Hindawi
Security and Communication Networks
Volume 2018, Article ID 4945152, 8 pages
https://doi.org/10.1155/2018/4945152

Research Article

A Clustering K-Anonymity Privacy-Preserving Method for Wearable IoT Devices

Fang Liu and Tong Li

College of Computer, National University of Defense Technology, Changsha, China

Correspondence should be addressed to Fang Liu; liufang06@gmail.com

Received 6 October 2017; Accepted 7 December 2017; Published 28 January 2018

Academic Editor: Zhen Liu

Wearable technology is one of the greatest applications of the Internet of Things. The popularity of wearable devices has led to a massive scale of personal (user-specific) data. Generally, data holders (manufacturers) of wearable devices are willing to share these data with others to get benefits. However, significant privacy concerns would arise when sharing the data with the third party in an improper manner. In this paper, we first propose a specific threat model about the data sharing process of wearable devices' data. Then we propose a K-anonymity method based on clustering to preserve privacy of wearable IoT devices' data and guarantee the usability of the collected data. Experiment results demonstrate the effectiveness of the proposed method.

1. Introduction

Wearable technology is one of the greatest applications of the Internet of Things. For the past few years, wearable devices have seen an explosive growth of popularity [1]. Corresponding to such advancement, more sensors are available to record various aspects of our daily lives [2], influencing our lives in an unconscious way.

However, security problems appear with the wide deployment of wearable devices. The most severe threat would be the privacy leakage of wearable devices data. After collecting data from the smart terminals, data holders (manufacturers) of wearable devices are willing to share the data with application developers to enrich their services or obtain monetary benefits. Typically, the data collected by these devices contain abundant privacy information [3, 4]. In addition, when sharing the data recorded by human-carried wearable sensors, some personal information, such as age, height, and weight, may also be submitted under warrant [5]. Therefore, though the original intention of data sharing is always positive, the uncontrolled personal information may raise the risk of privacy disclosure.

To balance the benefit of data sharing and the risk of privacy disclosure, we emphasize that it is critical to share data in a privacy preserved way. The privacy issues of wearable devices have been raised. Previous researches attempt to limit the privacy disclosure mainly by establishing several rigid laws [6]. Further, researchers have formulated some restrict standards to share personal data collected from wearable devices. These rules advocate that users are authorized to determine with whom their data would be shared and supervise the application of their contributed data. The third party must be supervised during the disposal of collected data. Although these laws play an important role in preserving privacy, there are still some vulnerabilities. It is hard to expect these rigid laws to prevent tricky adversaries. On the other hand, encryption and identity authentication are usual ways to preserve privacy. Although these methods have been proved to be effective in many cases, they are impractical in sharing data with the third party whose identity is uncertain [4].

Sharing these data anonymously seems to be a better choice. K-anonymity is a successful method to share data for its briefness and effectiveness. The data collected by wearable devices are always identifiable [7], which would be a severe threat against K-anonymity. Fortunately, certain characters of the dataset could be processed with K-anonymity to enhance the security level of users' privacy. In this paper, we introduce a clustering based K-anonymity method as the building block of privacy preserving for wearable devices contributed data. The clustering K-anonymity would assign similar records to the same equivalent set, while the similarity among these

records makes it harder to discriminate different identities than before. The notable contributions of this paper could be summarized as follows:

(1) We analyze the potential vulnerabilities of existing privacy-preserving methods for shared wearable devices data.

(2) We propose a threat model to achieve deanonymity against K-anonymity technique. Besides, we point out the vulnerabilities of the technique and improve it by referring to the inherent characteristics of wearable devices data.

(3) We evaluate the effectiveness of the proposed method with simulation experiments.

The rest of this paper is organized as follows. In Section 2, we introduce current researches on privacy preserving of wearable devices data. In Section 3, we introduce the attack model against the vulnerabilities of previous anonymity methods. In Section 4, we describe the clustering K-anonymity to solve the privacy problem. In Section 5, we discuss the performance of our improved method. Finally, we discuss and conclude the paper in Sections 5.3 and 6.

