IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2019, accepted June 29, 2019, date of publication July 8, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927386

A New Weight and Sensitivity Based Variable Maximum Distance to Average Vector Algorithm for Wearable Sensor Data Privacy Protection

ZHENJIANG ZHANG[1], (Member, IEEE), BOWEN HAN[2], HAN-CHIEH CHAO[3], (Senior Member, IEEE), FENG SUN[2], LORNA UDEN[4], AND DI TANG[5]

[1]School of Software Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China
[2]School of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Jiaotong University, Beijing 100044, China
[3]Department of Electrical Engineering, National Dong Hwa University, Hualien 97401, Taiwan
[4]School of Computing, FACS, Staffordshire University, The Octagon, Beaconside, Stafford ST4 2DE, U.K.
[5]The Third Research Institute of Ministry of Public Security, Shanghai 200031, China

Corresponding author: Zhenjiang Zhang (zhangzhenjiang@bjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation under Grant 61371071 and Grant 61801125, and in part by the Academic Discipline and Post-Graduate Education Project of the Beijing Municipal Commission of Education.

ABSTRACT The problem of privacy protection of wearable devices when publishing data can be solved based on the variable-maximum distance average vector. This paper proposes a new weight and sensitivity based variable maximum distance average vector (WSV-MDAV) method aiming to solve the problems that may be contained in the existing privacy protection algorithm. The proposed approach considers the difference of the importance among all the identifiers by setting corresponding weight coefficient W. Given a specific weight to each attribute in the table, we can subsequently get a distance metric based on weight. Similarly, the different sensitivity constraint S for different sensitive attributes is also available for our proposed method. Using the WSV-MDAV algorithm we propose a new privacy protection model for the data publishing of wearable device by introducing the concept of the differential privacy. The numerical results show that the proposed WSV-MDAV algorithm improves the privacy protection performance and reduces the information loss compared to the traditional method.

INDEX TERMS Wearable device, privacy protection, micro aggregation, differential privacy.

I. INTRODUCTION

Wearable technologies are networked devices that can collect data, track activities, and customize experiences to users' needs and desires. A wearable device is a portable device that is worn directly on the body or integrated into the user's clothes or accessories. It is not only a hardware device, but also a powerful function through software support, data interaction and cloud interaction. A wearable device can be understood as a natural ability based on the environment and human hands, language, walking, eye rotation ability, heart pulse beat ability, brain nerve thinking ability, etc. with the help of some built-in sensors, integrated chips, etc. to achieve information intelligent interaction and corresponding business function devices.

However, there are many challenges in the data publishing of the wearable devices, the center of which is the privacy and security. With the rapid use of wearable devices, wearable data is increasing exponentially [1], greatly increasing the risk of personal privacy information leakage. In the current technology context, the rapid rise of data mining technology makes it easier for people to obtain the information they want in their lives, based on different needs such as health care, scientific research, business planning, data sharing,

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Hui Yeh.

trend forecasting, and policy making. To fully exploit the value of wearable device data, many data owners such as governments, businesses, wearable device manufactures, and wearable device users are willing to share wearable device data with third parties for analysis and mining. Publishing and utilizing this data (also known as microdata) is an important process for advancing social development. In fact, this data can be used for scientific research, and mining. Analysis of many user data or business data is beneficial to better marketing, better social planning, and better social services. This information may be collected in a voluntary manner from user, the original intention of which is to provide users with a more personalized service and experience. Nevertheless, if the raw data collected by the wearable sensing device is directly released without any process, it will face the problem of personal privacy leakage. As is known, the data release privacy protection technology refers to the operation and processing of the original wearable device data, thereby protecting the privacy of the data and ensuring the availability of the data. The process of data release often faces privacy threats. It turns out that these threats are enough to trigger privacy leak. If people with ulterior motives use illegal means to collect data, when there is enough data, the integration mining analysis, tampering and utilization will cause unimaginable consequences.

Many approaches of protecting privacy have been proposed to protect the privacy of users [2]. Micro aggregation algorithm is an important way to anonymity micro data in data publishing. The optimal multivariable micro aggregation algorithm has been proved to be an NP problem. The MDAV algorithm is known as the fixed length micro aggregation algorithm, while the V-MDAV algorithm is a heuristic algorithm for multivariable micro aggregation. Compared to the fixed length micro aggregation algorithm MDAV, the V-MDAV algorithm can provide variable length groupings, giving better result from higher group homogeneity and similar computational load. However, the distance metric in V-MDAV algorithm has problems such as the lack of explicit calculation method, the sensitive attribute being vulnerable to homogeneity attack, background knowledge attack and similarity attack. To solve these problems mentioned above, we introduce the concept of weight on the premise that different quasi identifiers may have different importance. Two different weights are added before the corresponding quasi identifier in the distance metric of two tuples, that is, the individualized distance metric method based on the weight W-PDM.

Since the traditional (α, k) anonymous phase is more resistant to homogeneity attack than the common K anonymous model, we introduce (α, k) anonymous ideas into our micro aggregation algorithm. Because different sensitive attributes have different distributions, we set different constraints S for different sensitive attributes. By combining previous distance metric scheme based on weight, we propose the improved WSV-MDAV based on the variable length micro aggregation. By introducing the differential privacy to this model help to improve the ability of the privacy protection. Lastly, the correctness and functionality of the algorithm and model are verified by testing data.

The specific research contributions are as follow:

(1) A new personalized distance metric method is proposed based on the weight W, taking the personalized needs of different data providers into account as well as the computation of information entropy which considers the distribution of quasi identifier attributes.

(2) We introduce the idea of traditional (α, k) anonymous algorithm into micro aggregation algorithm because the sensitive attribute of V-MDAV algorithm is vulnerable to homogeneity attack, background attack and similarity attack, by proposing the calculation based on the sensitivity S of sensitive attributes. We set different non-related constraints for different sensitive attribute values in each sensitive attribute.

(3) Differential privacy is introduced to the WSV-MDAV privacy protection model for data publishing. Privacy protection is achieved by adding appropriate noise to the query or analysis result. The noise added into the data processed by micro aggregation is smaller than that into the original data, thus can further reduce the information loss and improve the availability of the data set.

(4) Our simulation results prove that our proposed model shows improved performance in privacy protection compared to the existing model. we also study the influence of parameters on the model.

The rest of this paper is organized as follows. We introduce the references corresponding to the privacy protection in Section II. In Section III, we present a new micro aggregation algorithm based on weight and sensitivity. The differential privacy is also introduced in this section. The simulation result of our model is presented in section IV. Finally, we discuss our conclusion in section V.

II. RELATED WORK

Wearables are slowly becoming more and more mainstream. It is predicted that wearable technology sector to hit $100 billion by 2023.

There are several Advantages of Wearable Technologies. Firstly, wearables allow us to stay connected. They can alert us of messages, incoming calls, emails, and much more without having to constantly be checking your phone. It can help you locate your phone when it is lost, or even connect to IoT enables devices in your home (such as switching on the lights or controlling the a/c temperature). Secondly, wearables can give us more accurate data. Thirdly, Wearable are set to make our lives safer and more efficient.

