IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 23, 2017, accepted December 28, 2017, date of publication January 10, 2018,
date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2791588

Application of Local Differential Privacy to Collection of Indoor Positioning Data

JONG WOOK KIM, (Member, IEEE), DAE-HO KIM, AND BEAKCHEOL JANG, (Member, IEEE)
Department of Computer Science, Sangmyung University, Seoul 03016, South Korea

Corresponding author: Beakcheol Jang (bjang@smu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education under Grant NRF-2017R1D1A1B03028097.

ABSTRACT Big data, which is being explosively generated in various areas, is considered as a new growth engine for diverse industries. In recent years, analysis of big data has attracted attention because it exhibits the potential to generate high value. In addition, with the advent of the IoT era, wherein each object is connected to all the others in a system, the importance of big data is likely to continue to be emphasized, due to the availability of data generated from diverse devices. With the increasing importance of indoor space in which most city dwellers spend over 80% of daily life, big data containing users' indoor positioning information is a critical asset for understanding the indoor behavior pattern of users, such as the shopping behavior pattern of customers in a large department store. However, there is also a risk of leakage of personal information, because it is feasible to infer the users' sensitive information by tracking and analyzing the users' indoor positions. Local differential privacy (LDP) is the state-of-the-art approach that is used to protect individual privacy in the process of data collection. LDP ensures that the privacy of the data contributor is protected by perturbing her/his original data at the data contributor's side; thus, the data collector cannot access the original data, but is still able to obtain population statistics. This paper focuses on the application of LDP to the collection of indoor positioning data. In particular, we experimentally evaluated the utilization of indoor positioning big data collected by leveraging LDP for estimating the density of the specified indoor area. Experimental results with both synthetic and actual data sets demonstrate that LDP is well applicable to the collection of indoor positioning data for the purpose of inferring population statistics.

INDEX TERMS Indoor positioning, local differential privacy, big data privacy.

I. INTRODUCTION

Today, with the wide proliferation of smartphones and mobile devices and the increasing importance of indoor space in which most city dwellers spend over 80% of daily life, various types of indoor-location-based services have attracted considerable attention. In the case of large complex buildings such as shopping malls, transportation transfer centers, and large museums, which are characterized as wide indoor spaces as well as maze-like passages, it is highly common to provide a smartphone-based service to visitors in order to aid them in identifying feasible routes to move around the site. In order to make such indoor location-based service possible, it is essential to accurately estimate the indoor position of a user; thus, related technologies have been actively studied over the last decade [13], [14], [19], [26].

Big data, which is being explosively generated in various areas, is considered as a new growth engine for diverse industries. In recent years, the analysis of big data has attracted attention because it exhibits the potential to generate high value. In addition, with the advent of the IoT era, wherein each object is connected to the others in a system, the importance of big data is likely to continue to be emphasized owing to the availability of data generated from diverse devices. Big data composed of users' indoor location information can also be used as a critical asset for understanding the indoor behavior pattern of users. For example, in order to analyze the shopping behavior pattern of customers, a large department store may collect the indoor movement data of each customer by tracking the wireless internet signals generated by a customer's mobile device. In addition, by analyzing the customer's indoor movement big data, it is feasible to reduce the bottleneck in the store and the waiting time for customers by placing popular products in a significantly less-crowded place.

As there is a growing interest in utilizing big data for decision-making, the risk of personal information leakage is also increasing. For example, Netflix, a video streaming company, released 100 million movie evaluation data of 500 000 users in a contest held in order to improve the accuracy of the movie recommendation algorithm. Notwithstanding the fact that Netflix released the data after removing personal identifiers, the researchers at the University of Texas were successful in re-identifying sensitive personal information from the released Netflix data by using other movie evaluation data available on online movie sites [25]. As the collection and utilization of big data increases, the risk of leakage of personal information also increases; thus, it is necessary to prevent the leakage of sensitive information of individuals in big data [15], [30].

In the case of collection of users' indoor position information, there is also a risk of leakage of personal information, because it is possible to infer the user's sensitive information by tracking and analyzing the user's indoor position. For example, by tracking indoor location information, it is feasible to determine which hospitals a specific user visits and thus infer what disease she/he is suffering from. Furthermore, indoor location information can reveal the products of interest in shopping malls where customers visited, the films of interest that customers watched in complex theaters, and artifacts of interest in large museums. Therefore, most users are reluctant to provide their own indoor location information to companies and organizations that desire to collect and analyze indoor location data in order to improve products or services. Hence, the utilization of the user's indoor location information big data for decision-making is highly limited in the real world.

A. CONTRIBUTIONS OF THIS PAPER

In recent years, research on privacy preserving indoor positioning technologies, which aim to protect user's privacy when using user's indoor location information, has been actively conducted. Most of the existing approaches rely on a centralized trusted server located between the user mobile device and the big data collection server. That is, a user transmits her/his indoor location information to the centralized trusted server, which in turn perturbs the user's precise indoor location information to satisfy the privacy requirements and transmits it to the big data collection server [10]–[12], [24], [36]. However, this approach is disadvantageous in that the risk of personal information leakage is very high because the exact indoor location information of a user should be transmitted to the trust server. An adversary is likely to intercept the data being transmitted from a user mobile device to a trusted server. In addition, an adversary can hack a centralized trusted server to obtain precise indoor location information of the user.

Local differential privacy (LDP) is the state-of-the-art approach that is used to protect individual privacy in the process of data collection [6]. The main concept of LDP is that the user first perturbs her/his original data by adding carefully designed random noises and then directly transmits the noisy data to the data collection server without relying on the centralized trusted server. Then, a data collector is able to compute population statistics. LDP ensures that the privacy of data users is protected because the data collector cannot access to the original user data. In real-world environments, LDP-based data collection is first implemented in Google Chrome browser to collect and track client-side information such as users' browser configuration. Although LDP is a method exhibiting high potential in collecting client-side data without concerns on the privacy leakage of data contributors, it is a more or less new technology; thus, to date, its application in real-world scenario is limited to the specific application domain [6], [7], [29]. Therefore, this paper explores the application of LDP to the collection of indoor positioning data:

- To our knowledge, this is the first study to apply LDP to the domain of indoor positioning systems, which has a growing number of applications.
- We experimentally evaluate the utilization of indoor location big data collected by leveraging LDP for the most common task in indoor location-based services: estimation of the density of the specified indoor location.

The rest of this paper is structured as follows: In Section II, we provide background information. Section III presents the method to apply LDP to the domain of indoor positioning systems. In Section IV, we experimentally evaluate the proposed approach using real and synthetic data sets. In Section V, we present the related work not already covered in the paper, and the conclusions are presented in Section VI.

