Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8886255, 11 pages
https://doi.org/10.1155/2021/8886255

Research Article

Histogram Publication over Numerical Values under Local Differential Privacy

Xu Zheng[1,2], **Ke Yan**[1,2], **Jingyuan Duan**[1], **Wenyi Tang**[1], **and Ling Tian**[1,2]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*Trusted Cloud Computing and Big Data Key Laboratory of Sichuan Province, Chengdu 610000, China*

Correspondence should be addressed to Ke Yan; kyan@uestc.edu.cn

Received 17 September 2020; Revised 9 November 2020; Accepted 13 January 2021; Published 8 February 2021

Academic Editor: Yingjie Wang

Local differential privacy has been considered the standard measurement for privacy preservation in distributed data collection. Corresponding mechanisms have been designed for multiple types of tasks, like the frequency estimation for categorical values and the mean value estimation for numerical values. However, the histogram publication of numerical values, containing abundant and crucial clues for the whole dataset, has not been thoroughly considered under this measurement. To simply encode data into different intervals upon each query will soon exhaust the bandwidth and the privacy budgets, which is infeasible for real scenarios. Therefore, this paper proposes a highly efficient framework for differentially private histogram publication of numerical values in a distributed environment. The proposed algorithms can efficiently adopt the correlations among multiple queries and achieve an optimal resource consumption. We also conduct extensive experiments on real-world data traces, and the results validate the improvement of proposed algorithms.

1. Introduction

Integrating the IoT with strong intelligent capability has been one major trend of the IoT system design. However, one prominent prerequisite of AI-driven IoT is the ubiquitous support of sensing data [1]. Recently, the pervasive adoption of smart devices provides unprecedented opportunities for data collection, benefiting the development of AI-driven IoTs [2]. However, the severe concerns on privacy have thwarted the data sharing. Therefore, this paper introduces a novel framework for distributed data statistic collection in IoTs, especially the statistics over numerical data.

Within the privacy preserved data collection, local differential privacy [3] has dramatically extended the capability on the derivation of diverse statistic information in distributed manners [4, 5]. It jointly preserves the sensitive information for data contributors under strict privacy preservation, while allowing the absence of a trusted third party as the data coordinator. It is widely accepted that LDP will be the future design principle for distributed data query with strong privacy preservation. Corresponding techniques have already been adopted by popular systems including Google Chrome [6], where contents are collected to evaluate the frequently visited websites. Currently, the studies are majorly and pervasively conducted on the frequency estimation of categorical data [7] and the mean value estimation for numerical data [8]. In this work, we consider another important topic, the histogram publication for numerical data under LDP, which is both critical and not well-handled.

The histogram provides some essential information for numerical values and can facilitate multiple services [9]. For example, understanding the distribution of health status among populations will be pivotal for policy making, which could be achieved by knowing the scales of exercises through the fitness data. The histograms can also work as a reference for numbers of values in concerned subranges, as they can be estimated via several consecutive bars in histograms. Actually, the histogram provides more abundant knowledge compared to the mean value and summation, especially for its capability on providing the contouring for data distribution. However, to mindlessly share the data for histogram publication will severely breach the privacy for contributors

[10], leading to numerous threats. For instance, the fitness data will reveal many details of a person, resulting in the pushing of spam advertisements and the raising on insurance fees. Fortunately, the local differential privacy provides potential opportunities for privacy-preserved histogram construction among multiple data contributors, even with a malicious data curator [11, 12]. Contributors can publish perturbed values to the data curator, which will aggregate the values and share it with data consumers without gaining significant knowledge on real values.

However, the implementation of the data collection must be carefully designed, as contributors are less willing to consume too much bandwidth and privacy budgets [13]. Firstly, contributors may have to spend their network resources to upload the numerical values to the data curators [14]. This is extremely unwillingness when many consumers request histograms with heterogeneous intervals. The heterogeneous histograms are common for consumers, as they usually hold different granularities of partition on the range. Some of them expect a moderate granularity on the whole range, while others may be interested in fine-grained histograms on some subintervals. Contributors will perturb the data values multiple times since the results are usually not reusable. The data value could fall in different intervals for different queries. Secondly, multiple data consumers may collude with each other and share their results. Then, the privacy of contributors will be pushed to an unexpected risk, due to the compositional property of local differential privacy.

Considering both challenges, current studies on LDP fails to provide rational solutions. Existing works can be categorized into two folds: LDP for categorical values and numerical values. The first fold mainly focuses on the frequency estimation, and they differ in how they achieve a balance between the variance and the bandwidth consumption. However, they are incapable for the histogram publication for numerical values, as there is no inherent category for a numerical value. An encoding result generated for one histogram could be totally inapplicable for another one. The second fold of studies is mainly designed for the mean value estimation over numerical data, where original data are usually perturbed into one of two fixed values [15] under LDP. The perturbed values are gathered by the data curator for estimation. However, there are few studies designed for the histogram publication under LDP.

To mitigate the gap, this paper for the first time thoroughly studies the problem of histogram publication over numerical values under LDP. In our framework, multiple data contributors each hold one numerical data. One semihonest data collector acts as the data collector, and multiple data consumers post their queries with corresponding granularities on histograms. The data curator will distribute the queries to contributors, who will later upload their noisy contents to derive the requested histograms.

We first propose two algorithms design for single and multiple histogram publication. The algorithms apply the idea of random response and take advantage of the fact that intervals of different histograms can be overlapped. The proposed algorithms are proved to achieve the optimal bandwidth consumption and privacy preservation, thus improving the efficiency for histogram publication. This paper also theoretically analyzes the accuracy and variance for the derived histogram results, together with the satisfaction on local differential privacy. Finally, we evaluate the performance of all proposed algorithms on real-world datasets, and the results reveal both the high utility and improved efficiency for the published values. As far as we know, this is the first work focusing on the efficient histogram publication for numerical values under LDP. Our main contribution includes the following:

(i) A novel framework for histogram publication over numerical data in distributed manners

(ii) Two efficient algorithms for distributed histogram publication under LDP, where both the data curator and consumers are semihonest

(iii) Theoretical analysis on the accuracy, the efficiency, and the privacy preservation

(iv) Extensive experiments on real dataset to validate the performance of proposed algorithms

The rest of the paper is organized as follows. Section 2 reviews the literature works. Section 3 proposes the problem formulation and some preliminaries. Section 4 introduces two algorithms for histogram publication. The evaluation results are shown in Section 5. Section 6 concludes the paper.

