Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6648775, 11 pages
https://doi.org/10.1155/2021/6648775

Research Article

A Sampling-Based Method for Highly Efficient Privacy-Preserving Data Publication

Guoming Lu,[1,2] Xu Zheng,[1,2] Jingyuan Duan,[1] Ling Tian,[1,2] and Xia Wang[3]

[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*
[3] *School of Statistics and Data Science, Beijing University of Technology, Beijing, China*

Correspondence should be addressed to Xu Zheng; xzheng@uestc.edu.cn

Received 26 October 2020; Revised 19 February 2021; Accepted 3 March 2021; Published 15 March 2021

Academic Editor: Lin Wang

The data publication from multiple contributors has been long considered a fundamental task for data processing in various domains. It has been treated as one prominent prerequisite for enabling AI techniques in wireless networks. With the emergence of diversified smart devices and applications, data held by individuals becomes more pervasive and nontrivial for publication. First, the data are more private and sensitive, as they cover every aspect of daily life, from the incoming data to the fitness data. Second, the publication of such data is also bandwidth-consuming, as they are likely to be stored on mobile devices. The local differential privacy has been considered a novel paradigm for such distributed data publication. However, existing works mostly request the encoding of contents into vector space for publication, which is still costly in network resources. Therefore, this work proposes a novel framework for highly efficient privacy-preserving data publication. Specifically, two sampling-based algorithms are proposed for the histogram publication, which is an important statistic for data analysis. The first algorithm applies a bit-level sampling strategy to both reduce the overall bandwidth and balance the cost among contributors. The second algorithm allows consumers to adjust their focus on different intervals and can properly allocate the sampling ratios to optimize the overall performance. Both the analysis and the validation of real-world data traces have demonstrated the advancement of our work.

1. Introduction

Enabling the pervasive adoption of cutting-edge techniques of artificial intelligence usually requests the support of huge scales of data [1]. With the joint contribution of smart devices and easy network access, the emergence of volumes of data has been extended from those dominating enterprises to individual contributors like IoT devices. Therefore, the collection of contents in wireless manners has been considered a fundamental task for data processing and AI enhancing [2]. However, the concerns on privacy and resource consumption also rise accordingly. The ubiquitous availability of data sources has broken the boundary between cyber life and physical life. It is believed that every aspect of our life is recorded by some data, thus providing numerous challenges for privacy-preserving data sharing [3]. Meanwhile, data are usually uploaded via wireless or wired networks as they are stored on personal devices. Then, the publication of data is also resource-consuming, especially for those multimedia contents. Both factors have significantly hindered the pervasive collection and make it a nontrivial problem. Therefore, this paper proposes a sampling-based strategy for private data publication.

Actually, data consumers can accept or are even more interested in the statistics about data instead of the detailed contents from every contributor. Such statistics could be sufficient and more reliable for analysis and decision-making. For example, service providers may apply the scale of traffic loads for traffic prediction, while regular users can plan their routes based on such statistics [4]. Among these statistics, the histogram, which provides the distribution of underlying facts, is believed to be essential for data analysis [5]. It may act as an index showing the portions of users falling in the category. Meanwhile, the histogram also provides sights for privacy preservation, as individuals do not disclose their original contents to data brokers or consumers.

To formalize such insights, the local differential privacy [6] is proposed and considered to be a novel paradigm for privacy preservation under distributed manners. LDP is extended from the original differential privacy by removing the request of a trustable data curator. In a typical LDP framework, each participant locally holds her contents and reports encoded and obfuscated results to the data broker, which will aggregate the contents to generate statistics. In this way, the individual contents are preserved, while the noise can be reduced during aggregation. Existing works are conducted for different types of statistics. For categorical values, the heavy hitters, frequent itemsets, and many other statistics are investigated [7]. For numerical values, the mean value, the summation, and some other aggregation queries are studied [8]. Meanwhile, there is also a batch of studies focusing on efficient and fair data publication under various scenarios [9–11].

However, current works for LDP request the encoding of original contents, which are usually bandwidth-consuming. Take the random response as an example [12]. It encodes the value into K-folds, where each fold represents one category of value, like the visited website. As all K-folds will be encoded, the bandwidth for encoded contents will be huge. This is extremely difficult when the candidates of values are large or even infinite (i.e., numerical value). Although some works are conducted, they have not thoroughly considered the problem.

Fortunately, the data consumers, when dealing with statistics, can actually accept minor uncertainty in the results. This is due to the inherent bias underlying the collected data, where the contributors themselves are also samples of the whole population. Meanwhile, statistics with bounded minor errors will not affect decision-making. Therefore, it is interesting to study whether we can further reduce bandwidth consumption while maintaining utility and privacy preservation.

As a result, this work proposed a novel framework for the publication of histograms in distributed manners. In the framework, data contributors locally hold their contents, like their incoming data. The data consumers request histograms with different granularities. The data brokers act as the coordinators among them, which are assumed to be semihonest. They will collect the results from participants and try to infer the raw contents beneath them. A sampling-based algorithm is designed, where the raw data are first encoded with randomized perturbation, and then a bit-level sample strategy is applied for publication. The data brokers will decode the sampled results and respond to consumers with aggregated histograms.

In the framework, all participants are assured with local differential privacy, which is theoretically proved under the sampling strategy. Furthermore, we also propose a mechanism where participants can efficiently derive the encoding scheme under multiple histograms with heterogeneous queries. As for the sampling, two strategies are given, based on whether queries address the same focus on all intervals. We prove the unbiased results for the first sampling and the optimized allocation of bandwidth under the second strategy. Finally, we evaluate the performance under real-world datasets. The results demonstrate the efficiency of our methods. As far as we know, this is the first study on the sample-based histogram publication over numerical values under local differential privacy. The main contribution of this work includes the following:

(i) A novel framework for efficient and privacy-preserved histogram publication over multiple participants

(ii) Two sampling-based strategies for distributed histogram publication under LDP

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

(iv) Evaluation on real-world data traces to demonstrate the effectiveness of proposed methods

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 sampling-based algorithms for histogram publication. The evaluation results are shown in Section 5. Section 6 concludes the paper.