II. BACKGROUND

A. RANDOMIZED RESPONSE

Randomized response is a survey technique to reduce errors caused by false answers of respondents in the case of sensitive questions [32]. Essentially, the randomized response technique is a survey method, which aims to eliminate or reduce the concerns of respondents by providing them an opportunity to select a question at a certain probability. That is, given a sensitive survey question whose answer is either ''Yes''or ''No'' (such as ''Did you cheat during tests at school?''), a survey respondent is asked to flip a fair coin in secret. If the coin comes up heads, the respondent answers the survey question truthfully. Otherwise (if the coin comes up tails), the respondent flips another coin in secret, and answers ''Yes'' (if the coin comes up a head) or ''No'' (if the coin comes up a tail). Through this method, the survey respondent has a very strong denial for the ''Yes'' or ''No'' answer of a sensitive question. Random response technique enables effective estimation of population proportion for sensitive questions, while providing sufficient protection for the privacy of respondents. For example, in the above example, the proportion of survey respondents whose truthful answer is ''Yes'' is estimated as $2 \times (\frac{N'}{N} - 0.25)$. Here, N and N' denote the total number of survey respondents and the number of

"Yes" answers including both the truthful "Yes" and random "Yes" answers, respectively.

B. DIFFERENTIAL PRIVACY

Differential privacy ensures that an adversary cannot infer with high confidence whether a particular individual is participating in the query result or not. Formally, a randomized function A satisfies ϵ-differential privacy, if and only if for (1) all database tables D and D' differing by at most one tuple and (2) any output O of A, the following equation holds [4], [5]:

$$\frac{Pr[A(D) = O]}{Pr[A(D') = O]} \leq e^{\epsilon}$$

Intuitively, given any output of A, an adversary is not able to distinguish with high confidence (controlled by the privacy parameter ϵ) whether the input of A is D or D'; this provides a strong denial for individuals included in the data sets. The common mechanism to achieve ϵ-differential privacy is to add the random noise generated from a Laplace distribution to the true result. A smaller value of the privacy parameter ϵ enforces a stronger privacy guarantee, but introduces larger noise to the true result.

C. LOCAL DIFFERENTIAL PRIVACY

Differential privacy was originally designed for the data-sharing scenario in which a trusted data curator who can access the individual data in a database perturbs the true query result computed based on the actual data in a database and sends the perturbed result to a user. On the other hands, the concept of LDP is proposed for the setting in which data contributors are asked to report their local data (to which carefully designed random noise is added such that any data contributor's information cannot be inferred with high confidence) to a data collector. Specifically, in LDP, a randomized algorithm A satisfies ϵ-differential privacy, if and only if for (1) all pairs of data contributor's data v_i and v_j, and (2) any output O of A, the following equation holds [6]:

$$\frac{Pr[A(v_i) = O]}{Pr[A(v_j) = O]} \leq e^{\epsilon}$$

Intuitively, the above equation implies that irrespective of the data that a collector receives from a contributor, the collector cannot infer with high confidence whether the contributor has sent v_i or v_j.

In real-world environments, an LDP-based privacy preserving data collection mechanism, RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), is first implemented in Google Chrome browser to collect and track the client-side data [6]. RAPPOR obfuscates the client-side data by leveraging randomized response mechanism (which will be described in detail later).

III. LDP-BASED PRIVACY-PRESERVING INDOOR POSITIONING DATA COLLECTION AND ITS APPLICATION

In this section, we first present a detailed description of the implementation of the privacy-preserving data collection mechanism for individuals' indoor positioning data by leveraging LDP. Then, we describe the application of collected indoor positioning data for estimating the density of the specified indoor location.

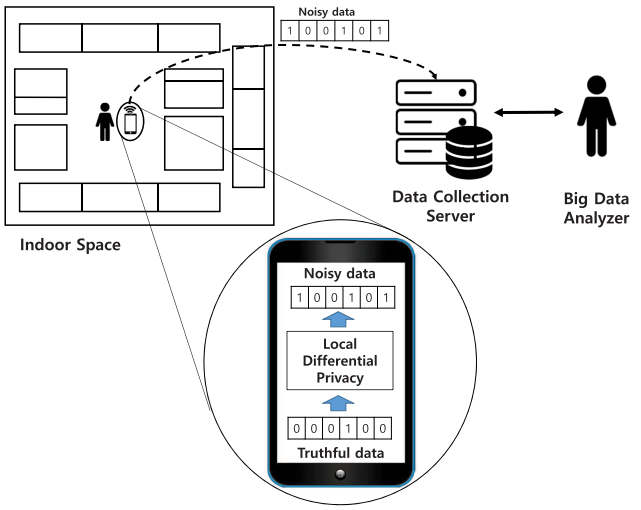


FIGURE 1. Overview of privacy-preserving indoor positioning data collection based on LDP.

A. PRIVACY-PRESERVING INDOOR POSITIONING DATA COLLECTION

Figure 1 shows an overview of the implementation of privacy-preserving indoor positioning data collection by using LDP. The method developed in this paper consists of two components: a client-side component, which is executed on a mobile device of the data contributors, and a server-side component, which operates on a data collector's server.

1) CLIENT SIDE

As described in Section I, over the last decade, there have been extensive efforts to estimate a user's indoor position. Although accurate and sophisticated methods that yield highly precise estimates of the user's indoor position are available, in this paper, we employ a straightforward beacon-based approach to estimate the indoor location of a user. This is because the purpose of this research is to approximately compute the distribution of users in the indoor space rather than precisely determine users' indoor locations. We now explain the client-side component in detail:

1) Let $B = \{b_1, b_2, \cdots b_n\}$ be a set of beacons installed in the indoor space, where each subscript represents a unique beacon ID. Here, n represents the number of beacons. For easy explanation, in this paper, we assume that the beacon IDs are from 1 to n. Received signal strength indicators from beacons are measured and sorted in descending order, and the beacon ID with the strongest signal is selected as the user's current indoor position. Let i $(1 \leq i \leq n)$ be the beacon ID with the strongest signal. Then, a n-bit array, L (which

denotes the current indoor location of a specific user) is defined as

$$L_k = \begin{cases} 1, & k = i \\ 0, & otherwise \end{cases}$$

Here, L_k represents the value of the k-th bit in L. In other words, the bit corresponding to the beacon ID with the strongest signal is set to 1, while the others are set to 0.

2) The next step is to perturb L, which is obtained from the previous step, by using the mechanism introduced by RAPPOR [6]. Each bit in L is first perturbed by randomized response as follows:

$$U_k = \begin{cases} 1, & with\ probability\ \frac{1}{2}f \\ 0, & with\ probability\ \frac{1}{2}f \\ L_k, & with\ probability\ 1-f \end{cases}$$

Here, f (whose possible value is between 0 and 1) is a system parameter that controls the level of privacy guarantee. That is, values close to 1 enforce a stronger privacy guarantee. In RAPPOR, this noisy U is referred to as the *permanent randomized response*, because it is used for all future noisy bit string of this specific L.