2. Related Work

2.1. Local Differential Privacy. Local differential privacy [3] has been currently treated as the standard principle of privacy preservation for distributed data publication. Existing works can be organized into two major categories: the privacy preservation for categorical values and numerical values.

As for the categorical values, Google designs RAPPOR and Basic RAPPOR methods [6] to collect the web logs from users in a private manner. In these methods, the detailed web logs will not be disclosed, while the service provider can still extract reliable information like frequently visited websites. Following the solutions, subsequent studies are conducted both to extend the capability of data collection and to reduce the requested bandwidth. The covered topics include the histogram distribution [16], the general graph structures [17], the outliers [18], range counting [19], and the frequent items [7, 20]. There are also some works [15] trying to conclude current studies on LDP and providing guidelines for applications. The efficiency of these methods is analyzed and discussed. However, these works are majorly designed for categorical values [21], where each data item has an inherent category. As for the numerical values, the extension is nontrivial. Either multiple encoded vectors or extremely large bandwidth is required.

The publication of numerical values [22] has also been studied by several works [23, 24]. Current trends of studies mainly focus on the differentially private estimation of the mean value [8, 25, 26]. Duchi et al. initially propose a

mechanism [8] for numerical data collection under LDP. The mechanism encodes each datum into one of two fixed values, which are later decoded and aggregated for analysis. Some other studies argue that the perturbed values fall out the original ranges, and designs improved mechanisms for better utilities [24]. The publication of other types of data, like the key-value data, is also studied [27]. However, all such methods are majorly designed for mean value estimation and incapable for the histogram publication.

2.2. Differential Privacy. The histogram publication is also a typical task for data publication under typical differential privacy [9, 28–31]. These works handle the differentially private releasing of histograms on different types of data structure, including the numerical values, hierarchical structures, general graphs, or other sophisticated schemas [32]. Corresponding mechanisms are proposed to reduce the scale of injected noise. All such methods request the existence of a trusted third party and ignore the bandwidth consumption during data collection.

2.3. Distributed Privacy-Preserved Data Publication. The distributed publication of private IoT data [33] has long been considered a primary task and focus. Typical techniques like K-anonymity are applied where the content held by each participant is at least indistinguishable among other $K-1$ participants. These works mainly achieve this by mixing the contents among a group of related participants [34–36]. For example, Palanisamy and Liu [37] propose a method for sensitive location concealing, by exchanging information with users in the same region. However, these studies mainly focus on the location data, which is just one domain of IoT data, the guarantees on privacy are also divergent from the differential privacy.

3. Problem Formulation

This section first provides the corresponding settings for the privacy-preserved histogram publication and then introduces some preliminaries on local differential privacy.

3.1. Problem Formulation. N data contributors are involved in the system, denoted as $\{u_1, u_2, \cdots, u_N\}$. Each contributor u_i holds one content d_i to be published. As our framework considers the publication of numerical values, each content d_i is assumed to fall in the range of $[D_L, D_U]$, where D_L and D_U stand for the minimum and maximum values.

One *data curator* collects the contents from contributors and publishes them to *data consumers*. Specifically, each data consumer provides l_j as the length of intervals in histogram queries, constituting the query set $l_1, l_2, \cdots, l_M$. The data curator first publishes the request to contributors. Upon receiving feedbacks, the data curator aggregates and derives the results for different queries. It allocates each received content d_i into the kth interval C_{jk} for jth query, where $D_L + (C_{jk}-1)l_j \le d_i \le D_L + C_{jk}l_j$. The aggregated counting $R_{jk}s$ are returned to different data consumers as the final outputs. As for each data contributor, the total bandwidth spent on the uploading of d_i is denoted as B_i.

3.1.1. Adversarial Model. In our framework, the data curator and the consumers are both malicious. They are honest-but-curious, which means they will infer the true values upon receiving the results. In this work, we adopt the local differential privacy as the measurement for privacy preservation. LDP allows the arbitrary background knowledge of adversaries while preserving the private contents for data owners. With LDP, a data contributor u_i will publish a noisy version of the content d_i to the data curator, which could be either a value or some relative data structure. The formal definition of local differential privacy is shown in Definition 1.

Definition 1 (local differential privacy). An algorithm $Q(\cdot)$ satisfies ϵ-local differential privacy$(\epsilon - \mathrm{LDP})$ where $\epsilon \ge 0$, if and only if for two arbitrary contents T_i and T_j,

$$\forall y \in \mathrm{Range}(Q):\ \Pr\left[Q(T_i) = y\right] \le e^{\epsilon} \Pr\left[Q(T_j) = y\right], \quad (1)$$

where $\mathrm{Range}(Q)$ denotes the set of all possible outputs of $Q(\cdot)$.

Intuitively, the local differential privacy ensures no significant information will be disclosed to the data receivers with arbitrary background knowledge. The parameter indicates the degree of privacy, where a larger means data contributors are less sensitive and will produce more accurate results.

3.1.2. Design Object. In our framework, the data contributors try to minimize their bandwidth consumption during the data uploading, while their privacy preservation is guaranteed. The data curator and consumers try to maintain the high utility for the derived histograms. Corresponding results should be both accurate and stable. Generally, the optimization goal is formulated as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} B_i \\ \text{s.t.} \quad & E\left(R'_{jk}\right) = R_{jk}, \quad \forall l_1, l_2, \cdots, l_M \\ & d_i \text{ is preserved under LDP}, \quad \forall i \in \{1, 2, \cdots, N\}. \end{aligned} \quad (2)$$

3.2. Preliminaries. Local differential privacy has been applied as the fundamental method for distributed content collection with strong privacy preservation. The random response method provides some basic idea for the implementation of this property. We take the Basic RAPPOR as an example, which is designed for the publication of categorical data.