2. Related Work

The rapid growth of wearable devices provides a massive scale of personal data, which would usually be gathered by the data holder. In some cases, data holders need to share data with others without compromising the privacy and keep its usability at the same time. In this section, we summarize existing methods about data sharing from two aspects, namely, privacy preserving for wearable devices data and anonymous sharing, respectively.

2.1. Privacy Preserving for Wearable Devices Data. There have been several studies about privacy preserving of wearable devices data. Current data holders of wearable devices protect the users' privacy mainly by some rigid rules. As Figure 1 shows, the data collected by wearable devices are typically stored in a database owned by the data holder. The third party who wants to acquire the users' data must get the permission of users at first. Users of wearable devices determine whether to share their personal data, and they are authorized to trace the use of their personal data. The third party must conform to the users' willing, and they could not violate these rules. The intention of the third party must be honest.

These rules play an important role in preserving the privacy, but there are several vulnerabilities. On one hand, we cannot guarantee that authentication works well. If the Access Control is bypassed by someone malicious, the users' privacy would be disclosed. On the other hand, most users of wearable devices are unprofessional, they could not understand the significance of their data, and their consciousness about privacy preserving is poor, making them vulnerable to potential attacks.

2.2. Anonymous Sharing. Encryption is a widely adopted and traditional way to preserve privacy, while it is designed for

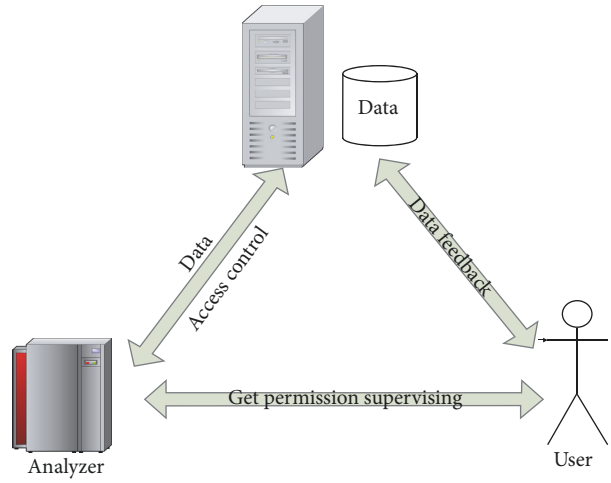

Figure 1: An overview of the privacy-preserving rules.

data sharing. The drawback of encryption in data sharing is secret key distribution as we cannot guarantee the reliability of the third party. Hence, preserving privacy by encryption is impractical. It is critical to find a way to preserve the privacy even when the malicious third party has acquired the data.

Differential privacy is an excellent method to preserve the privacy even when the overall background knowledge has been disclosed. The conception of such method is to preserve the privacy by adding moderate noises [8]. However, we note that this method would influence the usability of data with the involvement of noises.

In 2002, Sweeney proposed K-anonymity [9]. K-anonymity requires each record to be indistinguishable from at least $K - 1$ other records in quasi-identifiers domain. In the K-anonymity model, three compositions of each record are defined: (i) Attributes that clearly identify individuals, such as Social Security Number and ID Number, are defined as identifiers. (ii) Insensitive attributes that are combined to jointly identify individuals, such as name, sex, age, and Zip, are defined as quasi-identifiers. (iii) Attributes that are considered sensitive, such as salary and illness, are called sensitive attributes. In this paper, the sensitive attribute is the time serial collected by wearable devices. For convenience, we use ID, QI, and SD to represent identity, quasi-identity, and sensitive data, respectively, in this paper.

K-anonymity is an appropriate approach to share data anonymously. According to the principle of K-anonymity, data holders cut the linkage between ID and SD before sharing, in which case the ID information could not match with the SD information accurately. Further, such processing would cause no information loss in the SD information. However, inherent vulnerability of K-anonymity determines that the naive K-anonymity could not meet our requirement. In the next section, we discuss the threat model against K-anonymous data sharing of wearable devices.

