Hindawi
Journal of Electrical and Computer Engineering
Volume 2017, Article ID 2970673, 10 pages
https://doi.org/10.1155/2017/2970673

Research Article

The Anonymization Protection Algorithm Based on Fuzzy Clustering for the Ego of Data in the Internet of Things

Mingshan Xie,[1,2,3] Mengxing Huang,[1,2] Yong Bai,[1,2] and Zhuhua Hu[1,2]

[1]*State Key Laboratory of Marine Resource Utilization in South China Sea, Haikou 570228, China*
[2]*College of Information Science & Technology, Hainan University, Haikou 570228, China*
[3]*College of Network, Haikou College of Economics, Haikou 571127, China*

Correspondence should be addressed to Mengxing Huang; huangmx09@163.com

Received 10 February 2017; Accepted 5 March 2017; Published 8 June 2017

Academic Editor: Jit S. Mandeep

In order to enhance the enthusiasm of the data provider in the process of data interaction and improve the adequacy of data interaction, we put forward the concept of the ego of data and then analyzed the characteristics of the ego of data in the Internet of Things (IOT) in this paper. We implement two steps of data clustering for the Internet of things; the first step is the spatial location of adjacent fuzzy clustering, and the second step is the sampling time fuzzy clustering. Equivalent classes can be obtained through the two steps. In this way we can make the data with layout characteristics to be classified into different equivalent classes, so that the specific location information of the data can be obscured, the layout characteristics of tags are eliminated, and ultimately anonymization protection would be achieved. The experimental results show that the proposed algorithm can greatly improve the efficiency of protection of the data in the interaction with others in the incompletely open manner, without reducing the quality of anonymization and enhancing the information loss. The anonymization data set generated by this method has better data availability, and this algorithm can effectively improve the security of data exchange.

1. Introduction

In the information age, many things such as software, websites, and Internet of things are providing data, and there are also various interactive processes of data, for example, collaborative analysis of data, classification of data, integration of heterogeneous data, big data analysis, and trading. But in these interactions of the data many units or agencies are reluctant to open or publish their data collection. In fact, the data can be published after the appropriate treatment and can also play a positive role in our data analysis and mining. Data on the one hand need to be open and, on the other hand, need to be conservative. How to adjust? We need to recognize the ego characteristics of the data. We need to pay attention to the ego characteristics of the data. If we do not pay attention to it, the data sovereignty is not clear. Many disputes will be triggered. The insufficient cooperation will arise in data collaboration.

The ego of the data in the IOT has its own characteristics. It has the sensitivity of time and space and contains a lot of privacy information. In the process of data interaction, in order to achieve the common goal, the data need to share some information, but some sensitive information cannot be exposed. We should fully understand and respect the ego of data.

This paper is structured as follows. In Section 2, we present related work. Section 3 introduces the definition of the ego of data. In Section 4, we introduce the characteristics of the ego of data in the IOT. Section 5 provides the concepts involved in the anonymization protection of incompletely open cooperation for the ego of data in the IOT. Section 6 provides the design and analysis of an anonymization protection algorithm in the data interaction process in an incompletely open manner for the ego of data in the IOT. In Section 7 we have the experiment to validate the algorithm. We conclude our work and lay out future research in Section 8.

2. Related Work

The protection of data, especially the protection of privacy information, has been studied by experts and scholars. The K-anonymous model is proposed by [1], which requires that each data in the data set has at least k-1 data that can not be distinguished from the quasi identifier. According to the literature [2], the k-anonymity model can not resist the attacks of homogeneous attacks and background knowledge. The literature [3] proposed a l-diversity model, which requires that the sensitive values in the cluster have l different values. The (α, k)-anonymization model has been proposed in [4] to realize the diversity of the sensitive value, so as to improve the security of the data. The k-anonymous method based on clustering is proposed in [5].

The method of anonymization is often used in data privacy protection. The literature [6] proposed an anonymous proof protocol based on property certificate. By the trusted ring signature scheme based on property certificate, the authors have achieved the anonymity of computing nodes and prevented the leakage of platform configuration information. Muftic et al. have combined this method with the characteristics of business data to build the business information exchange system with security, privacy, and anonymity which provides innovative features of privacy and full anonymity of users in the paper [7].

Data privacy information protection has accumulated a lot of scholars' work in [8–10]. However, with the explosion of data and the arrival of the era of big data, the interaction between the data is more and more frequent. The coordination between the openness and the protection of data has attracted more and more attention. For the Internet of things, this problem is particularly prominent. The data of the Internet of things has the characteristics of mass and spatiotemporal sensitivity. Literature [11] introduced the data characteristics of the Internet of things. A lot of people's privacy information exists in the Internet of things. Each of its records has the specific location and time attributes that represent the place at which the tag of the Internet of things is located at a certain moment. In addition, the layout of several objects adjacent to each other, which has certain stability, makes it easier for the IOT to leak privacy information. There are few studies on the protection of data for the Internet of things. Reference [12] studied the privacy information protection of the Internet of things and proposed the concept of the data set distribution sequence to optimize the generation of cluster seeds. The authors studied the problem of privacy information protection in the IOT and proposed the concept of data set distribution sequence to optimize the generation of cluster seeds. They clustered the data in parallel so that the equivalent class contains the data of multiple nodes; thus the specific location information of the data can be blurred, and the layout characteristics of tags could be eliminated. A privacy preserving k-anonymous algorithm has been designed for the Internet of things.

