Protecting Privacy for Big Data in Body Sensor Networks: A Differential Privacy Approach

Chi Lin[1,2], Zihao Song[1,2], Qing Liu[1,2], Weifeng Sun[1,2](✉), and Guowei Wu[1,2]

[1] School of Software, Dalian University of Technology, Dalian, China
{c.lin,wfsun}@dlut.edu.cn
[2] Key Laboratory for Ubiquitous Network
and Service Software of Liaoning Province, Dalian, China

Abstract. As a special kind of application of wireless sensor networks, Body Sensor Networks (BSNs) have broad perspectives especially in clinical caring and medical monitoring. Big data acquired from BSNs usually contain sensitive information, which are compulsory to be appropriately protected. However, previous methods overlooked the privacy protection issue, leading to privacy violation. In this paper, a differential privacy protection scheme for big data in body sensor network is proposed. We introduce the concept of dynamic noise thresholds which makes our scheme more suitable for processing big data. It can ensure privacy during the whole life cycle of the data, which makes privacy protection for big data in BSNs promising. Extensive experiments are conducted to outline the merits of our scheme. Experimental results reveal that our scheme has higher level of privacy protection. Even in the case where the attacker has full background knowledge, it still provides sufficient ambiguity, which ensures being unable to match people based on the ECG data characteristic so as to preserve the privacy.

Keywords: Body sensor networks · Big data · Differential privacy

1 Introduction

As a special application system tailored for body health caring, the widespread use of Body Sensor Networks (BSNs) make it possible for realizing real-time monitoring [1–4]. In BSNs, big data which directly or indirectly reveal a person's physical condition are collected, aggregated and transmitted. For example, continuous ECG data are used in diseases tracking [13] and physiological condition monitoring [14].

Data transferred throughout BSNs usually share two critical features that should be paid attention to: sensitivity and vulnerability. Sensitivity means that data can reflect the status of the body, location or other sensitive information, which are needed to be appropriately preserved. Vulnerability indicates that data are forwarded in an open environment, therefore, data contents are easily to be trapped by the enemies or attackers. Once the data in BSNs are not well preserved, the privacy will be destroyed, leading the unexpected exposure of

S. Guo et al. (Eds.): CollaborateCom 2015, LNICST 163, pp. 163–172, 2016.
DOI: 10.1007/978-3-319-28910-6_15

physical information. Hence, preserving the privacy of sensitive data in BSNs is necessary and urgent.

In recent years, great efforts have been made for Big Data Privacy Protection (BDPP) in BSNs. In general, previous approaches fall into three kinds: (1) data collection protection [5–7], (2) data releasing protection [8–10] and (3) data analysis protection [11,12]. These approaches indeed provided effective methods for strengthening the privacy protection level in BDPP. However, they still suffer from some limitations:

- With respect to data accuracy, traditional methods added noise and implemented anonymous scheme, which sacrificed the accuracy of the data. Data transmitted under such conditions will be no longer accurate, which thus directly influences the data availability.
- With respect to data privacy protection, in traditional methods, such as k-anonymity, l-diversity or m-invariance etc., the values of k, l or m are required, which are not easy to be quantified or determined.
- With respect to data analysis [15], for example, in data water marking technology, only the static data set are well preserved, which are not feasible to be used for protecting dynamic and massive data. Therefore, such technology is not suitable for protecting data privacy in BSNs.

Researching the BDPP mechanisms in BSNs has significant contributions to our human life. Once big sensitive data are not properly preserved, users' privacy will be greatly damaged. Our motivation is to develop a method for solving the privacy protection problem based on differential privacy techniques. In our scheme, we proposed a protection model which covers the entire life cycle of data based on differential privacy technique.

The contributions of this paper can be summarized as follows:

1. To the best knowledge of the authors, this is the first time that differential privacy scheme is applied in protecting sensitive big data in BSNs. We propose a differential privacy protection model, which significantly reduces the risk of privacy exposure while ensuring data availability and accuracy.
2. To ensure the feasibility of the ECG big data, we introduce the concept of dynamic noise thresholds. It is used to interpret the relationships between noise size and data set size, which makes our scheme more suitable for big data.
3. To verify the advantages of our scheme, extensive experiments are conducted. Experimental results demonstrate that our scheme has a better performance in privacy protection. Even when the attacker grasps full background knowledge, our scheme still provides enough interference, making it unable to target a certain person by eavesdropping sensitive big data.

