Article

Effective Privacy-Preserving Collection of Health Data from a User's Wearable Device

Jong Wook Kim, Su-Mee Moon, Sang-ug Kang and Beakcheol Jang *

Department of Computer Science, Sangmyung University, Seoul 03016, Korea; jkim@smu.ac.kr (J.W.K.); sumeedi@naver.com (S.-M.M.); su kang@smu.ac.kr (S.-u.K.)

* Correspondence: bjang@smu.ac.kr

Received: 20 July 2020; Accepted: 11 September 2020; Published: 14 September 2020

Abstract: The popularity of wearable devices equipped with a variety of sensors that can measure users' health status and monitor their lifestyle has been increasing. In fact, healthcare service providers have been utilizing these devices as a primary means to collect considerable health data from users. Although the health data collected via wearable devices are useful for providing healthcare services, the indiscriminate collection of an individual's health data raises serious privacy concerns. This is because the health data measured and monitored by wearable devices contain sensitive information related to the wearer's personal health and lifestyle. Therefore, we propose a method to aggregate health data obtained from users' wearable devices in a privacy-preserving manner. The proposed method leverages local differential privacy, which is a de facto standard for privacy-preserving data processing and aggregation, to collect sensitive health data. In particular, to mitigate the error incurred by the perturbation mechanism of location differential privacy, the proposed scheme first samples a small number of salient data that best represents the original health data, after which the scheme collects the sampled salient data instead of the entire set of health data. Our experimental results show that the proposed sampling-based collection scheme achieves significant improvement in the estimated accuracy when compared with straightforward solutions. Furthermore, the experimental results verify that an effective tradeoff between the level of privacy protection and the accuracy of aggregate statistics can be achieved with the proposed approach.

Keywords: health data; wearable device; data privacy; health data aggregation

1. Introduction

The recent growth of individuals' interest in their personal health and wellness has prompted the use of smart healthcare services, which combine information and communications technologies with medical services. One of the key technologies that enable smart healthcare services is a recommendation method that provides individual users with customized healthcare-related services. Generally, these recommendation techniques require considerable health data to be collected from diverse users over a long period of time to enhance the recommendation quality by extracting aggregate statistics.

Wearable devices are equipped with a variety of sensors capable of measuring the users' environmental conditions (e.g., temperature, humidity, pressure, and ultraviolet radiation). In addition, the user's health status (e.g., their heart rate, sleep status, and blood pressure) as well as their lifestyle (e.g., daily step-count and calories burned per day) can be measured by wearable devices. For example, Figure 1 displays the change in daily cumulative step-count per hour measured by the built-in accelerometer of a wearable device, such as a commercial smart band, fitness band, and smartwatch. The widespread use of wearable devices that are able to measure the user's health status and lifestyle has made it possible for healthcare service providers to utilize these wearable devices as a primary means to collect considerable health data from diverse users.

Figure 1. An example of the change in daily cumulative step-count per hour that is measured by a wearable device.

The health data collected from wearable devices are useful for the provision of smart healthcare services; however, the indiscriminate collection of individuals' health data may raise serious privacy concerns because of the personal and sensitive nature of information pertaining to a person's health and lifestyle. For example, a user's private lifestyle could be inferred by analyzing the change in their daily cumulative step-count. From Figure 1, the activities associated with a user's lifestyle, such as the time they left home for work and the time they spent working in the office, can be deduced based on the change in the daily cumulative step-count as a function of time. The periods with little change in the step-count denote periods when the user is either remaining in a specific location or moving by using transportation vehicles, whereas periods during which the step-count rapidly increases denote periods when they are walking or exercising. Furthermore, combined with external data, such as information about the user's location measured by smartphone, an accurate impression of the user's lifestyle is feasible. Therefore, users are generally reluctant to provide their health-related data, measured by wearable devices, to service providers owing to privacy concerns. These concerns have been identified as the most significant challenge faced by smart healthcare services.

The literature on data management contains reports of extensive studies relating to data privacy protection. From the literature, differential privacy (DP) has emerged as a de facto standard for privacy-preserving data processing. DP is based on a formal mathematical definition that provides a probabilistic privacy guarantee against attackers with arbitrary background knowledge [1,2]. In a DP-based data collection scenario, a centralized trusted data aggregator (who is trusted to not reveal the original data received from data owners) collects the original data from the data owners, perturbs the collected original data by adding random noise to satisfy DP, and publishes the perturbed data for the purpose of data analysis.

Recently, local differential privacy (LDP), which is a variant of DP, has now attracted attention as a promising way of guaranteeing individual privacy during the process of data collection [3–5]. Unlike DP, LDP does not require the existence of the trusted data aggregator. In an LDP-based data collection scenario, each data owner, who does not fully trust a data aggregator, perturbs his/her original data in a manner that satisfies DP before sending it to the data aggregator. With its strong privacy guarantees, several successful LDP-based deployments have been implemented in industry by major technology companies, including Google [3], Samsung [6], Apple [7,8], and Microsoft [9].

In this study, we leverage LDP to collect sensitive health data from wearable devices (e.g., smart band, fitness band, smartwatch, etc.) in a privacy-preserving manner such that a data

owner's original data is not revealed at all (i.e., wearable device user's original health data is not exposed). In particular, our contributions can be summarized as follows:

- We propose an effective way to aggregate health data acquired from wearable devices under LDP. To mitigate the error incurred by the perturbation mechanism of LDP, the proposed method extracts (samples) a small amount of salient data from given health data and reports the identified salient data to the data collection server under LDP, instead of reporting all the health data. Then, the data collection server reconstructs the health data based on the small amount of salient data received from a wearable device.
- Through experiments with a real dataset, we demonstrate that the developed method can effectively collect health data from wearable devices, while preserving privacy.