The special research on the data protection for Internet of things is not too much to adapt to the requirements of the times. In the paper [13] k-anonymity notion and a method based on bottom-up clustering have been adopted to be used in wireless sensor networks (WSN) as a security framework with two levels of privacy. Samani et al. have modeled the privacy concepts and concerns in the IOT and proposed a privacy protection management framework for CDS at the interaction level in [14]. The application of the framework has been demonstrated by extending Contract Net Protocol (CNP) to support privacy protection for CDS.

The literature [12] is only for the privacy protection of the IOT. However, the research on the issue of privacy protection extension to the coordination problem between data openness and protection of the Internet of things is very low. In the process of data interaction in the Internet of things, on the one hand, we need to protect their privacy; on the other hand, we need to open their cooperation with other data. This paper is aimed at this problem, puts forward the concept of the ego of data, and then analyzes the characteristics of the ego of data in the Internet of things. The two parameters of acceptable node set and acceptable sampling period are given. We will build an incompletely open anonymous protection model for the ego of data which is suitable for the Internet of things environment. In this paper, anonymity protection algorithm based on the fuzzy clustering for the data in the IOT eliminates the layout characteristics of the Internet of things label in order to improve the security of the data and ensure the openness of the data.

3. Models and Definitions

3.1. The Ego of Data Definition. The idea of the ego of data is derived from the observation of data aggregation in the Internet of things: when the data from a variety of Internet of things aggregated to the cloud platform, it was always difficult to deal with the problem of balance between data opening and data protection for some data providers. Some of the data, in fact, did not affect the data provider but was completely shielded. In another case, some of data should be completely removed, because it was useless to the cloud platform for the Internet of things but has been occupying the transmission channel and storage space.

The ego of data characterizes the fact that the data assesses its own value and importance, requires a certain degree of protection of its own and restrictions on opening, and pursues the best balance between data opening and privacy protection in the process of data interaction.

Definition 1 (the ego of data). It is the sense of the value and importance of data. It quantifies the value of data using the method of maximum approximation and evaluates the effect of using the data. Let $E(d)$ represent the numerical expression of the ego of data. There is the following formula:

$$E(d) = \max\left(\|M(D) - M(D \ominus d)\|_q\right), \quad (1)$$

where d represents the concerned data. D represents the data set that contained d. $D \ominus d$ means the data set in which d is removed. M is the quantitative impact calculated by the algorithm; $\|M(D) - M(D \ominus d)\|_q$ denotes the q order norm distance of $M(D)$ and $M(D \ominus d)$.

The connection and difference between the ego of data and several concepts are shown in Table 1.

Table 1: The connection and difference between the ego of data and several concepts.

Concepts	Points of focus
Data sovereignty	Researching on the ownership of data
The value of data	Researching on the usefulness of data
The ego of data	Focusing on the study of the degree of data coordination, in the process of data interaction, including the estimation of the value of data, which is more extensive than the concept of data privacy
Data privacy	Focusing on data confidentiality and precautions

3.2. Characteristics of the Ego of Data

Interactivity. The ego of data reflects the ability of data to absorb other data or open itself to accomplish a specific work.

Various Forms. There are text, pictures, and other forms of data. In addition, there are all kinds of data structures.

Complex Relation of Subject. Each data has its own actors. The value and importance of data are determined by the behavior subject of the data. The complex relationship of the behavior subject of the data makes the data have the ego. The three main behavior subjects are as follows. The first is the owner of the data, namely, data collection platform owners. The second is the data producers who are concerned with and studied by the owner of the data collection platform. The third is the data user who can tap the value of the data. There are times when there is three-behavior subject unity, sometimes two-behavior subject unity, and sometimes the separation of three-behavior subject unity. Now, medical data can be an example to illustrate their relationship. The producers of the data are the patients, the users of the data are doctors, and the owner of the data is the data collection platform builder who is always a medical institution. In addition, farm data also can be an instance, the producer of the data are the crop, the user of the data and the owner of the data are farm managers. It is a case of two-behavior subject unity.

Nonintuitive Value. The value of data is not intuitive, which can be reflected by the data mining and data-processing technology.

Value Variability. The value of data is difficult to measure and can only be approximated to the real value. The number that can reflect the value of data is always varied with time and scene.

Different Domain. The value of data is varied with the users from different domains. The same data is very important for some of the actors, while it is not useful for any other actors.

Various Sensitive Information. Due to the complexity of the relationship of the three-behavior subject, the sensitivity of the data is not the same and the sensitivity of the record is not the same.

Uniqueness of Data Set Distribution. Different data sets have different distribution patterns. There are two forms of attribute distribution and record distribution.

3.3. Data Interactive Mode Based on the Ego of Data.

The research content of the ego of data is mainly focused on the interaction mechanism of the data and the adjustment of the algorithm complexity. The research methods of the ego of data are mainly a variety of quantitative analysis methods based on artificial intelligence technology and computational theory.

Different modes of cooperation are adopted by the ego of data, according to the evaluation of its own value, in cooperation with other data. There are two types of data cooperation mode.

The first type is to publish its own data. The ego of data provides its own data to other data users to complete a task. The second type is to absorb other data. Some data, because they are not complete, interact with other data in order to supplement and improve and modify its own data. Each type has the following data open modes.