3) Then, RAPPOR adds another randomness by perturbing each bit in the permanent randomized response U as follows:

$$P(S_k = 1) = \begin{cases} q, & if\ U_k = 1 \\ p, & if\ U_k = 0 \end{cases}$$

That is, the probability of setting the k-th bit in S (which is referred to as the *instantaneous randomized response* in RAPPOR) to 1 is controlled by a system parameter q (or p) and the value of U_k. According to RAPPOR, the above random encoding method satisfies ϵ-differential privacy guarantee.

4) The instantaneous randomized response, S, is transmitted to the data collector server.

We note that user indoor location data (i.e., the instantaneous randomized response S) is periodically transmitted to the data collector server either at fixed time intervals or when a user moves from one location to another, which can be determined by the data collector's requirement.

2) SERVER SIDE

Upon receiving user indoor location data, the data collector server stores it in a database for future analysis. Let $\mathtt{R}(pos, ts)$ be a table in the database where pos is a current indoor location and ts denotes a current timestamp. Upon receiving the indoor location data S at a timestamp ts_{cur}, the data collection server inserts a record corresponding to (S, ts_{cur}) into the table $\mathtt{R}(pos, ts)$.

B. ESTIMATING THE DENSITY OF THE SPECIFIED INDOOR LOCATION

In this subsection, we explain the application of indoor positioning big data collected according to the process described in Subsection III-A, to estimate the density of the specified indoor location, which is one of the most significant tasks in indoor location-based services. Particularly, in this paper, we first present a straightforward statistic-based approach and then introduce an EM-based approach.

1) STATISTIC-BASED APPROACH

Let us assume that we wish to estimate the density of a specific indoor area associated with the i-th beacon, b_i, in the time interval between ts_{start} and ts_{end}. Let $set(S)$ be a set of instantaneous randomized responses that the data collection server received in the time interval between ts_{start} and ts_{end} from the data contributors. Given $set(S)$, let us assume that $set(U)$ and $set(L)$ are the corresponding sets of the permanent randomized responses and the original location bit arrays, respectively. Let us further assume that $|set(S)|$ denotes the number of elements in $set(S)$. Similarly, $|set(U)|$ and $|set(L)|$ are defined. Then, apparently, we have $|set(S)| = |set(U)| = |set(L)|$.

First, based on step 3 in Subsection III-A, the number of the instantaneous randomized responses of which the i-th bit is expected to be set to 1, $num(S_i)$, is estimated as follows:

$$\hat{num}(S_i) = q \times num(U_i) + p \times (|set(U)| - num(U_i)),$$

where $num(U_i)$ represents the actual number of permanent randomized responses in $set(U)$ whose i-th bit is set to 1. Note that in general, the hat-notation, $\hat{}$, is used to denote that the value is estimated and thus, to distinguish an estimate from the true value.

Similarly, based on step 2 in Subsection III-A, $num(U_i)$ is estimated as follows:

$$\hat{num}(U_i) = (1-f) \times num(L_i) + \frac{1}{2}f \times |set(L)|,$$

where $num(L_i)$ represents the actual number of the original location bit arrays in $set(L)$ of which the i-th bit is set to 1.

By substituting an actual value of $num(U_i)$ in the first equation with the estimated one (i.e., $\hat{num}(U_i)$) in the second formula, $num(L_i)$ can be reexpressed as

$$num(L_i) = \frac{1}{1-f} \times \left(\frac{\hat{num}(S_i) - p \times |set(S)|}{q-p} - \frac{f \times |set(S)|}{2}\right).$$

Note that in the above equation, $|set(U)|$ and $|set(L)|$ are substituted by $|set(S)|$ because $|set(S)| = |set(U)| = |set(L)|$. Then, by using the actual values, which can be directly obtained from the data collection server, the estimator of $num(L_i)$ is formulated as follows:

$$\hat{num}(L_i) = \frac{1}{1-f} \times \left(\frac{N_i - p \times N_{total}}{q-p} - \frac{f \times N_{total}}{2}\right).$$

Here, N_{total} is the total number of instantaneous randomized responses, S, that the data collection server received from

the data contributors in the time interval between ts_{start} and ts_{end} and thus, it is equal to $|set(S)|$. We note that N_{total} is directly obtained by examining the values of the attributes ts (i.e., counting the number of records in the table R whose ts attribute's value is between ts_{start} and ts_{end}). Furthermore, N_i denotes the total number of instantaneous randomized responses, S, of which the i-bit is set to 1 as well as which is received between ts_{start} and ts_{end}. Similar to the case of N_{total}, N_i can be computed by scanning the values of the two attributes ts and pos.

Then, the density of an indoor location associated with the beacon, b_i, is estimated as follows:

$$Density_{est}(b_i) = \frac{\hat{num}(L_i)}{\sum_{y=1}^{n} \hat{num}(L_y)},$$

where n represents the number of beacons as defined in Subsection III-A.

2) EM-BASED APPROACH

In this subsection, we introduce the method to estimate the density of the specified indoor location by leveraging the EM algorithm, which is a highly popular method to obtain parameter estimates when a few of the data are missing or incomplete. Let us consider the indoor position associated with the i-th beacon, b_i. Let x_i be the corresponding representation of the n-bit array L for that indoor position (i.e., the i-th bit of x_i is set to 1, and the others are set to 0). Let us assume that N records with ts values between ts_{start} and ts_{end} are present in the table R. Let further assume that $POS = \{pos_1, pos_2, \cdots, pos_N\}$ is the set of values of the attribute pos of the records in the table R whose ts attribute values lie between ts_{start} and ts_{end}. Then, given the r-th observed value pos_r, the probability that pos_r is generated from $L = x_i$ is computed by Bayes' theorem as follows:

$$\begin{aligned} P(L = x_i|pos_r) &= \frac{P(L = x_i) \times P(pos_r|L = x_i)}{P(pos_r)} \\ &= \frac{P(L = x_i) \times P(pos_r|L = x_i)}{\sum_{y=1}^{n} P(L = x_y) \times P(pos_r|L = x_y)}, \end{aligned}$$

where n represents the number of beacons as defined in Subsection III-A. In this subsection, we intend to estimate $P(L = x_i)$ (where $1 \le i \le n$), based on POS which corresponds to a set of the N observed values of the attribute pos.