The remainder of this paper is structured as follows. We describe related work in the next section. Section 3 formalizes the problem. Section 4 presents the proposed method for collecting sensitive health data from wearable devices in a privacy-preserving manner. Section 5 contains details of the experimental evaluation of the proposed approach, and in Section 6, we conclude the paper.

2. Related Work

A considerable number of studies have implemented medical systems comprising wearable devices implanted with sensors to record data for health applications. Examples of the collected data include heart rate, electrocardiogram, blood pressure, sleep state, step counts, calories, etc. With intelligent systems, these lifelogging data are exploited to enable data owners to monitor individuals' health status during their daily activities, such as exercising, working, sleeping, etc. For example, CodeBlue [10] is a health care system that provides patient monitoring functionality by using body sensor data. AlarmNet [11] and Medical MoteCare [12] are patient monitoring systems that not only use body sensor data but also data gathered by environmental sensors.

Health monitoring is a field that has been actively studied. The exploitation of health data makes it possible to provide personalized medical and early disease diagnosis services. For instance, a doctor examining a patient whose health data are collected in a hospital could make an assessment based on the current state of the patient in comparison with the medical records of other patients, which enables improved assessment [13]. An electrocardiogram was used to monitor the state of heart disease and predict the degree of stress by observing changes in heart rate. Taelman et al. [14] investigated the relationship between the stress level and changes in the heart rate. In their experiments with 28 students, the average heart rate rose from 73.52 to 75.94 when the Mensa exam was presented to the students compared to images of pastoral scenes. This means that a mentally stressful situation stimulates the heart rhythm. Fisher et al. [15] introduced a method whereby a wearable device can be used to track the heart rate of a patient with congestive heart failure and the condition monitored remotely. Several studies have demonstrated a significant relationship between health and exercise, indicating that those who are active in their leisure time are more likely to live longer than those who are inactive [16,17].

With respect to human behavior recognition, recognizing patients' state by analyzing their health data is a viable option. For example, an emergency alarm system can be used to automatically request help when it detects abnormal movement of a disabled person or an elderly person with impeded mobility. These alarm systems analyze the data gathered by the inertial sensor embedded in a wearable device for monitoring purposes [18]. The analysis of daily behavior information in an office environment is a popular application in the area of human behavior recognition [19]. In this area, health data are actively being analyzed by classifying physical activities, using methods such as machine learning, and monitoring them for health management purposes. Altun et al. [20] conducted a study to classify 19 types of human behavior using small inertial and magnetic sensors. Behavioral classification using health data is mainly used to provide a service for patients or the elderly. Melillo et al. [21] introduced a method based on electrocardiograms for detecting and preventing

collapse owing to hypotension. In another work, Hamatani et al. [22] developed a method to provide customized healthcare services for individual users by collecting and analyzing wearable devices' biodata such as heart rates and body temperatures.

Much as the works presented in [10–22] aim at improving the well being and the general standard of living of humans, they pay limited attention to the user privacy concerns. In [10–15,18–22], the focus is mainly on methods for collecting and analyzing user data for the purpose of monitoring, without considering the user privacy violations involved. However, in [13], though the authors considered data access security in their design, user privacy still remains a challenge once an entity gains access to the data. For interested readers, a survey on privacy and security issues when using body implantable medical devices is presented in [23].

For practical implementation, privacy preservation is essential in healthcare and medical systems. To preserve privacy, researchers have proposed several techniques such as encryption, anonymization, and data perturbation (through DP and LDP) that protect the health data from intruders. However, data perturbation techniques have emerged as the future of privacy preservation because of the simplicity associated with their mechanisms [24]. Thus, a number of conducted researches have focused on the use of data perturbation techniques for e-health data privacy preservation. Beaulieu-Jones et al. [25] perturbed clinical data using differential privacy to efficiently and accurately train a privacy-preserving deep learning model. The authors integrated differential privacy with a cryptographic mechanism to enhance privacy. Mohammed et al. [26] proposed a lightweight framework for privacy-preserving data mining of cancer patients' data by using the Laplace noise of differential privacy. In [27,28], differential mechanisms are integrated with encryption techniques to preserve the privacy of genomic data. In [29], Tang et al. combined differential privacy with Boneh–Goh–Nissim cryptosystem and Shamir's secret sharing for privacy-preserving aggregation of data from different health devices. Guan et al. [30] combined differential privacy with machine learning to perform k-means clustering of differentially private health data recorded from medical IoT devices.

Apart from healthcare and medical systems, data perturbation techniques have found use in other applications, such as private heavy hitter identification, which employs locally differentially-private algorithms to find the top k items with the highest frequency along with the estimated frequency for each such item [5,31,32]. LDP has also been used for marginal distribution estimation, which computes the joint distribution of multiple variables [33,34] in a privacy preserving manner. Other applications that have integrated LDP for privacy preservation include: IoT acquired data analysis [6,35–38], deep learning [39,40], and data mining [41]. In the literature, the focus is mostly on the privacy of the data already in storage. However, for those that aggregate data periodically, they consider the entire data set which can be inefficient and inaccurate because of the increased error. We intend to address the challenge in this work.

3. Problem Definition, Background and Naive Solution

In this section, we first formally define the problem, provide background on LDP, and present a naive solution to the problem.