Totally Enclosed. The data is highly encrypted, so that it is unable to extract useful information in the process of the data mining. The most extreme case of this mode was that the data was not provided at all; that is, the data was deleted completely.

Incompletely Open. Only part of the information is published, in order to protect the sensitive information of the data. At the same time, the degree of influence of the protection method on the cooperation result is controlled in a certain range. The sensitive information can not correspond to the specific behavior subject.

Completely Open. Useful data information is readily available. The most extreme case of completely open mode is to provide raw data.

Definition 2 (the degree of opening). Set function $f : d \to I$. The data d is inputted; the useful information I can be output. Then

$$\mathrm{OP}(d) = \frac{f(d)}{d}. \tag{2}$$

$\mathrm{OP}(d)$ is known as the degree of openness of data d.

Both the totally enclosed mode and the completely open mode are the special cases of the incompletely open mode. When $\mathrm{OP}(d) = 1$, the data d is completely open. $\mathrm{OP}(d) \simeq 0$ indicates that the data d is close to the totally enclosed mode. $0 < \mathrm{OP}(d) < 1$ indicates that the data d take the incompletely open mode.

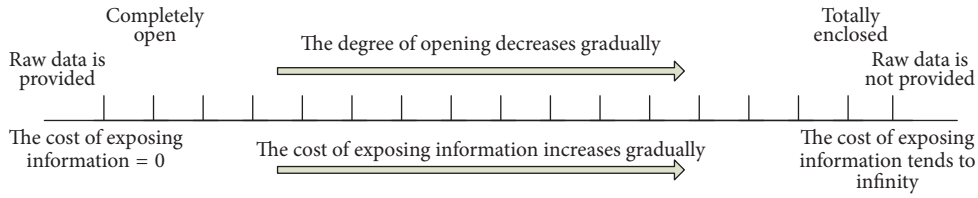

Figure 1: The relationship between OP(d) and COST(d).

The data that is incompletely open is evaluated. The lower the opening degree OP(d) is, the higher the requirement of the protection function and the cost would be. The relationship between OP(d) and COST(d) is shown in Figure 1.

Definition 3 (sensitive information). Some attributes or records that cannot be opened are always important. The range and sensitivity of sensitive information are determined by the behavior subject of the ego of data. Different sensitivity levels are set according to the degree of openness of sensitive information.

When the sensitive information is included in the data set, the interaction mode of the data set can be the mode of incompletely open.

The cost of data protection is paid in order to protect the sensitive information. The cost of exposing the sensitive information also has to be paid. "Unable to decrypt" indicates that the cost of decryption tends to infinity.

Definition 4 (the cost of exposing information). The cost was made up by the computing time and computing complexity of the data mining algorithm used to obtain useful information from the processed data. That is,

$$\text{COST}(d) = \begin{cases} 0, & \text{completely open} \\ \max \text{cost}(d) \times (1 - \text{OP}(d)), & \text{incompletely open} \\ \max \text{cost}(d), & \text{totally enclosed.} \end{cases} \quad (3)$$

COST(d) becomes the cost of exposing information exposure costs for the ego of data, where $\max \text{cost}(d)$ represents the maximum cost of exposing information. There is a decreasing function relationship between the cost of exposing information and the degree of opening. The relationship between OP(d) and COST(d) is shown in Figure 1.

4. The Characteristics of the Ego of Data in the Internet of Things

The ego of data needs to be more concerned with the Internet of things. For example, the producer of the data, which is the condition monitoring value like the patient's pulse, blood pressure, blood sugar, and so on, is the patients for the ego of data in the wireless body area network. Many patients and their families do not want these monitoring data to be leaked out. In the process of data interaction, the interaction of sensitive information is always needed. The incompletely open data cooperation mode is taken by the ego of data in the Internet of things in the process of data exchange.

The space attribute and time attribute of the Internet of things are taken into account, when the ego of data in the Internet of things is studied. Different IOT service system and application system have different requirements on the space attribute and time attribute.

The data collection of the Internet of things is completed by the acquisition nodes deployed in space. In order to study and use these nodes conveniently, they are usually numbered. The spatial information of nodes usually corresponds to the number of nodes. The data of the Internet of things are the data sampled at a certain time, so the data of the Internet of things corresponds to a node and a sampling time. These data collected by the adjacent nodes deployed in the same spatial range have intimate relationships with respect to spatial attributes. The sampling time of each node of the Internet of things also has correlation. Data sampled in the same time period have similar characteristics.

5. The Concepts Involved in the Anonymization Protection of Incompletely Open Cooperation for the Ego of Data in the Internet of Things

In this paper, anonymization protection algorithm has been adopted to complete the incompletely open cooperation for the ego of data. In some data interaction processes, the data processed by anonymization is needed. The data, which is replaced by the equivalence class in the anonymization protection algorithm, is more likely to be found in its own laws, is usually used to reflect the development of the object, and is conducive to data mining for partners. At the same time, it can also realize the sensitive information protection.

The anonymization model based on the fuzzy clustering method is adopted in the incompletely open cooperation for the ego of data in this paper. Considering the characteristics of the ego of data in the Internet of things, the two parameters of tolerable nodes set and tolerable sampling period are introduced, in order to ensure that anonymization results are available. Data in the Internet of things can be anonymous effectively, so as to achieve the incompletely open cooperation protection for the ego of data.