Although many researches are conducted on data publishing privacy protection among main stream devices, there is little research on privacy protection for data publishing of wearable devices. The key point of the privacy protection for data publishing of wearable devices is to limit the risk of disclosure to an acceptable level, while maximizing the efficiency, that is, the availability of data. This is difficult because there may exists a malicious or an attacker in

data receivers. A series of anonymity method for data publishing have been proposed in the field of statistical disclosure control (SDC) [3]. In privacy protection for data publishing, the problem is how to generate an anonymous data set suitable for public release is essentially a question of how to find a balance between information leakage risk and information loss. Publishing the original dataset can give the data user the highest availability, but may cause the leakage of the privacy for the dataset, or vice versa [4]. The k-anonymity, l-diversity and t-closeness are typical and commonly used among many privacy protection models for data publishing. The k-anonymity is the earliest privacy protection model, from which many other models are derived.

In k-anonymity model, even if an attacker gets some information about the target object from other channel from the background knowledge, he still can't determine the record of the target object in the data table, to achieve the privacy protection. Samarati and Sweeney were the first to propose the k-anonymity model in the literature. The main idea is that each record is the same as the other quasi identifier in the equivalence class. Later, Sweeney made a further revision for the anonymous implementation techniques (including generalization, concealment) in the literature [5]. At this point, k anonymity has become a simple and effective way of privacy protection and has received wide attention and research. From then on, many improved models or extension models have been proposed based on the k-anonymity. making up for the defects of k-anonymity. TM Truta, B Vinay, X Sun, eta, proposed the p-sensitive k-anonymity model in the literature [6], [7]; CW Wong proposed the (α, k) anonymous model [8], requiring that there are at least records in the equivalent class, with the frequency of each sensitive attribute in each equivalence class no more than α; and the (p, α)-sensitive k-anonymity is proposed in literature.

Sun *et al.* [9] proposed $((p_+, \alpha)$ -sensitive k-anonymity that consider the continuous exposure of location leasing to the privacy leak. Masoumzadeh and Joshi [10] proposed LBS (k, T)-anonymity; and Quoc and Dang [11] proposed eM^2 algorithm based on a Member Migration (MM) technique to satisfy the requirement of k-anonymity and reduce the information loss. Kisilevich et al. proposed KACTUS and KACTUS2 in two studies [12], [13] respectively. The two studies lead to the query model based on k-anonymity, expanding and improving the basic anonymous models as it reduces the risk of privacy leakage and improves the effectiveness of data. However, because k-anonymity is mainly based on the idea of the privacy protection that makes attackers undistinguishable to the target identifier in the same equivalent class, it cannot prevent similarity attacks or background knowledge attacks. To address this, Wang *et al.* [14], [15] proposed the confidence bounding model and JX Li *et al.* [16] proposed the (ε, m) model. However, all these models above are based on k-anonymity. These models only limit the quasi identities and pays no attention to limit the sensitive attributes. As a result, it is vulnerable to homogeneity attack and similarity attack.

Machanavajjhala put forward the l-diversity model, requiring that every sensitive attribute in each equivalence class has at least l ''very good'' attribute values, including: ① Distinct l-diversity; ② Probabilistic l-diversity; ③ Entropy l-diversity; ④ Recursive (c, l)-diversity [17]. However, l-diversity can't resist skew attacks, etc. while the optimal l-diversity is proved to be a NP hard problem. NH Li proposed the t-closeness model and (n, t)-closeness model [18], that have better effect on the protection of the sensitive attributes and stronger ability to resist homogeneity attacks with similar attack compared to the k-anonymity model and l-diversity model mentioned above. Based on the analysis and study of the t-closeness model and k-anonymity model in the literature, the (t, α)-closeness model is proposed for the lack of specific algorithm and the inability of semantic privacy to be customized for the t-closeness model [19]. Jordi proposed to achieve t-closeness through micro aggregation, improving the availability of data while reducing the impact of data outliers [20].

In addition, there are a number of studies on the release of wearable device location data, including privacy protection for single-site location and privacy protection for trajectories. Single-point location privacy protection technologies include area coverage, location spoofing, cryptograph, etc. Trajectory privacy protection includes coverage, mixed area, path confusion, and dumb trajectory. Among them, the research on regional coverage technology is most extensive. The most primitive area override generates an area that contains itself, making it impossible for the server to get the exact location of the user. However, the privacy protection provided by this approach is very limited. Therefore, the current mainstream area coverage technology uses k-anonymity as the principle to find an area that can cover at least k users, so that the server cannot distinguish which of the current requests is sent by which user. The idea of location spoofing to hide the real location in several fake locations is straightforward, but the communication overhead is high. The cryptography method protects the location privacy by using mathematical tools. In particular, the sever can be ignorant of the user's location. However, the high computational overhead and communication cost make this method impossible to use in large scale in practice.

III. A NEW VARIABLE LENGTH AND MICRO AGGREGATION ALGORITHM BASED ON WEIGHT

Traditional micro aggregation algorithm does not consider differences between attributes and is vulnerable to homogeneity attack. In this paper the authors propose a new variable length micro aggregation algorithm based on weight and sensitivity to solve the problem. Firstly, we put forward the concept of weight according to different attributes, setting different weight attributes. Next, using weight, we take different attributes to determine distance metric based on weight. Because the requirements for sensitivity differ from those in different attributes, we set different constraints for different

sensitivities to meet the requirements of different attributes for different sensitivities.

Lastly, we develop a new privacy protection model for data publishing of wearable devices using the variable length micro aggregation algorithm based on weight and sensitivity by introducing the concept of differential privacy. Evaluation of the model shows that it improves users' ability to protect their privacy.

A. A NEW WEIGHTED BASED ON PERSONALIZED DISTANCE METRIC METHOD

In k clustering, the data should be grouped according to the distance between the quasi identifier attributes of different records. Considering that the unit of measures for different attributes may lead to different result, some scholars present a method to normalize the distance. The main step is shown as follows:

(1) Compute the absolute deviation of the mean S_c.

$$S_C = \frac{1}{n}\left(|X_{1c} - m_c| + |X_{2c} - m_c| + \ldots + |X_{nc} - m_c|\right) \tag{3-1}$$

where n represents the amount of data, and $X_{1c}, X_{2c}, \ldots X_{nc}$ are the value of n data on attribute C respectively. m_c is the mean of the value of attribute C, that is,

$$m_c = \frac{1}{n}\left(X_{1c} + X_{2c} + \ldots + X_{nc}\right) \tag{3-2}$$

(2) Compute the normalized unit of measure.

$$Z_{ic} = \frac{X_{ic} - m - c}{S_c} \tag{3-3}$$

It shows greater robustness for isolated point when we take the absolute deviation of the mean into consideration rather than the standard deviation. Meanwhile, the Z-score of the isolated point won't be too small to detect.

After the normalization of the value of attributes, we can calculate the distance between different records. In the process of micro aggregation of the data table containing multiple attributes, Euclidean distance can be used to get the distance between the records with the weight we set to each attribute. The specific formula is as follows:

$$\text{dist}(X, Y) = \sqrt{\omega_1 |X_1 - Y_1|^2 + \omega_2 |X_2 - Y_2|^2 + \ldots + \omega_m |X_m - Y_m|^2} \tag{3-4}$$

where is m the number of quasi-identifier attributes.

Notice that the weight of the default attributes is equal in the traditional distance calculation scheme, and the individual requirements of different data providers for different quasi identifier attributes is not considered. In real life, different data publishing and different data distribution of different quasi identifier attributes often result in different importance of each attribute. Therefore, we propose a weight W based personalized distance metric for different quasi identifier attributes, according to their different data distribution and the corresponding weight defined by users themselves.