3. Threat Model

In order to preserve wearable devices' data privacy, we should learn about the threat of the privacy first. In this section, we first discuss the link attack; then we introduce the process of achieving deanonymity of the sensitive data. We introduce the detail process of privacy disclosure in Section 3.3 and narrate the whole threat model with an illustrative example in Section 3.4.

3.1. Link Attack. Leaving alone the privacy requirement, datasets that need to be shared are always composed of several quasi-identifiers and sensitive data, without the ID information. Therefore, we could define the structure of all the records in the form $\{QI_1, QI_2, \ldots, QI_n, SD\}$, while the background knowledge could be denoted as $\{ID, QI_1, QI_2, \ldots, QI_n\}$. Such kinds of information could be acquired by gathering other insensitive information. We combine these pieces of information through the quasi-identifier domain $\{QI_1, QI_2, \ldots, QI_n\}$ and then obtain information in the form of $\{ID, QI_1, QI_2, \ldots, QI_n, SD\}$. This result indicates the privacy is disclosed. Figure 2 gives an example of link attack while $n = 3$. In Figure 2, every identifier points to a sensitive data unit, so some privacy information within the sensitive data could be relinked back to a specific identity. As a result, sensitive data of specific identities are disclosed.

3.2. Deanonymity. In Section 3.1, we introduce the link attack briefly. The link attack could be well addressed by the K-anonymity. However, the threat would be more severe if sensitive data were identifiable. In the process of link attack, the adversary combines the background knowledge $\{ID, QI_1, QI_2, \ldots, QI_n\}$ with the shared dataset $\{QI_1, QI_2, \ldots, QI_n, SD\}$ by quasi-identifiers $\{QI_1, QI_2, \ldots, QI_n\}$. Data holders could generalize quasi-identifiers $\{QI_1, QI_2, \ldots, QI_n\}$ according to the principle of K-anonymity to prevent privacy disclosure against link attack. In cases where sensitive data are identifiable, K-anonymity would be hard to preserve privacy. K-anonymity is designed with little consideration about this form of privacy disclosure.

Wearable devices' data might be identifiable (e.g., GPS data, or data collected by triaxis accelerators). It is obvious that the data such as GPS is identifiable. According to the different traces of people, it would be easy to infer a user's identity. The data collected from triaxis accelerators seem insensitive, but they could be applied to discriminate the identity by means of machine learning. There have been several researches about recognizing one's identity by the data collected from triaxis accelerators [9–11].

Recognizing identities with the machine learning methods is a critical threat to the privacy. One may argue that if there are a large number of people, such attempt would be too complex to be practical. However, the link attack could be used here to shrink the data scope.

3.3. Whole Threat Model. In Sections 3.1 and 3.2, we introduce the link attack and describe deanonymity of the wearable device data. The whole process would be described as follows:

(1) Collect the shared data $\{QI_1, QI_2, \ldots, QI_n, SD\}$.

(2) Gather the background knowledge $\{ID, QI_1, \ldots, QI_n\}$.

(3) Link $\{QI_1, QI_2, \ldots, QI_n, SD\}$ and $\{ID, QI_1, QI_2, \ldots, QI_n\}$ to find the equivalent set ES which contains the ID of objective.

(4) Recognize the identity of each person in ES by ID through machine learning method.

(5) Rebuild the correspondence between ID and SD.

After the processing, the correspondence between users' identities and their sensitive data is rebuilt, unavoidably resulting in the disclosure of privacy. The threat model is shown in Figure 3.

3.4. An Example of Privacy Disclosure. For example, as Table 1 shows, Alice is an owner of a wearable device, and the manufacturer of the device collects the data produced by this device and the information about her age, height, and weight. Then the data holder shares a dataset (as Table 1 shows) which contains Alice's data. The adversary Evil gets this information, and he knows that Alice is 181 cm and 71 kg and of the age 24, so that Evil could get Alice's sensitive data readily by combining the dataset with the background knowledge.

The data holder cuts the linkage between identity and sensitive data by generalizing the quasi-identifiers before sharing according to K-anonymity. Table 2 shows the 2-anonymity result of Table 1. In Table 2, it would be hard to recognize Alice's identity with link attack. However, the data contained in SD could still disclose the identity of Alice. Specifically, if we extract proper feature of these data and put it into a suitable classifier, the identity could be recognized.