In the Internet of things, the set of data sampling nodes can be divided into many subsets according to the spatial deployment, and the sampling time of the node also can be included in a series of short continuous time fragments. The availability of data will not be affected by the case that the space position and sampling time attributes in each record from the nodes could be replaced separately by the abstract spatial attributes of the set and intermediate value of sampling period, as long as the placement of the sampling point set is covered to a sufficiently small space and the difference of sampling time is small enough.

Definition 5 (NODESP (x, y, z, λ)). It means the spatial attributes of equivalent node. Consider an equivalence class T $(d_1, \ldots, d_n)$, where d_i $(1 \leqslant i \leqslant n)$ means the ith data.

s_i describes the spatial attributes of nodes. NODESP (x, y, z, λ) describes the subset of nodes formed by spatial adjacency clustering in the Internet of things. The subset whose threshold of adjacency is λ is the minimal subset of nodes which meets $\forall s_i \asymp \text{NODESP}(x, y, z, \lambda)$, where $x = \sum_{i=1}^{i=n} x_i/n$, $y = \sum_{i=1}^{i=n} y_i/n$, and $z = \sum_{i=1}^{i=n} z_i/n$. x_i, y_i, and z_i represent, respectively, the abscissa, ordinate, and height of s_i in the three-dimensional space.

Definition 6 (SAMPTIME(t)). It denotes the equivalent sampling time of the record. Let an equivalence class $T(d_1, \ldots, d_n)$, where d_i $(1 \leqslant i \leqslant n)$ is the ith data. t_i denotes the sampling time attribute of d_i. SAMPTIME(t) is the time which meets $\forall t_i \asymp \text{SAMPTIME}(t)$, where $t = \text{mid}(t_1, \ldots, t_n)$.

Definition 7 ($C(\lambda_s)$). It means the subset of nodes determined by clustering threshold λ_s.

Definition 8 ($T(t_s)$). It represents the sampling period in which the time interval is t_s.

Definition 9 ($\text{TNC}(\lambda_{\text{th}})$). It is the acceptable subset of nodes in which clustering threshold is λ_{th}.

It will not affect the availability of data where the spatial attributes of $T(d_1, \ldots, d_n)$ have been replaced by the spatial attributes of NODESP (x, y, z, λ).

Definition 10 ($\text{TST}(t_{\text{th}})$). It means acceptable sampling period. It means the period in which the duration is t_{th}. It will not affect the availability of data where the time attributes of $T(d_1, \ldots, d_n)$ have been replaced by SAMPTIME(t).

Definition 11 (a clustering anonymization algorithm for realizing the ego of data protection in the data interaction process in an incompletely open manner in the Internet of things). Given the group of the ego of data in the Internet of things $D(d_1, \ldots, d_n)$, $\text{TNC}(\lambda_{\text{th}})$, and $\text{TST}(t_{\text{th}})$, any data d_i $(d_i \in D)$ has at least another k-1 records $d_1, \ldots, d_j$ $(j \geqslant k - 1)$ with the same identifier as d_i whose equivalent node has these attributes that the equivalent space attribute is $\text{NODESP}(x, y, z, \lambda)$ and the equivalent sampling time is SAMPTIME (t). If $\lambda_s \leqslant \lambda_{\text{th}}$, $t_s \leqslant t_{\text{th}}$, the set $D(d_1, \ldots, d_n)$ meets the Internet of things' ego of data anonymization protection.

Table 2: The spatial location information of the sampling node.

Node index	x-axis	y-axis	z-axis
p_1	x_1	y_1	z_1
p_2	x_2	y_2	z_2
⋮	⋮	⋮	⋮
p_{Nd}	x_{Nd}	y_{Nd}	z_{Nd}

6. The Design and Analysis of an Anonymization Protection Algorithm in the Data Interaction Process in an Incompletely Open Manner for the Ego of Data in the Internet of Things

Because the ego of data has temporal and spatial characteristics in the Internet of things, the data which has the different sampling time and different node indexes is assigned to the same equivalence class by means of fuzzy clustering of nodes according to the layout of nodes and fuzzy clustering of sampling time. It is required that each equivalence class has to contain 2 or more nodes in the process of fuzzy clustering of nodes and sampling time. The data of multiple nodes is contained in an equivalence class. After anonymization generalization, the data in the equivalence class has the same attributes of node position sampling time, so as to hide the correspondence between the data records and the nodes. The corresponding relationship between the data record and the sampling time is hidden. The location information and time information of the data are blurred.

6.1. The Analysis of Node Fuzzy Clustering. Let Nd represent the number of nodes in the networking environment; p represents the node. Those nodes that are close to each other are divided into a cluster.

The First Step (node coordinates standardization). The spatial location information of the sampling node is denoted by Spactab. It is shown in Table 2.

