicle Project ” begins with different kinds of electric vehicles, for examples, pure electric vehicle, hybrid electric vehicle and fuel cell electric vehicle.

2004

In Beijing, electric vehicle industry association is established by National Development and Reform Commission for the sake of electric vehicle standards unification and stakeholders information sharing between each other. 14.7$ billion is expected to provide by companies to develop boosting electric vehicle industry.

2007

300$ million is invested to exploit new energy vehicles in 2007.

2008

Compared to last half year, there is a 107.9% rapid growth. Also, 500 high efficiency electric vehicle are provided by vehicle manufacturer for Beijing 2008 Olympics. In 2009, 13 cities which are Beijing, Shanghai, Chongqing, Changchun, Dalian, Hangzhou, Jinan, Wuhan, Shenzhen, Hefei, Changsha, Kunming and Nanchang will become trial cities to put electric vehicles to use.

2009

1.5$ billion are supported by the State council for Auto Industry Restructuring and Revitalization Plan purpose to build new electric vehicle industry. Also, they provides 3$ billion to sustain technical exploitation. Furthermore, two-year trial project is proposed about allowance for electric vehicle purchaser in several cities with

RMB60,000 for battery electric vehicles and RMB50,000 for plug-in hybrid vehicles[24].

2010

Government drafts an auto industry exploitation program between 2011 and 2020.

From January to September in 2016, there are 289,000 electric vehicle has been sold, which has a 100.6% growth compared to last nine months in 2015, including 216,000 pure electric vehicles and 73,000 plug-in hybrid vehicles.[26][27]. Fig 15 displays sales of new electric vehicles in China from 2011 to first quarter 2016.

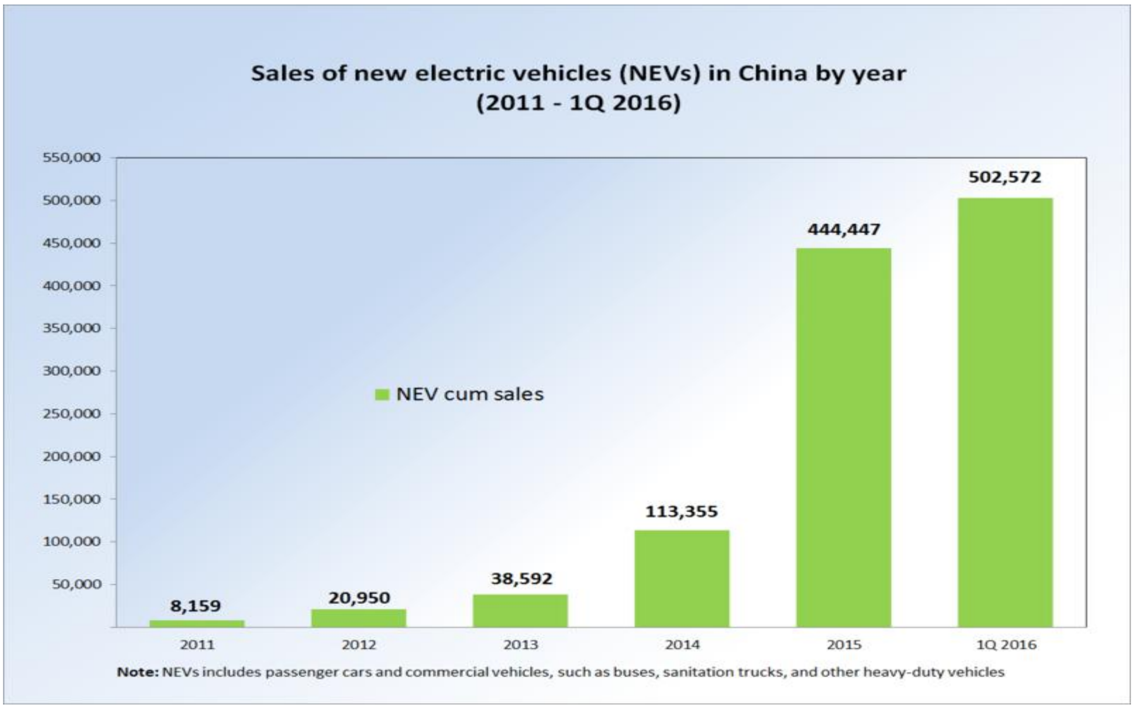
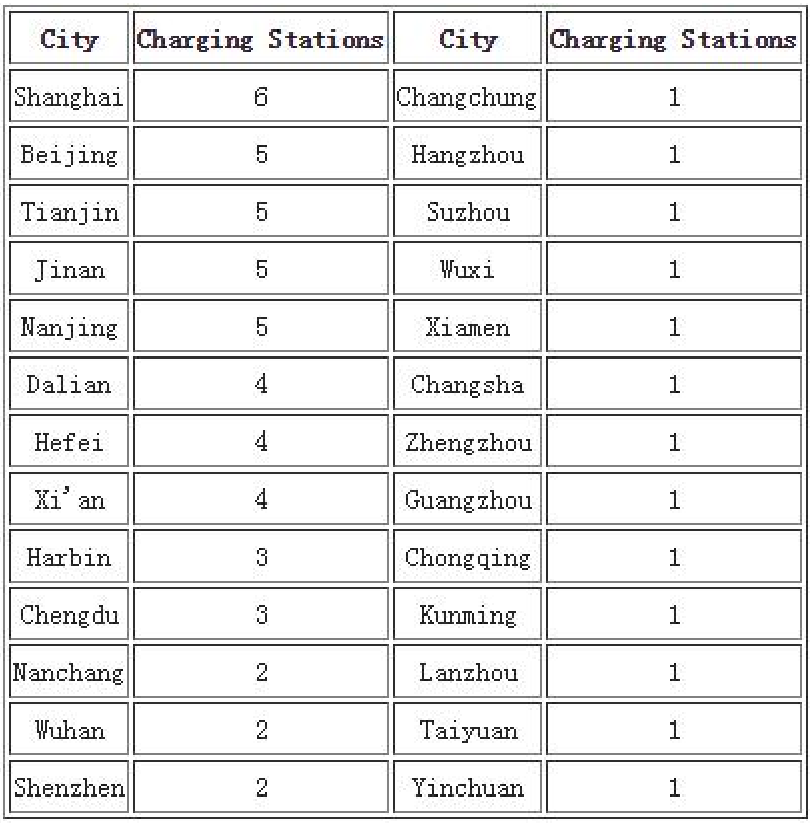


Fig 15. Sales of new electric vehicles in China by year(2011-1Q 2016)[25]

2.5.3 Electric vehicles charging equipment in China

[28]The first commercial electric vehicle charging station which is called Caoxi electric vehicle charging station has been built and used in August 2009. In 2010, there are 76 electric vehicle charging stations that have been established in 41 cities in china. [31]Table 3 is the concrete amount of charging stations in different cities in

China as of 2010. Because of the advantages of electric cars and the development of the smart grid, the target of the government is to possess at least 500,000 available hybrid or pure electric cars before 2015. Meanwhile, 5 million hybrid or pure electric cars are the goal by 2020.[29][30]



City	Charging Stations	City	Charging Stations
Shanghai	6	Changchung	1
Beijing	5	Hangzhou	1
Tianjin	5	Suzhou	1
Jinan	5	Wuxi	1
Nanjing	5	Xiamen	1
Dalian	4	Changsha	1
Hefei	4	Zhengzhou	1
Xi’an	4	Guangzhou	1
Harbin	3	Chongqing	1
Chengdu	3	Kunming	1
Nanchang	2	Lanzhou	1
Wuhan	2	Taiyuan	1
Shenzhen	2	Yinchuan	1

Table 3. Charging stations in different cities as of 2010[31]

2.6 Benefits of the smart grid

As for the smart gird, there are seven principal features as shown below[33]:

1. Residents have more options as well as bidirectional communications between residents and electricity plants or utility provider may improve the enthusiasm of customers. At the same time, active interaction between each other will benefit not only the smart grid but also our environment.

