Location Privacy Protection Based on Differential Privacy Strategy for Big Data in Industrial Internet of Things

Chunyong Yin, Jinwen Xi, Ruxia Sun, and Jin Wang, *Member, IEEE*

***Abstract*—In the research of location privacy protection, the existing methods are mostly based on the traditional anonymization, fuzzy and cryptography technology, and little success in the big data environment, for example, the sensor networks contain sensitive information, which is compulsory to be appropriately protected. Current trends, such as "Industrie 4.0" and Internet of Things (IoT), generate, process, and exchange vast amounts of security-critical and privacy-sensitive data, which makes them attractive targets of attacks. However, previous methods overlooked the privacy protection issue, leading to privacy violation. In this paper, we propose a location privacy protection method that satisfies differential privacy constraint to protect location data privacy and maximizes the utility of data and algorithm in Industrial IoT. In view of the high value and low density of location data, we combine the utility with the privacy and build a multilevel location information tree model. Furthermore, the index mechanism of differential privacy is used to select data according to the tree node accessing frequency. Finally, the Laplace scheme is used to add noises to accessing frequency of the selecting data. As is shown in the theoretical analysis and the experimental results, the proposed strategy can achieve significant improvements in terms of security, privacy, and applicability.**

***Index Terms*—Differential privacy, Internet of Things (IoT), location privacy protection, location privacy tree (LPT).**

I. Introduction

THE development of information technology has accumulated a great deal of data for today's digital systems. Big data is a very important research and development resource, and the demand for data publishing, sharing, and analysis is also growing rapidly. People pay more and more attention to the security of data protection, and the government, enterprises, and individuals are also improving their understanding of privacy protection.

Body sensor networks (BSNs), as a special application of wireless sensor networks [1], are deployed on the surface of bodies for periodically monitoring physical conditions. For a typical social participatory sensing application, it is important to motivate participatory, at the same time, the participatory sensing process should not disclose the privacy information of any participating party (the private data) or the community (patterns, distribution, etc.). Therefore, it is essential to develop a privacy protection data strategy for the protection of the users and the community [2].

Devices in the Internet of Things (IoT) generate, process, and exchange vast amounts of security and safety-critical data as well as privacy-sensitive information and, hence, are appealing targets of various attacks. To ensure the correct and safe operation of IoT systems, it is crucial to assure the integrity of the underlying devices, especially their code and data privacy, against malicious modification [3].

The privacy threats of Industrial IoT (IIoT) [4] can be simply divided into two categories: privacy threats based on data and privacy threats based on location. Data privacy issues mainly refer to the leakage of secret information in the process of data acquisition and transmission in IoT.

Location privacy is an important part of privacy protection of the IoT. It mainly refers to the location privacy of each node in the IoT and the location privacy of the IoT in providing various location services, especially including the radio frequency identification (RFID) reader location privacy, RFID user location privacy, sensor node location privacy, and location-based privacy issues based on location services [3], [5].

Usually, data collected, aggregated, and transmitted in sensor networks contain personal and sensitive information, which directly or indirectly reveals the condition of a person. If the data cannot be properly preserved, once exposed to the public, the privacy will be destroyed. Therefore, protecting the privacy of sensitive data is greatly important [6].

Location data implies moving objects, spatial coordinates, current time, and unique features different from other data,

Manuscript received August 27, 2017; revised October 15, 2017; accepted November 3, 2017. Date of publication November 15, 2017; date of current version August 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61772282, Grant 61373134, Grant 61772454, and Grant 61402234, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX17_0901), and in part by the Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology. Paper no. TII-17-1942. *(Corresponding author: Jin Wang.)*

C. Yin, J. Xi, and R. Sun are with the School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China (e-mail: yinchunyong@hotmail.com; javenxi@yeah.net; src@nuist.edu.cn).

J. Wang is with the College of Information Engineering, Yangzhou University, Yangzhou 225127, China (e-mail: jinwang@yzu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2017.2773646

which is discrete and of high value. Before the concept of big data, most of the privacy protection methods focus on a small number of nonpositional data. There are some limitations for the protection of location data privacy in big data. There are two main reasons for these limitations, which are as follows: 1) Multiple data fusion by big data makes traditional anonymity [7] and fuzzification technology difficult to take effect in location privacy protection [8]; 2) traditional cryptography technology takes little effect on the real-time analysis required by big data.

In summary, location privacy protection faces great challenges for big data in IIoT. Therefore, we propose a more rigorous LPT (location privacy tree) DP-k (differential privacy-k) location privacy protection method based on differential privacy strategy for big data in sensor networks.

Our specific contributions mainly include the following.

1) We introduce a tree structure to represent the location data in sensor networks, which called LPT, according to the characteristics and retrieval difficulty of location data.
2) The differential privacy strategy is suitable for location privacy protection because it is insensitive to the background knowledge, and the DP-k model has better protection effect. The Laplace and index mechanisms are the main implementation mechanisms of differential privacy, which can show the degree of privacy protection by allocating the privacy budget and is relatively reliable and rigorous.
3) We conduct extensive experiments on real-world datasets which show that the proposed location privacy protection method can protect users' location privacy without significantly affecting the privacy, availability, security, and effectiveness.

Section II presents necessary background and related work. Section III introduces the preliminary knowledge. Section IV details the proposed location privacy protection method. The real-world datasets and experimental results are presented in Section V, and conclusions are given in Section VI.

II. Related Work

In the process of data mining and data publishing, the privacy protection of location data mainly involves two aspects: privacy model and utility.

A. Privacy Model

In recent years, with the widespread application of location services and data mining and publishing, the location privacy protection model in location service can be divided into two categories: One is the traditional anonymous model based on grouping; and the other is the differential privacy model that ignores the attacker's background knowledge.