The remainder of this paper is structured as follows. Section 2 gives a brief overview on the preliminaries of this paper. Section 3 presents our scheme in detail. In Sect. 4, extensive experiments are conducted to show the advantages of our scheme. Finally, we conclude this paper and point out the future work in Sect. 5.

2 Preliminary

In this section, related preliminaries are introduced for comprehending our idea.

2.1 Problem Description

Traditional BDPP methods (as shown in Fig. 1) focused on data releasing and data mining issues. Once adversaries successfully eavesdrop sensitive data, data will be completely exposed, even worse, totally lose their effects. Simple encryption mechanisms cannot cover the entire life of data protection including data generating, data collecting, data transmitting and data publishing and so on. An effective alternative is to encrypt data during the transmission progress. However, in that case, data will be exposed to the public, which are prone to attacks. Once data are hacked, the real identity of the owner will be recognized, leading to privacy leakage. Although obfuscation scheme is widespread regarded as a useful method in privacy protection, the availability and accuracy are compromised.

Therefore, we develop a privacy protection scheme based on differential privacy, which appropriately solves the above concerns.

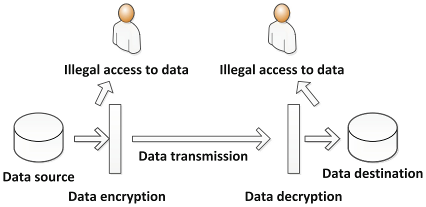

Fig. 1. Traditional privacy protections for BSNs

2.2 Identification and Data Characteristics of Electrocardiogram

Identification technology of ECG (electrocardiogram, ECG) has two main directions: one is extracting identification based on ECG feature points, the other is analyzing ECG waveform. As shown in Fig. 2, for a typical ECG, a cardiac cycle is made up of a P-wave, QRS wave, T wave and U waves. Usually, U waves are relatively small and inconspicuous.

In extracting identification, ECG feature points, such as P-wave width, QRS wave width, the period of PQ and QT, the R-wave peak and T-wave peak are regarded as important sources of information.

3 Our Scheme

In this section, we develop a secure way to minimize the cost to hide the ECG data. Here, we introduce the idea of differential privacy into privacy protection for big sensitive data in BSNs.

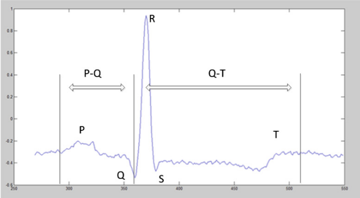

Fig. 2. Identification technology for ECG

A strict privacy protection model, defined by the differential privacy, will be considered as a standard of this program. To the best knowledge of the authors, this is a new direction that applies differential privacy technique into periodic stream data for safeguarding against malicious attacks. In our work, we choose the peak and valley data as characteristics points and put forward a simple example to clarify our method.

At first, we need to pre-process ECG data. We assign weights to each characteristic based on its importance for identification, for example, assuming that the identification significance of the peak value is greater than the valley value, thereby, peak value's weight will be bigger. Then bigger weights will be set to the parent nodes to construct the feature classification tree. This classifying method ensures that, ECG data can be divided into different equivalence classes. Suppose that the number of eigenvalue in ECG is n, then the final number of equivalence class in ECG features will be 2^n. Data that are divided into the same equivalence class will share approximate characteristic.

According to the concept of differential privacy, we introduce Eq. (1) as a strict standard.

$$\Pr[K(D1) \in S] \leq \exp(\varepsilon) * \Pr[K(D2) \in S] \tag{1}$$

Here, $D1$, $D2$ are two adjacent data sets, K is privacy protection algorithm for our design, $Pr[]$ is denoted as the risk of privacy exposure.

Taking big data identification into considerations, in our work, data are processed sequentially. There is no doubt to use the counting query with the sensitivity of count 1. We choose to take the variance reflecting data fluctuations as Pr. Assuming that the overall variance is influenced heavily, then apparently this set of data is clearly unexpected to this group of data. We thereby choose the variance of the data as a reflection of private exposure risk, which is intuitive and convenient.

$$\Pr[x]_n = s^2 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2}$$

So far, we have already illustrated the basis for achieving protection based on differential privacy. We believe that an ideal noise generation must be able to change as the data changes. For example, when the number of data is $2n$, the number of noise data, in fact, must be associated with the existing $2n$ data.