2. Related Work

2.1. Privacy-Preserved Data Publication. The publication of private data has been extensively studied during the past decades. Typical techniques including K-anonymity are proposed [13–16], where the sensitive contents are mixed and obfuscated before publication. However, these studies usually request the limited background knowledge of adversaries and are vulnerable to specific types of attacks. The differential privacy, as a novel index for privacy preservation, allows the existence of arbitrary attackers. There are also some studies investigating histogram publication under differential privacy, focusing on different types of data [17, 18]. They also apply the underlying properties of these data to further reduce the degrees of noise [5, 19, 20]. However, they assume the existence of a trustable data curator to coordinate the data publication, which is usually infeasible for distributed data publication. Finally, there are also some studies considering the fairness and other issues within the private data publication [21–23]. However, they fail to properly reduce the bandwidth consumption and are not compatible with the histogram publication.

2.2. Local Differential Privacy. Local differential privacy [6] provides a novel paradigm for distributed data publication under differential privacy. It allows multiple data contributors to privately aggregate their contents when the data broker is semihonest. Multiple types of statistics are studied, including the publication of graph structures [24], the range counting [25], and the histogram distribution [26]. There are also some studies investigating the efficiency of data publication, ranging from the RAPPOR and Basic RAPPOR methods [12] proposed by Google to sophisticated methods where more mathematical solutions are applied. Reference [27] reviews current studies on LDP providing guidelines for applications. Meanwhile, there is also a batch of studies on the publication of numerical values, and statistics like mean values are considered [8, 28, 29]. However, these studies tend to encode the numerical value into several fixed values, and the perturbed contents may fall out of the original range. This may reduce the utility of published contents [30, 31]. Therefore, the design of an efficient mechanism for histogram publication under local differential privacy is still a challenging topic.

2.3. Sampling-Based Data Collection. Finally, the sampling-based strategy has also been studied for data collection from multiple contributors. Maybe one typical domain where the sampling strategy functions well is the Internet of Things. The wireless and battery-powered sensors and actuators are usually limited in resources. The sampling-based data collection can balance the accuracy of the results and the devoted resources. Corresponding studies are conducted on statistics like Top-K values and data sketching.

As for the combination of sampling strategies and privacy preservation, there are also some studies arguing that the sampling component can strengthen the degree of differential privacy. However, these studies request content-level sampling, which means the sampled contributors can save no resources. Applying the sampling strategy while flexibly balancing the bandwidth consumption is still not well addressed.

3. Problem Formulation

This section first provides the problem formulation for distributed histogram publication, including the network and attacking models. Then, preliminaries on local differential privacy are given.

3.1. Network Model. The whole platform consists of three parties: the data brokers, the data consumers, and data contributors. Initially, data consumers post their histogram queries to data brokers, denoted as $l_1, l_2, \cdots, l_M$. To simplify our model, l_i indicates both the ith query and the interval length of the histogram requested by the query, i.e., the granularity of the histogram. Different consumers may request different diverse queries with different granularities, as they can hold heterogeneous purposes. For example, taxi companies request a coarsened level of traffic loads to guide the deployment of their services, while the navigation apps expect fine-grained histograms to generate a fast route. Upon receiving the requests, the data brokers will generate a data collection plan among participants. The plan consists of a set of consecutive intervals $[D_0, D_1), [D_1, D_2), \cdots, [D_K - 1, D_K]$, together with other parameters like the sampling ratio. As for the intervals, $D_0 = D_L$ indicates the minimum value of all contents, while $D_K = D_U$ refers to the maximum value of all contents. In this framework, we will focus on the snapshot query, such that the data brokers will collect queries from all consumers before generating a plan. Therefore, the intervals are generated based on all queries.

By receiving the plan from data brokers, data contributors will encode and report their contents. Assume N contributors exist in the system, noted as $\{u_1, u_2, \cdots, u_N\}$. Each of them holds one content d_i, and $D_L \leq d_i \leq D_U$. All contents belong to one dimension or can be applied for the same type of query, like the humidity level of a local area or the various sensing data capturing the traffic congestion. We assume the total bandwidth used for reporting d_i to the data brokers as B_i .

Finally, the data broker collects results from all contributors. It will first decode the reported contents and then aggregate them into different intervals. In a final step, the data broker will generate and distribute corresponding histograms to different consumers and charge them accordingly.

3.1.1. Adversarial Model. Due to the latent value of contents held by contributors, both data brokers and consumers are assumed to be semihonest. It means they will not break into contributors' devices to steal the contents but will follow the framework and try to infer the original contents. Therefore, contributors should carefully encode their contents to thwart such inference attacks. The local differential privacy has been considered an extension of differential privacy by removing the requirement of a trusted data broker. LDP still allows arbitrary background knowledge from the adversaries and can preserve individual contents within the statistics. To achieve the LDP property, data contributors can publish perturbed contents d_i, which are either noise values or some relative data structures. The definition of local differential privacy is shown in Definition 1.

Definition 1 (local differential privacy). An algorithm $Q(\cdot)$ satisfies ε-local differential privacy (ε-LDP), where $\varepsilon \geq 0$, if and only if for two arbitrary contents T_i and T_j:

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

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

Based on its definition, the local differential privacy ensures that no significant information will be disclosed to the data receivers. The parameter ε indicates the degree of privacy, where a larger ε means data contributors are less sensitive and will produce more accurate results.