1) Anonymous Model Based on Grouping: The traditional anonymous model based on grouping plays an important role in location privacy protection. Samarati *et al.* [9] proposed the k-anonymity method and also a large number of methods based on k-anonymity [10]–[12]. Previous research work [12], [13] shows that only using anonymous methods does not provide good protection to a wide range of data. According to Cristofaro De *et al.* [14], the encryption privacy protection method is proposed, which can completely protect the privacy of data and prevent the leakage of data in the process of location service, but the availability of data is insufficient. The development of traditional location privacy protection technology has gone through three stages: the "informed and consent" method proposed by Beresford *et al.* [15] has been developed to deal with anonymous processing in a single location and, then, to deal with anonymous processing of user's trajectory data. Heuristic privacy measurement, probability deduction, and private information retrieval based technologies are common methods to protect location privacy. The heuristic privacy measurement method mainly protects the users who are not in the strict privacy protection environment, such as k-anonymity, t-closeness [16], m-invariance [17], and l-diversity. However, these three kinds of methods are proposed as a unified attacking model, which protect the location data under the premise of accumulating certain background knowledge. As attackers grasp more background knowledge, these methods cannot effectively protect the user's location data privacy, and the lack of privacy protection about the type of relational privacy protection method is proposed in [18].

2) Differential Privacy Model: Differential privacy model is a strong privacy concept that is completely independent of attacker's background knowledge and computing ability and has become a research hotspot in recent years. Compared to the traditional privacy protection model, differential privacy has its unique advantages. First, the model assumes that the attacker has the maximum background knowledge. Second, differential privacy has a solid mathematical foundation, a strict definition of privacy protection, and a reliable quantitative evaluation method. In recent years, the differential privacy model has been widely applied in privacy protection. Especially, the functional mechanism is introduced in [19], in which an objective function of the ε-differential privacy disturbance optimization problem is used to protect privacy. The privacy-preserving mining (Diff-PPM) algorithm is proposed in [20] and combined with Markov Chain Monte Carlo, which provides privacy protection and maintains the high data availability while satisfying $(\varepsilon - \delta)$-differential privacy conditions. The PrivBasis and SmartTrunc method proposed in [21] adopt differential privacy model in the mining process of frequent itemsets, ensuring the privacy and utility of data analysis and anonymity. The DiffP-C4.5 and DiffGen algorithm introduced in [22] combines differential privacy with decision trees and other data structures to maintain a balance between data privacy and availability.

In a word, differential privacy is an effective privacy protection mechanism, which protects the user's location privacy while keeping enough useful information for data analysis.

B. Utility Maximization

In the process of location service, data mining and data publishing need to protect location privacy and provide enough information for data analysis. Therefore, data utility is the core issue that needs to be paid attention to. Friedman and Schuster

[23] use a compressive sensing theory to propose a perception mechanism to solve the published issue of statistical results, which can effectively solve the insufficiency problem of data utility, but it destroys the link among data. The DP-topk method is proposed in [24], which has a more rigorous definition of effectiveness. But this method ignored the relation among transaction data, and because of its poor efficiency in data processing individually, the availability of algorithm is not high. To maximize the utility of the results and meet the requirement of location privacy protection, this paper proposes an LPT-DP-k algorithm, which is a more rigorous differential privacy protection method, that can not only guarantee the high availability of data, but can also enhance the usability of the algorithm.

III. Preliminary Knowledge

Differential privacy is a common privacy protection framework that supported by the solid mathematical theory, which can provide privacy protection for data in case the attacker grasps the largest background knowledge.

Definition 1 (Differential privacy): Suppose there is a random algorithm M, P_M is all possible output sets for M, for any given adjacent dataset D and D' (there is at most one different record between them), $|D\Delta D'| \leq 1$, and S_M is any subset of P_M. If the algorithm M satisfies the following inequality, the algorithm M will satisfy ε-differential privacy protection

$$P_r\left[M\left(D\right) \in S_M\right] \leq \exp\left(\varepsilon\right) \times P_r\left[M\left(D'\right) \in S_M\right] \tag{1}$$

where $P_r[.]$ represents the randomness of the algorithm M on the datasets D and D'.

Definition 2 (Sensitivity): The differential privacy protection method defines two kinds of sensitivity, called global sensitivity and local sensitivity. Suppose there is a query function $f : D \rightarrow R^d$, the input is a dataset and the output is a d-dimensional real vector. For any adjacent dataset

$$\Delta f = \max_{D,D'} \left\| f\left(D\right) - f\left(D'\right) \right\|_1 \tag{2}$$

where f is the global sensitivity of function, and Δf represents the maximum change value of the output results. $\|f(D) - f(D')\|$ is the 1-order norm distance between $f(D)$ and $f(D')$.

Definition 3 (Implementation mechanism): The Laplace mechanism and exponential mechanism [25] are two of the most basic implementation mechanisms of differential privacy protection. In this paper, the Laplace mechanism is used to add the noise that obeys the Laplace distribution to realize the differential privacy. Assuming the privacy protection algorithm f based on the Laplace mechanism, the noise follows the Laplace distribution with the variance $\frac{\Delta f}{\varepsilon}$ and mean 0. Then, the probability density function is

$$P_r\left(x, \lambda\right) = \frac{1}{2\lambda} \exp\left(-\frac{|x|}{\lambda}\right) \tag{3}$$

where x represents the specific variable and $\lambda = \frac{\Delta f}{\varepsilon}$.