Figure 4 shows the discriminating rate. We divide 14 subjects according to the sequence directly. Each equivalent set contains 2 records. Obviously, the discriminating rate could indicate the severe threat of privacy disclosure.

4. The Proposed Clustering K-Anonymity Scheme

In our work, we try to adjust the division of records, making it hard to discriminate the identity within each equivalent set, thus preserving privacy. We find that, in the dataset of wearable devices, quasi-identities are always relevant to the sensitive data. For example, the dataset about GPS contains the quasi-identities about address, and the dataset about the gait contains the quasi-identities such as height, age, and weight. In this section, we try to assign such records with similar quasi-identifiers to the same equivalent set. Because of the relevance between SD and QI, it would be harder to recognize a specific identity in equivalent set than before.

We clarify the meaning of clustering K-anonymity in Section 4.1 and describe the details about clustering K-anonymity in Sections 4.2 and 4.3.

4.1. Meaning of Clustering K-Anonymity. K-anonymity is a general conception to share data in a privacy-preserving way. Dataset could be divided into several equivalent sets

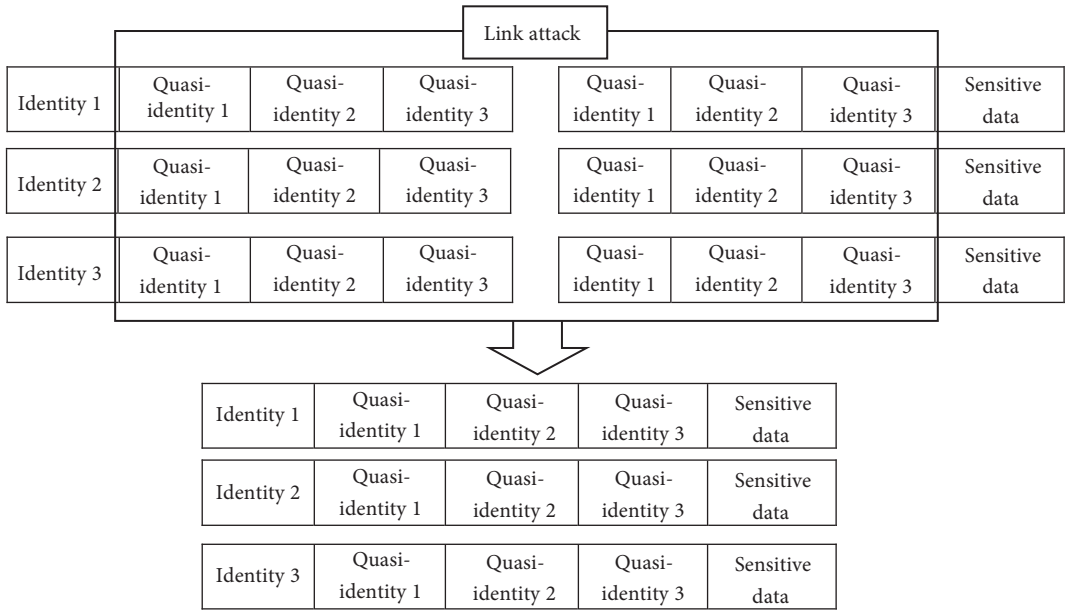

Figure 2: A general view of the link attack.

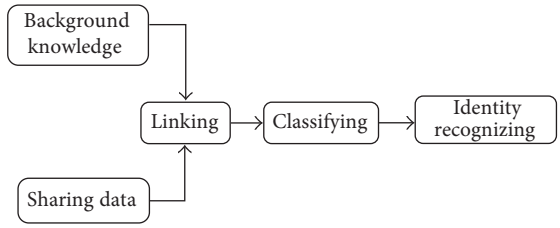

Figure 3: An overview of the threat model.