3.1. Problem Definition

Generally, wearable devices periodically measure a user's health status within a predefined fixed time interval. Let $U = \{U_1, U_2, \dots, U_w\}$ be the set of users in which w represents the total number of users. In this study, we model the health data of the i -th user U_i (which are measured by a wearable device) as a sequence $S_i = ((t_1, x_1^i), (t_2, x_2^i), \dots, (t_n, x_n^i))$ of length n . Note that (t_d, x_d^i) in a sequence denotes the value, x_d^i , measured by the wearable device at timestamp, t_d . Let us further assume that x_d^i is within the predefined range $[x_{min}, x_{max}]$. Especially, in this paper, we focus on a scenario that generates health data with a monotonically increasing property (i.e., $x_1^i \leq x_2^i \leq x_3^i, \dots, \leq x_n^i$), such as

the daily cumulative step-count (Figure 1) and daily cumulative calorie consumption, which are among the most common forms of health data. We also consider a scenario in which each user reports the daily health data measured by their wearable device to an aggregator. In this case, the user's health data are sent to the aggregator every 24 h.

Let $S = \{S_1, S_2, \dots, S_w\}$ be a set of sequences the aggregator receives from w users. Then, the problem addressed in this paper can be stated as follows: Given

- the set of users $U = \{U_1, U_2, \dots, U_w\}$, and
- the set of health data $S = \{S_1, S_2, \dots, S_w\}$ received from users,

compute the aggregate statistics $AS = ((t_1, ax_1), (t_2, ax_2), \dots, (t_n, ax_n))$. Here, the value of ax_d at timestamp t_d is computed as follows:

$$ax_d = \frac{1}{w} \times \sum_{h=1}^w x_d^h.$$

In other words, once receiving $S = \{S_1, S_2, \dots, S_w\}$ from w users, the aggregator wants to compute the aggregate statistics $AS = ((t_1, ax_1), (t_2, ax_2), \dots, (t_n, ax_n))$ where the value of ax_d at timestamp t_d is computed by averaging all values of x_d^h where $1 \leq h \leq w$.

3.2. Local Differential Privacy

The basic concept of LDP involves perturbation of the original data by the data owner by adding random noise, and reporting the perturbed data to a data aggregator. This mechanism guarantees that the data owner's original data are not exposed to an external entity. LDP is defined as follows: A randomized algorithm $\mathcal{A}$ satisfies ϵ -LDP, if and only if for (1) all pairs of the data owner's local data v_a and v_b , and (2) any output O of $\mathcal{A}$, the following equation is satisfied [3–5]:

$$Pr[\mathcal{A}(v_a) = O] \leq e^\epsilon \times Pr[\mathcal{A}(v_b) = O].$$

Here, $Pr[\mathcal{A}(v_a) = O]$ denotes the probability that the output of running a randomized algorithm $\mathcal{A}$ with v_a is O . The meaning of the aforementioned definition is that, regardless of the data an aggregator receives from a data owner, the aggregator cannot infer with high confidence (which is controlled by the privacy budget ϵ), whether the data owner has sent v_a or v_b . This provides the data owner with plausible deniability. The parameter ϵ is a privacy budget that controls the level of privacy. That is, small values of ϵ ensure strong privacy, guaranteed by adding a comparatively larger amount of noise to the original data. By contrast, large values of ϵ provide weak privacy protection by adding less noise to the original data.

An important property regarding the privacy budget is the sequential composition property, which LDP follows to attain differential privacy. That is, the available privacy budget, ϵ , can be partitioned into n smaller privacy budgets, $\epsilon_1, \epsilon_2, \dots, \epsilon_n$, such that $\epsilon = \sum_{h=1}^n \epsilon_h$ and the data owner uses each small privacy budget to report their local data to an aggregator.

3.3. Naive Solution

In this subsection, we introduce a straightforward privacy-preserving solution based on the use of LDP which consists of wearable device processing (i.e., data owner) and server processing of the collected data (i.e., data aggregator).

Wearable device processing: For explanatory purposes, we focus on the i -th user, $U_i \in U$. By the sequential composition property of LDP, the privacy budget, ϵ , is first partitioned into n privacy budgets. Then, given original health data $S_i = ((t_1, x_1^i), (t_2, x_2^i), \dots, (t_n, x_n^i))$ represented as a sequence of length n , each smaller privacy budget is used to generate a perturbed sequence $PS_i = ((t_1, px_1^i), (t_2, px_2^i), \dots, (t_n, px_n^i))$. Here, px_d^i is obtained using the LDP mechanism as follows:

$$px_d^i = x_d^i + \text{Lap}\left(\frac{\Delta s}{\epsilon/n}\right),$$

where Δs corresponds to the local sensitivity defined as $\Delta s = x_{\max} - x_{\min}$ and $\text{Lap}\left(\frac{\Delta s}{\epsilon/n}\right)$ denotes a random noise sampled from a Laplace distribution with mean $\mu = 0$ and scale $sf = \frac{\Delta s}{\epsilon/n}$. Note that this satisfies the ϵ -differential privacy because of $\epsilon = \sum_{h=1}^n \epsilon/n = n \times (\epsilon/n)$. Then, the user U_i reports the perturbed sequence PS_i instead of the original data S_i , which guarantees that the original data of the data owner is not exposed to the (untrusted) data aggregator.

Server processing of the collected data: Once it has been received the perturbed sequences from all the data owners $PS = \{PS_1, PS_2, \dots, PS_w\}$, the data aggregator computes the aggregate statistics $AS = \left((t_1, ax_1), (t_2, ax_2), \dots, (t_n, ax_n)\right)$. Here, the value of ax_d at timestamp, t_d , is estimated by using the perturbed values as follows:

$$ax_d = \frac{1}{w} \times \sum_{h=1}^w px_d^h.$$

The expected error incurred by this estimation is known to be linearly proportional to the sequence length n [6]. Thus, this scheme is not suitable when the sequence length, n , is large.