If the algorithm f is proportional to the probability of $\exp(\frac{\varepsilon\lambda(D,r)}{2\Delta\lambda})$ to select from O and export r, then the algorithm f provides ε-differential privacy protection, the concrete formula is as follows:

$$A\left(D, \lambda\right) = \left\{ r : \mathrm{P_r}\left(r \in O\right) \propto \exp\left(\frac{\varepsilon\lambda\left(D, r\right)}{2\Delta\lambda}\right) \right\} \tag{4}$$

where $\Delta\lambda$ is the global sensitivity of the scoring function $\lambda(D, r)$, (4) shows that the higher the score is, the larger the probability of the selected output will be.

IV. LPT-DP-k Location Privacy Protection

This section introduces the overview of the LPT-DP-k algorithm and the specific implementation details of the algorithm.

A. LPT-DP-k Algorithm

First, we choose the LPT construction to maintain the relation among the location data. Second, we select the sensitive location information that is most likely to disclose privacy to add noise. The algorithm summarizes as follows:

Algorithm 1: LPT-DP-k algorithm.

1: **Initialize:** data set D^*, location data set D and $D \in D^*$, differential privacy parameter ε, parameter k, the number of nodes that are accessed by the data is count and empty sets A, B.
2: Construct a location privacy tree (LPT).
3: The node data will be divided into two categories, and the frequent patterns of accessing frequency not less than $count$ are recorded in the set, see Section III-B.
4: Calculate the output probability, the formula is $P_r(a_i) = \frac{a_i.w}{\sum_{j=1}^{n} a_j.w}$, see Section III-C.
5: Add noise and combine with $LPT_D^{D^*}$ to construct the noisy location privacy tree $LPT_D^{D^*} - k$, see Section III-D.
6: Finally, publish the $LPT_D^{D^*} - k$ after step 4.

B. Construction of Location Information Tree Corresponding to Location Data

Assuming the location privacy tree $\mathrm{LPT}_D^{D^*}$ is corresponding to the dataset D, which records each person's visiting records within a month. According to the location information, the LPT-DP-k algorithm uses the tree structure with better correlation. The location correspondence is given in Table I.

As is shown in Table I, the user's location information and accessing frequency of the places within a month is listed, we can get the location correspondence table.

According to the content in Table II, we can construct the LPT, as shown in Fig. 1.

As shown in Fig. 1, the total number of nodes in $\mathrm{LPT}_D^{D^*}$ is $\sum_{i=1}^{D^*} C_{D^*}^1 = 2^{D^*} - 1$.

C. Weighted Selection Based on Exponential Mechanism

We use the exponential mechanism in the differential privacy protection model to do weighted selection. The formula is as

TABLE I
LOCATION CORRESPONDENCE

Number	Location Information	Accessing Count
1	Computer building	28
2	Remote Sensing Institute	13
3	All-in-one card management center	2
4	Mingde Building	6
5	Training Gym	5
6	Dormitory 3	30
...	...	...

TABLE II
LOCATION DATA INFORMATION

Number	Location Label	Accessing Count
1–28	1	28
29–41	2	13
42 and 43	3	2
44–49	4	6
50–54	5	5
55–84	6	30
85–97	1, 2	13
..	...	...

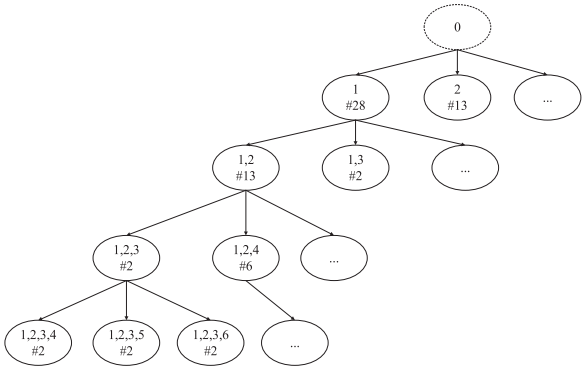


Fig. 1. LPT.

follows:

$$P_r\left(a_i\right)=\frac{a_i.w}{\sum_{j=1}^{n} a_j.w}. \tag{5}$$

Here, $P_r(a_i)$ represents the probability of being selected and $a_i.w$ represents the weight of the pattern a_i. The selection algorithm based on the exponential mechanism is as follows.

Algorithm 2: Selection algorithm.

1: **Initialize:** records in the frequent pattern set A.
2: Input a set A of frequent pattern records and score each record a_i.
3: Calculate the weight $a_i.w$ of each pattern record.
4: Take the weight $a_i.w$ of each pattern in the set A into the formula $P_r(a_i)=\frac{a_i.w}{\sum_{j=1}^{n} a_j.w}$, select k frequent pattern records a_i and record the set as B.

The scoring function is given as

$$M\left(A,a_i\right)=Q\left(a_i\right). \tag{6}$$

Here, $Q(a_i)$ represents the accessing frequency of pattern a_i. Calculate the weight $a_i.w$ of each pattern record

$$a_i.w=\exp\left(\frac{\varepsilon_1 * M\left(A,a_i\right)}{2\Delta M}\right). \tag{7}$$

This paper sets up the scoring function $M(A,a_i)=Q(a_i)$, $Q(a_i)$ represents the accessing frequency of pattern a_i. $M(A,a_i)$ represents the scoring value of a_i. The formula to calculate ΔM is as follows:

$$\Delta M=\max_{i,j\in N}\left\|Q\left(a_i\right)-Q\left(a_j\right)\right\|_1. \tag{8}$$

where ΔM represents the maximum value of the difference in the accessing frequency among N data record modes.

D. Noise Enhancement Based on Laplace Mechanism

This paper proposes the method of node data processing to improve the efficiency of data processing and the effectiveness of the algorithm. The concrete steps are as follows.