When the amount of data is n, the minimum value of the intensity available noise will be determined, which fluctuates with the change of existed data. We take peak values of characteristics as examples, the calculation method of dynamic characteristics of interference threshold is computed by Eq. (3).

$$H(x) = \left| \frac{\sum_{i=1}^{n-1} x_{n-1}}{n-1} - x_n \right| \times \ell \tag{3}$$

Here, $H(x)$ stands for the interference threshold when the nth data arrive. x is denoted as the data in the data set, ℓ is the interference correction value of this characteristic, which depends on specific features. Its purpose is to enable the real-time updates of data and provides a standard for dealing with the next arrival data.

Under the premise of a great amount of data, we can avoid duplicating the operations of the aforementioned data to increase the viability of ECG protection scheme. Based on the above definition of thresholds and limitations of differential ideas, we can find a sufficiently large enough one to (1) prevent the loss of privacy, and (2) carry out a devastating impact on data availability. Meanwhile, we add values at the point in feature noise, which is normally distributed with a typical value of $H(x)$ and the variance of 1.

Due to the limits of the private difference protection model, only a small amount of noise will be generated. As for the data in the same equivalence class, order processing cannot be carried out with the regular order, we choose the data from the beginning of the n data. Since we need the frontal $n-1$ data to initialize the feature interaction threshold value. Therefore, the processing order is shown in Fig. 3.

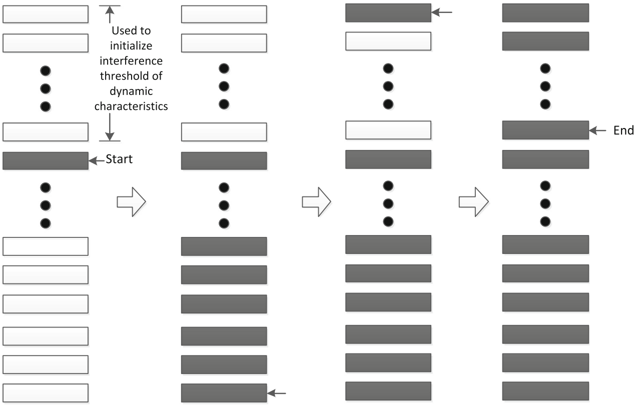
Used to initialize interference threshold of dynamic characteristics
Start
End

Fig. 3. Data processing order

In summary, our ECG privacy algorithm proceeds as Fig. 4. The first step is to scan ECG data set and extracting feature vectors. Then, characteristics classification tree is constructed and we divide ECG features to equivalence classes according to a feature vector. After that, we partition equivalence into classes and initialize feature interference thresholds. Next, we process the data one by one and add noise according to the important features. As the purpose of differential privacy is to maintain the available data as much as possible. In case of failure in adding noise, we additionally put new violence in the data. After completing categorizing data into an equivalence class, our algorithm is completed and the privacy of data is protected.

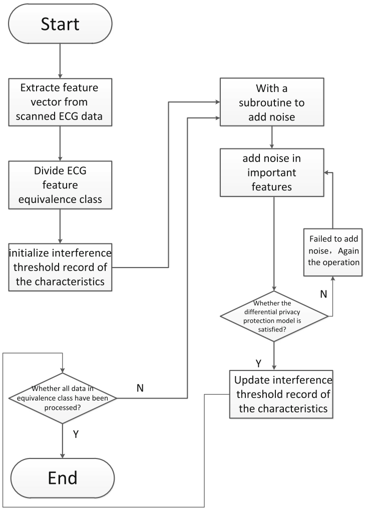
Start
Extracte feature vector from scanned ECG data
Divide ECG feature equivalence class
initialize interference threshold record of the characteristics
With a subroutine to add noise
add noise in important features
Failed to add noise，Again the operation
N
Whether the differential privacy protection model is satisfied?
Y
Update interference threshold record of the characteristics
Whether all data in equivalence class have been processed?
N
Y
End

Fig. 4. ECG privacy algorithm process

4 Experimental Results

4.1 Experimental Setup

We collected experimental data using Shimmer [16] as shown in Fig. 5.