2. Smart grid is suitable for various electricity generation processes and storage modes. Electricity plants can utilize multiple energy resources such as the wind

energy, solar energy, nuclear energy, fuel energy or other cleaner energy production modes.[32] For storage, several popular methods are be used, for example, compressed air energy storage, high-speed flywheels, pumped hydro, vehicle-to-grid,rail energy storage, solid electrochemical batteries, flow batteries, thermal energy storage and molten salt storage.

3. Smart grid benefits the appearance of novel products, services and markets. Customers choose new green power vehicles, for instance, the electric car, electric bus and hybrid car because of the the open market. Also, efficient electricity markets can decrease the transmission jam.

4. Offers a stable digital economy through high power quality. In order to reduce production and productivity losses particularly in digital-equipment circumstance, higher power quality and stability is needed to supply.

5. Optimizes the asset usage and handles efficiently. Optimized utilization of the asset and effective operation may reduce the cost in smart grid. Frequent and targeted maintenance minimizes facility faults and increases safety of operations.

6. Predicts and responds to the smart grid interference. Smart grid has the persistent self-evaluation function for the detection, analysis, replying, recovering elements and network parts.

7. Smart gird can effectively resist against hackers' attacks and natural disasters in order to increase the social security.

Seven principal features can be implemented through the proposition and development of technology solutions for smart grid. These solutions guide and affect the planning, designing, operating and maintaining to some extent. Several technology solutions which displays below can be taken into account while developing the execution plan of the smart grid[33]:

- Advanced Metering Infrastructure (AMI)
- Customer Side Systems (CS)

- DR (DR)
- Distribution Management System/Distribution Automation (DMS)
- Transmission Enhancement Applications (TA)
- Asset/System Optimization (AO)
- Distributed Energy Resources (DER)
- Information and Communications Integration (ICT)

Technology solutions that are listed above can bring six vital values which can benefit smart grid, environment and residents[33]:

- Reliability—reliable power supply and power quality can decrease the possibility of large-scale blackouts that will cost a lot of losses especially for digital device circumstances. Also, Smart grid has the high ability to withstand interruptions and disturbances.
- Economics—with the advantages of the DR, residents can save money through changing their daily habits and avoid the peak period. Compared to the normal gird, electricity prices of the smart grid is much more cheaper for customers. Meanwhile, smart grid can offer a large number of new job opportunities and incent the gross domestic product.
- Efficiency—new technologies will improve the efficiency and decrease the cost for production, transmission and distribution process.
- Environmental—cleaner and renewable resources occupy a large proportion compared to normal grid and more reasonable. Efficient ways of the generation, transmission and consumption can decrease the exhaust gas or harmful gas emission in a way.
- Security—smart gird is capable of efficiently detering the cyber attacks and natural disasters. At the same time, a large amount of losses can be avoided accordingly.
- Safety—grid-related harms and deaths can be decreased to some extent.

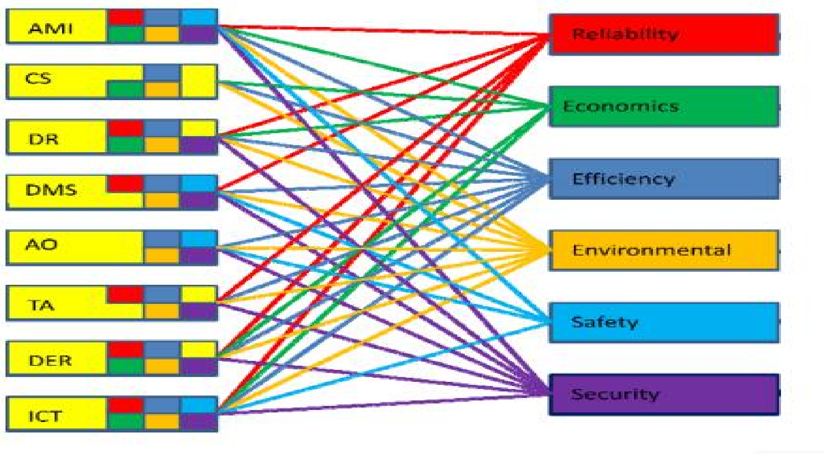


Fig 16. Relationship between technology solutions and six key values.[33]

Fig 16 embodies the "many-to-many" relationships between technology solutions and six key values. Also, this figure shows the mutual promotion of technology solutions in the smart grid.

2.7 Research question for DR in smart grid

[34]DR which is a demand-side management program reduces and controls the peak period electricity consumption through altering prices of electricity and changing residents' consumption patterns by some stimulations.

[34]Because of the deregulation policy which is already carried out in the United Kingdom and the United States and starts in Japan from 2016, different households have their own rights to choose the electricity providing company depending on their electricity consumption patterns. For the DR purpose, different electricity power companies have to share the customers' electricity consumption data.

ID	0:00-0:30 AM	0:30-1:00 AM	1:00-1:30 AM	1:30-2:00 AM	2:00-2:30 AM	2:30-3:00 AM	....	7:00-7:30 AM	7:30-8:00 AM	8:00-8:30 AM	......	5:00-5:30 PM	5:30-6:00 PM	....	11:00-11:30 PM	11:30-0:00 PM
1	1.3	1.3	1.3	1.3	1.3	1.3	1.3	2.6	2.4	1.3	1.3	3.4	3.0	2.8	1.3	1.3
2	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2.7	2.9	1.5	1.5	3.6	3.3	2.4	1.5
3	2.4	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	3.6	1.2	1.2	1.2	1.2	3.0	2.8

Table 4. Example of electricity consumption data for 3 households in 24 hours in housing estate A (unit:kwh)

Table 4 is an example of the electricity consumption data for 3 households in 24 hours and the interval of records is half an hour. By the analysis of this table, we can get the general information that household 1 waked up between 7:00 and 7:30 in the morning and left home before 8:00 AM. From 8:00 to 17:00, nobody stayed at home because the electricity consumption data did not change. After 17:00, someone returned home according to the electricity consumption data increasing from the 1.3kwh to 3.4kwh. After 23:00, they went to sleep because the electricity consumption data decreasing to the 1.3kwh. Similarly, we can also roughly analyze the timing of daily activities and patterns of the household 2 and household 3. If the raw data is disclosed on Internet, lawbreakers or criminals may analyze the electricity consumption data of some specific households which have the obvious characteristics. Then, they may match the ID of the electricity consumption data table with the real location in the housing estate A and thefts may occur. Meanwhile, criminals can infer their job attributes by residents' daily patterns.

Name	Job	Sex	Age	Phone number	Working place
Jiang	student	male	23	0414814278	University of Jyvaskyla
Tom	lawyer	male	30	0446546456	Cygnaeuksenkatu 10 Jyväskylä
Jane	salesperson	female	28	0414534534	Forum

Table 5. Personal information in housing estate A

Throng the long-term observation of criminals, crimes can also match the specific person with table 5 that is a general personal information form in the housing estate A based on their job, sex and age attributes. Not only thefts but also cyber fraud will take place.

According to the above concentrate analysis, disclosure of the raw electricity consumption data will bring a serious of severe consequences. So, the research question of this paper is that how to share the electricity consumption data accurately and securely without the raw data disclosure between electricity power companies. For next chapter, I will introduce several privacy-preserving algorithms.

3 Privacy-preserving algorithms

3.1 K-Anonymity

3.1.1 Introduction of the k-anonymity

Anonymization is an effective and straightforward method to protect the customers' privacy and achieve the privacy-preserving purpose. Presently, k-anonymity[35], l-diversity[36], and t-closeness[37] are widely investigated. In this paper, only K-Anonymity will be explained.

[35]K-anonymity is an algorithm to protect the customers' privacy information. For a released table, k-anonymity makes sure that there are at least k identical rows in case of re-identification. [43]Suppression and generalization are two common methods to be used to achieve k-anonymity. As for the suppression method, the most common way is to replace several table's values by asterisk or other punctuation marks. For generalization, an specific value will be replaced by a wide range. For example, the specific value 12 may be replaced by a range (10.15) or '<15'. [34] introduces several advantages and disadvantages of the k-anonymity. K-anonymity is easier to understand and the anonymized released table looks intuitive. However, for a small table, k-anonymity may lead to a lot of useful information loss to some extent. Meanwhile, [42] illustrates that k-anonymity is not suitable for high dimensional tables. Moreover, as seen in [44], k-anonymity may lead to the anonymized table meaningless and skewed if data holders can not suppress and generalize values proportionately or characteristics are not chosen classically. Fortunately, if suppression and generalization methods are used balanced to achieve k-anonymity, the released table may not seen such skewed and meaningless[45].