Algorithm 3: Laplace noising algorithm.

1: **Initialize:** the privacy budget ε, k frequent pattern set B after weighted selection, query function f.
2: Add noise into the frequent pattern set B to keep to the Laplace distribution and destabilize the real accessing frequency of k records. The added noise keeps to Laplace distribution that the probability density function is $\mathrm{P_r}(x,\lambda)=\frac{1}{2\lambda}\exp(-\frac{|x|}{\lambda})$.
3: According to (9), the noise set C can be calculated. Finally, the noisy location privacy tree $LPT_D^{D^*}-k$ is generated by combining the original data tree $LPT_D^{D^*}$ with the set C.

Therefore, it should be satisfied in every round of noise adding process

$$f\left(b\right)=f\left(b\right)+\mathrm{lap}\left(\frac{\Delta f}{\varepsilon}\right). \tag{9}$$

B is the set of k frequent pattern b. $f(b)$ is the query function on the record b and $\mathrm{lap}(\frac{\Delta f}{\varepsilon})$ represents the noise that keeps to Laplace distribution, as shown in Fig. 2.

True positive (TP), false Positive (FP), and accuracy [26] can be used to measure the utility of noisy location privacy tree $\mathrm{LPT}_D^{D^*}-k$. False rejection rate (FRR) is used to analyze the effectiveness of the algorithm. The smaller the FRR is, the higher the effectiveness of the algorithm will be.

TP: The number of frequent patterns that appear in $f_k(D)$ and in $f_k(D')$

$$\mathrm{TP}=\left|f_k\left(D\right)\cap f_k\left(D'\right)\right|. \tag{10}$$

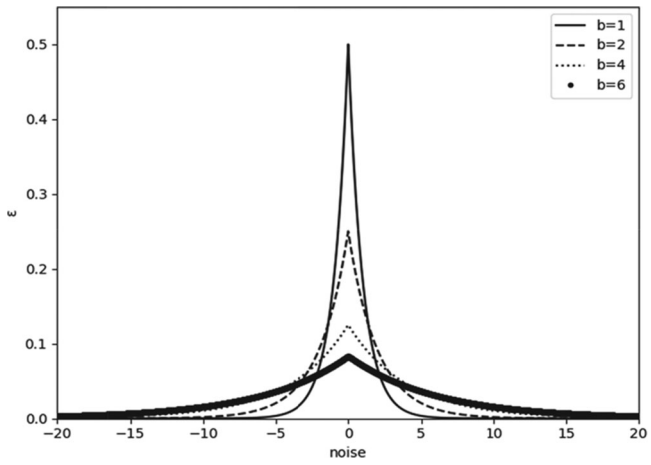


Fig. 2. Relationship between privacy budget and noise size.

TABLE III
EXPERIMENTAL DATASET

User Id	Check-In Time	Latitude	Longitude	Location Id
0	2010-10-19T23:55:27Z	30.23590912	–97.79513958	22847
0	2010-10-18T22:17:43Z	30.26910295	–97.74939537	420315
0	2010-10-17T23:42:03Z	30.25573099	–97.76338577	316637
0	2010-10-12T19:44:40Z	30.26910295	–97.74939537	420315

FP: The number of frequent patterns that appear not in $f_k(D)$, but in $f_k(D')$

$$\mathrm{FP} = \left| f_k\left(D'\right) - f_k\left(D\right) \cap f_k\left(D'\right) \right|. \tag{11}$$

Accuracy

$$\mathrm{Accuracy} = \frac{\left| f_k\left(D\right) \cap f_k\left(D'\right) \right|}{\left| f_k\left(D'\right) \right|}. \tag{12}$$

FRR:

$$\mathrm{FRR} = \frac{\left| f_k\left(D\right) \cup f_k\left(D'\right) - f_k\left(D'\right) \right|}{k}. \tag{13}$$

Error

$$\mathrm{Error} = \frac{\left\| \mathrm{LPT}_D^{D^*} - \mathrm{LPT}_D^{D^*} - k \right\|_2}{\left\| \mathrm{LPT}_D^{D^*} \right\|_2}. \tag{14}$$

V. EXPERIMENTAL ANALYSIS

A. Experiment Setting

1) Dataset: This paper uses the check-in dataset from the Gowalla (Gowalla total-check-in dataset). This dataset has a large amount of data and records the information of the users. In Table III, the experimental dataset is presented.

2) Operating Environment and Parameter Configuration: The experimental environment in this paper is MATLAB(R2013b) and PyCharm (Part of the program using python). Hardware environment is 2.60 GHz Core(TM) i7-6700HQ CPU, 20.00 GB, Win10 system of 64 bit. The parameters that are used during the experiment are given in Table IV.

TABLE IV
PARAMETER SETTING

Parameter	Value
k	20, 40, 60, 80, 100, 150, 200, 300, 500, 1000, 2000, 5000, 10 000, 20 000
ε	0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.1, 0.5, 0.75, 1, 1.1, 1.25, 1.5, 1.75, 5, 10, 15
Location id	1, 2, 3, 4, 5, 6
Location accessing frequency	28, 13, 2, 6, 5, 30

TABLE V
TIME OF BUILDING AND ADDING NOISE TO RECONSTRUCT THE LOCATION INFORMATION TREE

Pattern n	Time.build (s)	Time.update (s)	Pattern n	Time.build (s)	Time.update (s)
16	0.0000015	0.000103010	16	0.0000014	0.000102190
32	0.0000016	0.000103110	32	0.0000024	0.000102180
64	0.0000029	0.000103324	64	0.0000035	0.000102195
128	0.0000057	0.000103010	128	0.0000061	0.000102120
256	0.0000116	0.000104032	256	0.0000110	0.000102095
512	0.0000149	0.000103000	512	0.0000150	0.000102095
1024	0.0000262	0.000105032	1024	0.0000280	0.000102030
2048	0.0000370	0.000106340	2048	–	–
4096	0.0000443	0.000108970	4096	–	–

TABLE VI
EXTRACTION EFFICIENCY OF LOCATION DATA PRIVACY

ε	0.01	0.03	0.05	0.1	0.5	1.0	1.5
k	3432	3691	4254	42 103	8932	53 245	79 310

B. Analysis of Algorithm Timeliness

The timeliness analysis is mainly analyzed in three aspects.