In Basic RAPPOR, assume there is a L-bits vector with binary entry, denoted as $V = (v_1, v_2, \cdots, v_L)$. $v_i = 1$ indicates the data item d belongs to the ith category; otherwise, $v_i = 0$.

Then, V^0 can be generated by randomized response:

$$\Pr\left[V'[i] = 1\right] = \begin{cases} 1 - \dfrac{1}{2}f, & \text{if } V[i] = 1, \\ \dfrac{1}{2}f, & \text{if } V[i] = 0. \end{cases} \quad (3)$$

Finally, $V's$ will be sent to the data curator for subsequent analysis. Actually, this mechanism of perturbation achieves LDP property for vector V, which is proved by a previous work [15]:

Theorem 2. *For an arbitrary vector $V = (v_1, v_2, \cdots, v_L)$, the Basic RAPPOR achieves $\epsilon - LDP$ for $\epsilon = \ln(((1 - (1/2)f)/(1/2))^2)$.*

The data sampling, where contributors only partially upload their contents, is also a major strategy for resource saving in distributed data collection. It is believed that this can further reduce the disclosure of information. There are also some works arguing the amplification of the privacy preservation over data sampling. Li et al. have theoretically proved the effect [38], as is given in Theorem 3.

Theorem 3. *Assume $F(\cdot)$ to be an ϵ-differentially private algorithm and $S(\cdot)$ to be a sampling method algorithm. Then if $S(\cdot)$ is first applied to a dataset, which is later perturbed by $F(\cdot)$, the derived result satisfies $\ln(1 + P_0(e^{\epsilon} - 1))1$-differential privacy, where P_0 is the sampling probability.*

Finally, the compositional property of differential privacy can also be merged with the LDP.

Theorem 4 (sequential composition [39]). *Let $\{F_1(\cdot), F_2(\cdot), \cdots, F_k(\cdot)\}$ be a set of functions satisfying differential privacy and the privacy budgets to be $\epsilon_1, \epsilon_2, \cdots, \epsilon_k$, respectively. Then applying all $F_i(\cdot)s$ to one data item d_0 will provide a $\sum_{i=1}^{k} \epsilon_i$ -differential privacy.*

4. Distributed Histogram Publication under Local Differential Privacy

This section provides the algorithms for histogram publication. It first introduces the algorithm designed for single query; then an efficient algorithm designed for multiple queries is proposed.

4.1. Baseline Algorithm for Single Query. The first algorithm helps the data curator collect data from contributors for one single query. The main idea of this algorithm is to first convert the numerical value into categorical version and then applies typical mechanisms like the Basic RAPPOR. This conversion is feasible as a single query will provide a fixed partition on the whole range. We name the algorithm as *Single Histogram Publication* to distinguish it with subsequent methods, SHP for short.

In SHP, the data curator initially receives the query from data consumers, i.e., the width l_0 for each interval in histogram and the privacy budget ϵ_0. The data curator pushes the parameters to all data contributors, together with the range $[D_L, D_U]$.

4.1.1. Local Encoding. Upon receiving the message, each contributor u_i first encodes her value d_i into vector

$$D_i = (0, \cdots, 0, 1, 0, \cdots, 0), \tag{4}$$

where the jth entry equals 1 when

$$D_L + (j-1)l_0 \le d_i \le D_L + j \cdot l_0. \tag{5}$$

With the vector D_i, SHP applies the typical perturbation mechanisms like Basic RAPPOR, where

$$f = \frac{2}{e^{\epsilon/2} + 1}. \tag{6}$$

Assume the perturbed vector to be D_i', and contributor u_i uploads this vector to the data curator.

4.1.2. Decoding and Publishing. The data curator will first collect the vectors from all contributors. Then, it decodes and aggregates the vectors to derive the estimated counting for values in each interval. For each interval C_k, the number of contents that fall in this slot is calculated as

$$R_k = \frac{\left\| \left\{ D_i' \middle| D_{ik}' = 1 \right\} \right\| - 1/2 \cdot f \cdot N}{1 - f}, \tag{7}$$

where $\|\cdot\|$ indicates the number of elements in the set.

Finally, the data curator publishes the estimated results $(R1, R2, \cdots)$ to the data consumer.

4.1.3. Analysis. Several properties should be analyzed for the proposed algorithm, including the accuracy for the derived results, the guarantee on privacy preservation, and the efficiency.

Firstly, SHP provides an unbiased estimation for the histogram when applying Basic RAPPOR as the perturbation mechanism. The analysis is as follows: according to the property of the perturbation mechanism, the data curator can aggregate the vectors and derive an unbiased estimation on the histogram, as

$$E(R_k) = R_k^0, \tag{8}$$

where R_k^0 is the original result derived from all vectors $\{D_1, D_2, \cdots, D_N\}$. Meanwhile, SHP generates each vector D_i by projecting d_i into a corresponding interval. Then, the only 1 in D_i exactly refers to the index of interval d_i belonging to. Therefore, SHP can provide an unbiased estimation in each interval.

Theorem 5 (unbiased estimation). *The published result of SHP is an unbiased estimation for the real histogram, i.e., $E(R_k) = \|\{d_i \mid D_L + (k-1)l_0 \le d_i \le D_L + k \cdot l_0, \forall_i \le N\}\|$.*

Furthermore, the variance of SHP is determined by the applied perturbation mechanism, as generating D_i will introduce no extra randomness.

Now, we discuss the capability of privacy preservation of SHP. The analysis is also straightforward. The information in D_i is preserved with local differential privacy, where the privacy budget is ϵ_0. Furthermore, D_i provides identical information with d_i in the histogram publication, according to the encoding phase. Therefore, SHP can preserve the private content for each contributor with expected differential privacy.

Theorem 6 (local differential privacy). *SHP can preserve the numerical content of each contributor with ϵ_0 local differential privacy.*

Finally, we briefly discuss the efficiency of SHP. The bandwidth spent on content uploading is $O((D_U - D_L)/l_0)$.