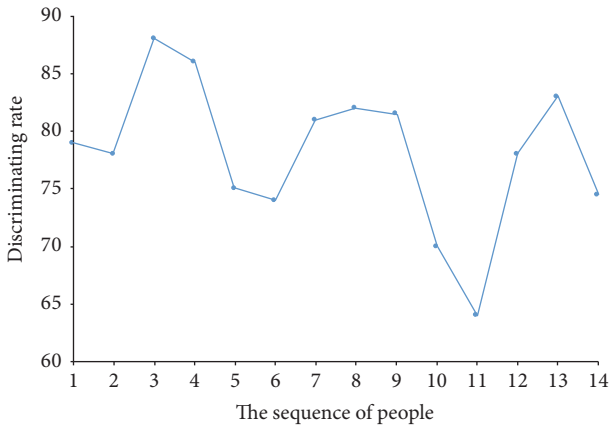

Figure 4: The discriminating rate of identity.

according to K-anonymity, and each set contains at least K and no more than $2K$ records. However, different division produces different effects on security.

In this paper, for preserving users' privacy, we expect the records in the same equivalent set to be as similar as possible.

Table 1: Original data.

Height	Weight	Age	Sensitive data
181	71	24	Time serials
183	75	23	Time serials
170	61	24	Time serials
175	70	31	Time serials

Table 2: Anonymity result of original data.

Height	Weight	Age	Sensitive data
18*	7*	2*	Time serials
18*	7*	2*	Time serials
17*	**	**	Time serials
17*	**	**	Time serials

* represents an anonymous character. ** represents two anonymous characters.

We find that QI of the shared datasets are usually closely related to the SD. For example, the dataset that contains GPS data may share address information, and the dataset about the data collected from triaxis accelerators shares the information such as age, height, and weight together. In these kinds of datasets, the quasi-identifiers are Zip code, age, height, and weight. We process the dataset with clustering and group records with similar quasi-identifiers in the same equivalent class. The rationale is that it is easy to discriminate the identity of people for the huge trace differences among different people, while if we cluster similar records by quasi-identifiers (e.g., address information), the differences would be reduced. Given more similarity between records in one equivalent class, there are fewer risks of privacy disclosure.

4.2. Distance Metric. The similarity between two records determines the division of datasets directly. There are detail descriptions about all kinds of data in dataset in [12–14]. All of these works try to transfer nonnumeric data into numerical value for further processing. Without loss of generality, we consider the case that all the data are numerical values.

In this paper, the similarity of two records is calculated by measuring the distance between two records. Intuitively, a larger distance indicates a smaller similarity, and vice versa. Let the quasi-identity domain of records X and Y be $\{X_1, X_2, \ldots, X_n\}$, and $\{Y_1, Y_2, \ldots, Y_n\}$, respectively. X_i, Y_i denote the ith quasi-identity. $\{K_1, K_2, \ldots, K_n\}$ denote the weights, where K_i is the weight of the ith quasi-identifier. The distance $d(X, Y)$ between records X and Y can then be defined as

$$d(X, Y) = \sum_{i=1}^{n} \left((X_i - Y_i) \times K_i \right)^2. \quad (1)$$

4.3. Details of the Clustering K-Anonymity. In this section, we discuss the details of the clustering K-anonymity in Algorithm 1 and then analyze its time complexity.

At first, we cluster the records in private table which need to be published, and assign similar records to the same equivalent set. Then, we unify the quasi-identifiers in the same clusters by generalizing and suppressing operations. The output of this algorithm is a table that satisfies the principle of K-anonymity. All the records in the same equivalent set are similar to each other. In this way, it would be harder to recognize the users' identities in one equivalent set; the privacy of these subjects would be more secure. We show the effectiveness of our method in Section 5.2.

The process of clustering K-anonymity algorithm is shown in Algorithm 1.

4.4. Time Complexity Analysis. Although the clustering K-anonymity well preserves the users' privacy, its feasibility should be further verified.

In the clustering phase, the time complexity of the operation that selects the nearest tuple should be $O(n)$, and the time complexity of the operation that selects the farthest tuple should be $O(n)$, so the overall complexity of clustering operation should be $O(n^2)$. In unifying phase, we check all equivalent sets in the dataset at first and then check each tuple in the equivalent set. Obviously, the overall time complexity in the unifying phase should be $O(n^2)$. So the time complexity of this algorithm should be $O(n^2)$. Such time complexity demonstrates that clustering K-anonymity could be achieved within finite time.