1) Timeliness of constructing LPT.
2) Timeliness of weighted selection and updating of LPT.
3) The efficiency of extracting location data.

As given in Table V, the efficiency of constructing and reconstructing LPT is higher when the mode n is constantly changing.

As given in Table VI, within a certain range, the extraction efficiency of location data increases with the increase in ε. As the ε increases, the degree of privacy protection decreases and the number of extraction in the unit time increases.

C. Analysis of Data Utility

This paper extracts data from multiple groups of patterns with 500, 1000, and 2000 data. The blue dots are used to represent the insensitive locations and the red ones indicate the sensitive locations. The experimental results are shown in Figs. 3–5.

The red sensitive locations are 103 and the blue insensitive locations are 397 in Fig. 3(a), whereas the red sensitive locations are 137 and the blue insensitive locations are 353 in Fig. 3(b).

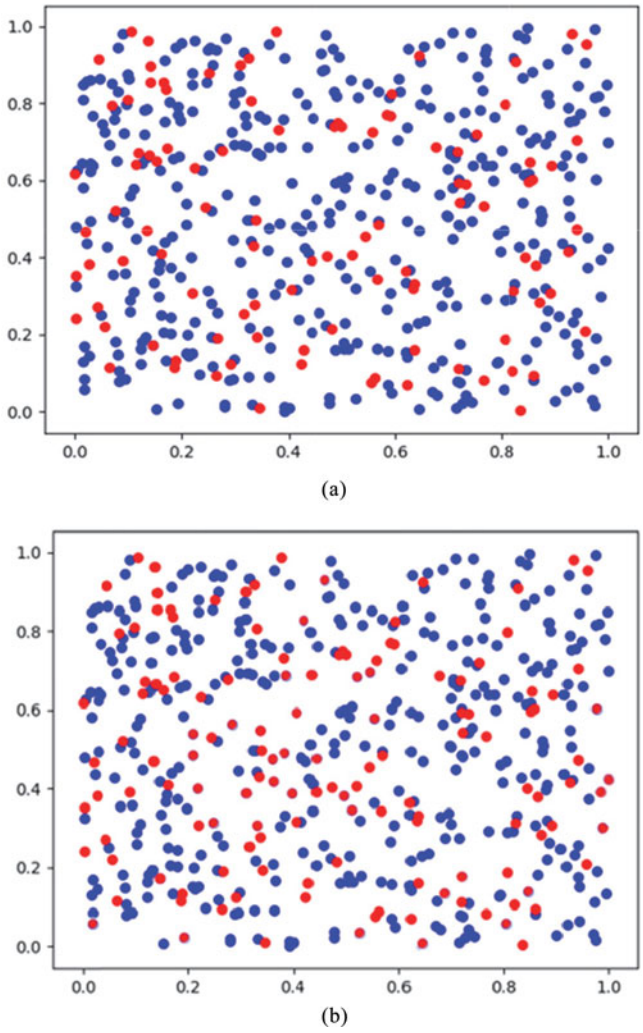


Fig. 3. Data distribution before and after protection (k is 500). (a) Locations are shown before adding noise. (b) Ones after adding noise.

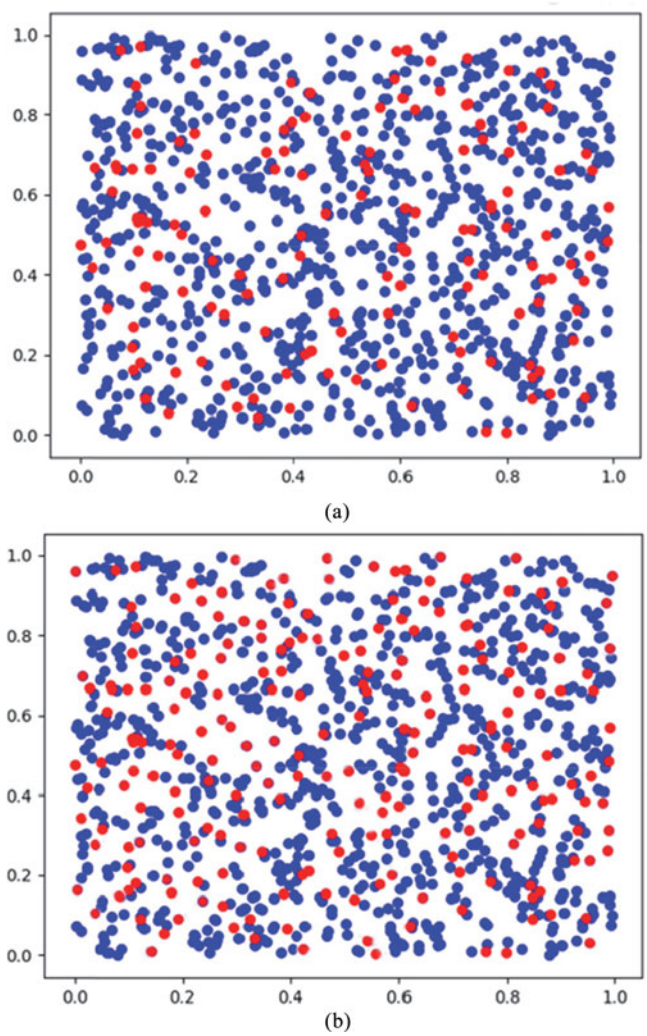


Fig. 4. Data distribution before and after protection (k is 1000). (a) Locations are shown before adding noise. (b) Ones after adding noise.