K-anonymity also be widely used in different research fields. [46] proposed a tool to measure the quantity of retained anonymity in data mining process. At the same time, this method can be used in many aspects of the data mining, for example, classification, clustering and association. (k-p)-Anonymity[47] is another anonymity method based on the K-Anonymity. This algorithm is quite useful for time series tables' anonymization. Firstly, generalization method is applied to achieve the k-anonymity. Pattern representation is a way to represent the increasing or decreasing from one time period to another period. For any record r in a k-group, if there exist at least P − 1 other records which have the same pattern representation as r, we say that P- anonymity is enforced for this k-group. As a result, we can partition the k-group further into subgroups. Global positioning system has motivated the development of location-based services. However, users' location-based information should be managed appropriately. If user's location-based private information disclosed by those services casually, some severe problems may occur. [48] uses k-anonymity to protect users' location-based information.

3.1.2 K-anonymity algorithm

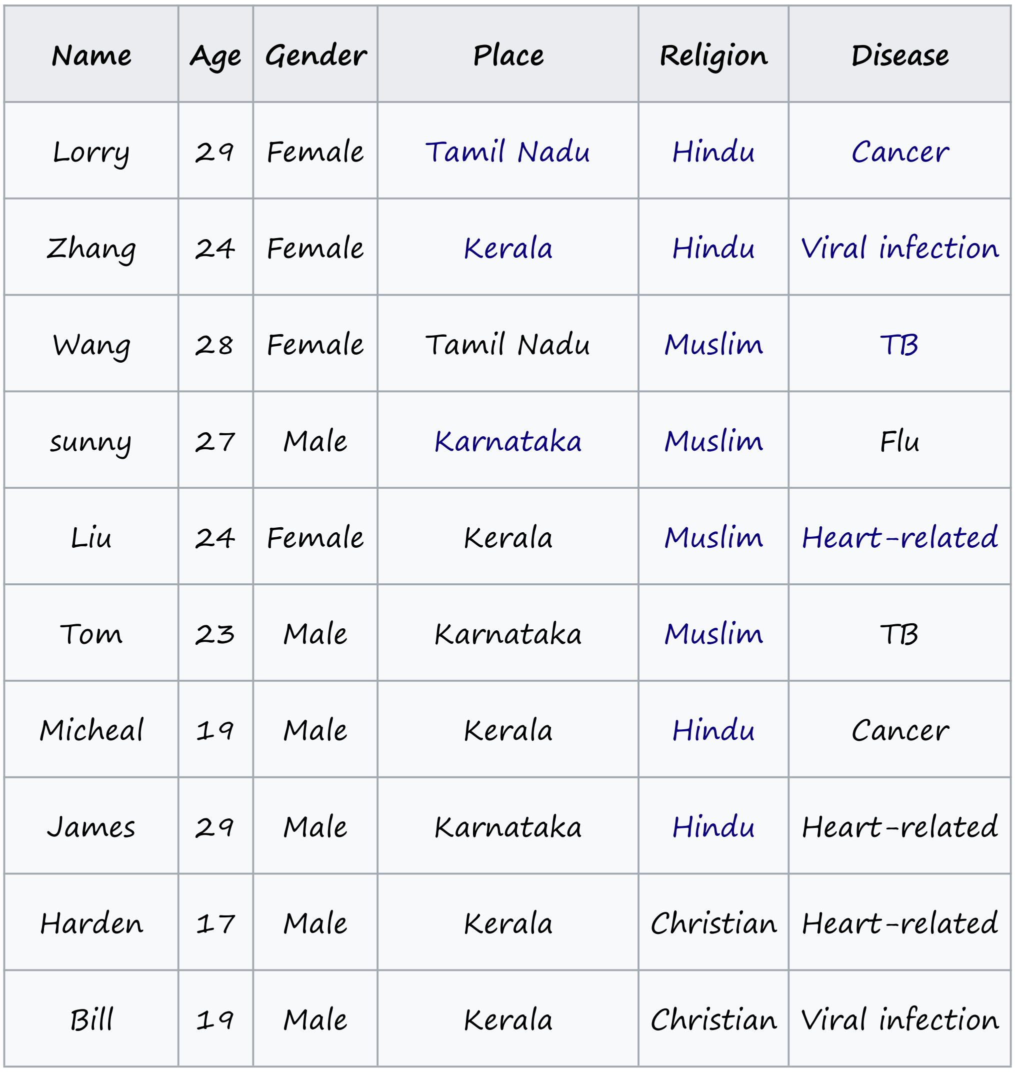
[35]Definition 1. Attributes

B(A1,…,An) is a table with a finite amount of rows. The finite set of attributes of B are {A1,…,An}.

[35]Definition 2. Quasi-identifier

Given a population of entities U, an entity-specific table T(A1,…,An), fc: U ®T and fg: T ® U', where U Í U'. A quasi-identifier of T, written QT, is a set of attributes {Ai,…,Aj} Í {A1,…,An} where: $piÎU such that fg(fc(pi)[QT]) = pi.

As seen in the table 6, there are 6 attributes which are name, age, gender, place, religion and disease respectively. Meanwhile, the table contains 10 patients' detailed data. In order to achieve k-anonymity, suppression and generalization methods which are mentioned before will be applied.