Next, we consider continuous data as an example. According to the importance of different quasi identifier attributes in the process of partitioning, we propose a micro aggregation algorithm based on weight W. The setting of the weight for each quasi identifier attribute should consider these two aspects:

(1) Firstly, it is individual oriented, that is, different data providers or wearable device users can assign values to the importance of the quasi identifier attributes and the degree of influence of their sensitive attributes. Thus, this relatively subjective weight assignment takes the personalized requirements of different data providers into account.

(2) Secondly, it is oriented to the publication data itself. The data distribution of the quasi identifier attribute in the data set often determines the amount of information it contains and the degree of influence on the sensitive attributes. The greater the entropy of the quasi identifier, the greater the uncertainty, and the greater the impact on the sensitive attributes, the greater the weight given to the attribute. Therefore, the weights assigned by the second method should be related to the distribution of the quasi identifier property and directly proportional to the information entropy. The resultant weight assignment is relatively objective.

When calculating the distance metric, we should consider these two factors simultaneously for the weight setting of a quasi-identifier attribute.

This paper mainly adopts two variables, ω_{ip}, representing the weight of a quasi-identifier attribute set by the data provider or the user and the weight set by the data publisher respectively. ω_{iu} is given the weight based on the understanding of different users for the importance and privacy of this attribute as well as whether the value of this attribute could lead to a sensitive attribute value leak. At the same time, the weight of all the quasi identifier attributes should be added up to 1; ω_{ip} is given the weight based on the entropy calculated by the data publisher according to the data distribution of this quasi identifier attribute. For any quasi identifier attribute Q_i, a formal definition of its weight ω_{Q_i} is given as shown in formula 3-1:

$$\omega_{Q_i} = \alpha\omega_{Q_{iu}} + \beta\omega_{Q_{ip}} \tag{3-5}$$

where $\omega_{Q_{iu}}$ represents the weight the user or the data publisher set for the quasi identifier attribute Q_i, given by calculating the entropy of, while α and β represent the weight set by the user and by the data publisher respectively. Both are set by default and can be calculated as shown in formula 3-2:

$$\omega_{Qiu} = \sum_{j=1}^{n} \frac{\omega_{Qiuj}}{n} \tag{3-6}$$

where n represents the number of data tables, that is, the number of the users or the data providers, and $\omega_{Q_{iuj}}$ represents the weight for the quasi identifier attribute Q_i set by the user.

Therefore, for the quasi identifier attribute Q_i and its information entropy H_{Q_i}, the weight set by data publisher needs to

meet the condition 3-3 based on personalized distance metric described above:

$$\begin{cases} 0 \le \omega_{\mathrm{Q_{ip}}} \le 1 \\ \sum_{i=1}^{m} \omega_{\mathrm{Q_{ip}}} = 1 \\ \omega_{\mathrm{Q_{ip}}} \propto H_{\mathrm{Q_i}} \end{cases} \tag{3-7}$$

They can be calculated through the formula 3-4:

$$\omega_{\mathrm{Q_{ip}}} = \frac{H_{\mathrm{Q_i}}}{\sum_{i=1}^{m} H_{\mathrm{Q_i}}} \tag{3-8}$$

where $\omega_{Q_{\mathrm{ip}}}$ is the weight of the quasi identifier attribute Q_i, and H_{Q_i} represents the information entropy of this at tribute while represents the number of the quasi identifier attributes.

As for the computation for the $\omega_{Q_{ip}}$, we introduce the concept of information entropy. We consider that if the source symbol has n values $V_1, V_2, \ldots, V_n$ with the probability $P_1, P_2, \ldots, P_n$ respectively. The appearance of various symbols is independent of each other. At this time, the average uncertainty of the source should be the statistical mean (E), that is the information entropy:

$$H(V) = E\left[-\log P_i\right] = -\sum_{i=1}^{n} P_i \log P_i \tag{3-9}$$

In the formula, the logarithm generally takes 2 as the base and the unit is the bit. Therefore, for the quasi identifier attribute Q_i, its information entropy H_{Q_i}, and the weight $\omega_{\mathrm{Q_{ip}}}$ set by the data publisher should satisfy the following condition:

$$\begin{cases} 0 \le \omega_{\mathrm{Q_{ip}}} \le 1 \\ \sum_{i=1}^{m} \omega_{\mathrm{Q_{ip}}} = 1 \\ \omega_{\mathrm{Q_{ip}}} \propto H_{\mathrm{Q_i}} \end{cases} \tag{3-10}$$

Thus, we can get $\omega_{Q_{ip}}$:

$$\omega_{\mathrm{Q_{ip}}} = \frac{H_{\mathrm{Q_i}}}{\sum_{i=1}^{m} H_{\mathrm{Q_i}}} \tag{3-11}$$

where $\omega_{Q_{ip}}$ is the weight of the quasi identifier Q_i, and H_{Q_i} is its information entropy. m is the number of the quasi identifier attribute.

B. VARIABLE LENGTH MICRO AGGREGATION ALGORITHM BASED ON SENSITIVITY

There are two main goals for privacy protection in data publishing. First is to protect identity information to prevent identity information leakage. The second is to protect sensitive attributes and prevent the leakage of sensitive attributes. The (α, k) anonymity model is based on the problems that may exists in the traditional k anonymity model, that is, traditional model is vulnerable to homogenous attacks and background knowledge attacks. The (α, k) anonymity model is defined as follows:

Definition 1 (α Non-Correlated Constraint): Given a dataset D, an attribute set Q, and a sensitivity attribute s. The value of s doesn't belong to the range S of the sensitivity attribute set Q. Assume that (E, s) represents the set containing s in equivalence class E. α is the threshold set by users, where $0 < \alpha < 1$. If s has a frequency less than or equal to α in each equivalence class, then the dataset D is α non-correlated to the attribute set Q and s. That is, for all the equivalence class E, we have $|(E, s)| \big/ |E| \le \alpha$.

Define 2 [(α, k) Anonymity Model):] If a data sheet is (α, k) Anonymity, then the data sheet is k Anonymity while satisfies α non-correlated constraint.

In the work [8] that proposed the (α, k) Anonymity model, it is achieved by local coding in generalization and coding, where local coding can disturb the value of the element of the element ancestor in dataset. For multidimensional data, the loss caused by generalization is too large and may lead to dimensional disasters. This paper proposes a new data publishing algorithm base on the sensitivity αi, combining the variable length micro aggregation algorithm V-MDAV and (α, k) Anonymity.

The limitation with Traditional (α, k) anonymity model is that there may exist in the k anonymity model, the problem of leakage of sensitive attributes caused by homogeneous attacks and background knowledge attacks. In traditional (α, k) anonymity, α is the threshold constraints set by the user. It is not easy to satisfy the same frequency constraint $\alpha$for all sensitive attribute's values. At the same time, it is important to consider that different sensitive attributes may lead to different requirements for the sensitive attributes after anonymity. Therefore, in real data publishing, it is more reasonable to set sensitive attribute constraints S for different sensitive attribute values in each sensitive attribute according to the distribution of different sensitive attributes. Next, we introduce the computation process of sensitivity constraint S of sensitive attribute in the data publishing table of single dimension sensitive attribute.