3.1.2. Design Object. Based on the network model and attacks from the adversaries, data contributors aim to both reduce their bandwidth consumption and preserve their raw contents during data uploading. The data brokers are concerned with coordinating the trading between other parties, so their focus is to generate a proper plan for data collection. The plan should be efficient and provide rational performance. Generally, assume that the accumulated variance for each query l_i is $\mathrm{Var}(\mathrm{l_i})$. The design object is given as follows:

$$
\begin{aligned}
&\min \quad \sum_{i=1}^{M} l_i \\
&\text{s.t.} \quad E\left(R_{jk}{'}\right) = R_{jk}, \quad \forall l_1, l_2, \cdots, l_M \\
&\qquad \sum B_i \le B_0 \\
&d_i \text{ is preserved under LDP}, \quad \forall i \in \{1, 2, \cdots, N\},
\end{aligned} \tag{2}
$$

which means the derived result should be unbiased, the total bandwidth should be constrained, and the privacy for each contributor should be preserved.

3.2. Preliminaries. This part introduces preliminaries for LDP. It first introduces one basic encoding-decoding-based method for data uploading and then addresses the compositional properties of LDP.

The random response method provides some basic ideas for the implementation of LDP. We take the Basic RAPPOR proposed by Google as an example.

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

Then, V' can be generated by the randomized response:

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

Finally, V' will be sent to the data curator for subsequent analysis. Actually, this mechanism of perturbation achieves the LDP property for vector V, which is proved by a previous study [27].

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

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. Li et al. have theoretically proved the effect [32].

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 $\ln(1 + P_0(e^{\varepsilon} - 1))$-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 (differentially sequential composition). *Let $\{F_1(\cdot), F_2(\cdot), \cdots, F_k(\cdot)\}$ be a set of functions satisfying differential privacy and privacy budgets be $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_k$, respectively. Then, applying all $F_i(\cdot)$ to one data item d_0 will provide $\sum_{i=1}^{k} \varepsilon_i$-differential privacy.*

4. Sampling-Based Histogram Publication

This part first provides a scheme applied for plan generation. Then, we argue that the efficiency could be further improved among all contributors by uploading partial results. The second part gives a bit-level sampling algorithm, where the bandwidth consumption among contributors is reduced and balanced. The third part proposes a biased sampling mechanism, where the sampling ratios can be adjusted and optimized according to the requests of consumers.

4.1. Data Plan for Multiple Queries. Within the framework, the data broker receives multiple queries $\{l_1, l_2, \cdots, l_M\}$ from consumers. It will then generate a corresponding plan $[D_0, D_1), [D_1, D_2), \cdots, [D_K - 1, D_K]$ from these queries. The main objective is to reduce the number of intervals in the plan, as each interval may reflect extra bandwidth consumption. The data broker applies the following strategy for plan generation.

The data broker iteratively generates intervals according to each query, which is determined by

$$
[D_0, D_0 + l_i \cdot 1), [D_0 + l_i \cdot 1, D_0 + l_i \cdot 2), \cdots, [D_0 + l_i \cdot K_i, D_K], \tag{4}
$$

where

$$
K_i = \left\lfloor \frac{D_U - D_L}{l_i} \right\rfloor. \tag{5}
$$

Then, the data broker combines all intervals from all queries. Specifically, every two consecutive checkpoints $D_0 + l_i \cdot j$ and $D_0 + l_m \cdot n$ from one or two queries will compose a new interval $[D_0 + l_i \cdot j, D_0 + l_m \cdot n)$. The newly generated intervals will be used by the data plan, as $[D_0, D_1), [D_1, D_2), \cdots, [D_K - 1, D_K]$. Therefore, the histogram for each query can be derived from such intervals by iteratively merging the results in several conjunctive intervals. The whole procedure will take $O(M \cdot (D_U - D_L))$ time. In the baseline method, data contributors may follow the typical random response to locally encode and obfuscate their contents. The results will be uploaded to data curators, which will further decode the results and publish the aggregated histograms to consumers.

4.2. Bit-Level Sampling for Histogram Publication. This part introduces the sampling-based algorithm for histogram publication. Intuitively, the data curator can randomly pick a group of contributors for histogram publication, or the contributors can locally determine whether to participate in data processing with some sampling ratios. However, the sampled contributors have to apply the encoding mechanisms and fully upload the vectors in both cases, which is unbalanced and unwilling. Therefore, this part proposes an improved algorithm to sample the contents from another dimension.

The algorithm is named as Bit-Sampling Histogram Publication (BSHP for short).

The main idea of BSHP is to implement bit-level sampling among contributors instead of one-time participant selection. Initially, BSHP follows the same steps as the baseline method, where the data curator processes and distributes the queries to contributors. After locally encoding their data values, contributors follow the request of data collection from the data curator in BSHP. In the jth iteration, the data curator announces a sampling ratio P_j to all contributors u_i, where $0 \leq P_j \leq 1$. Then, all u_i locally execute Bernoulli sampling with probability P_j. Contributors who sampled 1 as the result will upload the corresponding bit $D_{ij}{'}$ to the data curator. This request phase repeats K_0 rounds until all bits are requested.

In the decoding phase, the data curator first estimates the counting of data values in each interval of the integrated partition, where

$$R_k = \frac{\left(\left\|\left\{D_{i'} \mid D'_{ik} = 1\right\}\right\| / p_k\right) - 1/2 \cdot f \cdot N}{1 - f}, \quad \forall k \leq K_0. \tag{6}$$

Then, the data curator derives the results for different queries in the same way with the baseline method, and the whole algorithm terminates.