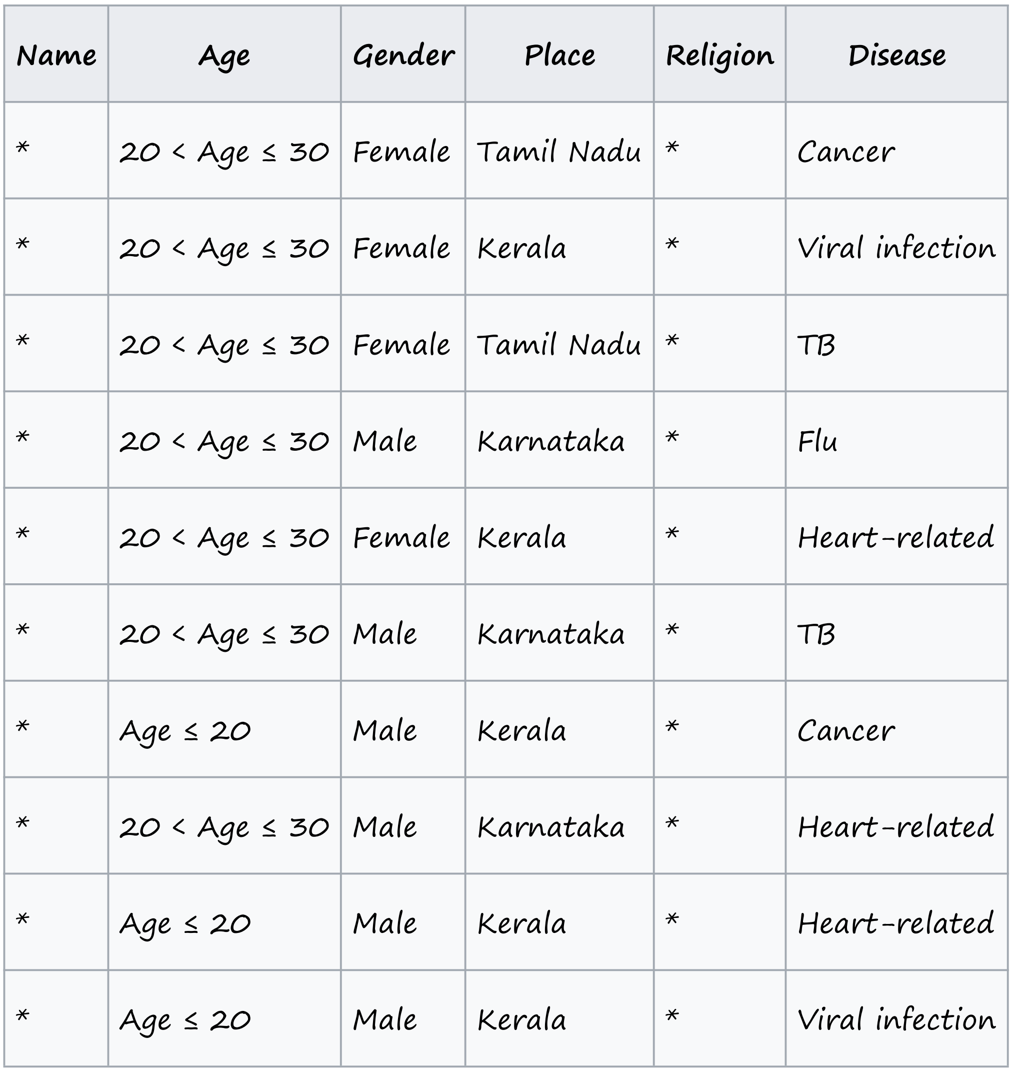


Name	Age	Gender	Place	Religion	Disease
Lorry	29	Female	Tamil Nadu	Hindu	Cancer
Zhang	24	Female	Kerala	Hindu	Viral infection
Wang	28	Female	Tamil Nadu	Muslim	TB
sunny	27	Male	Karnataka	Muslim	Flu
Liu	24	Female	Kerala	Muslim	Heart-related
Tom	23	Male	Karnataka	Muslim	TB
Micheal	19	Male	Kerala	Hindu	Cancer
James	29	Male	Karnataka	Hindu	Heart-related
Harden	17	Male	Kerala	Christian	Heart-related
Bill	19	Male	Kerala	Christian	Viral infection

Table 6. Patients' records in one hospital

Table 7 is an anonymized patients' records based on table 6 and k equals 2. Name and religion attributes are suppressed to the asterisk. Attribute age is generalized to a range.

In the table 7, there are only three attributes left which are age, gender and place. Furthermore, quasi-identifier of table 7 is {age, gender, place}. From table 7, the same quasi-identifier appears at least 2 times.



Name	Age	Gender	Place	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	Flu
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

Table 7. Anonymized patients' records in one hospital

3.2 SOM

3.2.1 Introduction of the SOM

Teuvo Kohonen, a Finnish researcher, proposed the definition of the SOM which is a kind of artificial neural network[38]. The main purpose of the SOM is to decrease the data from multi-dimensional to one or two dimensions. Neurons are the fundamental components that form a SOM. Each neuron is a weight vector which has the same dimension as the input data. A two-dimensional normal spacing in a rectangular grid is the common layout for neurons. Based on the input data, SOM will continuously and automatically find the closest neuron in the neuron layer. Then, the weight of the nearest vector will be changed according to predefined parameters and the distance between the input and nearest weighted vector. The same process will be carried out based on the predefined training times. In the end, a higher-dimensional input space will be mapped to a lower-dimensional space.

[49] displays several advantages and disadvantages of the SOM algorithm. As for advantages, firstly, the data which is processed by SOM algorithm is easier for us to comprehend. SOM can decrease the dimension from high to low rapidly and effectively, which also provides the convenience for us to find the similarities of the big data sets. Secondly, SOM has the ability to deal with different kinds of classification issues if the data summary is helpful and interactive. Thirdly, huge and complicated data sets can be handled easily by the SOM algorithm. Meanwhile, training SOM neurons can be done in a few time period without complex optimization formulas. Moreover, the whole training procedures are simple enough to comprehend and change because of the SOM algorithm's simplicity. However, SOM training process needs enough sample data for the sake of creating a significant map, otherwise the trained SOM may not classify the input data effectively. Data shortage or uncorrelated data may lead to bad effects for clustering. Furthermore, neighbouring neurons should be performed similarly in SOM.

Same as the k-anonymity, SOM algorithm is also applied in different types of areas. WEBSOM[50] project proposed an impressive method which is based on the SOM algorithm. This new method is used for information retrieval. The similar texts are mapped to the SOM closely, just like the similar bowls are placed closely in the kitchen cabinet. Meanwhile, the SOM provides an underlying name of the grouping. If users want to read the detailed information of this grouping, they just need to click the figure using the computer's mouse. While users find an area where they are interested, they can also use arrows to choose nearby areas and similar documents will be found. [51] is another paper about the breast cancer diagnosis based on the SOM algorithm. Breast cancer is the biggest cancer problem for females in developed countries. How to diagnose the benign and malignant tumor effectively and accurately is vital currently. Due to the superiority of the SOM algorithm, a high negative predictive value, 98.5%,can get. The SOM algorithm brings an excellent performance for distinguishing the types of tumors. [52]SOM algorithm can also be applied to predict bankruptcy. SOM algorithm can classify companies as the robust group or bankrupt-prone group. Each weight vector contains input and output vectors, only the input vector will be used for finding the closest unit. However, both input and output vectors are updated during training process. Similar companies are placed in nearby areas of the SOM and common attributes can be acquired easily. Therefore, a new company's attributes can be described reliably according to the mapped location.

3.2.2 SOM algorithm

The detailed algorithm description will be shown as follows:[34]

1) Initialize parameters and randomize neurons' weight vectors

At first, we have to decide the values of original parameters α,σ. Next, we need to randomize the neurons' weight vectors which are located in the Kohonen layers by referring to the input data weight vectors.

2) Find the best matching unit

After finishing the initialization process, we will try to find one neuron which has the closest distance with input data than any other neurons. The closest neuron called the best matching unit. The formula is:

$$c = \arg\min_i \left\| x(t) - w_t(t) \right\| \quad i = 0,1,2...N \tag{1}$$

c is the best matching unit; t is the count of training; x(t) is the input data; $w_i(t)$ is the weight vector of i neuron in current iteration t and N is the total number of neurons.

3) Update the values of the best matching unit and neighboring neurons

Update the neighboring neurons which are closer to the best matching unit and best matching unit itself in order to pull neurons closer to the input data.

$$w_i(t+1) = w_i(t) + h_{c,i}(t)(x - w_i(t)) \tag{2}$$

$$h_{c,i}(t) = \alpha(\mathrm{t})\exp(-\frac{d_{c,i}^2}{2\sigma(\mathrm{t})^2}) \tag{3}$$

Where α(t) is the learning rate; m is the maximum iteration time; $h_{c,i}$ is a parameter that is changed based on time, usually is named as the neighborhood function; $d_{c,i}$ is the distance between the best matching unit and corresponding neuron i; σ(t) is the radius of a neuron. For equation 2 and 3, the neurons that are closer to the best matching unit will be affected to a great extent because of the relationship between $d_{c,i}$ and $h_{c,i}$. Through updating process, it can pull neighboring neurons closer to input data. Step 2 and step 3 will be repeated while t<m.

4 Comparison of some privacy-preserving algorithms for DR purpose

4.1 K-Anonymity in DR

For the smart grid DR, Quasi-identifier is the time serious of one day, which is (0:00-0:30), (0:30-1:00),...,(23:30-0:00). K-Anonymity means that there are at least k identical records in a table. Generalization and masking quasi-identifiers will be used to satisfy K-Anonymization. Table 8 and Table 9 are an simple example to explain k-Anonymity for the DR purpose.

ID	0:00-0:30 AM	0:30-1:00 AM	1:00-1:30 AM	1:30-2:00 AM	2:00-2:30 AM	2:30-3:00 AM	....	7:00-7:30 AM	7:30-8:00 AM	8:00-8:30 AM	......	5:00-5:30 PM	5:30-6:00 PM	....	11:00-11:30 PM	11:30-0:00 PM
1	1.3	1.3	1.3	1.3	1.3	1.3	1.3	2.6	2.4	1.3	1.3	3.4	3.0	2.8	1.3	1.3
2	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2.7	2.9	1.5	1.5	3.6	3.3	2.4	1.5
3	2.4	1.2	1.2	1.2	1.2	1.2	1.2	1.2	1.2	3.6	1.2	1.2	1.2	1.2	3.0	2.8
4	2.5	2.5	1.7	1.7	1.7	1.7	1.7	1.7	2.7	2.6	1.7	1.7	2.3	2.0	2.1	2.2
5	2.54	2.54	2.54	2.54	2.54	2.54	2.54	2.9	2.8	2.3	2.3	2.3	2.3	2.3	2.9	2.52
6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	2.6	3.2	3.4	2.5	2.5	3.5	3.0	2.7	2.6

Table 8. Electricity consumption data for 6 households in 24 hours in housing estate A (unit:kwh)

Table 8 displays the detailed electricity consumption data for 6 households in 24 hours in housing estate A. In order to protect the customers' privacy information, table 8 is anonymized to the format of table 9 and k equals 3. As can be seen from the table 9, household 1, household 2 and household 3 are an anonymized group. Household 4 to household 6 are another anonymized group. For each group, the electricity consumption data in 24 hours is identical. As for traditional time serious k-Anonymity, the way to form a value range Ri=$[r_i^-, r_i^+]$ is that r_i^- is the minimum k record in one specific time serious(one quasi-identifier) and r_i^+ is the maximum k record in the same time serious. For example, the electricity consumption data for the household 1,2,3 between 0:00 to 0:30 is 1.3kwh, 1.5kwh and 2.4kwh respectively. With regard to Ri, r_i^- is the 1.3kwh and r_i^+ is the 2.4kwh. So, the electricity consumption data for household 1,2,3 between 0:00 to 0:30 is anonymized to range(1.3-2.4).