For data tables of single dimensional sensitive attribute values, assume that the data set should contain the following attributes $\{Q_1, Q_2, \ldots, Q_n, \mathrm{SA}\}$. Among them, $Q_1, Q_2, \ldots, Q_n$ is quasi identifier attribute, is sensitive attribute, and $\mathrm{SA}_1, \mathrm{SA}_2, \ldots, \mathrm{SA}_m$ is the value of sensitive attribute and is ordinal attribute. The value of sensitivity S should meet the following principles:

(1) Considering the influence of semantic factors, for $\mathrm{SA}_x, x \in [1, m]$, the closer to the extreme value of the sensitive attribute SA, the higher the sensitivity of SA_x, the smaller the uncorrelated frequency constraint S_i for this sensitive attribute. Because of the extreme values of attributes, especially sensitive attribute in data publishing tables, it tends to contain relatively large amounts of information.

(2) When the frequency of SA_x is higher in all the values of the attribute SA, the lesser the sensitivity of SA_x is, the greater the noncorrelation frequency constraint S_i for the sensitive

attribute. This is based on the distribution of data in sensitive attributes to set up S_i, just like the previous section which sets the weight according the entropy. The higher the frequency of SA_x, the less information contained in this value. Therefore, a large noncorrelated constraint S_i can be set.

(3) In each equivalence class, the product of S_i and k must be guaranteed to be greater than or equal to 1, which is the size of each equivalence class in the fixed length micro aggregation algorithm. Taking V-MDAV algorithm for example, it is necessary to ensure that $S_i \times k_i \geq 1$, where $k_i = |E_i|$, as well as $S_i \geq f_{\mathrm{SA_x}}$, which guarantees that the noncorrelated constraints S_i of the sensitive attribute value SA_x in each equivalent class are not less than the frequency $f_{\mathrm{SA_x}}$ of the sensitive attribute value SA_x in the whole data set, otherwise the (S_i, k) anonymity that satisfies the condition cannot generated.

Prove: assume that the number of equivalent classes divided by micro aggregation is n, and the number of sensitive attribute values SA_x is 1, which means that sensitive values SA_x in sensitive attributes exist only in one of the equivalent classes E_j, and the rest of the equivalent classes are not, that is, $\left|\left(E_{i=j}, \mathrm{SA}_x\right)\right| = 1$, $\left|\left(E_{i\neq j}, \mathrm{SA}_x\right)\right| = 0$.

For the equivalence classes $E_i, \forall i \in [1, n]$, there are definitions 3-5 based on noncorrelation constraints:

$$S_i \geq |(E_i, \mathrm{SA}_x)| \,/\, |E_i| \tag{3-12}$$

So, we can get the formula 3-6:

$$\begin{aligned} S_i \times k_i &\geq |(E_i, \mathrm{SA}_x)| = 0 \quad i \neq j \\ S_i \times k_i &\geq |(E_i, \mathrm{SA}_x)| = 1 \quad i = j \end{aligned} \tag{3-13}$$

Because the noncorrelation constraints S_i set for the same sensitive attribute value SA_x in different equivalent classes are the same. If $S_i < f_{SA_x}$, then $|\mathrm{SA}_x| \leq S_i \times |E| < f_{\mathrm{SA}_x} \times |E| = |\mathrm{SA}_x|$, which is contradictory, so $S_i \geq f_{\mathrm{SA}_x}$ should be guaranteed.

According to these three principles above, the values should meet 3-7:

$$\begin{cases} S_i \propto \dfrac{\min\{\mathrm{SA}_x - \mathrm{SA}_{\min}, \mathrm{SA}_{\max} - \mathrm{SA}_x\}}{\mathrm{SA}_{\max} - \mathrm{SA}_{\min}} \\ S_i \propto f_{\mathrm{SA}_x} \\ S_i \times k_i \geq 1 \\ S_i \times k_i \geq 1 \end{cases} \rightarrow \begin{cases} S_i \propto \dfrac{\min\{\mathrm{SA}_x - \mathrm{SA}_{\min}, \mathrm{SA}_{\max} - \mathrm{SA}_x\}}{S_{\max} - S_{\min}} \\ S_i \propto f_{\mathrm{SA_x}} \\ \max\left\{f_{\mathrm{SA_x}}, 1/k_i\right\} \leq S_i \leq 1 \end{cases} \tag{3-14}$$

Thus, the definition of can be given as:

When

$$\mu \frac{\min\{\mathrm{SA}_x - \mathrm{SA}_{\min}, \mathrm{SA}_{\max} - \mathrm{SA}_x\}}{\mathrm{SA}_{\max} - \mathrm{SA}_{\min}} + \nu f_{\mathrm{SA}_x} \geq \max\left\{f_{\mathrm{SA}_x}, \frac{1}{k_i}\right\},$$

then,

$$S_i = \mu \frac{\min\{\mathrm{SA}_x - \mathrm{SA}_{\min}, \mathrm{SA}_{\max} - \mathrm{SA}_x\}}{\mathrm{SA}_{\max} - \mathrm{SA}_{\min}} + \nu f_{\mathrm{SA}_x} \tag{3-15}$$

When

$$\mu \frac{\min\{\mathrm{SA}_x - \mathrm{SA}_{\min}, \mathrm{SA}_{\max} - \mathrm{SA}_x\}}{\mathrm{SA}_{\max} - \mathrm{SA}_{\min}} + \nu f_{\mathrm{SA}_x} < \max\left\{f_{\mathrm{SA_x}}, \frac{1}{k_i}\right\}$$

then,

$$S_i = \max\left\{f_{\mathrm{SA_x}}, \frac{1}{k_i}\right\} \tag{3-16}$$

Usually, μ and ν can be set to 1/2.

IV. A NEW DIFFERENTIAL PRIVACY MODEL FOR WEARABLE DATA PUBLISHING

According to the new variable length micro aggregation algorithm based on weight and sensitivity proposed above, a new privacy protection model for data publishing is proposed in combination with the related concepts of differential privacy.

A. DIFFERENTIAL PRIVACY MODEL

For all data sets D_1 and D_2, which differ only one piece of data (also known as adjacent datasets), Range (κ) is the range of random function κ, then this random function κ can provide ε differential privacy or ε-DP, and if all D_1, D_2 and $S \subset \mathrm{Range}\,(K)$ is satisfied [21]:

$$\Pr\left[\kappa\left(D_1\right) \in S\right] \leq \exp\left(\varepsilon\right) \Pr\left[\kappa\left(D_2\right) \in S\right] \tag{4-1}$$

It can be seen from the definition that differential privacy restricts the influence of any record on the output of the function κ. ε represents the level of privacy protection that the function can provide, which is usually very small, such as 0.1 or ln2. The smaller the ε, the closer the probability that the function κ gets the same output on the two data sets, the greater the level of privacy protection can be provided. Based on data distortion, the differential privacy add noise to the query result to protect the privacy. It is critical to decide the noise added to the result. Too much noise added to the result may lead to good protection performance with less data availability, or vice versa. Thus, we use sensitivity to evaluate the noise. We divide the sensitivity into Global Sensitivity, Local Sensitivity and Smooth Sensitivity.

Assume the function $f : D \rightarrow R^k$ with the dataset D as input. Its output is a k–dimensional vector. And the Global Sensitivity is defined as:

$$\triangle f = \max_{x,y \in D, d(x,y)=1} \| f\left(x\right) - f\left(y\right) \|_1 \tag{4-2}$$

where $d\,(x, y) = 1$ represents the situation that only one piece of data is different between dataset x and dataset y.