4.2.1. Analysis. This part analyzes the performance of BSHP. We first investigate the accuracy of histogram publication and then discuss issues on privacy preservation.

The following theorem indicates that BSHP provides an unbiased estimation for the counting of each interval in the integrated partition.

Theorem 5 (unbiased estimation). *For each interval $[W_k, W_{k+1}]$, BSHP can provide an unbiased estimation under given P_k and ε_0, indicating $E(R_k) = R_k^0$.*

Proof. Within BSHP, the data curator estimates R_k as

$$R_k = \frac{\left(\left\|\left\{D_{i'} \mid D'_{ik} = 1\right\}\right\| / p_k\right) - 1/2 \cdot f \cdot N}{1 - f}. \tag{7}$$

We denote R_k^0 as the total size of contents belonging to interval $[W_k, W_{k+1}]$. According to the definition of the random response, we have

$$R_k = R_k^0 \cdot V_s \cdot V_r + \left(N - R_k^0\right) \cdot V_s \cdot V_r{'}, \tag{8}$$

where V_s refers to the sampling variable whether a contributor is selected, and V_r and $V_r{'}$ indicate whether the corresponding bit is retained or reversed.

We have the following notation:

$$\Phi = R_k - R_k^0. \tag{9}$$

Then,

$$\begin{aligned}
E(\Phi) = E\left(R_k - R_k^0\right) &= E\left(\frac{\left\|\left\{D_{i'} \mid D'_{ik} = 1\right\}\right\| - N \cdot P_k \cdot 1/2f}{(1-f) \cdot P_k}\right) \\
- R_k^0 &= \frac{E\left(\left\|\left\{D_{i'} \mid D'_{ik} = 1\right\}\right\|\right) - N \cdot P_k \cdot 1/2f}{(1-f) \cdot P_k} \\
- R_k^0 &= \frac{E\left(R_k^0 V_s V_r + \left(N - R_k^0\right) V_s V_{r'}\right) - N P_k 1/2f}{(1-f) \cdot P_k} \\
- R_k^0&.
\end{aligned} \tag{10}$$

As V_s, V_r, and $V_r{'}$ are independent variables,

$$\begin{aligned}
&E\left(R_k^0 V_s V_r + \left(N - R_k^0\right) V_s V_{r'}\right) \\
&\quad = R_k^0 E(V_s) E(V_r) + \left(N - R_k^0\right) E(V_s) E(V_{r'}) \\
&\quad = R_k^0 P_k \left(1 - \frac{1}{2} f\right) + \left(N - R_k^0\right) P_k \frac{1}{2} f.
\end{aligned} \tag{11}$$

Therefore,

$$E(\Phi) = \frac{R_k^0 P_k (1-f)}{(1-f) \cdot P_k} - R_k^0 = 0. \tag{12}$$

Then, we have

$$E(R_k) = R_k^0, \tag{13}$$

which means R_k is an unbiased estimator for the counting of data values in $[W_k, W_{k+1}]$.

According to Theorem 5, BSHP provides an unbiased estimation for each interval in the integrated partition. Therefore, the final outputs for each query will also be an unbiased estimation, as the final results are derived from the combination of these unbiased countings.

The variance of the estimated result for an interval is calculated in Lemma 6. The main idea of the lemma is to combine the variance from two steps of sampling and derive the correlation between the variance and the two parameters P_k and ε_0.

Lemma 6. *For each interval $[W_k, W_{k+1}]$, with parameters P_k and ε_0, the variance follows $Var(R_k) \leq (((R_k^0)^2 + N^2 1/2f)/((1-f)^2 P_k)) + (1/4)(Nf^2/(1-f)^2) + (2NR_k^0 f/(1-f)^2)$.*

Proof. First of all,

$$\operatorname{Var}(\Phi) = \operatorname{Var}\left(R_k - R_k^0\right) = \operatorname{Var}(R_k). \tag{14}$$

Assume $R_k{}' = \|\{D_{i'} \mid D'_{ik} = 1\}\|$. Then,

$$\begin{aligned}\mathrm{Var}(\Phi) &= E\left(\Phi^2\right) - E(\Phi)^2 = E\left(\Phi^2\right)\\ &= E\left[\left(\frac{R_{k'} - NP_k 1/2f}{(1-f)\cdot P_k} - R_k^0\right)^2\right]\\ &= E\left[\left(\frac{R_{k'} - NP_k 1/2f}{(1-f)\cdot P_k} - R_k^0\right)^2\right]\\ &= \left(R_k^0\right)^2 + E\left(\frac{\left(R_{k'} - NP_k 1/2f\right)^2}{(1-f)^2\cdot P_k^2}\right)\\ &\quad - 2R_k^0 \cdot E\left(\frac{R_{k'} - NP_k 1/2f}{(1-f)\cdot P_k}\right).\end{aligned} \tag{15}$$