The likelihood $P(pos_r|L = x_i)$ is calculated as following: Based on step 2 in Subsection III-A, given L, the probabilities that the k-th bit of the corresponding permanent randomized response, U, sets to 1 and 0 are respectively computed as follows:

$$P(U_k = 1|L_k = 1) = 1 - \frac{1}{2}f, \quad P(U_k = 1|L_k = 0) = \frac{1}{2}f.$$
$$P(U_k = 0|L_k = 1) = \frac{1}{2}f, \quad P(U_k = 0|L_k = 0) = 1 - \frac{1}{2}f.$$

Furthermore, based on step 3 in Subsection III-A, given $L_k = 1$, the probabilities that the k-th bit of the instantaneous randomized response, S, sets to 1 and 0 are respectively computed as follows:

$$\begin{aligned} P(S_k = 1|L_k = 1) &= P(U_k = 1|L_k = 1) \times q \\ &\quad + P(U_k = 0|L_k = 1) \times p \\ &= (1 - \frac{1}{2}f)q + \frac{1}{2}fp \\ P(S_k = 0|L_k = 1) &= P(U_k = 1|L_k = 1) \times (1 - q) \\ &\quad + P(U_k = 0|L_k = 1) \times (1 - p) \\ &= (1 - \frac{1}{2}f)(1 - q) + \frac{1}{2}f(1 - p) \end{aligned}$$

Similarly, given $L_k = 0$, the probabilities that the k-th bit of the instantaneous randomized response, S, sets to 1 and 0 are respectively computed as follows:

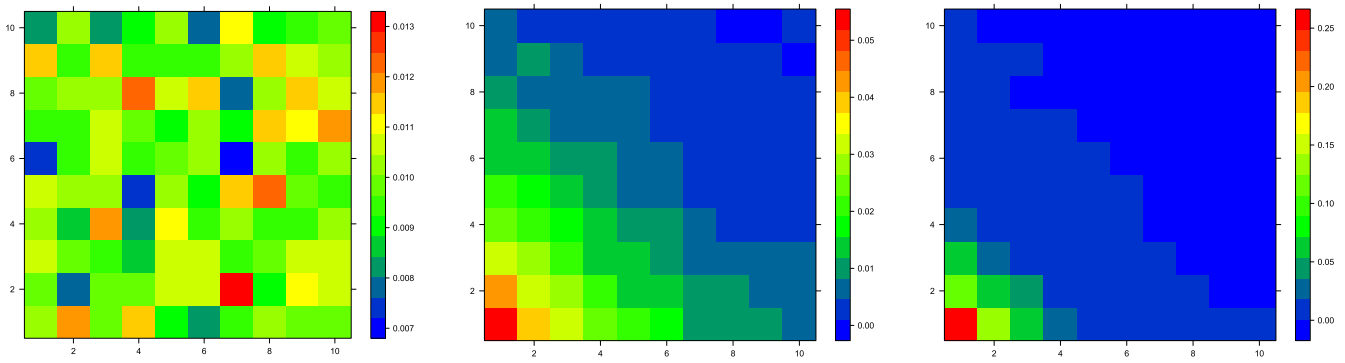
$$\begin{aligned} P(S_k = 1|L_k = 0) &= P(U_k = 1|L_k = 0) \times q \\ &\quad + P(U_k = 0|L_k = 0) \times p \\ &= \frac{1}{2}fq + (1 - \frac{1}{2}f)p \\ P(S_k = 0|L_k = 0) &= P(U_k = 1|L_k = 0) \times (1 - q) \\ &\quad + P(U_k = 0|L_k = 0) \times (1 - p) \\ &= \frac{1}{2}f(1 - q) + (1 - \frac{1}{2}f)(1 - p) \end{aligned}$$

Given the r-th observed value pos_r, let us assume that $pos_{r,u}$ denote the value of the u-th bit of pos_r (note that $pos_{r,u}$ is either 0 or 1 and $1 \le r \le n$). Then, given x_i, in which the i-th bit of x_i is set to 1 and the others are set to 0, the likelihood $P(pos_r|L = x_i)$ is defined as follows:

$$\begin{aligned} &P(pos_r|L = x_i) \\ &= \left(P(S_1 = 1|L_1 = 0)^{pos_{r,1}} \times P(S_1 = 0|L_1 = 0)^{(1-pos_{r,1})}\right) \\ &\quad \times \left(P(S_2 = 1|L_2 = 0)^{pos_{r,2}} \times P(S_2 = 0|L_2 = 0)^{(1-pos_{r,2})}\right) \\ &\quad \times \cdots\cdots\cdots \\ &\quad \times \left(P(S_i = 1|L_i = 1)^{pos_{r,i}} \times P(S_i = 0|L_i = 1)^{(1-pos_{r,i})}\right) \\ &\quad \times \cdots\cdots\cdots \\ &\quad \times \left(P(S_n = 1|L_n = 0)^{pos_{r,n}} \times P(S_n = 0|L_n = 0)^{(1-pos_{r,n})}\right) \end{aligned}$$

Example 1: Let $pos_r = 0101$ and $L = x_2 = 0100$. Then, the likelihood $P(pos_r|L = x_2)$ is computed as follows;

$$\begin{aligned} &P(pos_r|L = x_2) \\ &= \left(P(S_1 = 1|L_1 = 0)^{pos_{r,1}} \times P(S_1 = 0|L_1 = 0)^{(1-pos_{r,1})}\right) \\ &\quad \times \left(P(S_2 = 1|L_2 = 1)^{pos_{r,2}} \times P(S_2 = 0|L_2 = 1)^{(1-pos_{r,2})}\right) \\ &\quad \times \left(P(S_3 = 1|L_3 = 0)^{pos_{r,3}} \times P(S_3 = 0|L_3 = 0)^{(1-pos_{r,3})}\right) \\ &\quad \times \left(P(S_4 = 1|L_4 = 0)^{pos_{r,4}} \times P(S_4 = 0|L_4 = 0)^{(1-pos_{r,4})}\right) \end{aligned}$$



(a) (b) (c)

FIGURE 2. Three different distributions of density used in the experiments with synthetic data: in the case of skewed data sets, the distribution is skewed toward the lower-left corner. (a) Uniform distribution. (b) Medium-skewed distribution. (c) High-skewed distribution.

Since $pos_{r,1} = 0$, $pos_{r,2} = 1$, $pos_{r,3} = 0$ and $pos_{r,4} = 1$, the above equation is rewritten as

$$\begin{aligned} &P(pos_r|L = x_2) \\ &\quad = P(S_1 = 0|L_1 = 0) \times P(S_2 = 1|L_2 = 1) \\ &\qquad \times P(S_3 = 0|L_3 = 0) \times P(S_4 = 1|L_4 = 0) \end{aligned}$$

Let further assume that $f = 0$, $q = 0.75$, and $p = 0.25$. Then, the likelihood $P(pos_r|L = x_2)$ is computed as

$$\begin{aligned} &P(pos_r|L = x_2) \\ &\quad = (1-p) \times q \times (1-p) \times p \\ &\quad = (1-0.25) \times 0.75 \times (1-0.25) \times 0.25 = 0.105 \end{aligned}$$

The EM algorithm that computes $P(L = x_i)$, $1 \le i \le n$ proceeds as follows:

1) Initialization: Specify an initial parameter $\theta_i^{(0)}$ as follows:

$$\theta_i^{(0)} = P(L = x_i)^{(0)} = \frac{1}{n}, 1 \le i \le n$$

2) E-step: Compute the posterior probability $P(L = x_i|pos_r)$ using the current parameter

$$\begin{aligned} &P(L = x_i|pos_r; \theta_i^{(t)}) \\ &\quad = \frac{P(L = x_i) \times P(pos_r|L = x_i)}{\sum_{y=1}^{n} P(L = x_y) \times P(pos_r|L = x_y)} \\ &\quad = \frac{\theta_i^{(t)} \times P(pos_r|L = x_i)}{\sum_{y=1}^{n} \theta_y{}^{(t)} \times P(pos_r|L = x_y)}, \end{aligned}$$

Here, the likelihood $P(pos_r|L = x_i)$, $1 \le i \le n$ is computed as explained earlier.