The time complexity for each contributor is also $O((D_U - D_L)/l_0)$ during the encoding phase and $O(N \cdot ((D_U - D_L)/l_0))$ for the data curator during the decoding phase.

4.2. An Efficient Algorithm for Multiple Queries. This part gives the algorithm for histogram publication towards multiple queries. These queries could be heterogeneous on their widths of intervals, making the data publication nontrivial. To simply apply SHP for each query separately will consume huge bandwidths and privacy budgets. Therefore, the main idea of the proposed algorithm is to implement a single-time publication meeting all queries, to improve the efficiency for contributors. The algorithm is named as *Composited Histogram Publication*, CHP for short.

Initially, the data curator receives the queries from multiple data consumers, each with a set of parameters (l_i, ε_i). CHP extracts the minimum privacy budgets $\varepsilon_0 = \min_i \varepsilon_i$, which will provide a most rigorous privacy preservation for contributors. Then, CHP adopts all l_is. It first derives intervals for all queries and records the boundaries for these intervals as

$$\begin{aligned}&\{\{W11, W12, \cdots, W1K1\}, \{W21, W22, \cdots, W1K2\}, \cdots,\\ &\cdot\{WM1, WM2, \cdots, WMKM\}\}.\end{aligned} \tag{9}$$

Then, CHP arranges all boundaries on one single line in an ascending order. The start point of the line is D_L, and the end point of the line is D_U. CHP merges multiple boundaries when they refer to the same value. After the arrangement, CHP derives an integrated partition on $[D_L, D_U]$, denoted as

$$\{W_1, W_2, \cdots, W_{K0}\}, \tag{10}$$

where $W_1 = D_L$ and $W_{K0} = D_U$. At the end of this phase, CHP distributes the partition together with the privacy budget ϵ_0 to all contributors.

4.2.1. Local Encoding. Upon receiving the partition on whole range, each contributor u_i first encodes her value d_i into the vector similar with SHP:

$$D_i = \left(D_{i1}, \cdots, D_{iK_0-1}\right), \tag{11}$$

where the $D_{ij} = 1$ when

$$W_j \le d_i \le W_{j+1} \tag{12}$$

and $D_{ij} = 0$ otherwise.

With the vector D_i, CHP also applies the typical perturbation mechanisms, for example, Basic RAPPOR, with the following perturbation probability:

$$f = \frac{2}{e^{\epsilon_0/2} + 1}. \tag{13}$$

Finally, the perturbed vector D_i' will be sent to the data curator.

4.2.2. Decoding and Publishing. In this phase, the data curator will fuse the vectors collected from contributors and estimate the accumulated contents within each interval. Then, the data curator generates and publishes results for consumers, respectively.

In the first step, CHP estimates the counting R_k for each interval $[W_k, W_{k+1}]$ in the integrated partition as

$$R_k = \frac{\left\| \left\{ D_i' \middle| D_{ik}' = 1 \right\} \right\| - 1/2 \cdot f \cdot N}{1 - f}, \quad \forall k \le K_0.. \tag{14}$$

In the second step, for each consumer, CHP estimates the counting in each interval by accumulating the corresponding intervals in the integrated partition.

$$R_{ij} = \sum_{h=p}^{q} R_h, \tag{15}$$

where R_{ij} indicates the number of numerical values falling in range $[W_{ij}, W_{ij+1}]$ and $W_p = W_{ij}$, $W_q = W_{ij+1}$.

Finally, the data curator distributes the corresponding result set R_i to each consumer.

4.2.3. Analysis. Now, we analyze the performance for CHP. This part first proves that CHP can derive unbiased estimation of histograms for all data consumers. Then, the guarantee on differential privacy is given. Finally, this part shows the efficiency of CHP on bandwidth consumption.

The estimation in CHP includes three major steps: the generation of the integrated partition, the vector encoding and decoding, and the counting of outputs for each consumer. In the first step, CHP guarantees that there will be exactly a continuous set of intervals $W_i, W_i + 1, \cdots, W_{i+p}$ covering the same range for every $[W_{jk}, W_{jk} + 1]$. As for each of the interval in the set, the encoding and decoding in CHP will provide an unbiased estimation for the counting of numerical values inside. Finally, CHP estimates the result for $[W_{jk}, W_{jk} + 1]$ by adding up the results for intervals in $W_i, W_i + 1, \cdots, W_{i+p}$. This accumulation is a combination of unbiased estimation covering the same range, and thus, the final output is unbiased. The following theorem gives the corresponding conclusion.

Theorem 7 (unbiased estimation). *In CHP, the data curator provides unbiased histograms for all data consumers.*

Similar with SHP, the variance of the estimated result for CHP is also determined by the adopted perturbation mechanism. The major difference is the composition of multiple intervals during the final step, which will not change the scale of the variance.

Now, we discuss the property of differential privacy for CHP. It is obviously that CHP can provide ϵ_0-local differential privacy for each contributor. The analysis is the same with SHP as CHP applies the similar idea for data encoding and perturbation.

Furthermore, CHP allows each contributor to publish only once to respond for all queries. This is different from the baseline solution where SHP has to be applied M times, due to the heterogeneous partitions on the range. In the later case, it should be noticed that multiple data consumers could be malicious, and they will collude by sharing their results. Then, the actual privacy budget could be larger than $M \cdot \varepsilon_0$, which is much worse and usually unacceptable for data contributors. Theorem 8 states this property.

Theorem 8 (local differential privacy). *CHP preserves the numerical content of each contributor with ϵ_0-local differential privacy, even if the data consumers are malicious and comprehensively share their results.*

Finally, we discuss the efficiency of CHP. The bandwidth consumption for each contributor is no more than $O(\sum_{i=1}^{M} (D_U - D_L)/l_i)$. Accordingly, the time complexity for each contributor is also $O(\sum_{i=1}^{M}(D_U - D_L)/l_i)$, while the time complexity for the data curator in deriving the results is $O((\sum_{i=1}^{M}(D_U - D_L)/l_i)^2)$.