4. Proposed Method

In this section, we introduce the proposed method to collect sensitive health data from wearable devices using LDP. The general outline of the proposed scheme is similar to that of the naive solution introduced in the previous section. However, to mitigate the high expected error incurred by the perturbation mechanism of LDP when collecting the health data via wearable devices, we develop a novel scheme which reports sampled salient data instead of the entire health data. In particular, the proposed method first extracts (samples) a small amount of salient data from the entire health data, perturbs the identified salient data using the perturbation mechanism of LDP, and sends the perturbed salient data to a data collection server. Then, after receiving the perturbed salient data, the data collection server reconstructs the health data based on them. In the next subsections, we explain each of these steps in detail.

4.1. Wearable Device Processing

Wearable device processing consists of two phases: searching for a salient data set and reporting the perturbed salient data set to the data collection server.

Searching for a Set of Salient Data

As described in Section 3.3, the expected error incurred by using LDP is linearly proportional to the length of the sequence of health data. To mitigate the expected error caused by the perturbation mechanism of LDP, the proposed method first identifies a small amount of salient data, after which it uses LDP to process these data instead of processing the entire health data. Thus, given a health data, the objective of the first phase of wearable device processing is to search for an optimal salient data set that best represents the original health data.

For easy explanation, let us focus on the i -th user U_i and their health data, $S_i = \text{Big}((t_1, x_1^i), (t_2, x_2^i), \dots, (t_n, x_n^i))$, represented as a sequence of length n . Note that in the following, for ease of notation, we omit the superscript i whenever possible. As indicated in Section 3.1, in this study, we consider a scenario in which the collected health data increase monotonically (i.e., $x_1 \leq x_2 \leq x_3, \dots, \leq x_n$). Algorithm 1 contains the pseudo-code for computing a salient data set, $FS = \{(t_{f_1}, x_{f_1}), (t_{f_2}, x_{f_2}), \dots, (t_{f_m}, x_{f_m})\}$, which is composed of m elements (i.e., salient data), extracted from S_i . The inputs of the algorithm include the health data sequence S_i and the predefined maximum number of salient data α . In the initialization step, the best salient data set, FS_{best} , is initialized such

that it contains the first (i.e., (t_1, x_1)) and the last (i.e., (t_n, x_n)) data value of S_i (lines 1–3). Then, the best error, err_{best} , incurred by using LDP, is estimated with the initial best salient data set by the function *EstimateError_by_LDP()*, which is explained later in this subsection.

Algorithm 1: Pseudo-code for extracting a set of salient data from a given health data sequence

```

input:  $S_i = ((t_1, x_1), (t_2, x_2), \dots, (t_n, x_n)), \alpha$ 
1  $FS_{best} \leftarrow \emptyset;$ 
2  $FS_{best} \leftarrow FS_{best} \cup (t_1, x_1);$ 
3  $FS_{best} \leftarrow FS_{best} \cup (t_n, x_n);$ 
4  $err_{best} \leftarrow EstimateError\_by\_LDP(S_i, FS_{best});$ 
5  $FS_{cur} \leftarrow FS_{best};$ 
6 while  $|FS_{cur}| < \alpha$  do
    /* Step 1: Find the next best salient data */
7      $dist_{min} \leftarrow \infty;$ 
8     for  $h \leftarrow 2$  to  $(n - 1)$  do
9          $(t_h, x_h) \leftarrow GetElementAt(S_i, h);$ 
10        if  $(t_h, x_h) \notin FS_{cur}$  then
11             $dist \leftarrow ComputeDistance(S_i, FS_{cur} \cup (t_h, x_h));$ 
12            if  $dist < dist_{min}$  then
13                 $h_{min} \leftarrow h;$ 
14                 $dist_{min} \leftarrow dist;$ 
15            end
16        end
17    end
18     $(t_{h_{min}}, x_{h_{min}}) \leftarrow GetElementAt(S_i, h_{min});$ 
19     $FS_{cur} \leftarrow FS_{cur} \cup (t_{h_{min}}, x_{h_{min}});$ 
    /* Step 2: Update the best salient data set if necessary */
20     $err \leftarrow EstimateError\_by\_LDP(S_i, FS_{cur});$ 
21    if  $err < err_{best}$  then
22         $FS_{best} \leftarrow FS_{cur};$ 
23         $err_{best} \leftarrow err;$ 
24    end
25 end
26 return  $FS_{best}$ 

```

The main parts of the algorithm consist of two steps: finding the next best salient data and updating the best salient data set if necessary. In the first step (lines 7–19), the algorithm finds the next salient data by scanning all data in S_i and then adding these data to FS_{cur} . Here, the next salient data is selected in such a way that the distance (which is computed by *ComputeDistance()* in line 11) between the original health data S_i and the linear line obtained on the basis of salient data set FS_{cur} is minimized. For example, Figure 2 illustrates the computation of the distance (represented by black dotted lines) between the original health data (represented by a blue curve) and a salient data set (represented by red circles) containing four salient data. In this example, we assume that α is set to 6 and the salient data set contains 4 salient data. Here, each linear line segment (represented by an orange line) is obtained by connecting two adjacent salient data.

In the second step (lines 20–24), the algorithm compares the estimated error incurred by using LDP with FS_{cur} to the one with FS_{best} . If err is less than err_{best} , the best salient data set is updated with FS_{cur} . The algorithm iteratively repeats this process until the number of salient data in FS_{cur} reaches the predefined α (line 6), which implies that the number of salient data in the best salient data set FS_{best} is no greater than α .

Figure 2. Example of computing the distance between the original health data and the salient data set.