Global Sensitivity is determined by the function f itself, which measures the maximum change in the output of a function between adjacent datasets. Differential privacy can

be achieved by adding noise proportional to global sensitivity. The noise distribution such as Laplacian distribution and discrete Laplacian distribution can satisfy the condition.

Similarly, considering a specific dataset D and its arbitrary adjacent dataset, we define function $f : D \to R^k$ with the dataset D as input. The output is also a k-dimensional vector. The Local Sensitivity is:

$$\mathrm{LS}_f(D) = \max_{y:d(y,D)=1} \|f(y) - f(D)\|_1 \tag{4-3}$$

Apparently, the Global Sensitivity is the upper bound of Local Sensitivity. The Local Sensitivity is determined by the function f and dataset D.

Since the Local Sensitivity utilizes the dataset to obtain the distribution of the dataset to some extent, adding noise proportional to the Local Sensitivity to achieve differential privacy may reveal sensitive information in the dataset. At the same time, the Local Sensitivity of the adjacent dataset should also take into consideration, leading to the concept of Smooth Sensitivity to determine the amplitude of the noise along with the Local Sensitivity.

Given $\beta > 0$ and datasets D and D', the Smooth Sensitivity of the function f on D is defined as:

$$S_{f,\beta}(D) = \max_{D'} \left(\mathrm{LS}_f\left(D'\right) \exp\left(-\beta \cdot d\left(D, D'\right)\right)\right) \tag{4-4}$$

Notice that most of the research of the differential privacy concentrate on the function with smaller Global Sensitivity such as count, sum, and so on.

Besides, there is an interesting attribute in differential privacy which k-anonymity model doesn't have. The combination of several differential privacy model given different ε value still satisfies the requirement of differential privacy.

Lemma 3 (Serial Combination): Assume that κ_1 is a differential privacy algorithm satisfying ε_1-DP while κ_2 with parameter ε_2-DP. Then, the combination of the two algorithm satisfy the $(\varepsilon_1 + \varepsilon_2)$-DP.

Lemma 4 (Parallel Combination): Assume that κ_1 and κ_2 satisfy the requirement of ε-DP. If κ_1 and κ_2 is used to the datasets or the subsets without intersection, then the output of the combination (κ_1, κ_2) satisfies the requirement of ε-DP.

B. IMPLEMENTATION MECHANISM OF DIFFERENTIAL PRIVACY

In terms of numerical data, the mechanism of differential privacy can be regarded as the process of noise addition. The differential privacy returns the result with noise instead of the real query value. How much noise needs to be added depends on the change in the query function between adjacent datasets. The additional noise may be:

(1) independent of the dataset. The amplitude of the noise should be adjusted through the maximum change in the value of the function between adjacent datasets.

(2) dependent on the dataset. The additional noise should be adjusted based on the change of the query function in the actual datasets.

There are two main implementation mechanism of the differential privacy, that is, Laplace mechanism and exponential mechanism.

1) LAPLACE MECHANISM

In terms of numerical data, ε differential privacy should be achieved by add noise with Laplace distribution to the query result. we denote x-Laplace (μ, b) as the Laplace distribution with location parameter $\mu$and scale parameter b.

The process of adding noise is usually as follows: we set f as the query function with the actual output $f(X)$. $Y(X)$ is the random noise prepared to be added into $f(X)$. That is to say, $\kappa(X) = f(X) + Y(X)$ is the query result need to return to the user. Then, $Y(X)$ need to satisfy the Laplace distribution with location parameter 0, and scale parameter $\Delta(f)/\varepsilon$, where ε is the budget of differential privacy protection and $\Delta(f)$ is the Global sensitivity of the function f. Thus, the noise density function with Laplace distribution prepared to be added is:

$$p(x) = \frac{\varepsilon}{2\Delta(f)} e^{-|x|\varepsilon/\Delta(f)} \tag{4-5}$$

We can find that with the same ε, the higher $\Delta(f)$ is, the more noise with Laplace distribution need to add.

2) EXPONENTIAL MECHANISM

Limited by the fact that the Laplace mechanism is suitable for the numerical result only, exponential mechanism has significance when it comes to the entity object. We set that the input of the algorithm or the function f is the dataset D. The output is an entity object $r \in Range$ with output range multiplied by $f(D, r)$ is the usability function, where $\Delta(f)$ is the sensitivity of the function $f(D, r)$. If the algorithm or the function select to output r with the probability proportional to $\exp\left(\frac{\varepsilon f(D,r)}{2\Delta(f)}\right)$, then we consider that the function f provide ε differential privacy.

The essence of differential privacy is that "the presence or absence of a person's data in a dataset should not significantly change the analysis of the dataset." The limits on the risk of the information disclosure are very sufficient. If a person's data has a significant impact on the results of the analysis, then the person's privacy is likely to be at risk. Therefore, in essence, differential privacy ensures that data is adequately protected.

C. DATA PUBLISHING MODEL WITH DIFFERENTIAL PRIVACY

Differential privacy is an anonymized privacy model that provides more reliable privacy protection than previous models (such as k-anonymity and its extended model). Although differential privacy is popular among researchers and there are advances in privacy protection, in real-world applications, the application of differential privacy is still limited. The most basic reason is the poor accuracy and usability of differential privacy results. There are also many extended privacy models that seek to improve the accuracy of results by sacrificing their privacy guarantees. It is often overlooked that the avail-

ability of differential privacy output is quite limited. On the one hand, it is because of the need to achieve differential privacy protection to add noise to the data, and on the other hand, it is because only limited data types or limited queries available for performance. In contrast, a k-anonymous data publishing model has no assumption of any use of the data to be published, and doesn't limit the number and type of analysis that can be performed.

Research [22] shows that data availability can be improved without abandoning the strong privacy protection of differential privacy (from the perspective of distortion reduction). Based on this, the reduction of sensitivity depends on the number of k anonymity groups in the published data table: this value is related to the value and the cardinality of the data set. The larger the grouping size k, the lower the sensitivity of the class centroid generated by the micro aggregation; on the other hand, the smaller the data set, the smaller the number of different classes of centroids in the micro aggregated dataset. As the size of the grouping increases or the size of the dataset decreases, the sensitivity will be reduced. The smaller the noise added to the differential privacy, the higher utility the resulting differential privacy data has.

Thus, the differential privacy can be introduced into the privacy protection model for data publishing based on micro aggregation to reduce the noise which we need to add into the model to realize the differential privacy, making it more flexible and practical. The effectiveness of differential privacy of data publishing can be improved. So the model proposed in this paper introduces the micro aggregation algorithm WSV-MDAV based on weight W and sensitivity S to reduce the sensitivity of query function before performing differential privacy.

The process of differential privacy is to add noise to each record in the grouped data set $\overline{D}$ after the data table is processed by the variable length micro aggregation algorithm WSV-MDAV based on weight W and sensitivity S. The added noise conforms to the Laplace noise with a position parameter of 0 and a scale parameter of $\Delta f / \varepsilon$, in which Δf is the sensitivity of the query function, which is the distance between the two records of the farthest distance in the whole data table.

A complete privacy preserving model framework for wearable data publishing based on differential privacy is shown as follows:

TABLE 1. Privacy protection model for wearable data publishing based on differential privacy.