In our experiment, we collect 30 sets of ECG data by using Shimmer. The average size of each data is 4.3 MB, the size of the post-processing data is 2.2 MB. Each data contains sufficient cardiac cycles for determining characteristic point values. Two features, peaks and valleys, are selected and the requirement of privacy budget is set 0.6. The result of equivalence class partitioning is shown in Table 1.

Fig. 5. Collecting data by Shimmer [16]

Table 1. Equivalence class partitioning

Equivalence class	Data bulk
Equivalence class I	6
Equivalence class II	8
Equivalence class III	11
Equivalence class IV	5

4.2 Privacy Protection Experiments

In this section, several experiments are conducted to show the performance of our scheme.

As shown in Fig. 6, it is obvious that the difference between data before and after processing is big. After applying our differential privacy protection scheme, such differences are migrated. Figure 7 shows the comparison of data 6 before and after treatment, which indicates that our scheme produces enough interferences. This ensures that feature points generated by our scheme will not reveal the real identity of a certain user so as to achieve privacy protection.

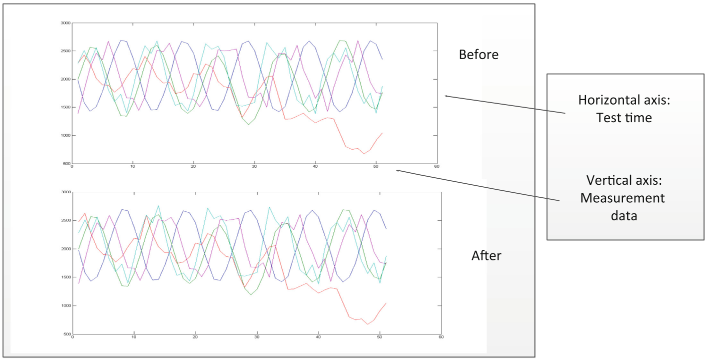

Fig. 6. Contrast before and after processing

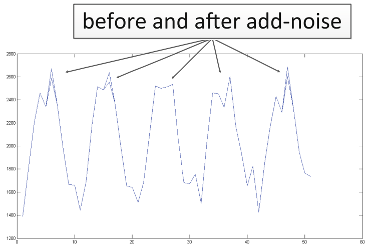

Fig. 7. Contrast before and after adding noises

To illustrate the effectiveness of our privacy scheme, we model and introduce an attacking model. We intend to demonstrate that our scheme is capable of defending an attack. In the attack, the attacker grasps full background knowledge such as the feature points of the ECG of a certain target user. He seeks for finding out a target user based on the data generated by our scheme. In our experiment, the attacker aims at linking the person's data from his background knowledge so as to find out the physical condition of the target victim. The attack is taken by two steps: (1) determine whether the personal data of the target user are contained in the current data set, and (2) mount an identity linking attack.

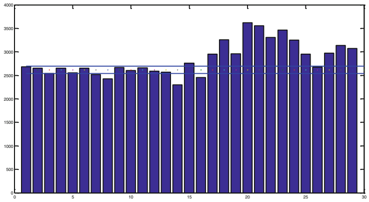

Fig. 8. Probability of exposure to privacy

As shown in Fig. 8, during the 24 experiments, when only one feature value is chosen, the probability of exposure of data privacy is 1/5. We assume that the actual amount of data is 30 × n. Therefore, for a single feature, the risk of exposure to privacy is $1/(5n)$. When we choose m feature values for preserving privacy. The probability of privacy exposure will be calculated as:

$$\frac{m}{5n} \le p' \le \left(\frac{1}{5n}\right)^m \tag{4}$$

Consider that the data itself are vague, therefore, eventual, the probability p of privacy exposure is:

$$\frac{m}{5n} \le p < \left(\frac{1}{5n}\right)^m \tag{5}$$

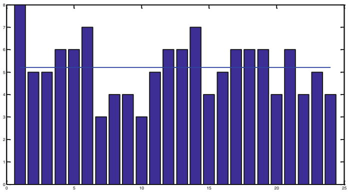

Fig. 9. Privacy level obtained by the user

As shown in Fig. 9, the value of m must be a single-digit, because the number of the feature points is not big. When the magnitude of data is higher, our algorithm will provide a reliable privacy protection for them. We note that, in our work, tricky data will provide better protections. Moreover, more data are input, better protection will be obtained.