Actually, CHP also guarantees the minimum number of bits in providing unbiased estimation for all queries. This property indicates CHP achieves optimal efficiency on bandwidth consumption. Theorem 9 shows the property and analysis.

Theorem 9 (efficiency). *With the unbiased perturbation mechanism, CHP achieves the unbiased estimation for all histograms with minimum number of encoding bits.*

Proof 1. We prove the theorem by contradiction. To derive an unbiased estimation for all queries, the boundaries in each of them must also appear in the integrated partition. This is exactly the same with the construction of the integrated partition.

Now, assume that some consecutive intervals can be merged to reduce the total bits, i.e., R_i and R_j, while the unbiased estimation is kept. Then, some boundaries in the integrated partition will be eliminated, i.e., the boundary between R_i and R_j. This is contradicted with the requirement where the boundaries for all queries should be retained for unbiased estimation, as the boundary between R_i and R_j is generated according to some queries.

Therefore, no intervals in CHP could be merged, and the minimum number of encoding bits is achieved.

Table 1: Statistics for datasets.

	Total contributors	Max salary	Min salary
Baltimore	13,683	250,000	1,800
New York	138,715	297,625	1
San Francisco	291,825	515,102	0

4.3. Discussion. Our framework assumes that each participant holds exactly one content. However, it can also fit participants with multicontents. The extension could be achieved by two strategies. In the first category, a participant can encode each of her content into one independent bit vector, and then uploads these vectors to the service provider. In this case, the total bandwidth of uploading is determined by the scales of contents and the bits for encoding. In the second category, a participant can first encode each of her content according to the first strategy. Then, these bit vectors will be accumulated, and each entry of the aggregated vector will record the total number of vector with "1" for the same entry in the vector. In this later case, the total number of bandwidth will be determined by the bits within one vector, which is significantly reduced when compared with the first strategy. We can also prove the results are still unbiased, which can be extended from Theorems 5 and 7.

5. Evaluation

This section evaluates the performance of the proposed algorithms. We adopt the salary data collected for normal citizens in the United States [40]. Specifically, we extract the information in New York city, San Francisco, and Baltimore, respectively. The statistics of the three cities are shown in Table 1. In our evaluation, multiple data consumers expect to derive the distribution of incoming levels in different granularities. Therefore, they will post their requests on histogram publication. The data contributors will publish their data to the consumers, and the privacy concerns and the bandwidth consumption should be treated. The data curator acts as the coordinator among the two sides.

As the extension of current studies on numerical values is nontrivial, we compare their performance with one baseline algorithm. In this algorithm, the data contributors respond to each consumer separately. To thwart the collusion among consumers, the baseline algorithm requests the consumers to share the privacy budgets among multiple responses, e.g., assume the total privacy budgets to be ϵ_0, then a contributor will apply ϵ_0/K budget to each of K queries. We also compare the performance with the sampling algorithm.

The metric applied for the evaluation is the mean square errors (MSE for short). Furthermore, we run each test group 20 times to alleviate the influence of randomness.

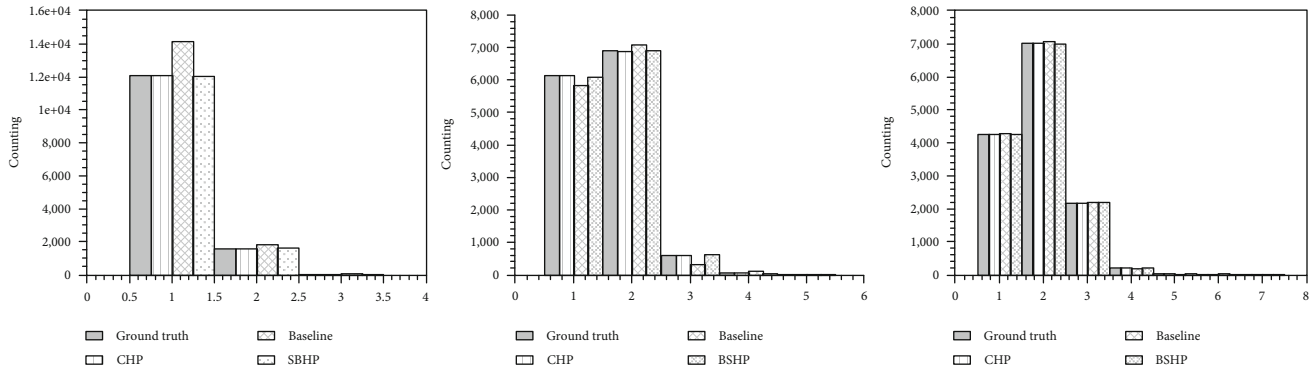

Figure 1: Multigranularity histograms for Baltimore.

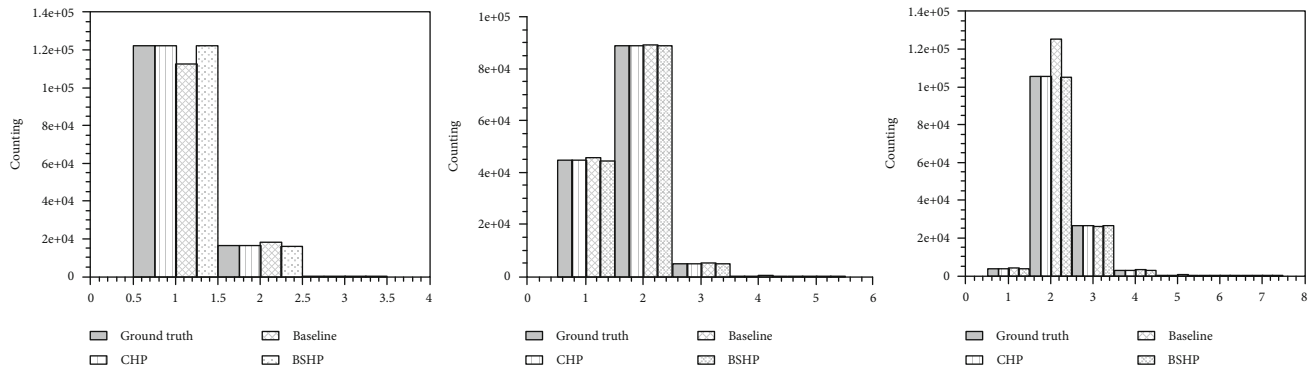

Figure 2: Multigranularity histograms for New York city.