ID	0:00-0:30 AM	0:30-1:00 AM	1:00-1:30 AM	1:30-2:00 AM	2:00-2:30 AM	2:30-3:00 AM	....	7:00-7:30 AM	7:30-8:00 AM	8:00-8:30 AM	......	5:00-5:30 PM	5:30-6:00 PM	....	11:00-11:30 PM	11:30-0:00 PM
1	(1.3-2.4)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-2.6)	(1.2-2.7)	(1.3-3.6)	(1.2-1.5)	(1.2-3.4)	(1.2-3.6)	(1.2-3.3)	(1.3-3.0)	(1.3-2.8)
2	(1.3-2.4)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-2.6)	(1.2-2.7)	(1.3-3.6)	(1.2-1.5)	(1.2-3.4)	(1.2-3.6)	(1.2-3.6)	(1.3-3.0)	(1.3-2.8)
3	(1.3-2.4)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-1.5)	(1.2-2.6)	(1.2-2.7)	(1.3-3.6)	(1.2-1.5)	(1.2-3.4)	(1.2-3.6)	(1.2-3.6)	(1.3-3.0)	(1.3-2.8)
4	(2.5-2.6)	(2.5-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.9)	(2.7-3.2)	(2.3-3.4)	(1.7-2.5)	(1.7-2.5)	(2.3-3.5)	(2.0-3.0)	(2.1-2.9)	(2.2-2.6)
5	(2.5-2.6)	(2.5-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.9)	(2.7-3.2)	(2.3-3.4)	(1.7-2.5)	(1.7-2.5)	(2.3-3.5)	(2.0-3.0)	(2.1-2.9)	(2.2-2.6)
6	(2.5-2.6)	(2.5-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.6)	(1.7-2.9)	(2.7-3.2)	(2.3-3.4)	(1.7-2.5)	(1.7-2.5)	(2.3-3.5)	(2.0-3.0)	(2.1-2.9)	(2.2-2.6)

Table 9. Anonymized electricity consumption data for 6 households in 24 hours in housing estate A (unit:kwh) and k=3

Through the comparison with the table 8 and table 9, k-Anonymity can prevent the customer's electricity consumption data from disclosing. For instance, from the table 8, at least one resident who lives in household 1 waked up between 7:00 and 7:30 am, because the electricity consumption data increasing from 1.3kwh to 2.6 kwh. But for the table 9, the value is in range format from 1.2kwh to 2.6kwh. Even though, we can analyze that there are at least one resident in first group(household 1,2,3) waked up because of data changed from (1.2-1.5) to (1.2-2.6). However, we can not acquire precise information of residents' activities. Meanwhile, the anonymized table increases the difficulty of identification. Nevertheless, DR needs accurate electricity consumption data. The process of k-Anonymity has a tremendous loss of accuracy.

4.2 Platform for sharing power consumption data based on SOM

4.2.1 Electricity consumption sharing platform using SOM

[34] proposes a new platform for DR purpose between different electric power companies. They suppose that more than one electric power companies exist and each company contacts with several households in one territory. Therefore, electricity consumption data sharing is needed when the electricity providers have a huge burden to provide enough electricity. To be honest, the easiest way to share is just sharing the total electricity consumption data because they can get rid of customers' privacy-preserving issues. However, sharing individual electricity consumption data will bring more advantages, for example, the third party can make a detailed, efficient and reasonable DR report if they know the precious electricity consumption data. For a formal DR report, the household who uses more electricity in one period will face a higher reduction rate. For instance, DR report may write, "households which consume over 800 Wh will decrease by 8%, households which consume between 500 and 800 Wh will decrease by 4%, households which consume less than 500 Wh will decrease by 2%." Without the detailed individual electricity consumption data, the precious reduction rate can not be done.

According to the above concept of the electricity consumption data sharing platform, the electricity power companies will make use of the SOM as a common data sharing framework. For each company, they just need to map their raw data to the SOM without the disclosure of raw data. N is the number of the electricity power companies; ep_i is the i th electricity power company(i=1,2,3...N); the electricity consumption data of one household which contracts with one company will be shown as follows:

$$H_{ep_i,j} = (h_{ep_i,j,00:00} \ , h_{ep_i,j,00:30} ,\ldots\ldots, h_{ep_i,j,23:30})$$

$$\mathrm{j}=1,2,3...,n_{ep_i}$$

Where $H_{ep_i,j}$ means that one day electricity consumption data with the half hour interval of the j household and j household has contracted with the electricity power company ep_i. $h_{ep_i,j,00:00}$ denotes the electricity consumption data between 00:00 and 00:30 of the j household which contracts with the electricity power company ep_i. In addition, n_{ep_i} is the number of households which have contracted with the electricity power company ep_i. The above model will be used to explain the proposed platform.

Algorithm 1 and Fig 17 display the electricity consumption data sharing platform using SOM. Meanwhile, the specific explanation will be described as follows:

1) Initialize

At first, we will initialize parameters α,σ of the SOM as we did in chapter 3.2.1. Neurons' weight vectors are also needed to be decided by referring to the input data pattern. All of those initialization processes are executed by electricity power companies before the training step.

Algorithm 1 Data sharing process based on SOM

```
#Initialize
Initialize(W)
#Train SOM
for t in range(m):
    for i in range(N):
        for j in range($n_{ep_i}$):
            C=find best matching unit($H_{ep,j}$,W),
            Update the neuron(W)
        Pass W to the next ep
#Map and Count
for i in range(N):
    for j in range($n_{ep_i}$):
        C=find best matching unit($H_{ep,j}$,W),
        Count[c]+=1
    Pass W to the next ep
Publish(W,count)
```

2) Train SOM

In the beginning, ep_1, the first electricity power company, inputs first electricity consumption data $H_{ep_1,1}$ to SOM and finds the best matching unit for $H_{ep_1,1}$ through the distance. Then, the values of neuron's weight vector will be updated by equations (2) and (3). Next, $H_{ep_1,2}$ will be inputted and does the same manner until the last data in ep_1 is been inputted. For the next step, trained SOM is passed to the second electricity power company ep_2 and the same procedure will be conducted until the last company ep_N. While one loop is finished, SOM is passed to ep_1 again and the second iteration starts. The whole training process is completed when the iteration parameter t equals the maximum value m which is mentioned in 3.2.1.