In conclusion, our experimental results show that the ECG privacy protection programs achieve desired privacy requirements.

5 Conclusions and Future Works

In this paper, we proposed a differential privacy protection model, which significantly reduces the risk of privacy exposure and greatly ensures the availability of data. We introduce the concept of dynamic noise thresholds to lift the direct relationship between the added noise and data set size, making our scheme more suitable to process big data. At last, several experiments are conducted to show the performance of our scheme. Experimental results reveal that our scheme can provide ideal protection effect. Even in the case that the attacker has full background knowledge, our scheme still can produce enough interference, and the attacker is unable to match the objective person from the ECG data.

As part of our future works, we will put forward theoretical analysis on how to apply differential privacy scheme for big sensitive data in BSNs.

Acknowledgements. This research is sponsored in part by the National Natural Science Foundation of China (No. 61173179, No. 61402078) and Program for New Century Excellent Talents in University (NCET-13-0083). This research is also sponsored in part supported by the Fundamental Research Funds for the Central Universities (No. DUT14RC(3)090).

References

1. Zheng, Z., Zhu, J., Lyu, M.R.: Service-generated big data and big data-as-a-service: an overview. In: IEEE International Congress on Big Data, pp. 403–410 (2013)
2. Huang, Z., Cao, F., Li, J., Chen, X.: Developing sea cloud data system key technologies for large data analysis and mining. J. Netw. New Media **1**(6), 20–26 (2012)

3. Bressan, N., Andrew, J.: Integration of drug dosing data with physiological data streams using a cloud computing paradigm. In: 35th Annual International Conference on Engineering in Medicine and Biology Society (EMBC), pp. 4175–4178. IEEE (2013)
4. Kai, E., Ashir, A.: Technical challenges in providing remote health consultancy services for the unreached community. In: 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 1016–1020. IEEE (2013)
5. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr. Cardiol. Rep. **16**, 441 (2013)
6. Kafali, O., Bromuri, S., Sindlar, M.: Commodity 12: a smart e-health environment for diabetes management. J. Ambient Intell. Smart Environ. **5**(1), 479–502 (2013)
7. Wu, J., Roy, J., Stewart, W.F.: Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care **48**(6), S106–S113 (2010)
8. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395–405 (2012)
9. Huang, Q.R., Qin, Z., Zhang, S., Chow, C.M.: Clinical patterns of obstructive sleep apnea and its co morbid conditions: a data mining approach. J. Clin. Sleep Med. **4**(6), 543 (2008)
10. Zrimec, T., Wong, J.: Improving computer aided disease detection using knowledge of disease appearance. In: Med Info 2007: Proceedings of the 12th World Congressing Health (Medical) Informatics; Building Sustainable Health Systems, p. 1324. IOS Press, Amsterdam (2007)
11. Melzer, T.R., Richard, W.: Arterial spinlabelling reveals an abnormal cerebral perfusion pattern in Parkinson's disease. Brain, awq377 (2011)
12. Xue, Y., Li, Q., Jin, L., Feng, L., Clifton, D.A., Clifford, G.D.: Detecting adolescent psychological pressures from micro-blog. In: Zhang, Y., Yao, G., He, J., Wang, L., Smalheiser, N.R., Yin, X. (eds.) HIS 2014. LNCS, vol. 8423, pp. 83–94. Springer, Heidelberg (2014)
13. Yoo, J., Yan, L., Lee, S.: A wearable ECG acquisition system with compact planar-fashionable circuit board based shirt. IEEE Trans. Inf. Technol. Biomed. **13**(6), 897–902 (2009)
14. Gargiulo, G., Bifulco, P., Cesarelli, M.: An ultra-high input impedance ECG amplifier for long-term monitoring of athletes. Med. Devices (Auckl) **3**, 1–9 (2010)
15. Yan, Y., Qin, X., Fan, J., Wang, L.: A review of big data research in medicine & healthcare. E-Sci. Technol. Appl. **5**(6), 3–16 (2014)
16. Shimmer. http://www.shimmersensing.com/