As V_s is the Bernoulli sampling,

$$E\left(V_s^2\right) = E(V_s). \tag{16}$$

The same conclusion also holds for V_r and $V_r{}'$. Therefore,

$$\begin{aligned}\mathrm{Var}(\Phi) &= \left(R_k^0\right)^2 + \frac{E\left(R_{k'} - NP_k 1/2f\right)^2}{(1-f)^2\cdot P_k^2} - 2R_k^0\cdot R_k^0\\ &= \frac{E\left(\left(R_{k'}\right)^2\right) - N^2P_k^2 1/4f^2 - NP_k f E\left(R_{k'}\right))}{(1-f)^2\cdot P_k^2} - \left(R_k^0\right)^2\\ &= \frac{NP_k f}{(1-f)^2\cdot P_k^2}\left(R_k^0 P_k\left(1-\frac{1}{2}f\right) + \left(N-R_k^0\right)P_k\left(\frac{1}{2}f\right)\right)\\ &\quad - \left(R_k^0\right)^2 - \frac{N^2P_k^2 1/4f^2}{(1-f)^2\cdot P_k^2} + \frac{\left(R_k^0\right)^2 P_k(1-1/2f) + \left(N-R_k^0\right)^2 P_k 1/2f}{(1-f)^2\cdot P_k^2}\\ &\quad + \frac{R_k^0\left(N-R_k^0\right)P_k^2(1-1/2f)f}{(1-f)^2\cdot P_k^2} \le \frac{\left(R_k^0\right)^2 + N^2 1/2f}{(1-f)^2 P_k} + \frac{1}{4}\frac{Nf^2}{(1-f)^2}\\ &\quad + \frac{2NR_k^0 f}{(1-f)^2}.\end{aligned} \tag{17}$$

As for each data consumer, the variance of the histogram is determined by the summation of variances in different intervals. We omit the conclusion here as the summation is straightforward.

Now we analyze the property of privacy preservation by BSHP. In BSHP, each contributor will only upload partial results of their vectors. Meanwhile, the perturbation and sampling could actually be applied in any order. According to the conclusions in [32], the sampling will strengthen privacy preservation. Therefore, we have the following conclusion.

Theorem 7 (local differential privacy). *BSHP preserves the data value for each contributor under ε'-local differential privacy, where $\varepsilon' \le \varepsilon_0$.*

4.3. Weighted Sampling for Histogram Publication. This part further studies the sampling method for histogram publication. Specifically, the data consumers may hold different requests on histograms. Taking the incoming data as an example, some consumers may prefer more accurate results for people with high salaries, while others may expect to derive the results in the middle of the population. Therefore, the algorithms for histogram publication should also handle such sophisticated utilities for consumers. The proposed algorithm is named as Weighted-Sampling Histogram Publication (WSHP for short).

Initially, WSHP allows each consumer to report their weights at different intervals. The weights for all intervals in the ith query are

$$\left\{\omega_{i1}, \omega_{i2}, \cdots, \omega_{iK_i-1}\right\}. \tag{18}$$

The data curator first derives the integrated partition on the whole range, i.e., $\{W_1, W_2, \cdots, W_{K_0}\}$. Then, WSHP counts the accumulated weights for each interval. For an arbitrary interval $[W_i, W_{i+1}]$, its weight ω_i is derived by adding up corresponding weights from all contributors. Assume the jth query has its kth interval covering $[W_i, W_{i+1}]$; then, the weight inherited from ω_{jk} is $\omega_{jk}\cdot((W_{i+1} - W_i)/(W_{jk+1} - W_{jk}))$. Following this strategy, WSHP traverses all contributors to derive all ω_i:

$$\omega_i = \sum_{j=1}^{M} \omega_{jk}\cdot\frac{W_{i+1} - W_i}{W_{jk+1} - W_{jk}}, \tag{19}$$

where $[W_i, W_{i+1}] \subset [W_{jk}, W_{jk+1}]$. Notice that $[W_i, W_{i+1}]$ will also belong to some $[W_{jk}, W_{jk+1}]$. Otherwise, $[W_i, W_{i+1}]$ will be further partitioned into subintervals.

Based on the weights, the data curator extracts corresponding sampling probabilities for different intervals. We consider a specific case where the incoming data are uniformly distributed over the whole range. Then, the sampling probabilities are determined by the following constraints. Firstly,

$$\frac{\sum_{i=1}^{K_0-1} P_i}{K_0 - 1} = P_0, \tag{20}$$

where P_0 is the overall ratio of collected bits. Secondly, for an arbitrary pair of P_i and P_j,

$$\frac{P_i}{P_j} = \frac{\omega_i(((W_{i+1} - W_i)/(D_U - D_L))\cdot N)^2 + \omega_i\alpha}{\omega_j\left(\left(\left(W_{j+1} - W_j\right)/(D_U - D_L)\right)\cdot N\right)^2 + \omega_j\alpha}, \tag{21}$$

where $\alpha = (1/2)N^2 f$.

Finally, WSHP follows the same strategies with BSHP to iteratively sample bits from contributors based on the corresponding P_i and derives the results for different consumers.

4.3.1. Analysis. The objective of WSHP is to derive improved utilities for data consumers with heterogeneous concerns. The following theorem indicates that the WSHP algorithm can maximize the overall utility for all requestors.

Theorem 8. *With fixed privacy budgets and bandwidths, WSHP can achieve optimal utilities for data consumers when the data values uniformly are distributed in the whole range.*

Proof. As the bandwidths and privacy budgets are fixed in this case, WSHP adjusts the sampling ratios to balance the accuracy among different intervals. Meanwhile, the general variance is applied to measure the accuracy as WSHP provides an unbiased estimation. The general variance is

$$\mathrm{Var}(\mathcal{Q}) = \sum_{i=1}^{M} \mathrm{Var}(Q_i) = \sum_{i=1}^{M} \sum_{j=1}^{K_i-1} \omega_{ij} \mathrm{Var}(R_{ij}). \tag{22}$$

Then, according to the correlations between ω_{ij} and ω_k and the relationships between intervals in the integrated partition and histograms, we have

$$\mathrm{Var}(\mathcal{Q}) = \sum_{j=1}^{K_0-1} \omega_j \mathrm{Var}(R_j). \tag{23}$$

Now we combine the analysis in equation (17) and derive

$$\mathrm{Var}(R_j) = \frac{\left(R_j^0\right)^2 + \alpha}{\beta P_j} + \gamma R_j^0 + \delta, \tag{24}$$

where α, β, γ, and δ are all constant, and $\alpha = (1/2)N^2 f$.