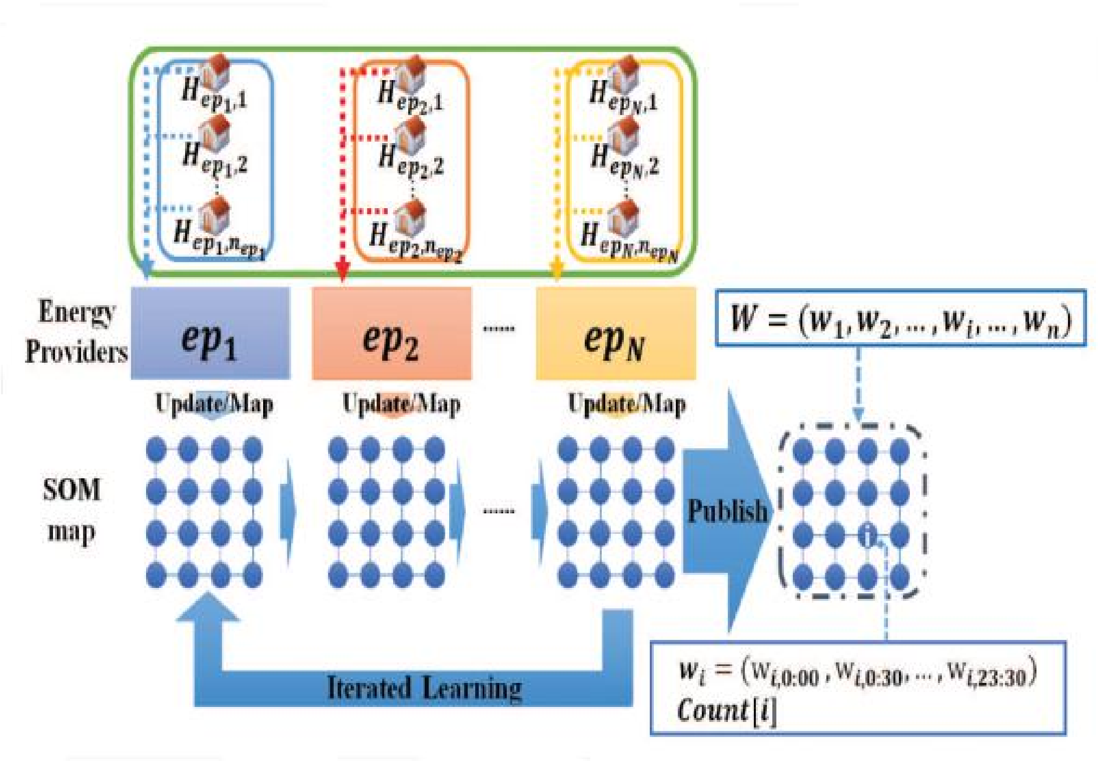


Fig 17. Data sharing between electricity power companies using SOM

3) Map data

The trained SOM is used for the mapping purpose in this step. Each electricity consumption data $H_{ep_1,j}$ which belongs to ep_1 will find its own best matching unit and is mapped to the corresponding neuron. However, the neurons' weight vector values keep the same in the whole map step. Furthermore, SOM does not save any raw electricity consumption data. Afterwards, the number of $H_{ep_i,j}$ data which is mapped to the neuron i will be counted and saved in the SOM for calculating the total amount of electricity consumption value in one territory. For example, neuron 1 counts 3 times, neuron 2 counts 2 times, neuron 3 counts 4 times and neuron 4 counts 6 times. The total electricity consumption value as follows:

Total electricity consumption=neuron1*3 + neuron2*2 + neuron3*4 + neuron4*6

To be honest, training SOM is a time-consuming and burdensome computation work to some extent. If SOM will be used in a momentary task or computation ability is restrained because of the performance of a computer, the mapping task can not be fulfilled successfully and preciously. So, creating a SOM using the data a day before to map the next day's electricity consumption data may be a feasible approach, which will reduce the computational burden to a great extent. But the error rate may increase compared to use the intraday electricity consumption data. However, training SOM in advance using previous data can bring another advantage, for instance, it will increase the degree of privacy-preserving because training data and mapping data are not the same.

4.2.2 Electricity consumption data collection and evaluation

In order to evaluate the data sharing framework which is proposed above, We collected 400 households' one day electricity consumption data in China from data repository. The electricity consumption data format is recorded as follows:

$$H_{ep_i,j}=(h_{ep_i,j,00:00} \ , h_{ep_i,j,00:30} ,......, h_{ep_i,j,23:30})$$

Which means that the electricity consumption data transmission interval was half an hour.

Evaluation of the data sharing framework is based on two parameters:

1. Accuracy

The original total amount of 400 household's one day electricity consumption data are calculated. Meanwhile, mapped neurons and corresponding times are aggregated in order to acquire the total electricity consumption value. The definition of accuracy is defined as follows:

Accuracy=total amount(SOM) / original total amount (1)

The concentrate values of 400 households in one area are described as in table 10.

total amount(SOM) unit:Kwh	original total amount unit:Kwh
2083.26899723452	2142.1194521039724

Table 10.Comparison between SOM and original total amount of electricity consumption value

Through the above values, the accuracy equals 97.25% which is a relatively satisfactory value.

2. Entropy

Entropy[39] is a parameter to measure the disorder of a system. In 1948, Shannon published a paper "A Mathematical Theory of Communication" and proposed the information entropy theory. Shannon pointed out that any information contains the redundancy and the size of the redundancy is associated with the probability of occurrence or uncertainty of every symbol (number, letter, or word) in the message. Shannon borrowed from the concept of thermodynamics, the average amount of information that excludes the redundancy called "information entropy". In other words, information entropy is the probability of occurrence of discrete random events. The more orderly a system is, the lower value the entropy of information has. On the contrary, the more chaotic a system is, the higher value the information entropy has. Therefore, the information entropy can also be a measure of the degree of systematic order. The information entropy's formula is as follows:

$$H(u) = -\sum_{i=1}^{n} p_i \log_b p_i \qquad (2)$$

Where H(u) is the value of information entropy; p_i is the probability of the number i; b is the base of the logarithm. In general, the value of b equals 2. When the probability of i equals 0, the summand $0\log_b(0)$ is treated as 0 because of the theorem $\lim_{p\to 0+} p\log(p) = 0$.

For example, there are 4 basketball teams, the corresponding probability to win the champion is 0.35, 0.15, 0.4, 0.1. The information entropy is:

H= -(0.35*log(2,0.35) + 0.15 * log(2,0.15)+0.4* log(2,0.4)+0.1 * log(2,0.1))

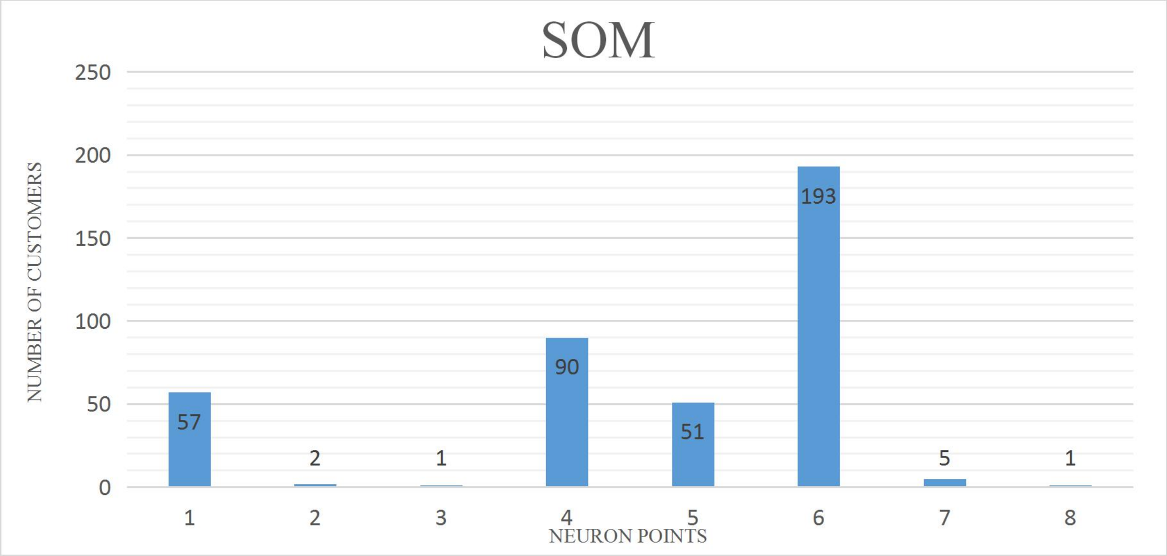


Fig 18. Mapped neurons and corresponding counts of 400 households based on SOM

According to the Fig 18, the value of entropy is 40.239295 which is a higher value which means that the system is chaotic to some extent. As for neuron 2, 3 and 8, there are only less than 4 households mapped. From privacy-preserving aspect, those households are easier to be recognized compared to others. For sake of improving the degree of privacy-preserving, an improved framework is proposed in the next chapter 4.3.