We now explain the function *EstimateError_by_LDP()* that estimates the error induced by the perturbation step of LDP (Algorithm 2). Algorithm 2 accepts the health data, S_i , and the salient data set, FS , as its input. In line 2, the algorithm generates β random noises sampled from a Laplace distribution with mean $\mu = 0$ and scale $sf = \frac{\Delta s}{\epsilon/m}$, where Δs is the local sensitivity explained in Section 3.3, m is the number of salient data in FS , and β is a predefined parameter. Then, in line 3, v_{noise} is computed by averaging the β random noises generated in the previous step. In the next steps, the noised salient data set FS_{noise} is generated by adding v_{noise} to each salient data in FS (line 4–7). Then, in line 8, the algorithm computes the nonlinear curve that best expresses the salient data in FS_{noise} by using polynomial regression. Finally, the distance (i.e., error) between the original health data S_i and the nonlinear curve obtained in the previous step is computed.

For example, Figure 3 illustrates the method to compute the estimated error by the function *EstimateError_by_LDP()*. The original health data (i.e., S_i) is represented by the blue curve and each salient data in FS is represented by the red circle (Figure 3a). Here, the x-axis corresponds to the timestamps (i.e., $t_1, t_2, \dots, t_n$ of S_i) and the y-axis represents the values (i.e., $x_1, x_2, \dots, x_n$ of S_i). In Figure 3b, each noised salient datum, represented by the green circle, is obtained by adding v_{noise} to each salient data. Then, in Figure 3c, by using polynomial regression which finds the best-fitting curve by minimizing the sum of the deviations from each given data point (i.e., each green circle in this example) to the curve, the nonlinear curve (represented by the green curve) is computed. We also note that the best-fitting curve computed by using polynomial regression does not need to pass through existing given data points. Finally, the distance between the original health data and the best-fitting curve is computed in Figure 3d.

Algorithm 2: EstimateError_by_LDP()

input: $S_i = ((t_1, x_1), (t_2, x_2), \dots, (t_n, x_n))$, FS

- 1 $m \leftarrow |FS|$;
- 2 $r_1, r_2, \dots, r_\beta \leftarrow \text{generate } \beta \text{ random number } \sim \text{Lap}(\frac{\Delta s}{\epsilon/m})$;
- 3 $v_{noise} \leftarrow \frac{1}{\beta} \times \sum_{h=1}^{\beta} r_h$;
- 4 $FS_{noise} \leftarrow \emptyset$;
- 5 **for each** $(t_{f_h}, x_{f_h}) \in FS$ **do**
- 6 $FS_{noise} \leftarrow FS_{noise} \cup (t_{f_h}, x_{f_h} + v_{noise})$;
- 7 **end**
- 8 $p \leftarrow \text{PolynomialRegression}(FS_{noise})$;
- 9 $dist \leftarrow \text{ComputeDistance}(S_i, p)$;
- 10 **return** $dist$

Figure 3. Example of computing the error by the function EstimateError_by_LDP().

Once the salient data set is identified from the health data, the next step entails reporting the perturbed set to the data collection server under LDP. Let $FS_i = \{(t_{f_1}, x_{f_1}), (t_{f_2}, x_{f_2}) \cdots (t_{f_m}, x_{f_m})\}$ be the salient data set, which consists of m salient data, extracted from S_i , as explained in the previous phase. Then, based on the sequential composition property of LDP, the privacy budget, ϵ , is equally partitioned into m privacy budgets, each of which is used to perturb each salient data in FS_i . Formally, given FS_i , let $PFS_i = \{(t_{f_1}, px_{f_1}), (t_{f_2}, px_{f_2}) \cdots (t_{f_m}, px_{f_m})\}$ be the corresponding perturbed salient data set. Here, px_{f_h} is obtained by adding random noise sampled from a Laplace distribution with mean $\mu = 0$ and scale $sf = \frac{\Delta s}{\epsilon/m}$:

$$px_{f_h} = x_{f_h} + \text{Lap}\left(\frac{\Delta s}{\epsilon/m}\right).$$

This satisfies the ϵ -differential privacy because $\epsilon = \sum_{h=1}^m \epsilon/m$.

The last step of wearable device processing is to send the perturbed salient data set, PFS_i , to the data aggregation server. Note that, unlike the naive solution in Section 3.3, the proposed method is able to reduce the error caused by the perturbation mechanism of LDP by using the privacy budget ϵ to perturb and report the small number of salient data instead of reporting the entire health data (i.e., $m < n$).

4.2. Server Processing of the Collected Data

After receiving the perturbed salient data set from a wearable device, the data collection server first reconstructs the health data based on the perturbed salient data set. Given PFS_i , for easy explanation, let us assume that the timestamp of a salient datum in PFS_i satisfies the condition $t_{f_1} < t_{f_2} < \cdots < t_{f_m}$.

Given two adjacent salient data, (t_{f_h}, px_{f_h}) and $(t_{f_{h+1}}, px_{f_{h+1}})$, the health data located between two timestamps, t_{f_h} and $t_{f_{h+1}}$, are estimated by the points that lie on the linear line connecting (t_{f_h}, px_{f_h}) and $(t_{f_{h+1}}, px_{f_{h+1}})$. Let a be the slope and b be the y-intercept of the linear line connecting (t_{f_h}, px_{f_h}) and $(t_{f_{h+1}}, px_{f_{h+1}})$. Then, a and b , respectively, are computed as follows:

$$a = \frac{px_{f_{h+1}} - px_{f_h}}{t_{f_{h+1}} - t_{f_h}}, \quad b = px_{f_h} - a \times t_{f_h}.$$

All health data that fall into the time interval between t_{f_h} and $t_{f_{h+1}}$ are represented with a linear line with a slope a and the y-intercept $(0, b)$.