The Euclidean distance formula is used to calculate the distance between nodes:

$$\text{dist}_{p_i,p_j} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2}. \quad (4)$$

Mutual distance matrix is

$$\begin{bmatrix} \text{dist}_{p_1,p_1} & \text{dist}_{p_1,p_2} & \cdots & \text{dist}_{p_1,p_{\text{Nd}}} \\ \text{dist}_{p_2,p_2} & \text{dist}_{p_2,p_2} & \cdots & \text{dist}_{p_2,p_{\text{Nd}}} \\ \cdots & \cdots & \cdots & \cdots \\ \text{dist}_{p_{\text{Nd}},p_1} & \text{dist}_{p_{\text{Nd}},p_2} & \cdots & \text{dist}_{p_{\text{Nd}},p_{\text{Nd}}} \end{bmatrix}. \quad (5)$$

The Second Step (calibration (the establishment of fuzzy adjacent matrix)). The distance matrix is standardized and the maximum range formula is used:

$$\gamma_{p_i,p_j} = \frac{\text{dist}_{p_i,p_j} - \text{dist}_{\min}}{\text{dist}_{\max} - \text{dist}_{\min}}. \quad (6)$$

Input: Spactab and R;
Output: Subndgrp; % clustered node set
BEGIN
(1) Subndgrp = ∅
(2) FOR each r in R
(3) $\lambda_s = r$; % Value from R is assigned to variable r
(4) IF $\lambda_s \le \lambda_{\text{th}}$ then
(5) Subndgrp = Subndgrp $\bigcup_{m\le \text{NUMGRP}(\lambda_s)} \{[\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m \mid m \in N\}$
% Subndgrp is a node set which spatial adjacent degree is λ_s. $\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i$ represents the subset of nodes, where, NUMNd(λ_s, m) denotes the node number of the subset of spatial adjacent nodes determined by λ_s. The number of nodes in each cluster subset is equal or unequal. $[\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m$ is shown as the mth subset determined by the intercept λ_s. $\bigcup_{m\le \text{NUMGRP}(\lambda_s)} [\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m$ denotes the union of all sub sets. NUMGRP(λ_s) represents the subset number determined by λ_s. %
(6) ENDIF
(7) ENDFOR
END

Algorithm 1: Node clustering.

Table 3: The original sampling data set.

DATA index	Sampling time	Node index	Sensitive information
1	t_1	p_i	data_1
2	t_2	p_j	data_2
⋮	⋮	⋮	⋮
Ndata	$t_{N\text{data}}$	p_k	$\text{data}_{N\text{data}}$

Fuzzy adjacent matrix R':

$$\begin{bmatrix} \gamma_{p_1,p_1} & \gamma_{p_1,p_2} & \cdots & \gamma_{p_1,p_{\text{Nd}}} \\ \gamma_{p_2,p_1} & \gamma_{p_2,p_2} & \cdots & \gamma_{p_2,p_{\text{Nd}}} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{p_{\text{Nd}},p_1} & \gamma_{p_{\text{Nd}},p_2} & \cdots & \gamma_{p_{\text{Nd}},p_{\text{Nd}}} \end{bmatrix}. \tag{7}$$

The Third Step. See Algorithm 1.

6.2. The Analysis of Fuzzy Clustering of Sampling Time. The original sampling data set is denoted by Sampdata. We can assume that the number of records in the set is denoted by Ndata. It is shown in Table 3.

Let the ego of data of the mth nodes subset $[\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m$ group be D_m. We can assume that there are n records of data in D_m.

$$\begin{pmatrix} \text{The index of } D_m & d_1 & d_2 & \cdots & d_n \\ \text{Sampling time} & t_i & t_j & \cdots & t_k \\ \text{Node index} & P_u & P_v & \cdots & P_w \end{pmatrix}, \tag{8}$$

where $\{P_u, P_v, \ldots, P_w\} = [\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m \subseteq$ Subndgrp, $D_m \subseteq$ Sampdata.

They are clustered according to the sampling time difference of the nodes. The nodes with small sampling time difference are divided into a subset, that is, to form an equivalence class.

The First Step. Using the following formula to calculate the sampling time difference between each other:

$$\begin{aligned} &\text{epoch}_{d_i,d_j} \\ &\quad = \left| d_i \cdot \text{Sampling time} - d_j \cdot \text{Sampling time} \right|, \end{aligned} \tag{9}$$

where $d_i \cdot$ Sampling time denotes the sampling time of the d_i record.

Mutual time difference matrix R' is

$$\begin{bmatrix} \text{epoch}_{d_1,d_1} & \text{epoch}_{d_1,d_2} & \cdots & \text{epoch}_{d_1,d_n} \\ \text{epoch}_{d_2,d_1} & \text{epoch}_{d_2,d_2} & \cdots & \text{epoch}_{d_2,d_n} \\ \cdots & \cdots & \cdots & \cdots \\ \text{epoch}_{d_n,d_1} & \text{epoch}_{d_n,d_2} & \cdots & \text{epoch}_{d_n,d_n} \end{bmatrix}. \tag{10}$$

The Second Step. See Algorithm 2.