4.3 Platform for sharing power consumption data based on SOM and K-Means

4.3.1 Introduction of K-Means

K-Means[40] is a kind of vector quantization method which is primary proposed in signal processing area and it also widely used in the data mining currently. The purpose of k-means clustering is to divide n vectors into k clusters where each vector is part of the cluster with the nearest distance. The Euclidean squared distance measure is used to calculate the distance between each vector and cluster centers. The detailed algorithmic steps for k-means are shown as follows[41]:

X = {x1,x2,x3,……..,xn} is the serious of points and V = {v1,v2,…….,vc} is the serious of k-means clustering centers.

1. Select the number of c cluster centers randomly.

2. Measure the distance between every points and cluster centers using the Euclidean squared distance measure.

3. Partition each point to the corresponding cluster center whose distance from center is minimum.

4. Calculate each cluster center again using the following formula:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i \qquad (1)$$

v_i is the recalculated cluster center; c_i is the amount of points in cluster i; x_i is the serious of points belong to cluster i.

5. Measure the distance between each point and new acquired cluster centers and do the step 3 again.

6. If all points belong to the same cluster center without repartition, let us terminate the whole process, if not then repeat from the step 4 again.

4.3.2 Electricity consumption data sharing platform using SOM and k-means

Algorithm 2 illustrates the detailed steps of the data sharing process based on the SOM and k-means. From step 1 to step 4, it is the same process as the algorithm 1 in order to train the SOM. Then instead of map and count steps, We use k-means to refine the SOM neuron network's weight vectors for the sake of decreasing entropy value. After that, map and count steps will be conducted as algorithm 1.

```
Algorithm 2 Data sharing process based on SOM and K-Means

1. Initialize weights of SOM
2. Input training set
3. Adjust the weight of SOM
4. SOM finished weight adjustment
5. Choose pre-defined numbers of cluster centers randomly
6. Use K-Means algorithm to refine the weights of SOM
7. Map electricity consumption data to refined neuron network
```

4.3.3 Evaluation of the SOM with k-means platform

We use the same data repository, 400 households' one day electricity consumption data in China, to evaluate the data sharing platform based on SOM and k-means. Meanwhile, efficiency and entropy are 2 parameters which are used to evaluate.

1. Accuracy

total amount(SOM and k-means) unit:Kwh	original total amount unit:Kwh
2083.2689972294406	2142.1194521039724

Table 11. comparison between SOM with k-means and original total amount of electricity consumption value

Through the above data, the accuracy equals 97.25%. Therefore, the platform using the SOM algorithm only or the SOM with k-means algorithm has the same accuracy.

Entropy

According to Fig 19, the entropy is 11.483019881650677 which is a lower value compared to the value 40.23929593357644. There is only one neuron, neuron 4, which has less than 10 mapped households.

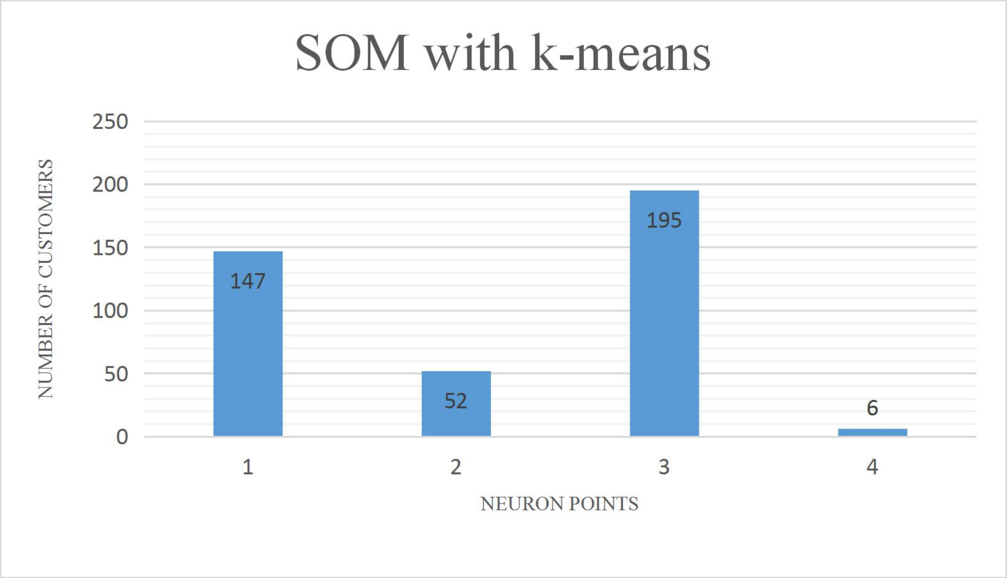


Fig 19. Mapped neurons and corresponding counts of 400 households in one area based on SOM and k-means

4.4 Contribution

In this paper, [34] proposed a feasible platform using the SOM algorithm for the DR. However, according to my experiment, the accuracy of this platform can achieve 97.25% which is a relatively satisfactory value. At the same time, the information entropy theory is also used to measure the safety performance of this platform. According to the Fig 18, for neuron 2, 3 and 8, there are only less than 4 households mapped, which means that the electricity consumption raw data may disclose. Also, through the calculation of the information entropy, the entropy is more than 40 which means that the difference between each neuron’s value is large. In order to increase the safety performance of this platform, clustering algorithm is applied to solve the problem. K-means algorithm is one of the most famous clustering algorithms.

Therefore, the combination between SOM and k-means algorithms may deal the problem. Through the experiment, the new platform performs excellently. In the case of the same accuracy, the information entropy decrease from more than 40 to near 11 which means that the degree of systematic order reduces obviously. Through the Fig 19, only neuron 4 is less than 10. In general, the main contribution is the usage of the k-means clustering algorithm to increase the customers' privacy-preserving.

5 Conclusion and future plan

Four Japanese researchers proposed a privacy-preserving data sharing framework that is based on SOM for DR purpose. The different electricity power companies can utilize this framework to share the customers' electricity consumption data without the raw data published in one area. Meanwhile, the accuracy achieves around 97% which is an ideal value. However through the experiment, another parameter,entropy, is higher. It means that several neurons only corresponding few mapped households than others. From privacy-preserving perspective, a disordered distribution may damage to the customers' information security. In order to decrease the entropy value, a new data sharing framework was proposed based on SOM and k-means algorithm which is a method for vector clustering. Through the refinement of trained neurons, the entropy value decreases obviously without accuracy loss. In conclusion, the new framework increases the security of data sharing process. At the same time, customers' privacy may not disclose easily.

K-means is only one of many clustering methods and performs well in this paper. However, there are also some other well-known clustering algorithms such as Fuzzy C-Means Clustering[53][54], Expectation-Maximization[55]. In the future, Fuzzy C-Means Clustering and Expectation-Maximization algorithms will be used with SOM. Meanwhile, the performance of these algorithms will also be measured.

Reference

[1]. M. Rathmair and J. Haase, "Load Identification and Management Framework for Private Households,"Proceedings of 39th Annual Conference of the IEEE Industrial Electronics Society 2013, pp. 5729-5734

[2]. M. Zeifman, M., "Nonintrusive appliance load monitoring (NIALM) for energy control in residential buildings" In International Conference on Energy Efficiency in Domestic Appliances and Lighting, EEDAL, 2011

[3]. S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting consumer privacy from electric load monitoring," in Proceedings of the 18th ACM conference on Computer and communications security - CCS '11, 2011, pp. 87-98

[4]. National Institute of Standards and Technology, NIST IR 7628 Guideline for Smart Grid Cyber Security, Vol. 2 Security Architecture and Security Requirement, Draft, July 2010

[5]"Origins and Evolution of the Electric Grid". Visited on October 27, 2016. http://www.brooksidestrategies.com/resources/origins-and-evolution-of-the-electric-grid/.

[6]"Smart Grids in Distribution Networks". Visited on October 27, 2016. https://www.iea.org/publications/freepublications/publication/TechnologyRoadmapHow2GuideforSmartGridsinDistributionNetworks.pdf.