There are 141 red sensitive locations and 859 blue insensitive locations in Fig. 4(a), after protection, there are 219 red sensitive position and 781 blue insensitive locations in Fig. 4(b), which shows that the sensitive locations have increased.

There are 235 red sensitive locations and 1765 blue insensitive locations in Fig. 5(a), after protection, there are 380 red sensitive locations and 1620 blue insensitive locations in Fig. 5(b). Table VII presents the statistical results of the data utility.

We compare the proposed method with three methods of TDPS_LP_Result, TDPS_Signal, and TDPS_EP, and the experimental results are shown in Fig. 6.

From Fig. 6, we can make the following observations.

1) The noise error of the proposed LPT-DP-k algorithm is smaller, when $\varepsilon > 0.015$, the error is beginning to stabilize, which shows that the utility of the proposed method is higher and better than the previous proposed methods.
2) As the privacy parameter ε increases, the Error will be smaller; when meeting the ε-differential privacy protection, the smaller the ε is, the more the noise needs and the higher the privacy protection level will be.
3) For TDPS_LP_Result, TDPS_LP_Signal, and TDPS_E methods, the privacy parameters are required to be larger to achieve the lower error. When the privacy parameters are small, the error is relatively larger. The experimental results are given in Tables VIII and IX.

As shown in Table VIII, with the value of k increasing, the accuracy of four methods decreases; and when $k > 200$, the accuracy of the previous three methods is less than 70%, whereas the accuracy of the proposed method still remains at about 95%, when $k > 500$, the accuracy decreases but still maintains at more than 80%.

As presented in Table IX, the smaller the privacy parameter is, the higher the privacy protection level will be.

Finally, this paper compares the utility of DP-topk and LPT-DP-k by FRR. The experimental results are shown in Figs. 7 and 8.

Assuming k is 100, the privacy parameter is changed, and the result is shown in Fig. 7: with the privacy parameter increasing, the value of FRR with DP-topk and the proposed method maintains low, but the value of the proposed method is lower. From

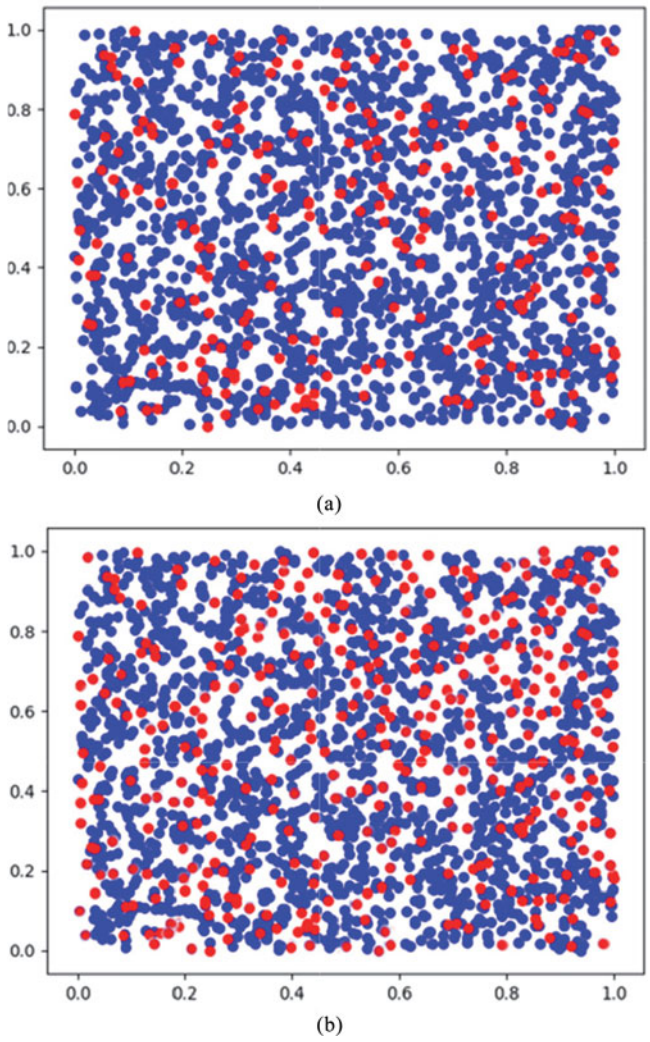


Fig. 5. Data distribution before and after protection (k is 2000). (a) Locations are shown before adding noise. (b) Ones after adding noise.

TABLE VII
UTILITY STATISTICS UNDER DIFFERENT DATA MODELS

Pattern Number	Before the Experiment		After the Experiment	
	Sensitive Locations (Red Dots)	Insensitive Locations (Blue Dots)	Sensitive Locations (Red Dots)	Insensitive Locations (Blue Dots)
500	103	397	137	353
1000	141	859	219	781
2000	235	1765	380	1620

this, we can see that the availability of the proposed method is better when the protection level is high.

Assuming ε is 1, the value of k is changed, and the result is shown in Fig. 8: with k increasing, the value of FRR increases, but the value of DP-topk is larger than that in this paper. To sum up, the proposed method has obvious advantages both in the level of location privacy protection and the effectiveness of the algorithm.