Generating set of data sets D_ε to meet ε-DP by micro aggregation

Function: Generation of an ε -DP data set D_ε via micro aggregation

input: D is the original dataset with n records,
M is the micro aggregation algorithm WSV-MDAV based on the weight W and sensitivity S_i.
$S_\varepsilon()$ is the process of adding noise to realize the differential privacy;
$Q_r()$ is a query function that returns the attribute value of r -th record in the data set.
output: the data set D_ε satisfying the ε -DP
Processing the original dataset:

$$\text{microaggregated data set } M(D) \to \overline{D}:$$

k **division process** :
Calculate the weight of each quasi identifier property ω_i ;
Calculate the distance metric $D_{n\times n}(i,j)$ of the dataset D, and calculate the sensitivity S_i of the sensitive attribute SA_i ;
Calculating the center point C of the data set;
while (If there are more than k records waiting to be allocated) **do**
Select a record e that is the farthest from the center of the data set;
With e as the center, select the record closest to e and start adding, forming a group g_i ;
while (the amount of group g_i is less than k) **do**
if the number of records with the same sensitive attribute value SA_i in g_i as χ
doesn't exceed $k \times S_i$ (round down) then
add χ into g_i ;
end if
end while
end while

Extension process:
For the remaining unallocated records, select the nearest grouping g_i from the record;
If the adding condition is satisfied, then, add it to group g_i . Otherwise, the group with the closest distance will continue to be searched except g_i .
If all the groups traversed are unable to satisfy the adding condition, then, it will be added into the nearest group.
Aggregation process:
For all groups, group is used to replace the quasi identifier of all data in the group;
Get the data table after the micro aggregation $\overline{D}$.
for $r = 1$ to n **do**
$S_\varepsilon\left(\mathrm{Q}\,\mathrm{r}\left(\overline{D}\right)\right) \to d_\varepsilon$
Insert d_ε into D_ε
end for
return D_ε
end function

V. SIMULATION RESULT AND ANALYSIS

A. COMMON SIMULATION PARAMETERS

In micro aggregation algorithm, the parameters used to measure the privacy protection performance of a model usually include information loss rate and privacy leakage risk [23].

1) INFORMATION LOSS RATE

In a micro aggregation algorithm, all records are divided into different groups (also called equivalent classes); the values of the quasi identifier attributes of the records in each group are replaced before the release of the group's centroid, by which the information loss is generated. Therefore, a good micro aggregation algorithm is to maximize the homogeneity of the

group to reduce information loss. That is to say, ''high in class aggregation and low coupling among classes''.

We set the original data set as X, and g_i is the group generated in the k partition process. The parameters that measure the equivalence of the inner class of the equivalence class represent the sum of squares of the distance between each record and the center of mass in the group, which can be measured by the intra class homogeneity GSE:

$$\mathrm{GSE}(g_i) = \sum_{j=1}^{ni} \mathrm{dist}\left(X_{ij}, \overline{X_i}\right) \tag{5-1}$$

where is the number of records in the group g_i, and $\overline{Xi} = \frac{1}{n}\sum_{j=1}^{ni} X_{ij}$. dist () uses the weight W based personalized distance measurement method W-PDM in the third chapter to calculate.

After micro aggregation, the homogeneity of all the data tables is the sum of all intra group homogeneity, which usually measured by SSE:

where g is the number of groups generated in the k partition process. The global homogeneity of data tables is measured by SST, representing the sum of global homogeneity of each group:

$$\mathrm{SST} = \sum_{i=1}^{g}\sum_{j=1}^{ni} \mathrm{dist}\left(X_{ij}, \overline{X}\right) \tag{5-2}$$

where $\overline{X}$ is the centroid of the original data table.

The information loss rate can be measured by IL (Information Loss):

IL (Information Loss) represents the rate of intra group homogeneity and intra group global homogeneity, reflecting the change before and after the micro aggregation.

2) RISK OF PRIVACY DISCLOSURE

The risk of privacy leakage can be measured by the ratio RL (Record Linkages) of the number of records in the original dataset to the number of records in the original dataset:

$$\mathrm{RL} = 100 \times \frac{\sum_{xj \in X} \Pr\left(x_j^{'}\right)}{n} \tag{5-3}$$

where n is the number of records in original dataset. And the probability RL (Record Linkages) of the linked records in the anonymous dataset can be calculated as follows:

$$\Pr\left(x_j^{'}\right) = \begin{cases} 0 & \text{if } x_j \notin G \\ \dfrac{1}{|G|} & \text{if } x_j \in G \end{cases} \tag{5-4}$$

where G is an anonymous data set after micro aggregation.

B. SIMULATION AND ANALYSIS

Because this paper is concerned with wearable device data, two wearable data sets are used. The first one is Statlog (Heart) Data Set in UCI Machine Learning Repository. This data set mainly records some data of the heart patient, including the numeric, ordinal, subtype, and two-value data, where we select 4 quasi identifier attributes, which are age, sex, resting blood pressure, maximum heart rate achieved, as well as a sensitive attribute chest pain type (4 values). The other is commonly used to simulate data publishing privacy protection model or data set, which is a part of Adult data set in UCI Machine Learning Repository. We take 4 quasi identifier attributes including numerical, ordinal and two value data, which are age, sex, education-num, hours-per-week, as well as a marital-status sensitive attribute, with a total of 7 values, Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. For the convenience of computing, the Adult data set is preprocessed before the simulation, and the sex (gender) attribute has two values. The female sex (gender) attribute value is set to 0, and the male sex (gender) attribute value is set to 1. The seven values of sensitive attributes are set to ordinal data with a value of 1-7 based on the frequency appearing in the data set. The system environment is Intel (R) Core (TM) i5-2450M CPU, 2.50GHz, 4.00G RAM, professional edition operating system, and the simulation tool is Mathworks Matlab R2014b.

Because the two original data sets are not evaluated by the importance of different attributes, Statlog (Heart) Data Set takes the average value of the weights set by volunteers for the four quasi identifier attributes as parameters in order to embody the user's personalized requirement in the process of calculating the weight of the paper. Which means $\omega_{Q_{1u}} = 0.2$, $\omega_{Q_{2u}} = 0.1$, $\omega_{Q_{3u}} = 0.3$, $\omega_{Q_{4u}} = 0.4$, For the Adult data set, the weights are $\omega_{Q_{1u}} = 0.2$, $\omega_{Q_{2u}} = 0.1$, $\omega_{Q_{3u}} = 0.4$, $\omega_{Q_{4u}} = 0.4$. In order to verify the effectiveness of the proposed algorithm and model, we simulated the traditional V-MDAV model and the WSV-MDAV model based on the weight W and sensitivity S of the differential privacy. When k takes on different values, the results of the information loss(IL) of two data sets are shown in figure 1 and 2.

As it can be seen from Figure 1 and Figure 2, whether it is a traditional V-MDAV algorithm or a micro aggregation algorithm WSV-MDAV based on weight W and sensitivity S, as the value increases, the loss of information will increase, that is, the decrease of data validity. From the formula 4-7, formula 4-8 and formula 4-9, we learn that SST is the same for the same original data table, and k represents the smallest number of records in each equivalent class in the process of micro aggregation. So the larger the k is, the larger the equivalence class is, the lower the homogeneity of the class, as well as the larger the SSE with the greater the information loss IL. When the k value is the same, it can be found in Figure 1 that when $k < 5$, the information loss of the traditional algorithm is slightly lower, and after $k > 5$, the greater the k, the more obvious the increase of the information loss of the traditional V-MDAV algorithm. The proposed WSV-MDAV algorithm based on weight W and sensitivity S has no obvious increase in information loss but tends to be stable.