3) M-step: Update θ_i as follows:

$$\theta_i^{(t+1)} = \frac{1}{N} \times \sum_{r=1}^{N} \left(P(L = x_i|pos_r; \theta_i^{(t)}) \right)$$

4) Iteration: step 2 and step 3 are repeated until the changes in parameters are within a predefined threshold as follows:

$$max_i|\theta_i^{(t+1)} - \theta_i^{(t)}| < \gamma.$$

Finally, the density of a specific indoor area associated with the beacon b_i is estimated as follows:

$$Density_{est}(b_i) = \theta_i^{(t+1)}$$

We note that [7] also uses the EM algorithm to estimate the joint probability between multiple variables based on the noised data collected by LDP. However, the approach presented in [7] exhibits a limitation in that it assumes the value of f to be equals to 0.

IV. EXPERIMENTAL EVALUATION

In this section, we describe the experiments we carried out to evaluate the effectiveness of the proposed approach. In order to evaluate the approach presented in Section III in a controlled manner, we first generated a large number of synthetic data sets with varying parameters and used these data in our initial experimental evaluation. Secondly, we also collected and used a real data set to verify the practical utility of the proposed method. In the experiments, we report the results for the statistic-based approach in Subsection III-B.1 and the EM-based approach in Subsection III-B.2. To compare these two schemes, we measure the error rate as follows:

$$error\ rate = \frac{1}{n} \times \sum_{i=1}^{n} |Density_{actual}(b_i) - Density_{est}(b_i)|.$$

Here, n represents the number of beacons, and $Density_{actual}(b_i)$ and $Density_{est}(b_i)$ correspond to the actual and estimated density, respectively, of the indoor area associated with the i-th beacon.

A. EXPERIMENTS WITH SYNTHETIC DATA

We first evaluated the proposed approach with synthetic data sets. For our experiments, we generated synthetic indoor positioning data sets as follows: We first fix the number of beacons to 100 and assume that each beacon is located at each grid of a 10×10 unit square of grids. In order to evaluate the proposed approach in various environments, we generated

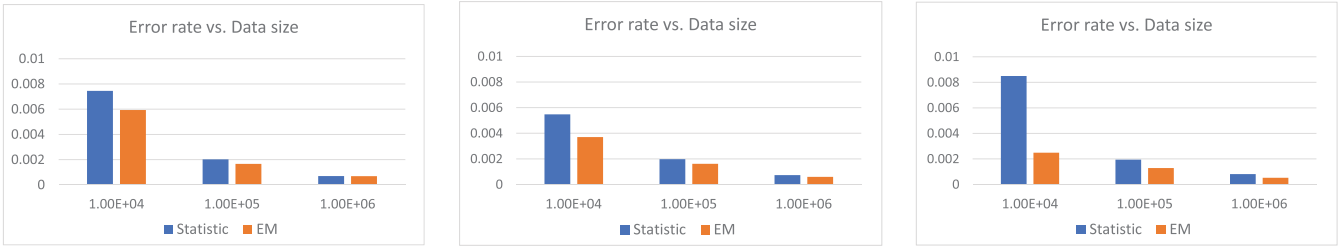


FIGURE 3. Error rate versus data size (with $f = 0$, $q = 0.75$ and $p = 0.25$). (a) Uniform distribution. (b) Medium-skewed distribution. (c) High-skewed distribution.