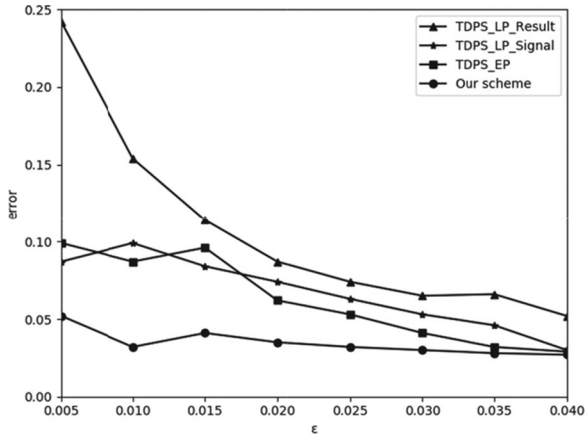


Fig. 6. Error comparison of adding-noise methods.

TABLE VIII
UTILITY COMPARISON OF TOP k FREQUENT ITEMSET MINING UNDER FIXED PRIVACY PARAMETER ($\varepsilon = 1$)

k	TP				FP				Accuracy			
	LPR	LPS	EP	DPk	LPR	LPS	EP	DPk	LPR	LPS	EP	DPk
50	46	47	26	49	4	3	4	1	0.93	0.95	0.53	0.98
100	80	78	39	96	20	22	61	4	0.80	0.78	0.39	0.96
200	126	126	70	190	74	74	130	10	0.63	0.63	0.35	0.95
500	305	290	150	455	195	210	350	45	0.61	0.58	0.30	0.91
1000	530	540	230	890	470	460	770	110	0.53	0.54	0.23	0.89
2000	680	940	40	1720	1320	1060	1960	280	0.38	0.47	0.20	0.86

TABLE IX
UTILITY COMPARISON OF TOP k FREQUENT ITEMSET MINING UNDER FIXED FREQUENT PATTERNS ($k = 200$)

ε	TP				FP				Accuracy			
	LPR	LPS	EP	DPk	LPR	LPS	EP	DPk	LPR	LPS	EP	DPk
0.01	136	184	70	188	64	16	130	12	0.68	0.92	0.35	0.94
0.05	178	182	64	188	22	18	136	12	0.89	0.91	0.32	0.94
0.1	174	180	90	190	26	20	110	10	0.87	0.90	0.45	0.95
0.5	180	190	160	190	20	10	40	10	0.90	0.95	0.80	0.95
1	182	188	168	188	18	12	32	12	0.91	0.94	0.84	0.94
1.5	184	190	174	190	16	10	26	10	0.92	0.95	0.87	0.95

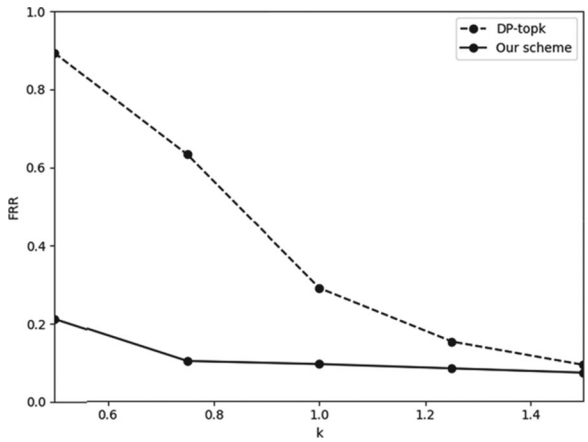


Fig. 7. FRR comparison when $k = 100$.

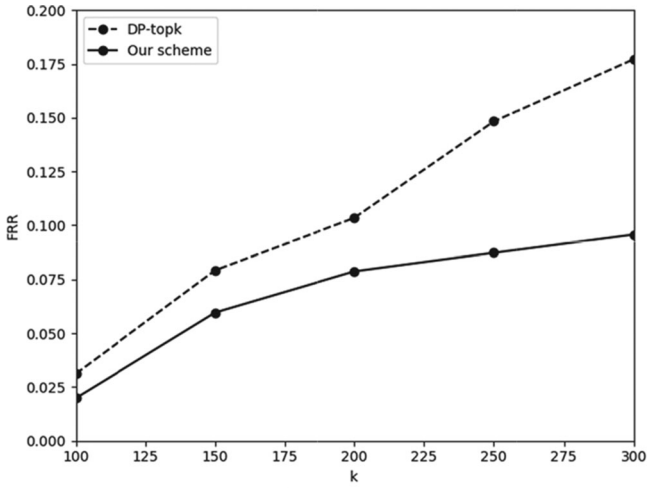


Fig. 8. FRR comparison when $\varepsilon = 1$.

VI. Conclusion

This paper proposes a location privacy protection method based on the differential privacy strategy for big data in sensor networks. The method expresses the position dataset by constructing the location information tree model, which solves the problem that the location data are difficult to be expressed because of its characteristics of high dispersion and low density and adds noise information to cover the original trajectory and position data. It is more effective in protecting the privacy of data and maintaining high availability of data and algorithm. The differential privacy protection model is applied to the protection of location privacy. Compared with the traditional location privacy protection algorithm, the proposed algorithm is more rigorous and has higher algorithm utility and processing efficiency. In the next step, we will discuss to explore the more efficient data structure to express location information in the process of location data expression and propose more utility target functions for different application scenarios.

References

[1] I. FAkyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw., Int. J. Comput. Telecommun. Netw.*, vol. 38, no. 4, pp. 393–422, 2010.