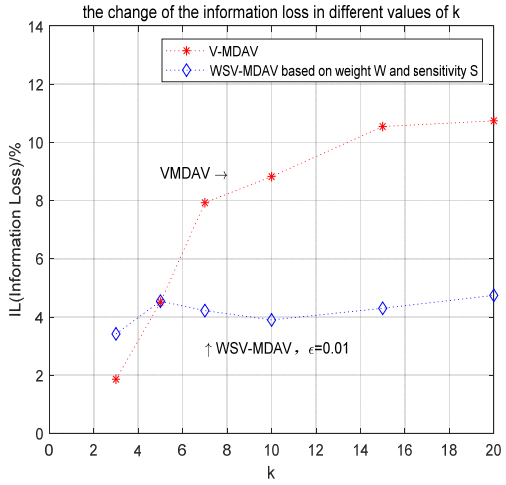


FIGURE 1. Comparison of information loss in Heart dataset.

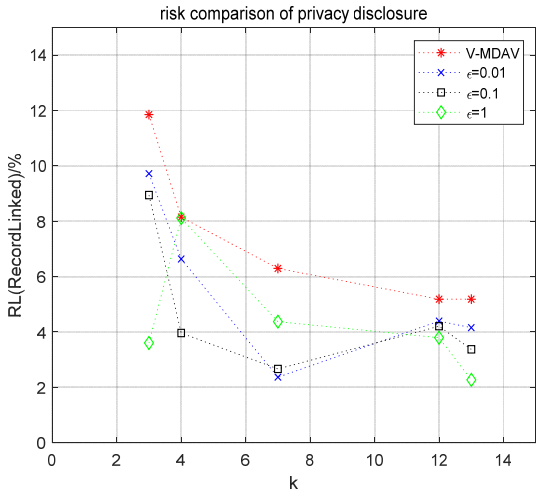


FIGURE 3. Comparison of privacy disclosure risk in Heart dataset.

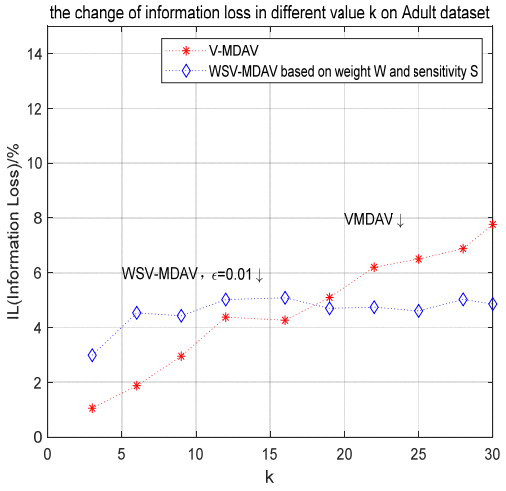


FIGURE 2. Comparison of information loss in Adult dataset.

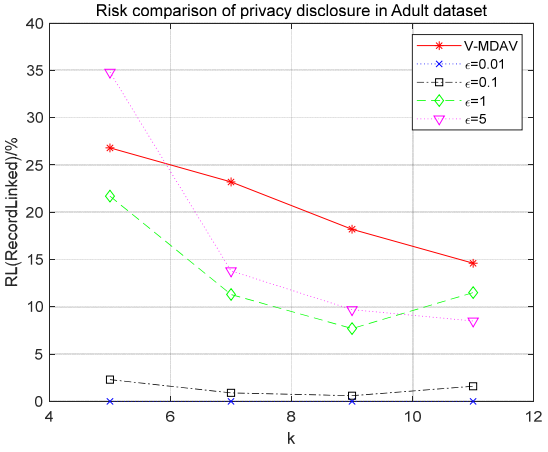


FIGURE 4. Comparison of privacy disclosure risk in Adult dataset.

Therefore, we can draw the conclusion that compared with the traditional V-MDAV algorithm, the weight W and sensitivity S based WSV-MDAV algorithm proposed in this paper has smaller information loss and more efficient data, especially when the k value is larger and tends to be stable.

The following is a comparison of the risk of privacy disclosure, which is mainly a simulation of the RL parameters that can be linked to the original data table after the data is processed, which measures the privacy protection capability of the model. For the differential privacy wearable data release model proposed in this paper, the parameter of the differential privacy in the first data set can be set $\varepsilon = 0.01$, $\varepsilon = 0.1$, $\varepsilon = 1$ respectively, while $\varepsilon = 0.01$, $\varepsilon = 0.1$, $\varepsilon = 1$, $\varepsilon = 5$ in the second data set. The traditional V-MDAV algorithm and the improved algorithm of the above ε values are simulated respectively. The comparison of privacy leakage risk is shown in figures 3 and 4.

As shown in Figure 3 and Figure 4, whether it is a traditional V-MDAV algorithm or a differential privacy based clustering algorithm WSV-MDAV based on weight W and sensitivity S, the record probability RL of linking to the original data table decreases with the increase of values, that is, the risk of privacy disclosure is reduced, and the privacy preserving ability of the model is improved. According to the principle of micro aggregation and $k$anonymity, as well as from formula 4-10, and formula 4-11, it is known that k represents the smallest number of records in each equivalent class in the process of micro aggregation. The larger the k, the larger the equivalent class is, the smaller the RL, the lower the risk of privacy disclosure, so the better the privacy protection ability of the model. At the same time, the sensitivity S proposed in this paper limits the distribution of sensitive attributes in each equivalent class and reduces the probability of the leakage of sensitive attributes. Therefore, the privacy protection capability of the model is further improved.

It is also demonstrated in Figure 3 and Figure 4 that the smaller the same $k$value, the smaller the ε, the smaller the probability of the RL link to the original table, the higher

the privacy disclosure risk of the model, the higher the privacy protection ability of the model. In our simulation, it is found that the sensitivity of query function is used as the global sensitivity of adding Laplace noise. Because of the micro aggregation processing, the weight based personalized distance measurement is introduced, which is more reasonable for distance measurement between records, thus reducing the sensitivity of the inquiry function and the noise added to the differential privacy.

Therefore, we can draw a conclusion: compared with the traditional V-MDAV algorithm, the weight W and sensitivity S based micro aggregation algorithm WSV-MDAV and the ε differential privacy data publishing model have lower privacy disclosure risk when the k value is the same, and the smaller the parameter ε of the differential privacy, the better the power of the model's privacy protection. With the increase of k value, the privacy protection ability of the model will also be improved.

VI. CONCLUSION

In this paper, a micro aggregation algorithm WSV-MDAV based on weight W and sensitivity S is proposed to solve the problem of the micro aggregation algorithm V-MDAV used in the data publishing process, and a complete privacy protection model for wearable equipment data release is formed by introducing differential privacy after the micro aggregation processing. The improved model and algorithm are simulated to verify its effectiveness.

Although there are improvement using the improved model and algorithm, there are problems still to be addressed. That need further research and improvement. The following are some of the shortcomings identified and suggestions for future research.

(1) The use of sensitivity S_i in the model proposed in this paper is limited to single dimension sensitive attributes. How to strengthen the extensibility of the model so that it can be applied to a variety of sensitive attributes is the main research work in the future.