6.3. The Analysis of the Ego of Data Anonymization Protection Algorithm in the Internet of Things. See Algorithm 3.

In the Internet of things, the physical location of the nodes in the subset whose intercept is λ_s is fixed, and the number of nodes in each cluster node is equal or unequal. Arithmetic average of the number is s. If the data is evenly distributed, the location information leakage (the probability

Input: R', TST(t_{th}) and $[\bigcup_{i\le \text{NUMNd}(\lambda_s,m)} P_i]_m$;
Output: Substgrp; % the set of equivalent class
BEGIN
(1) Substgrp = ∅;
(2) FOR each r in R'
(3) $t_s = r$;
(4) if $t_s < t_{\text{th}}$ then
(5) Substgrp = $\bigcup_{n\le \text{NUMGRP}(t_s) \text{ and } \max(\text{NUMd}(t_s))}\{[\bigcup_{j\le \text{NUMd}(t_s,n) \text{ and } \text{NUMNd}(n)\ge 2} d_j]_n \mid n \in N\}$
% The set of equivalence classes with sampling interval of t_s is obtained. $\bigcup_{j\le \text{NUMd}(t_s,n) \text{ and } \text{NUMNd}(n)\ge 2} d_j$ denotes the data subset whose intercept is t_s, namely, an equivalent class. The function of NUMd(t_s, n) is that the number of the nth equivalent class would be counted out. $[\bigcup_{j\le \text{NUMd}(t_s,n) \text{ and } \text{NUMNd}(n)\ge 2} d_j]_n$ denotes the nth subset, namely, the nth equivalent class. It is asked that the data for each equivalence class contains at least 2 nodes. The function of NUMNd(n) is that the number of nth equivalent class. In this paper, NUMNd(n) ≥ 2. $\bigcup_{n\le \text{NUMGRP}(t_s) \text{ and } \max(\text{NUMd}(t_s))}[\bigcup_{j\le \text{NUMd}(t_s,n) \text{ and } \text{NUMNd}(n)\ge 2} d_j]_n$ represents that the union of all subsets, namely the union of all equivalence classes. The function NUMGRP(t_s) is that total number of equivalence classes whose intercept is t_s is calculated. The maximum number of records would be covered in an equivalence class. We should choose the set of equivalence classes which can cover the maximum number of records.
(6) ENDIF
(7) ENDFOR
END

Algorithm 2: Sampling time clustering algorithm.

of the specific location of the data being guessed) is $1/s$. Similarly, if the sampling time of the data in the equivalence class is distributed in the period, the probability of causing the leakage of time information is $1/t_s$. The larger the t_s, the smaller the probability of leakage of the time information, but also the longer the sampling time of the equivalence class and the greater the loss of information caused by the anonymization result. However, when t_s is over, it will also affect the security of the data. People need to find a balance between security and information loss.

In this paper, the Internet of things' ego of data anonymization model sets acceptable subset of nodes and acceptable sampling period. Since the number of nodes in the fuzzy clustering is required to be at least 2, each equivalence class contains multiple nodes. The time fuzzy clustering algorithm ensures that the sampling time span of the equivalence class is in the appropriate range, so as to achieve the effect of sensitive information protection under the premise of ensuring the availability of data. The anonymization protection algorithm based on fuzzy clustering for the ego of data in the Internet of things in this paper ensures that each equivalence class contains the data of multiple computing nodes by means of fuzzy clustering method. The records in the equivalence class can not correspond to the sampling nodes, thus reducing the risk of leakage of location information.

7. Experiments

The data set used in this paper is the measured data provided by Intel Berkeley Laboratory [12]. The data is collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light, and voltage values once every 31 seconds. Data was collected using the Tiny DB in-network query processing system, built on the TinyOS platform.

15 nodes are randomly selected from the 54 nodes to perform fuzzy clustering, and 499 records of the 15 nodes are randomly selected. Due to the data of the Internet of things being cyclical, focusing on the transformation of the day, we can delete the date column in order to facilitate research. Because in the Internet of things time sampling is always in a continuous period and sampling interval is very short, the sampling time difference of this experiment is accurate to hours.

Set a fixed threshold of the acceptable subset of nodes for clustering. In this paper we randomly set $\lambda_{\text{th}} = 0.3$. The set of clusters of nodes is Subndgrp = {{1, 3, 6}, {8, 9, 12}, {15, 17}, {18, 21}, {37, 39, 44}, {47, 53}}. Fuzzy clustering is used to realize the clustering of all nodes. Figure 2 depicts the respective total number of records processed by anonymization at different acceptable sampling periods t_{th}. It is shown in Figure 2 that t_{th} is very small, resulting in the fact that a lot of data do not meet the conditions of clustering. With the increase of t_{th}, more and more records satisfy the clustering conditions. When $t_{\text{th}} = 18$, almost all the records meet the conditions of clustering.

The value of the spatial attributes of the node in an equivalent class is the average value of the spatial attribute value of the nodes in the acceptable subset, and the sampling time of nodes in an equivalent class is within the sampling period of the subset. The spatial and temporal attributes of the data after anonymous generalization are not greatly affected, and this algorithm does not affect the availability of the data. In the data exchange process, data