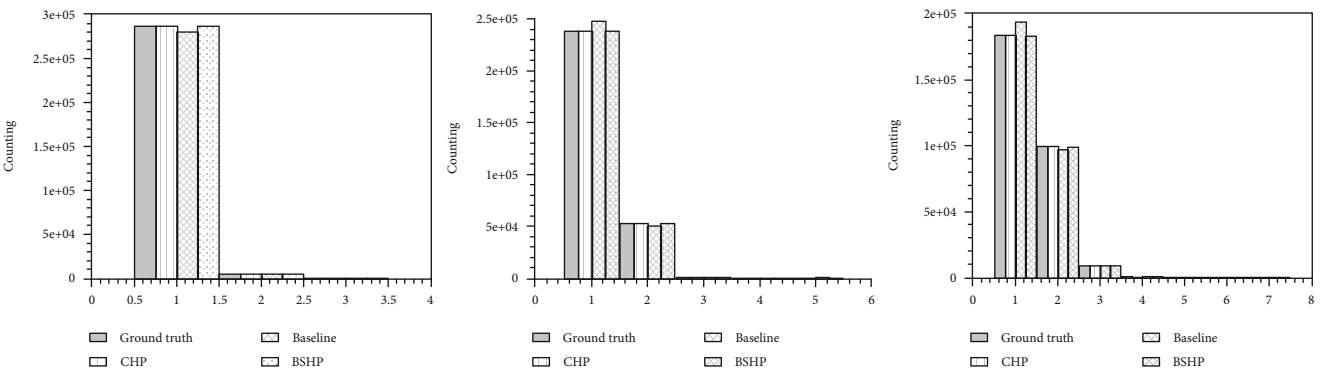

Figure 3: Multigranularity histograms for San Francisco.

5.1. Basic Performance. This part studies the basic performance for all proposed methods. We are interested in both the derived results and the overall effectiveness of the methods. The parameter settings are as follows: there are three consumers in the system, requesting 3-fold, 5-fold, and 7-fold histograms, respectively. They share the total budgets with $\varepsilon = 15$, where the baseline algorithm partitions the budgets among all three consumers. Our proposed algorithms, on the other hand, can allocate all budgets in one output.

The results are shown in Figures 1–3. As we see, the proposed algorithms provide better utilities. They outperform the baseline algorithms in all groups and achieve more accurate shapes for histograms. The difference is actually very significant, when considering there are many data values belonging to some intervals to reduce the influence of randomness.

We also compare the MSE performance of all algorithms with various privacy budgets. In this group, the privacy budgets vary from 3 to 18. According to the results in Figure 4, the proposed algorithms can reduce the MSE for histograms. The improvement is more significant when the privacy budgets is relatively large. We can see that CHP will introduce few errors when $\varepsilon = 18$, indicating the achievement of high accuracy. The reason is that these algorithms bypass the partition of budgets towards multiple queries, thus reducing the noise in published data.

Finally, the performance for San Francisco is worse than those for NYC and Baltimore. One potential reason is that

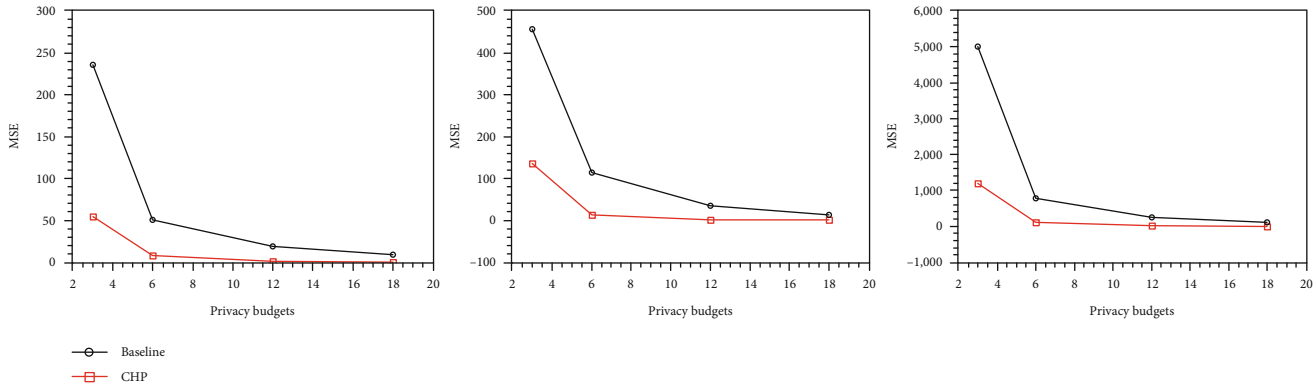

Figure 4: Mean square errors for histogram with various privacy budgets.