[7]He, D., Chen, C., Bu, J., Chan, S., Zhang, Y., & Guizani, M. (2012). Secure service provision in smart grid communications. IEEE Communications Magazine, 50(8), 53–61. http://doi.org/10.1109/MCOM.2012.6257527

[8] NISTNIST, "Framework and Roadmap for Smart Grid Interoperability Standards," Release 1.0, NIST Special Publication 1108, Jan. 2010

[9]Gerwen, R., Jaarsma, S., & Wilhite, R. (2006). Smart Metering. Leonardo Energy, 1(July), 1–9. http://doi.org/10.1016/j.joca.2008.06.011

[10]"what is a smart meter". Visited on October 29, 2016. https://www.smartenergygb.org/en/about-smart-meters/what-is-a-smart-meter

[11]Zhang, Y., Wang, L., Sun, W., Ii, R. C. G., & Alam, M. (2011). Distributed Intrusion Detection System in a Multi-Layer Network Architecture of Smart Grids. IEEE Transactions on Smart Grid, 2(4), 796–808. http://doi.org/10.1109/TSG.2011.2159818

[12]Kuzlu, M., Pipattanasomporn, M., & Rahman, S. (2014). Communication network requirements for major smart grid applications in HAN, NAN and WAN. Computer Networks, 67, 74–88. http://doi.org/10.1016/j.comnet.2014.03.029

[13]Nicanfar, H., Talebifard, P., Alasaad, A., & Leung, V. C. M. (2013). Privacy-preserving scheme in smart grid communication using enhanced network coding. IEEE International Conference on Communications, 2022–2026. http://doi.org/10.1109/ICC.2013.6654822

[14]Application, F., Data, P., & Group, P. E. (1995). United States Patent [1 9] [11] Patent Number : [45] Date of Patent :, 3–6. http://doi.org/10.1074 /JBC.274.42.30033.(51)

[15]Roth, Hans (March 2011). Das erste vierrädrige Elektroauto der Welt [The first four-wheeled electric car in the world] (in German). pp. 2–3.

[16] Guarnieri, M. (2012). "Looking back to electric cars". Proc. HISTELCON 2012 - 3rd Region-8 IEEE HISTory of Electro - Technology CONference: The Origins of Electrotechnologies: #6487583.

[17]Sperling, Daniel; Gordon, Deborah (2009). Two billion cars: driving towar d sustainability. Oxford University Press. pp. 22–26. ISBN 978-0-19-537664-7.

[18] David B. Sandalow, ed. (2009). Plug-In Electric Vehicles: What Role for Washington? (1st. ed.). The Brookings Institution. pp. 1–6. ISBN 978-0-8157-

0305-1.

[19]"Electro Automotive: FAQ on Electric Car Efficiency & Pollution". Electro auto.com. Retrieved 2010-04-18.

[20]" China car sales 'overtook the US' in 2009". Visited on November 6, 2016. http://news.bbc.co.uk/2/hi/8451887.stm

[21]"Freidman OpEd: China's 'Moon Shot' Versus America's". http://evworld.co m/news.cfm?newsid=2405

[22]Bradsher, Keith (2009-04-02). "China Vies to Be World's Leader in Electric Cars". The New York Times.

[23]"China electric vehicles to hit 1 million by 2020: report". Reuters. 2010-10-16. Retrieved 2011-05-02.

[24]Liza, Lin. "China to Subsidize Alternative Energy Car Purchases", Bloomberg, China, 6 January 2010.

[25]China Association of Automobile Manufacturers (2012-01-16). "5,579 electric cars sold in China in 2011". Wind Energy and Electric Vehicle Review. Retrieved 2014-01-12.

[26]Liu Wanxiang (2016-10-12). "Automobile Association: slowdown ends, new energy vehicle sales in September rose to 44 000"(in Chinese). D1EV.com. Retrieved 2016-10-12. Sales of new energy vehicles totaled 44,000 units in September 2016, consisting of 35,000 all-electric vehicles and 9,000 plug-in hybrids. Total car sales during the first nine months of 2016 totaled 19,360,000 units.

[27]"EV sales growth, production slow down". China Daily. China.org.cn. 2016-10-17. Retrieved 2016-10-25.

[28]Xiaoping, J., Keyi, J., & Bo, W. (n.d.). Electric Vehicles and Charging Networks in China.

[29]"Marketing Information". smart grid tec – china. 2010. Retrieved 9 April 2012.

[30]"China's largest electric car charging station opens in Beijing". xinhuannet. 2012. Retrieved 9 April 2012.

[31]"China Electric Vehicle Charging Station Market Report,2010". Research in China. 2010. Retrieved 10 April 2012.

[32]"9 Ways to Store Energy on the Grid". Visited on December 3, 2016. http://discovermagazine.com/2015/july-aug/26-power-stash

[33]"Understanding the Benefits of the Smart Grid". Visited on December 3, 2 016.https://www.netl.doe.gov/File%20Library/research/energy%20efficiency/smart%20grid/whitepapers/06-18-2010_Understanding-Smart-Grid-Benefits.pdf.

[34]Okada, K., Matsui, K., Haase, J., & Nishi, H. (2015). Privacy-preserving Data Collection for Demand Response using Self-organizing Map. Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on.

[35]L. Sweeney, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5. pp. 557-570, 2002.

[36]A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,"l-D iversity: Privacy Beyond k-Anonymity," ACM Transactions onKnowledge Discov ery, vol. 1, no. 1, 2007.

[37]N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in 2007 IEEE 23rd International Conference on Data Engineering, 2007, pp. 106-115.

[38]T. Kohonen, "The self-organizing map," Proc. IEEE, vol. 78,no.9, pp. 1464-1480, 1990.

[39]C. E. Shannon, "A mathematical theory of communication," Bell Syst.Tech. J., vol. 27, pp. 379-423, July, 1948.

[40]Hartigan, J. A., and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, 1979, pp. 100–108. www.jstor.org/stable/2346830.

[41]"k-means clustering algorithm". Visited on January 11,2017. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[42]Aggarwal, C. C. (2005). On K-anonymity and the Curse of Dimensionality. Proceedings of the 31st International Conference on Very Large Data Bases, 901–909. Retrieved from http://dl.acm.org/citation.cfm?id=1083592.1083696

[43]Sweeney, L. (2002). Achieving k-anonymity Privacy Protection using Gener alization and Suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 571–588. http://doi.org/10.1142/S021848850200 165X

[44] Angiuli, Olivia; Joe Blitzstein; Jim Waldo. "How to De-Identify Your Data". ACM Queue. ACM.

[45] Angiuli, Olivia; Jim Waldo (June 2016). "Statistical Tradeoffs between Generalization and Suppression in the De-Identification of Large-Scale Data Sets". IEEE Computer Society Intl Conference on Computers, Software, and Applications.

[46]Friedman, A., Wolff, R., & Schuster, A. (2008). Providing k-anonymity in data mining. VLDB Journal, 17(4), 789–804. http://doi.org/10.1007/s00778-006-0039-5

[47]Shang, X., Chen, K., Shou, L., Chen, G., & Hu, T. (2010). (k , P) -Anonymity : Towards Pattern-Preserving Anonymity of Time-Series Data. CIKM '10 Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 1333–1336.

[48]Gedik, B., & Liu, L. (2004). A Customizable k-Anonymity Model for Protecting Location Privacy, 620–629. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.6338

[49]"Self-organizing Maps". Visited on January 25,2017. https://www.cs.hmc.edu/~kpang/nn/som.html

[50]Timo Honkela, Samuel Kaski, Krista Lagus & Teuvo Kohonen. WEBSOM-S elf-Organizing Maps of Document Collections, Finland.

[51]Chen, D., Chang, R., & Huang, Y. (2000). Breast cancer diagnosis using s elf-organizing map for sonography. Ultrasound in Medicine & Biology, 26(3), 405–411. http://linkinghub.elsevier.com/retrieve/pii/S0301562999001568

[52]Kiviluoto, K. (1998). Predicting Bankrupticies with the Self-Organizing Map. Neurocomputing, 21(1), 191–201.

[53]J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57.

[54]J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York.

[55]A.P. Dempster, N.M. Laird, and D.B. Rubin (1977): "Maximum Likelihood from Incomplete Data via theEM algorithm", Journal of the Royal Statistical Society, Series B, vol. 39, 1:1-38.