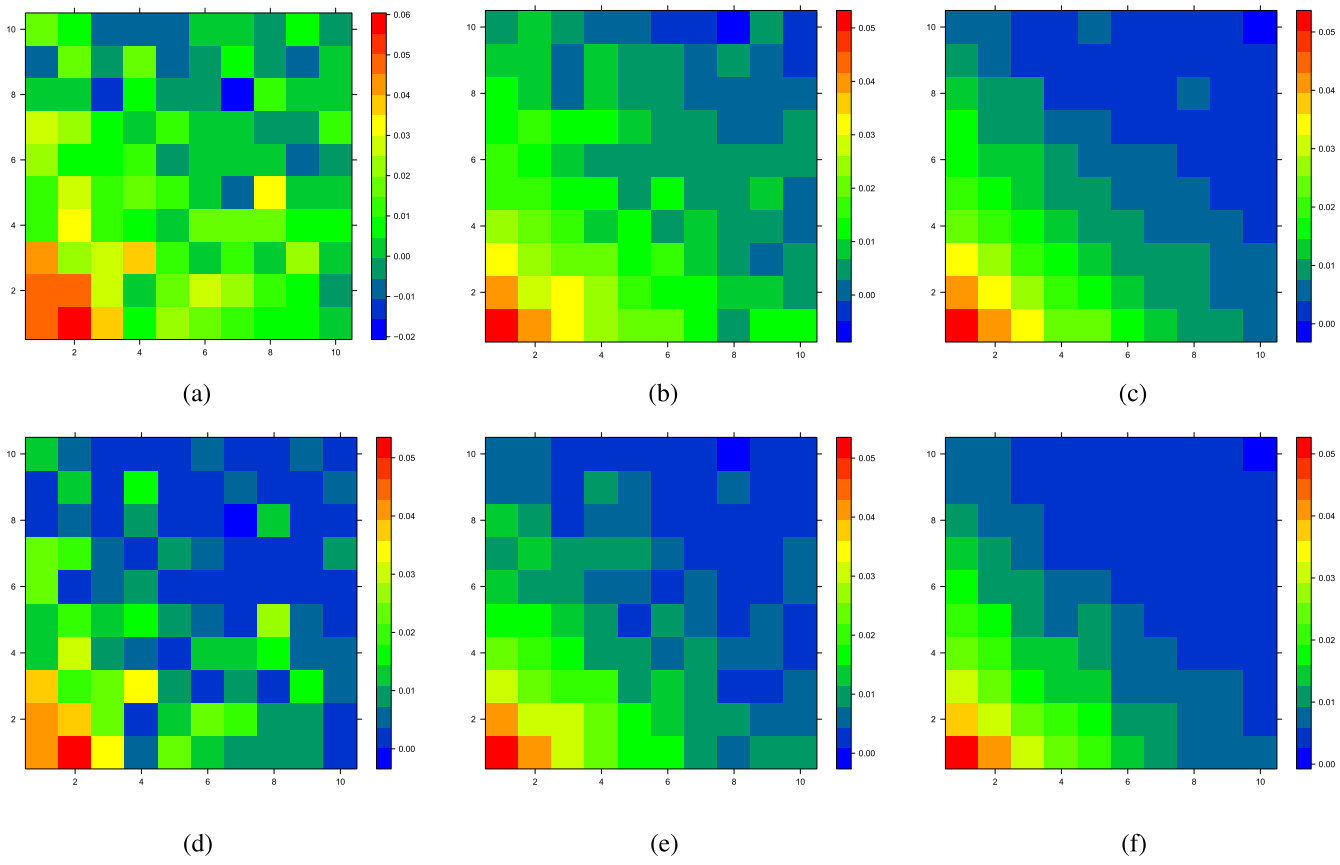


FIGURE 4. The estimated density distribution for the dataset of medium-skewed distribution in Figure 2(b). (a) Statistic-based approach (0.01M). (b) Statistic-based approach (0.1M). (c) Statistic-based approach (1M). (d) EM-based approach (0.01M). (e) EM-based approach (0.1M). (f) EM-based approach (1M).

three distributions of density: uniform distribution, medium-skewed distribution and high-skewed distribution (Figure 2). As can be seen in Figure 2 (b) and (c), in the case of skewed data sets, the distribution is skewed toward the lower-left corner. That is, the density of the skewed data sets becomes denser as one shifts toward the lower-left corner, while it becomes sparser as one shifts toward the upper-right corner. For each distribution, we generated 0.01M, 0.1M, and 1M synthetic data sets.

Figure 3 shows the error rate for varying data sizes. In this figure, the x-axis and y-axis represent the data size and the error rate, respectively. In this experiment, f, q and p are set to 0, 0.75 and 0.25 respectively, which corresponds to $\epsilon = ln(9)$. As can be seen in the figure, for all the data distributions, the error rate decreases as the data size increases. The error rate becomes closer to 0 as the data size increases, which verifies that LDP is adequately applicable to the collection of indoor positioning data. The EM-based approach outperforms the statistic-based approach in terms of error rates. However, the performance gaps between the statistic- and the EM-based schemes decreases as the size of data increases.

In order to further investigate the effect of size of data on the estimation accuracy, we plot (as shown in Figure 4)

FIGURE 5. Error rate versus f (with $q = 0.75$ and $p = 0.25$). (a) Uniform distribution (0.01M). (b) Uniform distribution (0.1M). (c) Uniform distribution (1M). (d) Medium-skewed distribution (0.01M). (e) Medium-skewed distribution (0.1M). (f) Medium-skewed distribution (1M). (g) High-skewed distribution (0.01M). (h) High-skewed distribution (0.1M). (i) High-skewed distribution (1M).

the estimated density distribution of both approaches for the data set of medium-skewed distribution that corresponds to Figure 2(b). As observed in the figure, with the small data sets (i.e., data size = 0.01M, 0.1M), the estimation accuracy is not high. In these cases, it is observed that the EM-based approach generates significantly better results than the statistic-based approach. Moreover, with the large data set (i.e., 1M), the estimated density distributions of both approaches become highly similar to that in Figure 2(b). The experimental results in Figure 3 and 4 indicate that the EM-based approach achieves higher precision in the density estimation than the statistic-based approach does, particularly when the data size is small.

Figure 5 shows the error rate for varying f in which the x-axis and y-axis represent f and the error rate respectively. In this experiment, q and p are set to 0.75 and 0.25 respectively, while f varies from 0.0 to 0.4 in increments of 0.1, which provides a level of privacy from $\epsilon = ln(3.449)$ to $\epsilon = ln(9)$. The key observations based on Figures 5 can be summarized as follows: First of all, as expected, the error rate decreases for all the data distributions as the data size increases. As can be seen in the figure, as f increases, the error rate increases. This is because as f increases, the random noise added by step 2 in Subsection III-A increases, which results in high estimation error, while ensuring high privacy level. The effect of f on the error rate of the EM-based approach is relatively marginal compared with that of the statistic-based approach. For most of the experiments, it is observed that as f increases, and thus the level of privacy increases, the performance gaps between the statistic- and the EM-based approaches become larger, which implies that the EM-based approach is suitable for applications that require high level of privacy. Furthermore, Figure 5 shows that as the degree of skewness increases, the performance gaps between the statistic- and the EM-based approaches increase.

Finally, Figure 6 shows the error rate for varying q and p. In this experiment, f is set to 0.2, while the values of (q, p) vary among (0.95,0.05), (0.85,0.15), (0.75,0.25) and (0.65,0.35), which provides a level of privacy from $\epsilon = ln(2.66)$ to $\epsilon = ln(37.73)$. The key observations based on Figures 6 can be summarized as follows: Once again, the error