Let $RS_i = ((t_1, rx_1^i), (t_2, rx_2^i), \dots, (t_n, rx_n^i))$ be a reconstructed health data sequence of the i -th user. Let $RS = \{RS_1, RS_2, RS_3, \dots, RS_w\}$ be a set of reconstructed health data of w wearable device users. Then, the data aggregator computes the aggregate statistics $AS = ((t_1, ax_1), (t_2, ax_2), \dots, (t_n, ax_n))$. Here, the value of ax_d at timestamp t_d is estimated using the reconstructed data as follows:

$$ax_d = \frac{1}{w} \times \sum_{i=1}^w rx_d^i.$$

That is, the average value of the d -th timestamp is obtained by averaging the values of the d -th timestamp for all the reconstructed health data.

5. Experimental Evaluation

In this section, we describe the experiments that were conducted to evaluate the effectiveness of the proposed approach. We first explain the experimental setup and then discuss the experimental results.

5.1. Data Set

To evaluate the effectiveness of the proposed method, we collected 290 cumulative step-count data from students (undergraduate and graduate) of Sangmyung University on daily basis as they carry on their activities. The recordings were done in one-minute time intervals between 10:00–21:00 using the Gear S3 smartwatch. The effect of the amount of data on the performance of the proposed method was investigated by generating 290×10^1 and 290×10^2 data sets. The 290×10^1 and 290×10^2 data sets were generated by replicating each data in 290 daily cumulative step-count data by 10^1 and 10^2 times.

5.2. Baseline Approaches

In the experiments, in addition to reporting results for the proposed approach(OS), which is based on the optimal salient data selection presented in Section 4, we also report results for the following alternatives:

- *NS* corresponds to the naive solution explained in Section 3.3.
- *RS* is the random selection method which first randomly selects a predefined fixed number of salient data from a given health data, and then reports the randomly selected salient data to the data collection server under LDP.
- *NS* corresponds to the non-optimal selection method which first selects a predefined fixed number of salient data from a given health data by using the first step in Algorithm 1, and then reports the selected salient data to the data collection server under LDP. We note that unlike *OS*, *NS* use the first step in Algorithm 1, but does not leverage the second step (which considers the error incurred by the perturbation mechanism of LDP) in Algorithm 1

We note that *RS* and *NS* are used to evaluate the effectiveness of the proposed method for identifying a salient data set, as presented in Section 4.1.

5.3. Experimental Setup

Three different levels of privacy, $\epsilon = 0.5$, $\epsilon = 1.0$, and $\epsilon = 2.0$, were used in the experiments. In the experimental data set, the difference between a possible minimum value and a possible maximum value of daily cumulative step-count is less than 8000, and thus, the local sensitivity Δs was set to 8000. Furthermore, the predefined parameters, α and β , in Algorithms 1 and 2 were assigned the values of 6 and 200, respectively.

To measure the estimation accuracy, we use an error rate, e , defined as follows:

$$e = \frac{1}{n} \times \sum_{d=1}^n |ax_d^{actual} - ax_d^{est}|.$$

Here, n denotes the length of the health data sequence, and ax_d^{actual} and ax_d^{est} , respectively, correspond to the actual and the estimated value of ax_d at the timestamp, t_d , in aggregate static $AS = ((t_1, ax_1), (t_2, ax_2), \dots, (t_n, ax_n))$.

5.4. Experimental Results

Table 1 lists the average error rates for varying privacy budget ϵ and datasets of different sizes. In the experiments, the privacy budget ϵ varied from 0.5 to 2.0 and the sizes of the datasets varied from 290×10^1 to 290×10^2 . Key observations based on the table can be summarized as follows: As the number of collected data sizes increases, the error rate decreases, which implies that as additional data are collected from users of wearable devices, the estimation accuracy becomes high. As ϵ decreases, the error rate increases. This is because a decrease in the privacy budget causes random noise, which is added to the original data by the perturbation phase of LDP, to increase, whereupon the level of privacy protection of the user's health data increases. As indicated by the results in the table, the method proposed in this paper significantly outperforms the naive solution for all privacy budgets and data sizes. These experimental results verify that the proposed method, which reports carefully selected salient data, is more effective than the naive solution, which reports the entire collection of health data.

Table 1. Comparison of average error rates for varying privacy budget ϵ and data sizes.

Data Size	NS			OS (Proposed Method)		
	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$
290×10^1	240,843.99	119,103.22	59,302.83	471.28	210.77	143.62
290×10^2	7604.89	3783.73	1890.87	17.20	9.18	5.57

Figure 4 plots the actual and estimated aggregate static for datasets of various sizes. In this figure, the y-axis corresponds to the daily step-count and the x-axis represents the time between 10:00 and 21:00. In the experiment, the privacy budget is set to 1.0, whereas the amount of data varies from 290 to 290×10^2 . As expected, as the amount of data increases, the aggregated health data sequence estimated by the proposed approach more closely approximates the actual data (which is represented by a blue curve). As can be seen in the figure, with the 290 dataset, an accurate estimation cannot be obtained due to the insufficient collected data. However, with the 290×10^2 dataset, a fairly good estimation can be achieved by the proposed approach in this paper. These experimental results indicate that the method proposed in this study effectively exploits the collected dataset.

Figure 4. Actual vs estimated aggregate health data sequences for various data sizes ($\epsilon = 1.0$).