The results in equation (24) could be merged into equation (23),

$$\mathrm{Var}(\mathcal{Q}) = \sum_{j=1}^{K_0-1} \omega_j \left(\frac{\left(R_j^0\right)^2 + \alpha}{\beta P_j} + \gamma R_j^0 + \delta \right). \tag{25}$$

Minimizing the variance $\mathrm{Var}(\mathcal{Q})$ requests the knowledge on R_j^0s, which is obviously unavailable for data curators. Instead, we assume that the underlying data values are uniformly distributed in the range. Then, R_j^0 can be approximated by $((W_{j+1} - W_j)/(D_U - D_L)) \cdot N$. Therefore, the variance Var $(\mathcal{Q})$ is determined by $\sum_{j=1}^{K_0-1}(\omega_j(((W_{j+1} - W_j)/(D_U - D_L)) \cdot N)^2 + \omega_j \alpha)/P_j$. It is obvious that equation (21) can minimize $\mathrm{Var}(\mathcal{Q})$ in this circumstance, which is given in Theorem 8.

5. Discussion

This section covers the situations where data consumers may post their queries asynchronously, and the data curator has to acquire the data once and respond to continuously emerged queries.

The basic settings are similar to the previous cases, where the privacy budgets ε_0 and the bandwidth budgets B_0 are both fixed. In this case, the data curator could do the following:

(1) Devote all resources to extract one single histogram for all queries

(2) Partition the resource into multiple histograms and combine them for multiple queries

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

We will show that the first strategy is actually preferred even if it is straightforward. Initially, the derived results should try to provide an unbiased estimation for forthcoming queries. However, this is usually infeasible due to the diverse partition of intervals in histograms. Then, an alternative objective is to minimize the difference between the ground truth and the estimated result. In this worst case, the distance could be all data values falling in two consecutive intervals in the published histogram. Therefore, minimizing this distance leads to the identical most fine-grained partition on intervals, which implies the adoption of all resources.

On the other hand, we can also achieve the same conclusion by considering the use of privacy budgets. Intuitively, partitioning the budgets into multiple folds will not reduce the overall variance, while extra bandwidths will be wasted for content uploading. Therefore, it is also preferred that the first strategy should be selected for the online querying model.

Our future study will investigate the design of methods toward online histogram publication. Maybe other advanced mechanisms besides the random response will be introduced for this case.

6. Evaluation

In this section, we adopt the salary data collected for normal citizens in the United States [33] to verify the performance of the sampling-based methods. New York City, San Francisco, and Baltimore are selected for our evaluation. Table 1 shows the overview of the datasets. We assume that data consumers request for the histogram of incoming levels with heterogeneous granularity. The data contributors will publish their data to the consumers, and the privacy concerns and bandwidth consumption should be treated. The data curator will coordinate the trading between two parties by generating the data plan and the final results.

The performance of the proposed algorithm is compared with a baseline method. In this method, 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. Each algorithm has been executed 20 times to mitigate the randomness. Finally, the mean square errors (MSE for short) are applied as the metric.

6.1. Basic Performance. This part studies both the numerical values and the overall performance. There are three

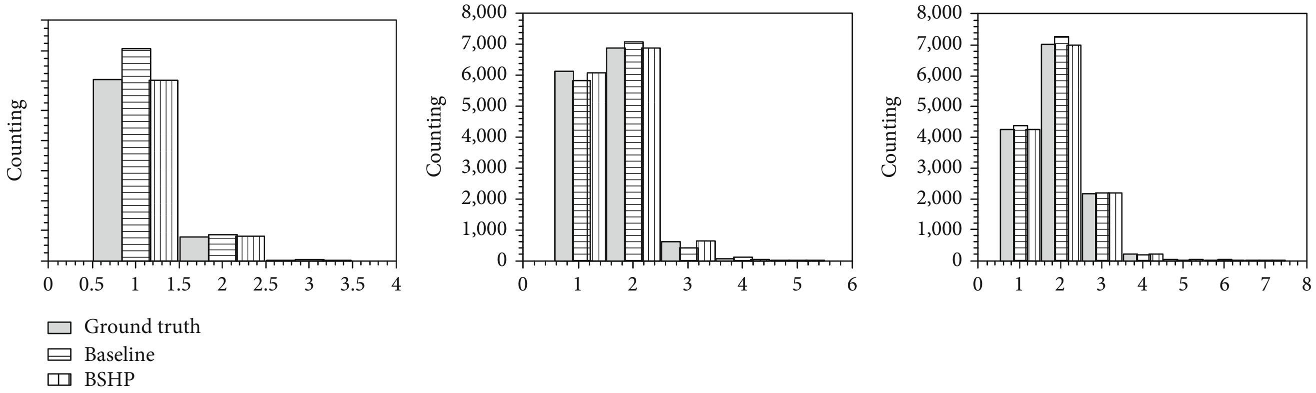

Figure 1: Multigranularity histograms for Baltimore.