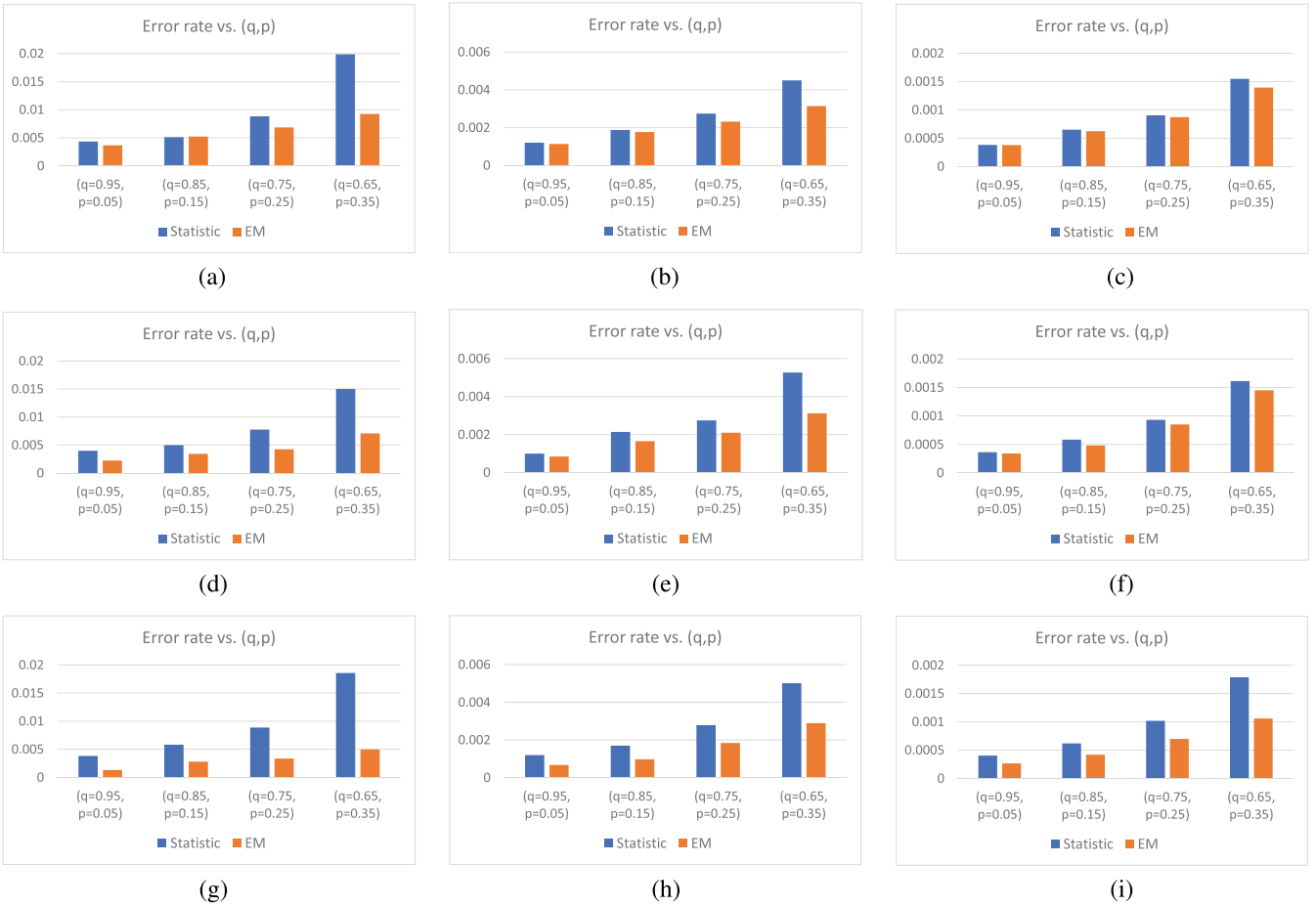


FIGURE 6. Error rate versus q and p (with $f = 0.2$). (a) Uniform distribution (0.01M). (b) Uniform distribution (0.1M). (c) Uniform distribution (1M). (d) Medium-skewed distribution (0.01M). (e) Medium-skewed distribution (0.1M). (f) Medium-skewed distribution (1M). (g) High-skewed distribution (0.01M). (h) High-skewed distribution (0.1M). (i) High-skewed distribution (1M).

rate decreases for all the data distributions as the data size increases; this is consistent with the previous experimental results. As p increases and q decreases (which is the scenario wherein the random noise added by step 3 in Subsection III-A increases), the error rate increases. As shown in the figure, the performance gaps between the statistic- and the EM-based approaches get larger, as the random noise added by step 3 in Subsection III-A increases; this indicates that the EM-based scheme is more robust against noises than the statistic-based method. Figure 6 also shows that as the degree of skewness increases, the performance variations between the statistic- and the EM-based approaches exhibit increasing trends.

In summary, the experimental results with the synthetic data sets indicate that the LDP is well applicable to the collection of indoor positioning data for the purpose of inferring population statistics. Among the two alternative methods that estimate the density of a specific indoor location based on the indoor positioning data sets collected by LDP, the EM-based approach is more robust for the privacy level, degree of data skewness, and size of collected data than the statistic-based approach.

B. EXPERIMENTS WITH REAL DATA

In this subsection, we evaluate the usefulness of the proposed method in a real-world environment setting. In order to collect real indoor positioning data in an LDP manner, we implemented the algorithm presented in Subsection III-A on the Android platform. Then, we installed 18 beacons at the Computer Science Department building of Sangmyung University, which consists of classrooms and laboratories, and collected 74008 indoor positioning data. In this experiment, f, q and p are set to 0.2, 0.75 and 0.25 respectively, which provides $\epsilon = ln(5.44)$ differential privacy.

Figure 7 shows the estimated densities of the indoor locations, which are associated with the beacon IDs, by the statistic- and EM-based approaches. In this figure, the x-axis and y-axis represent the beacon ID and the estimated density, respectively. Furthermore, for the comparison purpose, we plot the actual densities that are obtained from the original, and thus non-noisy indoor positioning data in Figure 7. As can be seen in the figure, a reasonable estimate of the densities can be obtained by both the approaches. Between the two alternative methods, the EM-based method yields

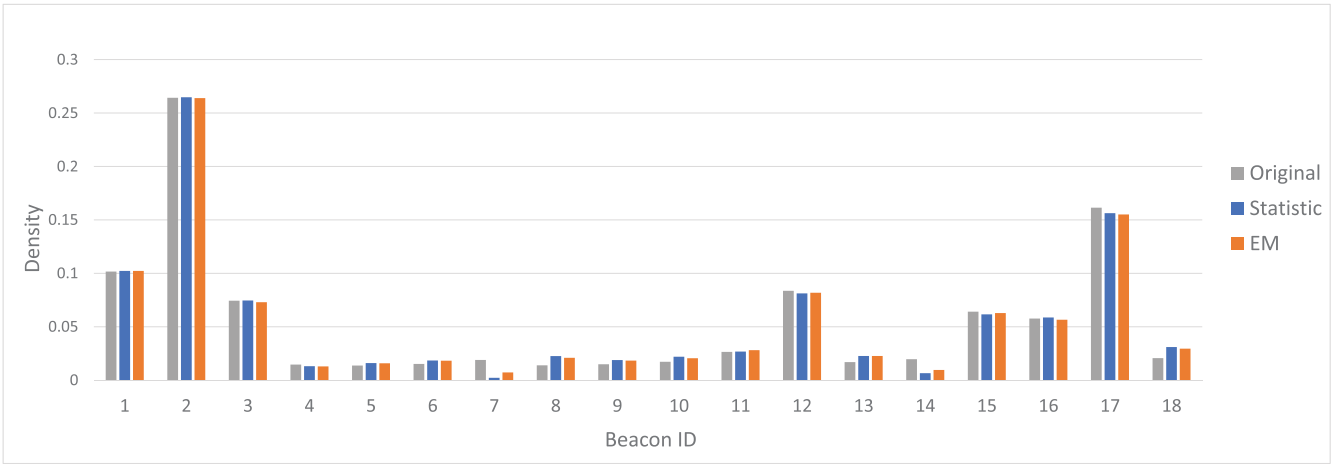


FIGURE 7. The estimated densities by the statistic- and EM-based approaches: for the comparison purpose, the actual densities that are obtained from the original indoor positioning data are plotted in the graph.

a more effective estimate of the densities than the statistic-based approach does, particularly as shown in the indoor locations associated with beacons 7 and 14. The experiment results with real data set verify that LDP is well applicable to the collection of indoor positioning data for the purpose of inferring population statistics.