We also evaluated the effectiveness of the proposed method for identifying a salient data set, as presented in Section 4.1. Figure 5 shows the average error rates for the three alternative methods, *RS*, *NS*, and *OS*. In this figure, the y-axis corresponds to the error rate and the x-axis represents the number of salient data that are extracted from the given health data. In the experiments, the privacy budget ϵ varied from 0.5 to 2, while the size of the dataset was fixed to 290×10^2 . Furthermore, for the random and non-optimal selection methods, which require the number of salient data that are extracted from the given health data to be predefined, the number of extracted salient data varies from 3 to 6. We note that, unlike these two methods, for the proposed method explained in Section 4, the optimal number of salient data, which minimize the expected error incurred by using LDP, is determined by Algorithm 1.

Figure 5. Average error rates for three alternative methods, *RS*, *NS*, and *OS* (data size = 290×10^2).

Key observations based on Figure 5 can be summarized as follows: as the privacy budget, ϵ , decreases, the error rate increases. This is because, as the privacy budget decreases, noises added by the wearable device-side increase, which, in turn, leads to decreased estimation accuracy at the data collection server-side. These evaluation results are consistent with the ones presented in Table 1. *NS*,

which leverages the first step of Algorithm 1, outperforms *RS* for all privacy budgets. These results indicate that the first phase in Algorithm 1 is able to effectively search the salient data sets that best represent the original health data. More importantly, the experimental results show that the proposed scheme (represented as red dotted lines) delivers the best performance for all privacy budgets. These results verify that the method presented in this paper is able to effectively identify a salient data set from given health data by considering the expected error caused by the perturbation mechanism of LDP.

6. Conclusions

In this paper, we developed methods to collect sensitive health data by using wearable devices in a privacy-preserving manner by leveraging LDP. The method proposed here first identifies a small number of salient data that best represents the original health data, and then reports the identified salient data to the data aggregation server under LDP instead of reporting the entire collection of health data. Compared with the naive approach, the proposed scheme is able to reduce the error induced by the perturbation phase of LDP. The experimental results show that the proposed method achieves a significant improvement in the estimated accuracy when compared with the naive solutions. Furthermore, the results verify that an effective tradeoff between the level of privacy protection and the accuracy of aggregate statistics can be achieved with the proposed approach.

Author Contributions: J.W.K. designed the algorithm and wrote the manuscript. S.-M.M. implemented the algorithm and conducted experiments. S.-u.K. and B.J. analyzed the experimental results and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00269, A research on safe and convenient big data processing methods) and the Basic Science Research Program through the National Research Foundation of Korea (NRF-2020R1F1A1072622).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dwork, C. Differential privacy. In Proceedings of the International Conference on Automata, Languages and Programming, Venice, Italy, 9–16 July 2006; pp. 1–12.
2. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In *Lecture Notes in Computer Science, Proceedings of the Third Conference on Theory of Cryptography*, New York, NY, USA, 4–7 March 2006; Springer: Berlin/Heidelberg, Germany, 2006.
3. Erlingsson, U.; Pihur, V.; Korolova, A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014; pp. 1054–1067.
4. Wang, T.; Blocki, J.; Li, N.; Jha, S. Locally differentially private protocols for frequency estimation. In Proceedings of the 26th USENIX Conference on Security Symposium, Berkeley, CA, USA, 16–18 August 2017.
5. Bassily, R.; Smith, A. Local, private, efficient protocols for succinct histograms. In Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, Portland, OR, USA, 14–17 June 2015.
6. Nguyen, T.T.; Xiao, X.; Yang, Y.; Hui, S.C.; Shin, H.; Shin, J. Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy. 2016. Available online: <https://arxiv.org/abs/1606.05053> (accessed on 14 July 2020).
7. Differential Privacy Team, Apple. Learning with Privacy at Scale. 2018. Available online: <https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf> (accessed on 14 July 2020).
8. Tang, J.; Korolova, A.; Bai, X.; Wang, X.; Wang, X. Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12. 2017. Available online: <https://arxiv.org/abs/1709.02753> (accessed on 14 July 2020).