Input: λ_{th}, t_{th}, Spactab, Sampdata;
Output: processed data set;
Begin
(1) Do Algorithm 1; % All nodes are implemented by fuzzy clustering according to the spatial location and get the node clustering set Subndgrp.
(2) FOR each $[\bigcup_{i\leq \mathrm{NUMd}(\lambda_s)} P_i]_m$ in Subndgrp % Get a subset of nodes from Subndgrp orderly.
(3) Do Algorithm 2; % The sampling time fuzzy clustering is performed on the data records of the nodes in the subset, and the equivalence class set is obtained. Anonymization processing for each equivalence class.
(4) FOR each $[\bigcup_{j\leq \mathrm{NUMd}(t_s,n) \text{ and } \mathrm{NUMNd}(n)\geq 2} d_j]_n$ in Substgrp % Get a subset of nodes from Substgrp orderly.
(5) Psum = 0; % Psum denotes the total number of nodes in an equivalence class.
(6) FOR each p_j in $[\bigcup_{i\leq \mathrm{NUMd}(\lambda_s)} P_i]_m$
(7) $x$sum = $x$sum + $p_j \cdot x$; % $p_j \cdot x$ represents the x-axis of p_j node in the Spactab.
(8) $y$sum = $y$sum + $p_j \cdot y$; % $p_j \cdot y$ represents the y-axis of p_j node in the Spactab.
(9) $z$sum = $x$sum + $p_j \cdot z$; % $p_j \cdot z$ represents the z-axis of p_j node in the Spactab.
(10) ENDFOR
(11) $x$temp = $x$sum/$p$sum; $y$temp = $y$sum/$p$sum; $z$temp = $z$sum/$p$sum;
(12) FOR each p_j in Spactab % Replace the spatial location information of the node number in the equivalent class data record with the spatial attribute NODESP (x, y, Z, lambda) of the equivalent node of the equivalence class.
(13) IF $p_j \in [\bigcup_{i\leq \mathrm{NUMd}(\lambda_s)} P_i]_m$ then
(14) $p_j \cdot x = x$temp; $y = y$temp; $p_z = z$temp;
(15) ENDIF
(16) ENDFOR
(17) t = mid (d_1·Sampling time,. . ., d_n·Sampling time) % Equivalent sampling calculated with the intermediate value of the sampling time of all data records of time is equivalence class.
(18) For each d_j in $[\bigcup_{j\leq \mathrm{NUMd}(t_s,n) \text{ and } \mathrm{NUMNd}(n)\geq 2} d_j]_n$
(19) $d_j \cdot$ Sampling time = t% Replace the sampling time attribute of the record with the sampling time in the equivalent class.
(20) ENEFOR
(21) ENDFOR
(22) ENDFOR
(23) Count the rest of the data to delete.
% A small number of records cannot be clustered, because the special distance or the duration of sampling time between those records and the most number of records. If the few records are putted into the equivalent class, the nodes' sampling duration of the equivalent class is greater than t_{th}, or the spatial contiguity of nodes in the equivalent class goes beyond λ_{th}. %
END

Algorithm 3: Anonymization protection of the ego of data based on fuzzy clustering in the Internet of things.

after anonymous generalization can be open and can be used by partners. The information of the data after anonymous generalization is approximately equal to the useful information. The open degree OP(D) of the whole data set is approximated as the ratio of the data after anonymous generalization to the whole data, where D represents the data set used in this experiment. At the same time, OP(D) also reflects the anonymity efficiency of anonymous protection algorithms. Figure 3 describes the OP(D) of the proposed algorithm in this paper under different acceptable sampling periods.

It is shown in Figure 3 that OP(D) is increased with the increase of t_{th}. OP(D) value is high, that is, a higher rate of data anonymization.

The node loss rate of clustering K depends on the number of lost nodes in the process of clustering. K = LNd/ND, where LNd represents the number of lost nodes in the clustering; Nd is the total number of nodes in the IOT. All records of lost nodes will not be classified into equivalent classes. K hasa direct impact on the efficiency of anonymous protection and the openness of data and is an important indicator of the data interaction of Internet of things.

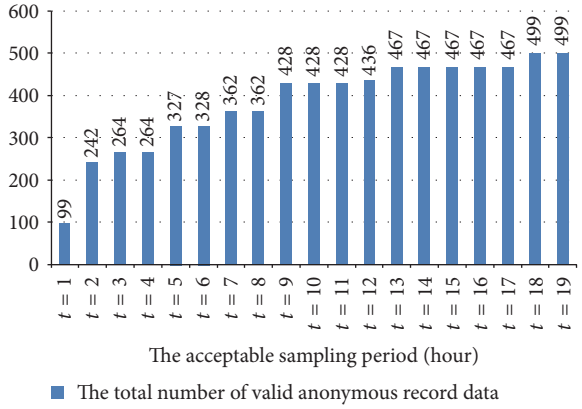

Figure 2: The respective total number of anonymous records at different acceptable sampling periods.

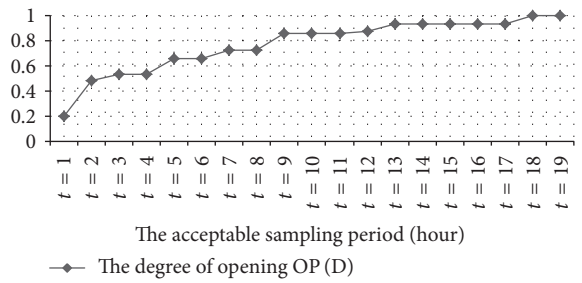

Figure 3: The OP(D) of the proposed algorithm under different acceptable sampling periods.

Figure 4 represents the two node loss rates of clustering K based on the two algorithms at different acceptable mutual distances between sampling nodes. The first is the anonymization protection algorithm based on fuzzy clustering proposed in this paper; the second is the anonymization protection algorithm based on seed clustering [12]. In this experiment, the value λ_{th} is converted to the acceptable mutual distance between nodes for convenient comparison. The second algorithm selects the node numbers for {6, 17, 37, and 47} as seed nodes, according to the law of the maximum number of nodes and the scattered position.