(2) The data publishing privacy protection in this paper is based on static data publishing. There exist many researched on the privacy protection of the centralized static data release, but researches on dynamic data publishing and joint data publishing are less [24]. Therefore, the existing privacy protection models and algorithms need to be changed. Further studies are needed to address the dynamic data publishing.

(3) This paper focuses only on data privacy in data-oriented privacy protection. With the development of wearable devices, the data of wearable devices are becoming more and more abundant. It is important in future to consider privacy protection of environment-oriented privacy such as location, routing information and so on.

(4) For the published data set, we have to know how to correctly select the attribute of the identifier for the published data set; how to guarantee the privacy of the data query, that is, to query the published data set, and to ensure that the privacy is not leaked in the query. Though many solutions have been proposed by some scholars, there are still many problems exist that need further research

(5). Many of the privacy protection technologies stay at the theoretical level. It is important that we look to apply these technologies effectively as part of future research.

ACKNOWLEDGMENT

The authors also thank Weihao Dong, Kun Liu for their contribution to this manuscript.

REFERENCES

[1] J. P. Daries, J. Reich, J. Waldo, E. M. Young, J. Whittinghill, D. T. Seaton, A. D. Ho, and I. Chuang, "Privacy, anonymity, and big data in the social sciences," *Commun. ACM*, vol. 57, no. 9, pp. 56–63, 2014.

[2] Z.-J. Zhang, C.-F. Lai, and H.-C. Chao, "A green data transmission mechanism for wireless multimedia sensor networks using information fusion," *IEEE wireless Commun.*, vol. 21, no. 4, pp. 14–19, Aug. 2014.

[3] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf, *Statistical Disclosure Control*. Hoboken, NJ, USA: Wiley, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/978-1118348239

[4] G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Metayer, R. Tirtea, and S. Schiffner, "Privacy and data protection by design—From policy to engineering," 2014, *arXiv:1501.03726*. [Online]. Available: https://arxiv.org/abs/1501.03726

[5] L. Sweeney, "Achieving K-anonymity privacy protection using generalization and suppression," *Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, 2002.

[6] T. M. Truta and B. Vinay, "Privacy protection: P-sensitive K-anonymity property," in *Proc. Data Eng. Workshops*, Apr. 2006, p. 94.

[7] X. Sun, H. Wang, J. Li, and D. Ross, "Achieving P-sensitive K-anonymity via anatomy," presented at the 6th IEEE Int. Conf. e-Bus. Eng. (ICEBE), Oct. 2009. [Online]. Available: http://works.bepress.com/xiaoxun_sun/21/

[8] 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," presented at the 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Philadelphia, PA, USA, 2006. [Online]. Available: http://eprints.usq.edu.au/2094/1/Alpha-KDD06.pdf

[9] X. Sun, H. Wang, J. Li, T. M. Truta, and P. Li, "(p^+, α)-sensitive k-anonymity: A new enhanced privacy protection model," in *Proc. IEEE Int. Conf. Comput. Inf. Technol. (CIT)*, Jul. 2008, pp. 59–64.

[10] A. Masoumzadeh and J. Joshi, "An alternative approach to κ-anonymity for location-based services," *Procedia Comput. Sci.*, vol. 5, no. 1, pp. 522–530, 2011.

[11] P. H. Van Quoc and T. K. Dang, "eM^2: An efficient member migration algorithm for ensuring κ-anonymity and mitigating information loss," in *Proc. 7th VLDB Conf. Secure Data Manage.,* in Lecture Notes in Computer Science, vol. 1. Singapore, 2010, pp. 26–40.

[12] S. Kisilevich, Y. Elovici, B. Shapira, L. Rokach, "kACTUS 2: Privacy preserving in classification tasks using κ-anonymity," in *Protecting Persons While Protecting the People*. Berlin, Germany: Springer, 2009, pp. 63–81.

[13] S. Kisilevich, L. Rokach, Y. Elovici, B. Shapira, "Efficient multidimensional suppression for K-anonymity," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 334–347, Mar. 2010.

[14] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, 2006, pp. 414–423.

[15] K. Wang, B. C. M. Fung, and P. S. Yu, "Handicapping attacker's confidence: An alternative to κ-anonymization," *Knowl. Inf. Syst.*, vol. 11, no. 3, pp. 345–368, 2007.

[16] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Vancouver, Bc, Canada, 2008, pp. 1–13.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond κ-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, Apr. 2006, p. 24.

[18] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 943–956, Jul. 2010.

[19] X. Z. Huang, *An Enhanced t-Closeness Privacy Protection Method*. Harbin, China: Harbin Engineering Univ., 2012.

[20] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "T-closeness through microaggregation: Strict privacy with enhanced utility preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 3098–3110, Nov. 2015.
[21] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megías, "Individual differential privacy: A utility-preserving formulation of differential privacy guarantees," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 6, pp. 1418–1429, Jun. 2017.
[22] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, "Improving the utility of differentially private data releases via κ-anonymity," in *Proc. IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, vol. 1, no. 1, Jul. 2013, pp. 372–379.
[23] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based κ-anonymity," *VLDB J.*, vol. 25, no. 5, pp. 771–794, 2014.
[24] S. Goryczka, L. Xiong, and B. C. M. Fung, "M-privacy for collaborative data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2520–2533, Oct. 2013.

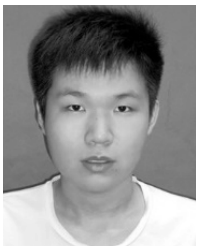
ZHENJIANG ZHANG received the Ph.D. degree in communication and information systems from Beijing Jiaotong University (BJTU), Beijing, China, in 2008, where he has been a Professor, since 2014. He is currently the Vice Dean of the School of Software Engineering, BJTU. He has published more than 70 professional research papers. His research interests include cognitive radio, communication protocols, and wireless sensor networks.

BOWEN HAN received the B.E. degree in electronic and information engineering and the master's degree from Beijing Jiaotong University, in 2015 and 2018, respectively. In 2016, he joined China Electric Power Research Institute. His major is in communication and information systems. His research interests include spectrum sensing technology in cognitive radio and privacy protection in wireless communication.

HAN-CHIEH CHAO (SM'04) received the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1989 and 1993, respectively.

He is currently a Professor with the Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan. He has been authored or coauthored five books, and has published about 400 refereed-professional research papers. His research interests include high-speed networks, wireless networks, IPv6-based networks, digital creative arts, e-Government, and digital divide.

Dr. Chao is a Fellow of the IET (IEE). He was an Officer of Award and Recognition from the IEEE Taipei Section, from 2010 to 2012.



FENG SUN received the bachelor's degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, in 2017, where he is currently pursuing the Ph.D. degree in communication engineering. His research interests include edge computing and deep learning.

LORNA UDEN received the Ph.D. degree from Staffordshire University, U.K. Her research interests include learning technology, web engineering and technology, activity theory, big data, innovation, knowledge management, semantic web, web services, big data, service innovation, social media, intelligent transport systems, the Internet of Things (IOT), and problem-based learning (PBL).

DI TANG received the B.E. degree in telecommunication engineering from Xidian University, and the Ph.D. degree in electrical engineering from Michigan State University. Since 2015, he has been with The Third Research Institute of the Ministry of Public Security, where he is currently an Associate Research Fellow. His research interests include wireless sensor network security, privacy-preserving communications, and vehicle network security.