9. Ding, B.; Kulkarni, J.; Yekhanin, S. Collecting telemetry data privately. In Proceedings of the International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3574–3583.
10. Missen, K.; Porter, J.E.; Raymond, A.; de Vent, K.; Larkins, J.A. Adult deterioration detection system (ADDS): An evaluation of the impact on MET and code blue activations in a regional healthcare service. *Collegian* **2018**, *25*, 157–161. [[CrossRef](#)]
11. El-Bendary, M.F.N.; Ramadan, R.; Hassanien, A. *Wireless Sensor Networks: A Medical Perspective*; CRC Press, Taylor and Francis Group: Boca Raton, FL, USA, 2013; pp. 713–732.
12. Navarro, K.F.; Lawrence, E.; Lim, B. Medical MoteCare: A distributed personal healthcare monitoring system. In Proceedings of the International Conference on eHealth, Telemedicine, and Social Medicine, Cancun, Mexico, 1–7 February 2009.
13. Manogaran, G.; Varatharajan, R.; Lopez, D.; Kumar, P.M.; Sundarasekar, R.; Thota, C. A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system. *Future Gener. Comput. Syst.* **2018**, *82*, 375–387. [[CrossRef](#)]
14. Taelman, J.; Vandeput, S.; Spaepen, A.; VanHuffel, S. Influence of mental stress on heart rate and heart rate variability. In *IFMBE Proceedings, Proceedings of the International Federation for Medical and Biological Engineering, Antwerp, Belgium, 23–27 November 2008*; Springer: Berlin/Heidelberg, Germany, 2009.
15. Fisher, R.; Smailagic, A.; Sokos, G. Monitoring Health Changes in Congestive Heart Failure Patients Using Wearables and Clinical Data. In Proceedings of the IEEE International Conference on Machine Learning and Applications, Cancun, Mexico, 18–21 December 2017.
16. Warburton, D.E.R.; Bredin, S.S.D. Health benefits of physical activity: A strengths-based approach. *J. Clin. Med.* **2019**, *8*, 2044. [[CrossRef](#)] [[PubMed](#)]
17. Rueggsegger, G.N.; Booth, F.W. Health benefits of exercise. In *Cold Spring Harbor Perspectives in Medicine*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2018.
18. Hong, Y.J.; Kim, I.J.; Ahn, S.C.; Kim, H.G. Activity recognition using wearable sensors for elder care. In Proceedings of the Second International Conference on Future Generation Communication and Networking, Hainan Island, China, 13–15 December 2008.
19. Jalal, A.; Quaid, M.A.K.; Hasan, A.S. Wearable sensor-based human behavior understanding and recognition in daily life for smart environments. In Proceedings of the International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 17–19 December 2018.
20. Altun, K.; Barshan, B.; Tuncel, O. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognit.* **2010**, *43*, 3605–3620. [[CrossRef](#)]
21. Melillo, P.; Castaldo, R.; Sannino, G.; Orrico, A.; Pietro, G.D.; Pecchia, L. Wearable technology and ECG processing for fall risk assessment, prevention and detection. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015.
22. Hamatani, T.; Uchiyama, A.; Higashino, T. HeatWatch: Preventing heatstroke using a smart watch. In Proceedings of the IEEE International Conference of Pervasive Computing and Communications Workshops, Kona, HI, USA, 13–17 March 2017.
23. Camara, C.; Peris-Lopez, P.; Tapiado, J.E. Security and privacy issues in implantable medical devices: A comprehensive survey. *J. Biomed. Inform.* **2015**, *55*, 272–289. [[CrossRef](#)] [[PubMed](#)]
24. Hassan, M.U.; Rehmani, M.H.; Chen, J. Differential privacy techniques for cyber physical systems: A survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 746–789. [[CrossRef](#)]
25. Beaulieu-Jones, B.K.; Yuan, W.; Finlayson, S.G.; Wu, Z.S. Privacy-preserving distributed deep learning for clinical data. *arXiv* **2018**, arXiv:1812.01484.
26. Mohammed, N.; Barouti, S.; Alhadidi, D.; Chen, R. Secure and private management of healthcare databases for data mining. In Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS), Sao Carlos, Brazil, 22–25 June 2015; pp. 191–196.
27. Raisaro, J.L.; Troncoso-Pastoriza, J.; Misbach, M.; Sousa, J.S.; Pradervand, S.; Missiaglia, E.; Michielin, O.; Ford, B.; Hubaux, J.-P. MedCo: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16*, 1328–1341. [[CrossRef](#)] [[PubMed](#)]
28. Raisaro, J.L.; Choi, G.; Pradervand, S.; Colsenet, R.; Jacquemont, N.; Rosat, N.; Mooser, V.; Hubaux, J.-P. Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *15*, 1413–1426. [[CrossRef](#)] [[PubMed](#)]

29. Tang, W.; Ren, J.; Deng, K.; Zhang, Y. Secure data aggregation of lightweight e-healthcare IoT devices with fair incentives. *IEEE Internet Things J.* **2019**, *6*, 8714–8726. [[CrossRef](#)]
30. Guan, Z.; Lv, Z.; Du, X.; Wu, L.; Guizani, M. Achieving data utility-privacy tradeoff in internet of medical things: A machine learning approach. *Future Gener. Comput. Syst.* **2019**, *98*, 60–68. [[CrossRef](#)]
31. Bassily, R.; Nissim, K.; Stemmer, U.; Thakurta, A. Practical locally private heavy hitters. In Proceedings of the International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2285–2293.
32. Bun, M.; Nelson, J.; Stemmer, U. Heavy hitters and the structure of local privacy. In Proceedings of the ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, 10–15 June 2018; pp. 435–447.
33. Fanti, G.; Pihur, V.; Erlingsson, U. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. In Proceedings of the Privacy Enhancing Technologies Symposium, Darmstadt, Germany, 19–22 July 2016; pp. 41–61.
34. Cormode, G.; Kulkarni, T.; Srivastava, D. Marginal Release Under Local Differential Privacy. In Proceedings of the International Conference on Management of Data, Houston, TX, USA, 10–15 May 2018; pp. 131–146.
35. Xu, C.; Ren, J.; She, L.; Zhang, Y.; Qin, Z.; Ren, K. EdgeSanitizer: Locally differentially private deep inference at the edge for mobile data analytics. *IEEE Internet Things J.* **2019**, *6*, 5140–5151. [[CrossRef](#)]
36. Kim, J.W.; Kim, D.H.; Jang, B. Application of local differential privacy to collection of indoor positioning data. *IEEE Access* **2018**, *6*, 4276–4286. [[CrossRef](#)]
37. Kim, J.W.; Jang, B. Workload-aware indoor positioning data collection via local differential privacy. *IEEE Commun. Lett.* **2019**, *23*, 1352–1356. [[CrossRef](#)]
38. Kim, J.W.; Lim, J.H.; Moon, S.M.; Jang, B. Collecting health lifelog data from smartwatch users in a privacy-preserving manner. *IEEE Trans. Consum. Electron.* **2019**, *65*, 369–378. [[CrossRef](#)]
39. Shokri, R.; Shmatikov, V. Privacy-preserving deep learning. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1310–1321.
40. Gong, M.; Feng, J.; Xie, Y. Privacy-enhanced multi-party deep learning. *Neural Netw.* **2020**, *121*, 484–496. [[CrossRef](#)] [[PubMed](#)]
41. Wang, T.; Li, N.; Jha, S. Locally differentially private frequent itemset mining. In Proceedings of the IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–24 May 2018; pp. 127–143.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).