5. Performance Evaluation

In this section, we evaluate the performance of clustering K-anonymity, mainly considering the performance on security. We focus on the data collected by the triaxis accelerator for its popularity [9–11, 15], and its insensitive impression. Experiment results would verify the effectiveness of the clustering K-anonymity.

5.1. Experiment Settings. In this experiment, to show the effectiveness of clustering K-anonymity, we compare 4 kinds of K-anonymity. The 4 methods are Partial Datafly K-anonymity, Overall Datafly K-anonymity [16], μ-Argus K-anonymity [17], and clustering K-anonymity, and they are different in the division of datasets. We want to demonstrate that the division of dataset could influence the security of privacy.

The measurement of distance $d(X, Y)$ between two records X and Y is a critical factor to influence the final result. Here, we define the distance $d(X, Y)$ as follows:

$$d(X, Y) = (X_1 - Y_1)^2 + \left((X_2 - Y_2) \times \frac{1}{5} \right)^2 + (X_3 - Y_3)^2, \quad (2)$$

where X_1, X_2, and X_3 denote the age, the height, and the weight information, respectively. These attributes are quasi-identifiers in the dataset. We determine these parameters according to several rounds of experiment results. This group of parameters is effective in influencing the final result. Note that if we adopt a more accurate model instead, we would get more accurate result.

In this experiment, we achieve deanonymity with the data collected from triaxis accelerator sensors. The goal of this method is to preserve privacy, so a lower discriminating rate within one equivalent set suggests a higher security performance. We show the experiment results in Section 5.2.

5.2. Comparative Results and Analysis. Figure 5 shows the discriminating rate of the identities in each equivalent set. The dataset is divided according to the principle of 2-anonymity. It is clear that the discriminating rate of clustering 2-anonymity is relatively lower than other 2-anonymity. We can thus claim that the clustering 2-anonymity is the most secure method among the four considered methods.

Figures 6, 7, 8, and 9 show the results of 3-anonymity achieved by 4 methods mentioned above, respectively. These figures show the discriminating rate of the identities. It is obvious that the discriminating rate distribution of clustering 3-anonymity tends to be lower than other methods. More than half of the discriminating rates of the clustering 3-anonymity are lower than 60%, while, for the other methods, most of discriminating rates are more than 60%. The result demonstrates that clustering 3-anonymity is more secure than others.

On the other hand, the clustering K-anonymity brings no change to the sensitive data domain, so the usability of sensitive data could be guaranteed.

Analysis. In this experiment, the SD of all the records stay invariant. Because of the different combination about the equivalent set, the discriminating rate in each equivalent set would be different. Reasonable assignment of records improves security level of clustering K-anonymity.

5.3. Discussion. In this section, we discuss some interesting open research issues.

INPUT:
the data table need to be published: PT;
the anonymous parameter: K;
Quasi-identifiers: $QI_1, QI_2, \ldots, QI_n$;
OUTPUT:
the data table GT satisfies K-anonymity;
the list of quasi-identifiers: LF;
the list of records with similar quasi-identifiers: F;
the list LO as the list of L;
(1) **Repeat**
(2) Calculate $d(c, r)$ between c and other records r;
(3) Cluster the $K - 1$ nearest records with c;
(4) Remove these records form LF, and add these records to L;
(5) Select the farthest records c' from c to be the new core;
(6) Add L to LO, then clear L.
(7) **until** $|LF| < K$
(8) The remained records in LF are assigned to the nearest cluster in LO;
(9) Unify each cluster in LO by Generalizing;
(10) Create GT, QI from LO, and create SD from PT;
(11) **return** GT

Algorithm 1: Clustering K-anonymity.