It is shown in Figure 4 that the two loss rates of node clustering reach 0, when the acceptable mutual distance between nodes increases to 20 meters. As the distance becomes larger, the loss rate decreases, down to 0. Figure 4 represents the fact that the node loss rate of clustering based on the fuzzy clustering algorithm is smaller than that based on the seed clustering algorithm. The loss rate is reduced to 0 when the distance is about 15 meters in our algorithm, while the distance is reduced to about 20 meters, which makes the loss rate reduce to about 0 in the seed clustering algorithm. The algorithm proposed in this paper makes the loss rate reach 0 faster than the seed clustering algorithm.

8. Conclusions

In the process of data exchange, data needs to be open on the one hand, and, on the other hand, it needs to be conserved. In order to solve this contradiction, this paper puts forward the concept of the ego of data. After analyzing the characteristics of the ego of data in IOT, we use anonymization protection model to make the data get a certain degree of security in the case of a certain degree of openness. In this paper, we introduce the acceptable subset of nodes and the acceptable sampling period, based on the analysis of the time attribute and spatial properties of the ego of data in the IOT. We can obtain the equivalent class by the fuzzy clustering method and guarantee that the equivalent class contains many nodes. Finally, an anonymization protection algorithm which is suitable for the data exchange in incompletely open manner for the ego of data in the IOT is designed in this paper. The anonymous data set generated by the algorithm can effectively protect the sensitive information of the Internet of things under the premise of ensuring the availability of the data. As a future work, we will continue to extend the concept extension of the ego of data. We are planning to solve the problem of data protection in the IOT by integrating differential privacy protection.

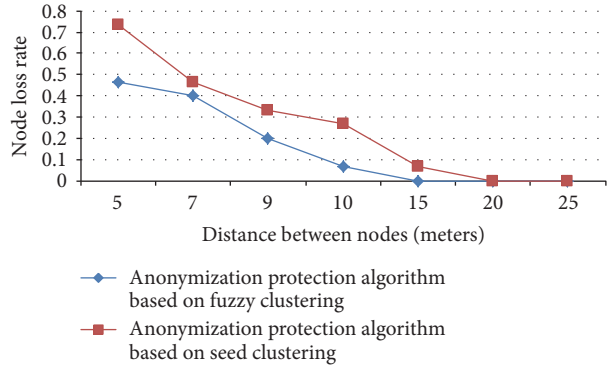

Figure 4: The two loss rates of node clustering.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was financially supported by the Project of Natural Science Foundation of Hainan Province in China (Grant nos. 20166232 and 617033), the National Natural Science Foundation of China (Grant nos. 61462022 and 61561017), and Open Project of State Key Laboratory of Marine Resource Utilization in South China Sea (Grant no. 2016013B).

References

[1] 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.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in

Proceedings of the 22nd International Conference on Data Engineering (ICDE '06), pp. 24–36, Atlanta, Ga, USA, April 2006.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke et al., "L-diversity: privacy beyond *k*-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, article 3, 52 pages, 2007.

[4] R. Wong, J. Li, A. Fu, and K. Wang, "(α, k)-anonymous data publishing," *Journal of Intelligent Information Systems*, vol. 33, no. 2, pp. 209–234, 2009.

[5] G. Aggarwal, T. Feder, K. Kenthapadi et al., "Achieving anonymity via clustering," in *Proceeding of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '06)*, pp. 153–162, New York-NY-USA, June 2006.

[6] Ning Z. H., W. Jiang, J. Zhan et al., "Property-based anonymous attestation in trusted cloud computing," *Journal of Electrical and Computer Engineering*, vol. 2014, no. 17, pp. 1–7, 2014.

[7] S. Muftic, N. B. Abdullah, and I. Kounelis, "Business information exchange system with security, privacy, and anonymity," *Journal of Electrical and Computer Engineering*, vol. 2016, no. 1, pp. 1–10, 2016.

[8] S. Landau, "Control use of data to protect privacy," *Science*, vol. 347, no. 6221, pp. 504–506, 2015.

[9] J. C. Doshi and B. Trivedi, "Hybrid intelligent access control framework to protect data privacy and theft," in *Proceeding of the International Conference on Advances in Computing, Communications and Informatics (ICACCI '15)*, pp. 1766–1770, 2015.

[10] Y.-S. Jeong and S.-S. Shin, "An efficient authentication scheme to protect user privacy in seamless big data services," *Wireless Personal Communications*, vol. 86, no. 1, pp. 7–19, 2016.

[11] Y. Ge and F. Li, "Data management in the internet of things," *Chinese society of computer communication*, vol. 6, no. 4, pp. 30–34, 2010.

[12] H. Wei and C. Zhong, "Information technology of Internet of things k-anonymous algorithm based on parallel clustering," *Information Technology*, vol. 12, pp. 6–10, 2013.

[13] H. Bah and A. Lev, "k-anonymity based framework for privacy preserving data collection in wireless sensor networks," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 18, no. 2, pp. 241–271, 2010.

[14] A. Samani, H. H. Ghenniwa, and A. Wahaishi, "Privacy in internet of things: a model and protection framework," *Procedia Computer Science*, vol. 52, no. 538, pp. 606–613, 2015.