[2] K. Xing, C. Hu, J. Yu, X. Cheng, and F. Zhang, "Mutual privacy preserving k-means clustering in social participatory sensing," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2066–2076, Aug. 2017.

[3] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017.

[4] F. Xiao, L. Sha, Z. Yuan, and R. Wang, "VulHunter: A discovery for unknown bugs based on analysis for known patches in industry internet of things," *IEEE Trans. Emerg. Topics Comput.*, vol. PP, no. 99, pp. 1–1, 2017.

[5] Z. Wang, F. Xiao, N. Ye, R. Wang, and P. Yang, "A see-through-wall system for device-free human motion sensing based on battery-free RFID," *ACM Trans. Embedded Comput. Syst.*, vol. 17, no. 1, pp. 1–21, 2017.

[6] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2688–2710, 2010.

[7] C. Yin, S. Zhang, J. Xi, and J. Wang, "An improved anonymity model for big data security based on clustering algorithm," *Concurrency Comput. Pract. Experience*, vol. 29, no. 7, pp. 1–13, 2017.

[8] C. Yin, J. Xi, and R. Sun, "Location privacy protection based on improved K-value method in augmented reality on mobile devices," *Mobile Inf. Syst.*, vol. 2017, pp. 1–7, 2017.

[9] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proc. 7th ACM SIGACT-SIGMOD-SIGART Symp. Principles Database Syst.*, 1998, pp. 188–202.

[10] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, "(α, k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishing," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 754–759.

[11] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd IEEE Int. Conf. Data Eng.*, 2007, pp. 106–115.

[12] E. Zheleva and L. Getoor, "Preserving the privacy of sensitive relationships in graph data," in *Proc. 1st ACM SIGKDD Workshop Privacy, Security, Trust KDD*, 2007, pp. 153–171.

[13] A. Korolova, R. Motwani, S. U. Nabar, and Y. Xu, "Link privacy in social networks," in *Proc. 24th Int. Conf. Data Eng.*, 2008, pp. 1355–1357.

[14] E. Cristofaro De, C. Soriente, G. Tsudik, and A. Williams, "Hummingbird: Privacy at the time of Twitter," in *Proc. IEEE Symp. Security Privacy*, 2012, pp. 285–299.

[15] A. R. Beresford, A. Rice, N. Skehin, and R. Sohan, "MockDroid: Trading privacy for application functionality on smartphones," in *Proc. 12th Workshop Mobile Comput. Syst. Appl*, 2011, pp. 49–54.

[16] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacy grid," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 237–246.

[17] L. Liu, "From data privacy to location privacy: Models and algorithms," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 1429–1430.

[18] C. Dwork, "Differential privacy in new settings," in *Proc. 21st Annu. ACM-SIAM Symp., Discrete Algorithms Soc. Ind. Appl. Math.*, 2010, pp. 174–183.

[19] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.

[20] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for v-support vector regression," *Neural Netw.*, vol. 67, pp. 140–150, 2015.

[21] E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2013, pp. 545–553.

[22] C. Zeng, J. F. Naughton, and J. Cai, "On differential private frequent itemset mining," in *Proc. 39th Conf. Very Large Database*, Trento, Italy, 2013, pp. 1087–1098.

[23] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Washington, DC, USA, 2010, pp. 493–502.

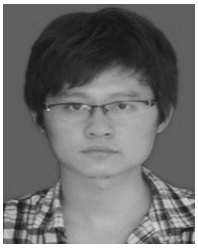
[24] Y. D. Li, Z. Zhang, M. Winslett, and Y. Yang, "Compressive mechanism: Utilizing sparse representation in differential privacy," in *Proc. 10th Annu. ACM Workshop Privacy Electron. Soc.*, 2011, pp. 177–182.

[25] X. J. Zhang, M. Wang, and X. F. Meng, "An accurate method for mining top-k frequent pattern under differential privacy," *J. Comput. Res. Develop.*, vol. 51, no. 1, pp. 104–114, 2014.

[26] C. Yin and J. Xi, "Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm," *Multimedia Tools Appl.*, vol. 76, pp. 16875–16891, Aug. 2017.

Chunyong Yin received the B.S. degree in computer science from the Shandong University of Technology, Zibo, China, in 1998, and the M.S. and Ph.D. degrees in computer science from Guizhou University, Guiyang, China, in 2005 and 2008, respectively.

He was a Postdoctoral Research Associate with the University of New Brunswick, Canada, in 2011 and 2012. He is a Professor and the Dean of the Nanjing University of Information Science & Technology, Nanjing, China. He has authored or coauthored more than 20 journal and conference papers. His current research interests include privacy preserving and sensor networking, machine learning, and network security. He is a member of ACM.



Jinwen Xi received the Bachelor's degree in computer science and technology from the Huaiyin Institute of Technology, Huai'an, China, in 2011. He is currently working toward the Master's degree in computer science and technology from the Nanjing University of Information Science and Technology, Nanjing, China.

His current research interests include applied cryptography and big data security and privacy.

Ruxia Sun received the B.E. degree in electrical engineering from the Shandong University of Technology, Zibo, China, in 1997.

She is an Associate Professor with the Nanjing University of Information Science & Technology, Nanjing, China. Her current research interests include cyber physical systems, machine learning, and mathematical modeling.

Jin Wang (M'11) received the B.S. and M.S. degrees in electronical engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2002 and 2005, respectively, and the Ph.D. degree in computer engineering from Kyung Hee University, Seoul, South Korea, in 2010.

He is currently a Professor with the College of Information Engineering, Yangzhou University, Yangzhou, China. His research interests mainly include routing algorithm design, performance evaluation, and optimization for wireless ad hoc and sensor networks. He is a Member of ACM.