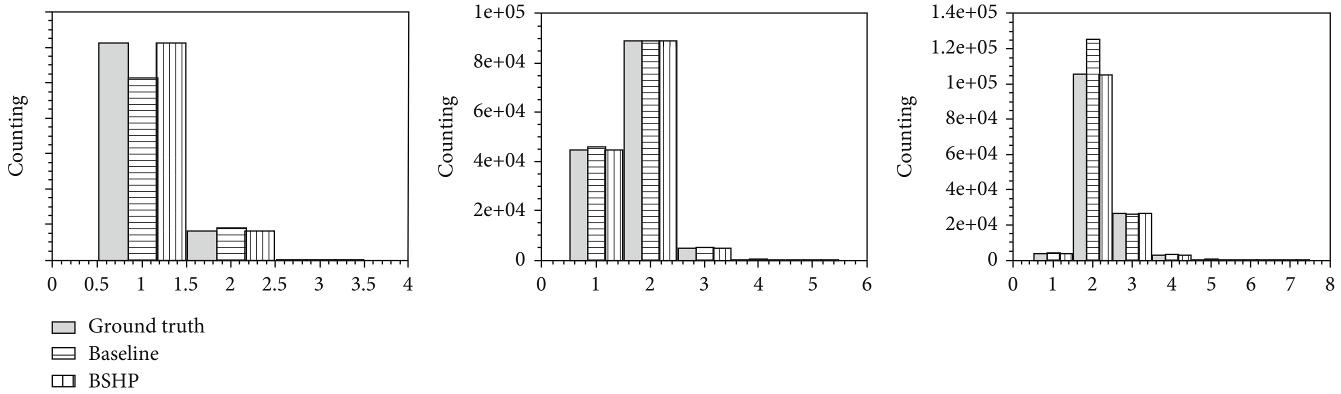

Figure 2: Multigranularity histograms for New York City.

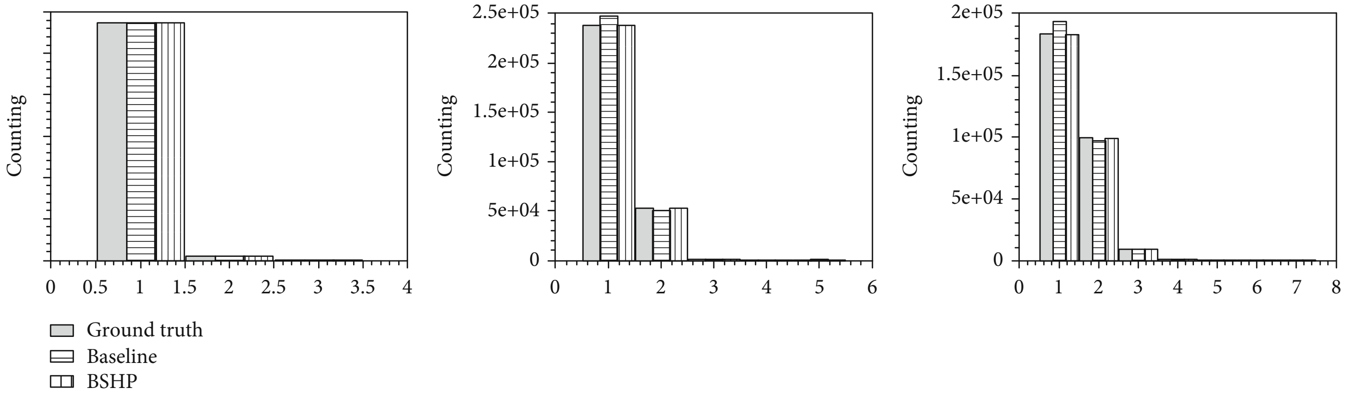

Figure 3: Multigranularity histograms for San Francisco.

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. Meanwhile, our sampling-based algorithms apply all budgets for one common query. The sampling probability is 0.8.

The results are given in Figures 1–3. As we see, the proposed algorithms provide better utilities. They outperform the baseline method and achieve more accurate shapes for histograms, even though only part of the bits is collected under sampling. The difference is actually very significant when considering there are many data values belonging to some intervals to reduce the influence of randomness. It indicates that the proposed method can achieve good utilities by reduced bandwidth consumption.

This part also studies the overall performance under different budgets. In this group, the privacy budgets vary from 3 to 18. Two sampling-based algorithms are evaluated, with sampling probabilities set as 0.8 and 0.4.

According to the results in Figure 4, the proposed algorithms can reduce the MSE for histograms. The improvement is more significant when the privacy budget is

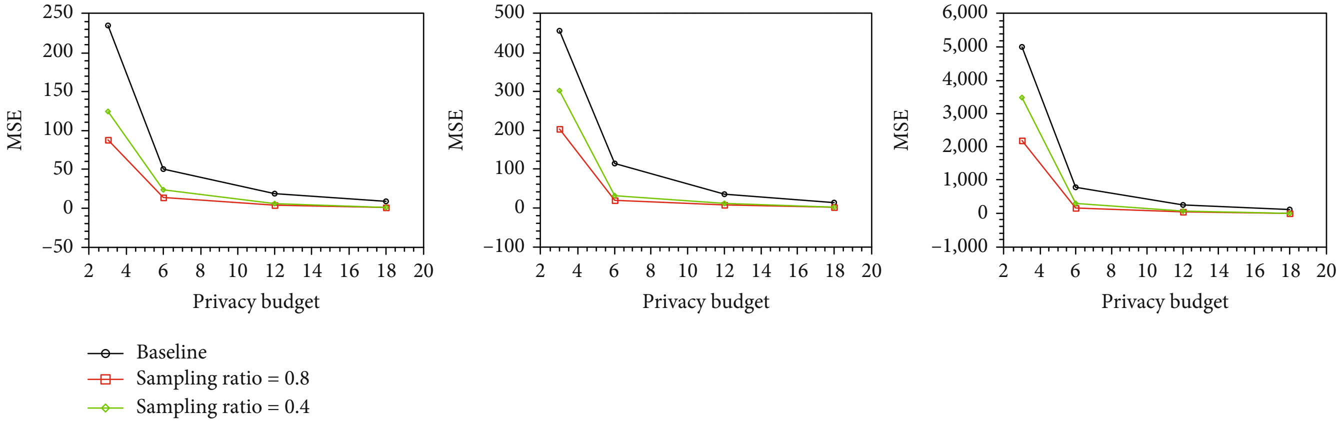

FIGURE 4: Mean square errors for histograms with various privacy budgets.