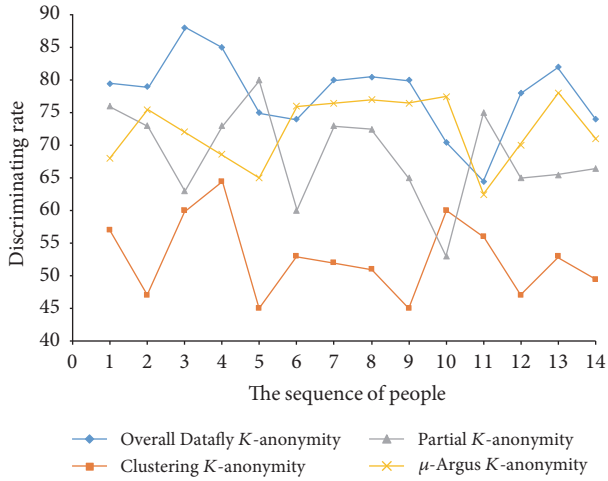

Figure 5: The discriminating rate of 2-anonymity.

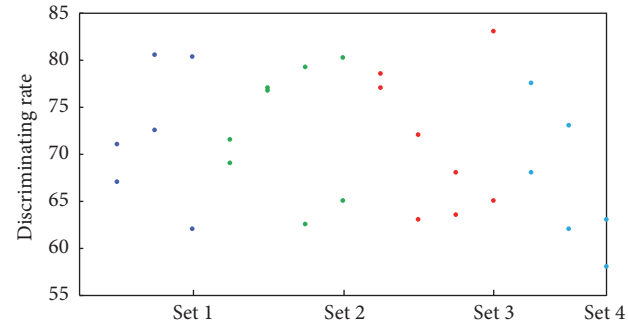

Figure 6: The discriminating rate of Overall Datafly 3-anonymity.

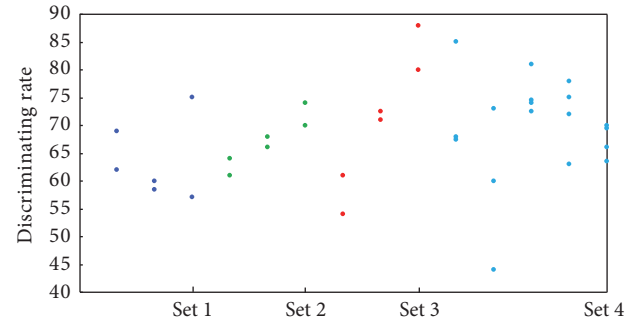

Figure 7: The discriminating rate of Partial Datafly 3-anonymity.

Relevance between the Sensitive Attributes and Quasi-Identifiers. In this paper, we propose the method based on the correlation between the quasi-identifiers and sensitive attributes. Nevertheless, in some cases, there is little correlation between the sensitive attributes and the quasi-identifiers; the method proposed in this paper would not be applicable.

Limitation of Distance Definition. The definition of distance is important as it would influence the final result. We simply use the Euclidean distance to calculate the distance between two records without accurate mathematical model. We could further introduce an accurate mathematical model to measure the distance.

Multiple Influence Factors. In this paper, we propose to evaluate the quasi-identifiers influence on sensitive attributes. In fact, there are many factors influencing the sensitive attributes beside quasi-identifiers, such as theindividuality,

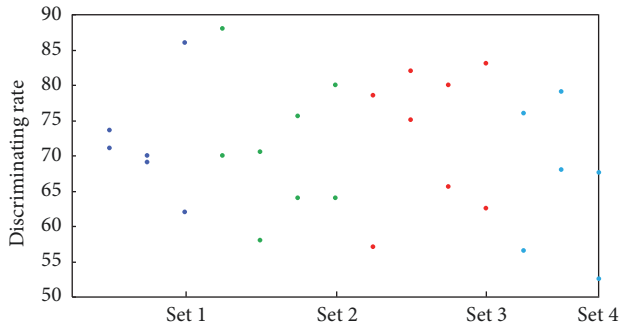

Figure 8: The discriminating rate of μ-Argus 3-anonymity.