V. RELATED WORK

Differential privacy [4], which is the strongest scheme for protecting individuals' privacy in released data, has been extensively studied in diverse areas, including data mining and medical analysis. Differential privacy ensures that an attacker cannot know whether a specific individual is included in the released data, regardless of any background knowledge attack. Differential privacy can be used in two different setting. The first one is offline setting where a statistical summary, such as histograms or a set of synthetic data that mimic the original data, is released for public use [5], [17], [21], [34]. The second one is online setting where the client issues a statistical query to the original database, and then a perturbed version of the query result is returned to the client [22], [27], [35]. Differential privacy can be applied to the domain of spatial data. Private Spatial Decomposition (PSD) releases differentially private spatial histograms which are generated by partitioning a spatial domain into several regions and adding carefully designed random noises to the number of objects belonging to each region in a DP-compliant manner [3], [28], [33].

Extensive studies have been conducted in the area of privacy-preserving data publishing (PPDP). The most popular anonymization algorithm, k-anonymity, was first formulated in [30]. Various algorithms have been proposed to achieve k-anonymity requirement. *LeFevre et al.* finds full-domain optimal k-anonymous generalizations with a bottom-up pruning approach [15]. *Wang et al.* proposed a bottom-up generalization algorithm to find a minimal k-anonymization for classification [31]. *Fung et al.* presented the top-down specialization scheme in which the specialization process terminates if further specialization on quasi-identifier attribute values violates k-anonymity requirement [8]. Mondrian [16] is a multidimensional generalization model that anonymizes data by recursively partitioning the space across the dimension. Clustering-based methods have been proposed to effectively find k-anonymous table. For example, [1], [2] group k similar records into a cluster and generalize each cluster to achieve k-anonymity. Besides k-anonymity, many privacy metrics have been proposed in the literature. Reference [20] introduced l-diversity that requires that each equivalence has at least l well represented values of a sensitive attribute. *Li et al.* proposed t-closeness that requires that the distribution of a sensitive attribute in each equivalence class is similar to the distribution of the entire table [18]. A comprehensive survey of privacy-preserving data publishing can be found in [9] and [23].

VI. CONCLUSION

LDP is the state-of-the-art approach that is used to protect individual privacy in the process of data collection. LDP ensures that the privacy of the data contributor is protected by perturbing her/his original data at the data contributor's side; thus, the data collector cannot access to the original data of the contributors. In this paper, we explored the application of LDP to the collection of indoor positioning data. Especially, we experimentally evaluated the utilization of indoor location big data collected by leveraging LDP for estimating the density of the specified indoor area. Experimental results with both synthetic and real data sets verify that LDP is well applicable to the collection of indoor positioning data for the purpose of inferring population statistics.

REFERENCES

[1] G. Aggarwal *et al.*, "Achieving anonymity via clustering," *ACM Trans. Algorithms*, vol. 6, no. 3, Jun. 2010, Art. no. 49.

[2] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," *Advances in Databases: Concepts, Systems and Applications*. Berlin, Germany: Springer, 2007, pp. 188–200.
[3] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2012, pp. 20–31.
[4] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Conf. Automata, Lang. Programm.*, 2006, pp. 1–12.
[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptogr.*, 2006, pp. 265–284.
[6] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1054–1067.
[7] G. Fanti, V. Pihur, and U. Erlingsson, "Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries," in *Proc. Privacy Enhancing Technol. Symp.*, 2016, pp. 41–61.
[8] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2005, pp. 205–216.
[9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010, Art. no. 14.
[10] B. Gedik and L. Liu, "Location privacy in mobile systems: A personalized anonymization model," in *Proc. 25th IEEE Int. Conf. Distrib. Comput. Syst.*, Jun. 2005, pp. 620–629.
[11] A. Gkoulalas-Divanis, V. S. Verykios, and M. F. Mokbel, "Identifying unsafe routes for network-based trajectory privacy," in *Proc. SIAM Int. Conf. Data Mining*, 2009, p. 12.
[12] A. Gkoulalas-Divanis, P. Kalnis, and V. S. Verykios, "Providing k-anonymity in location based services," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 1, pp. 3–10, 2010.
[13] R. Harle, "A survey of indoor inertial positioning systems for pedestrians," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1281–1293, 3rd Quart., 2013.
[14] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Comput.*, vol. 34, no. 8, pp. 57–66, 2001.
[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 49–60.
[16] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2006, p. 25.
[17] H. Li, L. Xiong, L. Zhang, and X. Jiang, "DPSynthesizer: Differentially private data synthesizer for privacy preserving data sharing," *Proc. VLDB Endowment*, vol. 7, no. 3, pp. 1677–1680, 2014.
[18] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.
[19] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.
[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, 2007, Art. no. 3.
[21] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2008, pp. 277–286.
[22] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp. 19–30.
[23] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. K. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 4, Oct. 2010, Art. no. 18.
[24] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new casper: Query processing for location services without compromising privacy," in *Proc. 32nd Int. Conf. Very Large Data Bases*, 2006, pp. 763–774.
[25] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy*, May 2008, pp. 111–125.
[26] K. Pahlavan, X. Li, and J. P. Makela, "Indoor geolocation science and technology," *IEEE Commun. Mag.*, vol. 40, no. 2, pp. 112–118, Feb. 2002.
[27] S. Peng, Y. Yang, Z. Zhang, M. Winslett, and Y. Yu, "Query optimization for differentially private data management systems," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2013, pp. 1093–1104.
[28] W. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2013, pp. 757–768.
[29] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren, "Heavy hitter estimation over set-valued data with local differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 192–203.
[30] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[31] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2004, pp. 249–256.
[32] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.
[33] Y. Xiao, L. Xiong, and C. Yua, "Differentially private data release through multidimensional partitioning," in *Proc. VLDB Conf. Secure Data Manage.*, 2010, pp. 150–168.
[34] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1200–1214, Aug. 2011.
[35] X. Xiao, G. Bender, M. Hay, and J. Gehrke, "iReduct: Differential privacy with reduced relative errors," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 229–240.
[36] P. Zacharouli, A. Gkoulalas-Divanis, and V. S. Verykios, "A k-anonymity model for spatio-temporal data," in *Proc. IEEE Workshop Spatio-Temporal Data Mining*, Apr. 2007, pp. 555–564.

JONG WOOK KIM (M'17) received the Ph.D. degree from the Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group at Teradata, from 2010 to 2013. He is currently an Assistant Professor of computer science with Sangmyung University. His primary research interest is in the area of data privacy, distributed databases, and query optimization. He is a member of the ACM.

DAE-HO KIM received the B.S. degree in computer science from Sangmyung University in 2017, where he is currently pursuing the master's degree with the Department of Computer Science. His research mainly focuses on data privacy and big data processing.

BEAKCHEOL JANG received the B.S. degree from Yonsei University in 2001, the M.S. degree from the Korea Advanced Institute of Science and Technology in 2002, and the Ph.D. degree from North Carolina State University in 2009, all in computer science. He is currently an Assistant Professor with the Department of Computer Science, Sangmyung University. His primary research interest is in wireless networking with an emphasis on ad-hoc networking, wireless local area networks, and mobile network technologies.