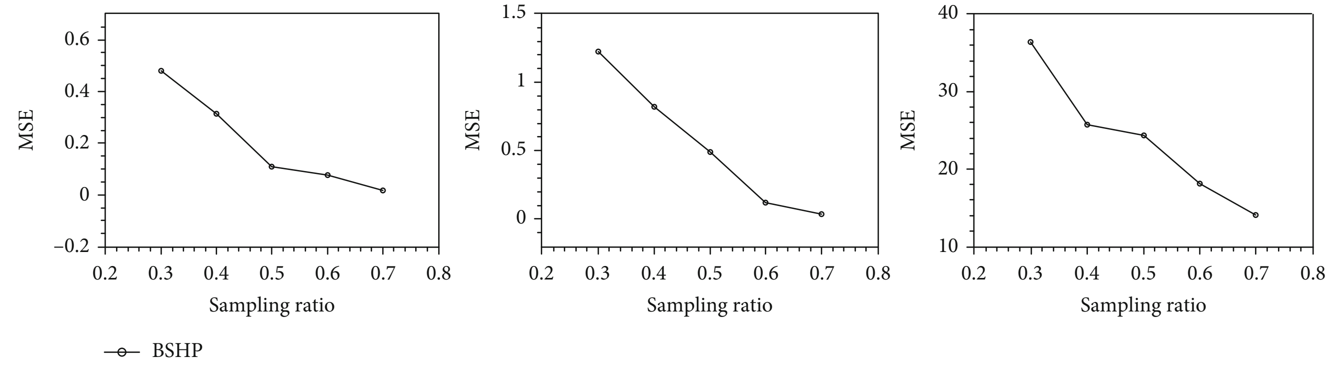

FIGURE 5: Mean square errors for histograms with various sampling ratios.

relatively large. The reason is that the saving on the privacy budget can overwhelm the effect of sampling so as to maintain a good utility. We also observe that a higher sampling ratio can improve the general performance. This is intuitively rational, as more samples could decrease the impacts of randomness.

6.2. Heterogeneous Sampling Ratios. Finally, we study the impact of various sampling ratios on histogram publication. In this group, the privacy budget is 15, and the data consumers still request 3-fold, 5-fold, and 7-fold histograms. The sampling ratios increase from 0.3 to 0.7 with the incremental step as 0.1. The results are depicted in Figure 5.

The general performance is improved for all three datasets according to the results, where the MSE value is reduced by at least 50% (San Francisco). However, the performance actually stays on the same scale for each dataset. This observation implies that increasing the sampling ratio (i.e., the bandwidth) will not always improve the data utility. In this case, the privacy budget will become the bottleneck for highly effective data publication. However, it also indicates that the bandwidth could be saved while the total utilities will not be reduced by too large.

Generally, both proposed algorithms can effectively and efficiently improve the performance of histogram publication.

7. Conclusion

To jointly preserve sensitive information and improve efficiency during data collection has long been considered a challenging task for data processing. The emergence of local differential privacy sheds light on this task. However, existing works fail to combine the sampling strategy with the mechanisms designed for LDP. Therefore, this work proposes a novel framework for privacy-preserved histogram publication in distributed manners. It first investigates a novel plan for data collection over numerical values and then designs two sampling-based algorithms for data encoding and decoding. These algorithms apply bit-level sampling to balance the cost among data contributors and can help consumers adjust their devotion on different intervals of the histogram. Extensive analysis is proposed, including unbiased results, privacy preservation, and optimization in allocating the bandwidth resources. Finally, we conduct an evaluation on one real-world dataset to show the superiority of proposed algorithms.

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

References

[1] 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.

[2] C. Lim, K.-J. Kim, and P. P. Maglio, "Smart cities with big data: reference models, challenges, and considerations," *Cities*, vol. 82, pp. 86–99, 2018.

[3] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in Internet-of-Things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, 2017.

[4] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, and M. J. Deen, "A tensor-based big-data-driven routing recommendation approach for heterogeneous networks," *IEEE Network*, vol. 33, no. 1, pp. 64–69, 2019.

[5] 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.

[6] 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, Portland, OR, USA, June 2015.

[7] M. Bun, J. Nelson, and U. Stemmer, "Heavy hitters and the structure of local privacy," in *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 435–447, San Francisco, California, USA, 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] 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.

[10] 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, Georgia, USA, 2017.

[11] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and M. J. Deen, "A tensor-based multiattributes visual feature recognition method for industrial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2231–2241, 2020.

[12] 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, 2014.

[13] K. Micinski, P. Phelps, and J. S. Foster, "An empirical study of location truncation on android," *Weather*, vol. 2, p. 21, 2013.

[14] R. Chow and P. Golle, "Faking contextual data for fun, profit, and privacy," in *Proceedings of the 8th ACM workshop on Privacy in the electronic society - WPES '09*, pp. 105–108, Chicago, Illinois, USA, 2009.

[15] 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.

[16] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li, "Achieving k-anonymity in privacy-aware location-based services," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pp. 754–762, Toronto, ON, Canada, April 2014.

[17] 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.

[18] 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, December 2012.

[19] 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.

[20] 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, April 2014.

[21] 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.

[22] 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.

[23] 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.

[24] 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, Texas, USA, October 2017.

[25] 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, July 2019.

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

[27] 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.

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

[29] 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.

[30] 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, April 2019.

[31] 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.

[32] 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*, pp. 32-33, Singapore, 2012.

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