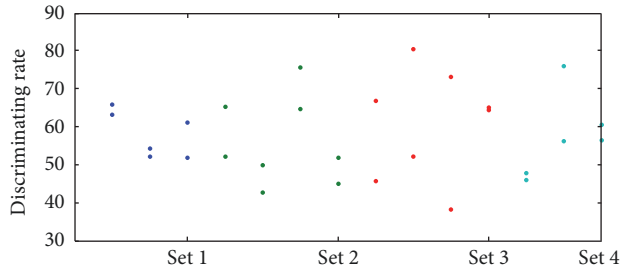

Figure 9: The discriminating rate of clustering 3-anonymity.

the characteristics, and the gender. These factors sometimes even have stronger influence on the sensitive attributes. In our method, we ignore these kinds of factors, but the accuracy would still be somewhat influenced.

6. Conclusion

While K-anonymity and its improvement projects protect against identity disclosure, it does not provide sufficient protection against sensitive attributes disclosure, especially when sensitive attributes are identifiable. This paper proposes clustering K-anonymity, which requires records with similar quasi-identifiers to be assigned to the same equivalent set. We take advantage of the relevance between the quasi-identifiers and sensitive attributes. The Euclidean distance with parameters is calculated to measure the similarity between two records, wherein the parameters are determined according to the actual requirement. The experiment results demonstrate the effectiveness of our method.

Disclosure

Fang Liu is currently affiliated to School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China under Grant no. 61379145 for their support of this work.

References

[1] Statista, "Forecasted value of the global wearable devices market from 2012 to 2018 (in billion U.S. dollars)," http://www.statista.com/statistics/259372/wearabledevice-market-value/.

[2] T. Yan, Y. Lu, and N. Zhang, "Privacy disclosure from wearable devices," in *Proceedings of the the 2015 Workshop*, pp. 13–18, ACM, Hangzhou, China, June 2015.

[3] P. Casale, O. Pujol, and P. Radeva, "Personalization and user verification in wearable systems using biometric walking patterns," *Personal and Ubiquitous Computing*, vol. 16, no. 5, pp. 563–580, 2012.

[4] C. Cornelius, R. Peterson, J. Skinner, R. Halter, and D. Kotz, "A wearable system that knows who wears it," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*, pp. 55–67, ACM, New York, NY, USA, June 2014.

[5] M. Zhang and A. A. Sawchuk, "USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of the International Conference on Ubiquitous Computing (UbiComp '12)*, pp. 1036–1043, ACM, September 2012.

[6] S. Safavi and Z. Shukur, "Conceptual privacy framework for health information on wearable device," *PLoS ONE*, vol. 9, no. 12, article e114306, 2014.

[7] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 6669, pp. 289–296, 2011.

[8] N. Mohammed, X. Jiang, R. Chen, B. C. M. Fung, and L. Ohno-Machado, "Privacy-preserving heterogeneous health data sharing," *Journal of the American Medical Informatics Association* , vol. 20, no. 3, pp. 462–469, 2013.

[9] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[10] H. K. Ali and D. I. G. Amalarethinam, "Activity recognition with multi-tape fuzzy finite automata," *International Journal of Modern Education and Computer Science*, vol. 5, no. 5, pp. 60–65, 2013.

[11] Ó. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[12] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[13] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra, "Comparing sdc methods for microdata on the basis of information loss and disclosure," in *Proceedings of the ETK-NTTS 2001*, vol. 2316, Luxemburg, Luxemburg, 2001.

[14] L. Fang, K. Dai, W. Zhiying, and C. Zhiping, "Research on the technology of quantitative security evaluation based on fuzzy number arithmetic operations," *Fuzzy Systems & Mathematics*, vol. 18, no. 4, pp. 122–126, 2004.

[15] A. Hundepool and L. Willenborg, "And-argus: software for statistical disclosure control," in *Proceedings of the Third International Seminar on Statistical Conference*, 2015.

[16] J. Peng, C.-J. Tang, W.-Q. Cheng, B.-M. Shi, and S.-J. Qiao, "A hierarchy distance computing based clustering algorithm," *Chinese Journal of Computers*, vol. 30, no. 5, pp. 786–795, 2007.

[17] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system," in *Proceedings of the AMIA Annual Fall Symposium*, vol. 4, pp. 51–59, The American Medical Informatics Association, 1997.