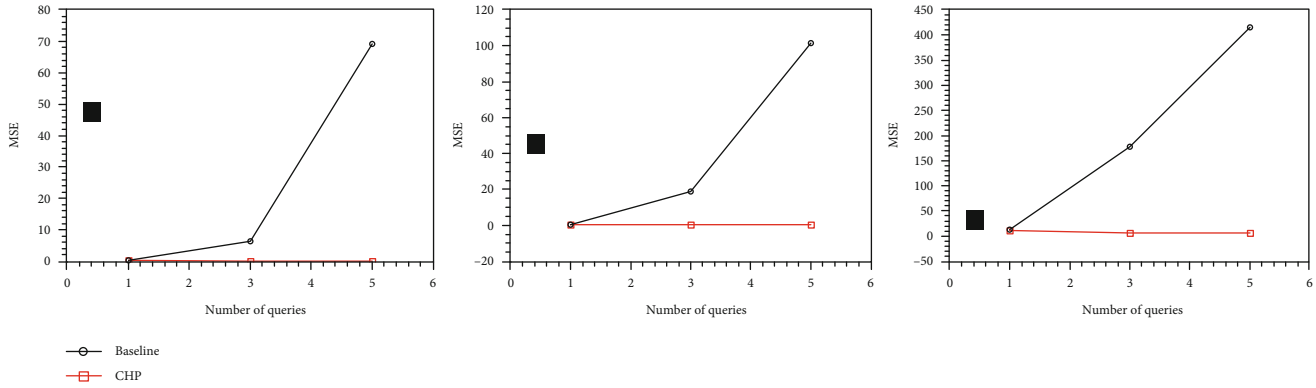

Figure 5: Mean square errors with different numbers of queries.

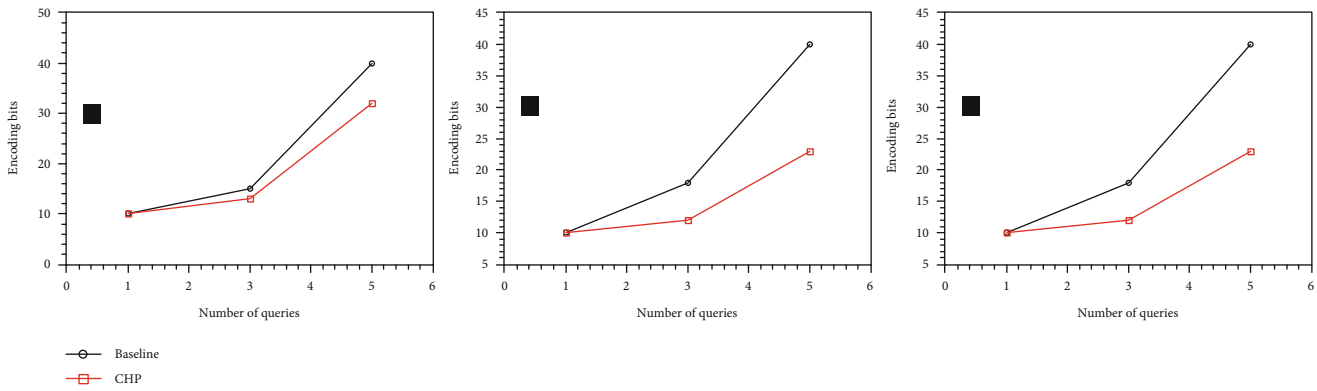

Figure 6: Encoding bits with different methods.

some intervals of histograms for San Francisco include few data values. However, the noise in the outputs still exists, which will be aggravated and lead to severe increase on MSE.

5.2. Heterogeneous Data Consumers. Within the real histogram publication, data consumers could be diverse on their behaviors. Therefore, the performance should be validated under different circumstances. In this part, three groups of data consumers are considered. The first group includes one single consumer, requesting a 10-fold histogram. The second group includes three data consumers requesting 3-fold, 5-fold, and 7-fold histograms. The third one has 5 consumers inside, whose requests are 3 folds, 5 folds, 7 folds, 10 folds, and 15 folds. The privacy budget is 15, and the sampling ratio is 0.8. The results are shown in Figure 5.

We observe that CHP and the sampling-based algorithm can maintain a similar performance among different groups, besides their low MSE. This is due to the fact that both algorithms request the data publication to be executed only once for multiple queries. Nevertheless, the baseline algorithm will execute the publication once for each queries. Then, the performance will suffer dramatic falling when the number of

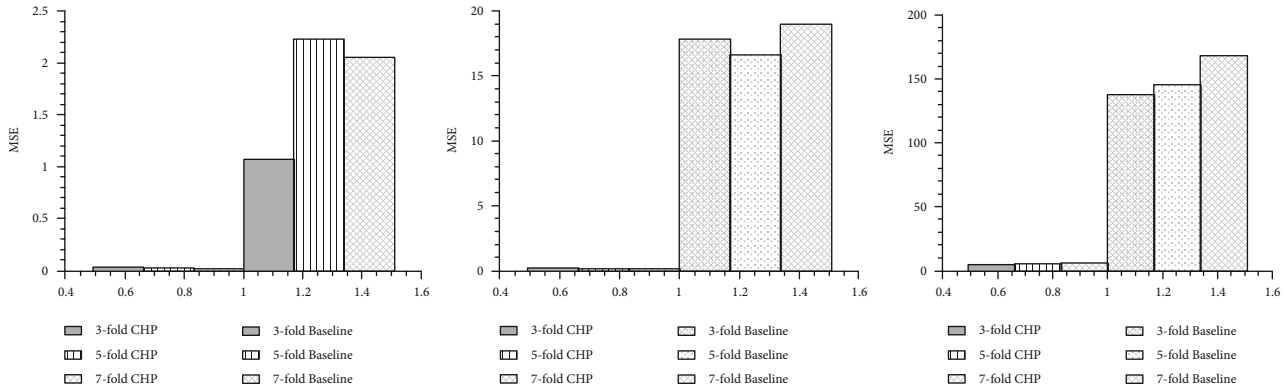

Figure 7: Comparison of MSE among different queries.

queries rises. We also observe that CHP reduces the number of total bits for encoding in Figure 6, which is already noticeable with very few queries.

We also compare the performance among different queries. Figure 7 shows the MSEs for 3-fold, 5-fold, and 7-fold histograms. According to the results, both CHP and the baseline algorithm can guarantee relatively similar performance on all queries. This again validates our analysis that the variance is determined by the privacy budgets, which are identical for different queries and their folds.

6. Conclusion

Local differential privacy provides novel paradigms for distributed and safety data queries. Various techniques have been designed towards heterogeneous categories of queries. However, the histogram, providing some essential information for the numerical data, has not been thoroughly considered. Existing methods are either incapable or lead to unwillingness resource consumption. As a result, this paper proposes a novel framework towards the differentially private publication of histogram over numerical values. A novel encoding mechanism is designed where the numerical data could be encoded once for multiple queries. It achieves highly efficient bandwidth consumption and can reduce the unnecessary waste on privacy budgets. The accuracy of derived results, the optimization on bandwidth consumption, and the strict privacy preservation are analyzed for all algorithms, and we also discuss the extension for online queries.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2018YFC0807500), by the National Natural Science Foundation of China (Nos. U19A2059 and 61802050), and by the Ministry of Science and Technology of Sichuan Province Program (Nos. 2018GZDZX0048 and 20ZDYF0343).

References

[1] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and J. Deen, "A tensor-based multi-attributes visual feature recognition method for industrial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 2231–2241, 2020.

[2] L. Ren, Z. Meng, X. Wang, L. Zhang, and L. T. Yang, "A data-driven approach of product quality prediction for complex production systems," *IEEE Transactions on Industrial Informatics*, no. 99, p. 1, 2020.

[3] R. Bassily and A. Smith, "Local, private, efficient protocols for succinct histograms," in *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pp. 127–135, New York, NY, USA, June 2015.

[4] W. Zhu, P. Kairouz, H. Sun, B. McMahan, and W. Li, "Federated heavy hitters discovery with differential privacy," 2019, http://arxiv.org/abs/1902.08534.

[5] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[6] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067, Scottsdale, Arizona, USA, November 2014.

[7] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," in *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 435–447, Houston, TX, USA, May 2018.

[8] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–201, 2018.

[9] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB Journal*, vol. 22, no. 6, pp. 797–822, 2013.

[10] J. Wang, Z. Cai, and J. Yu, "Achieving personalized k-anonymity-based content privacy for autonomous vehicles in cps," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4242–4251, 2020.

[11] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial iots," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[12] X. Zheng, G. Luo, and Z. Cai, "A fair mechanism for private data publication in online social networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 880–891, 2020.

[13] X. Zheng, Z. Cai, J. Li, and H. Gao, "Locationprivacy-aware review publication mechanism for local business service systems," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, 2017.

[14] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[15] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proc. of the 26th USENIX Security Symposium*, pp. 729–745, Vancouver, BC, Canada, 2017.

[16] S. Wang, L. Huang, P. Wang, H. Deng, H. Xu, and W. Yang, "Private weighted histogram aggregation in crowdsourcing," in *International Conference on Wireless Algorithms, Systems, and Applications*, pp. 250–261, Springer, 2016.

[17] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren, "Generating synthetic decentralized social graphs with local differential privacy," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 425–438, Dallas, TX, USA, 2017.

[18] X. Zheng, A. Chen, G. Luo, L. Tian, and Z. Cai, "Privacy-preserved distinct content collection in human-assisted ubiquitous computing systems," *Information Sciences*, vol. 493, pp. 91–104, 2019.

[19] Z. Cai and Z. He, "Trading private range counting over big iot data," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 144–153, Dallas, TX, USA, 2019.

[20] 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 *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 192–203, Vienna, Austria, 2016.

[21] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[22] L. Ren, Z. Meng, X. Wang, R. Lu, and L. T. Yang, "A wide-deep-sequence model-based quality prediction method in industrial process analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3721–3731, 2020.

[23] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, NW Washington, DC, USA, 2013.

[24] N. Wang, X. Xiao, Y. Yang et al., "Collecting and analyzing multidimensional data with local differential privacy," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 638–649, Macao, China, 2019.

[25] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy," *Information Sciences*, vol. 250, pp. 200–214, 2013.

[26] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1176–1184, 2015.

[27] Q. Ye, H. Hu, X. Meng, and H. Zheng, "Privkv: key-value data collection with local differential privacy," in *PrivKV: Key-Value Data Collection with Local Differential Privacy*, San Francisco, CA, USA, 2019.

[28] X. Zhang, R. Chen, J. Xu, X. Meng, and Y. Xie, "Towards accurate histogram publication under differential privacy," in *Proceedings of the 2014 SIAM international conference on data mining*, pp. 587–595, Philadelphia, Pennsylvania, USA, 2014.

[29] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1021–1032, 2010.

[30] G. Acs, C. Castelluccia, and R. Chen, "Differentially private histogram publishing through lossy compression," in *2012 IEEE 12th International Conference on Data Mining*, pp. 1–10, Brussels, Belgium, 2012.

[31] Y.-H. Kuo, C.-C. Chiu, D. Kifer, M. Hay, and A. Machanavajjhala, "Differentially private hierarchical count-of-counts histograms," *Proceedings of the VLDB Endowment*, vol. 11, no. 11, pp. 1509–1521, 2018.

[32] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, 2018.

[33] X. Wang, L. T. Yang, Y. Wang, L. Ren, and M. J. Deen, "Adtt: a highly-efficient distributed tensor-train decomposition method for iiot big data," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1573–1582, 2020.

[34] S. Amini, J. Lindqvist, J. Hong, J. Lin, E. Toch, and N. Sadeh, "Caché: caching location-enhanced content to improve user privacy," in *Proceedings of the 9th international conference on Mobile systems, applications, and services - MobiSys '11*, pp. 197–210, Washington, DC, USA, 2011.

[35] R. Lu, X. Lin, Z. Shi, and J. Shao, "Plam: a privacy-preserving framework for local-area mobile social networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pp. 763–771, Toronto, ON, Canada, April 2014.

[36] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Optimal geo-indistinguishable mechanisms for location privacy," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 251–262, Scottsdale, Arizona, USA, November 2014.

[37] B. Palanisamy and L. Liu, "Mobimix: protecting location privacy with mix-zones over road networks," in *2011 IEEE 27th International Conference on Data Engineering*, pp. 494–505, Hannover, Germany, April 2011.

[38] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy or, kanonymization meets differential

privacy," in *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security - ASIACCS '12*, pp. 32-33, Seoul, Korea, 2012.

[39] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, pp. 627–636, New York, NY, USA, 2009.

[40] "Data.world," https://data.world/datasets/salary.