Trading Private Range Counting over Big IoT Data

Zhipeng Cai
Department of Computer Science
Georgia State University
Atlanta, United States
zcai@gsu.edu

Zaobo He
Department of Computer Science and Software Engineering
Miami University
Oxford, United States
hez26@miamioh.edu

***Abstract*—Data privacy arises as one of the most important concerns, facing the pervasive commoditization of big data statistic analysis in Internet of Things (IoT). Current solutions are incapable to thoroughly solve the privacy issues on data pricing and guarantee the utility of statistic outputs. Therefore, this paper studies the problem of trading private statistic results for IoT data, by considering three factors. Specifically, a novel framework for trading range counting results is proposed. The framework applies a sampling-based method to generate approximated counting results, which are further perturbed for privacy concerns and then released. The results are theoretically proved to achieve unbiasedness, bounded variance, and strengthened privacy guarantee under differential privacy. Moreover, a pricing approach is proposed for the traded results, which is proved to be immune against arbitrage attacks. The framework is evaluated by estimating the air pollution levels with different ranges on 2014 CityPulse Smart City datasets. The analysis and evaluation results demonstrate that our framework greatly reduces the error of range counting approximation; and the optimal perturbation approach enables that the private counting satisfies the specified approximation degree while providing strong privacy guarantee.**

***Index Terms*—Differential Privacy, Range counting, Pricing**

I. Introduction

The Internet of Things (IoT) has been regarded as a new paradigm of big data platform. For example, smart city applications have been deployed to timely monitor, analyze and response upon volumes of physical data. As a fundamental data analyzing operation, *range counting* aggregation acts as a critical component for these applications. For instance, data analyzers compute range counting over massive particulate matter level, traffic volume or weather data to monitor pollution levels. These aggregates are not only valuable to data owners, but also attractive to other communities with business purposes. However, data in IoT are collected in a distributed manner and strongly correlated with users' sensitive status, aggravating the cost and privacy concerns for data analyzing operations. Therefore, this paper proposes a novel framework for range counting aggregation, which jointly considers utility, cost, privacy preservation, and charging for derived answers.

Actually, many information platforms have emerged to facilitate such data circulation from raw data to data consumers (*i.e.*, service requesters). In particular, these platforms usually prefer to trade statistics of raw data, like the range counting aggregation results, to data requesters. As shown in an FTC's survey on several data brokers [1], Acxiom, as an essential broker, collects personal information from more than 700 million users, and sells aggregation statistical information to big companies such as Oracle, Microsoft, AT&T, etc. Although attractive, conducting and trading range counting aggregation in IoT brings two major challenges, *i.e.*, the concerns on resource consumption and privacy disclosure, and the design of pricing mechanisms.

The first challenge lies in the performance concern and the privacy concern. If all IoT data are collected to compute the exact range counting aggregation, considerable communication and computation overhead will be incurred [2] [3]. However, in many cases, approximate range counting results with less overhead are actually sufficient enough for data customers to perform data analysis [4], [5], [6]. Meanwhile, data privacy is another serious concern, which obstacles the wide deployment and adoption of smart devices. The underlying reason is that smart devices collect, understand and interact with a user in a pervasive and intimate way [7]. Thus, aggregation results released to requesters should avoid considerable leakage of sensitive information. Although individual efforts have been made for each concern respectively, it remains unsettled to compute privacy-preserved approximate range counting aggregation efficiently.

The second challenge is to establish effective pricing mechanisms for trading approximate range counting aggregation results. This challenge arises from potential *arbitrage* opportunities in trading procedures. Generally, data consumers are usually allowed to specify their own expected approximation degrees, upon which data brokers compute an approximate aggregation result and perturb it for privacy preservation. In this case, a smaller approximation degree intuitively leads to a higher price. However, with a poorly designed pricing mechanism, malicious consumers may circumvent to pay the desired price of a query. These consumers turn to buy multiple cheaper results with high variance, and reduce the variance by averaging the returned results. Then this sophisticated trading practice is an *arbitrage* attack when the total price of high-variance aggregation is less than a single one with low-variance. As far as we know, the *arbitrage* attack has not been investigated for pricing mechanisms in IoT.

To mitigate the gaps, we propose a sampling-based algorithm for privacy-preserved approximate range counting aggregation. The algorithm presents an unbiased estimator for range

2575-8411/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDCS.2019.00023

counting with bounded variance. To derive privacy-preserved range counting, the state-of-the-art approach *Differential Privacy* [8] is adopted, which allows unlimited reasoning power and background knowledge of adversaries. To meet such a requirement, our algorithm introduces an additional noise to the original approximate result.

However, putting differential privacy into approximate range counting remains a challenging problem, since the two-stage compositive approximation should still meet the requirement of a customer. To address the above issues, we formulate an optimization problem that takes approximation degree as input, and outputs an optimal noise-adding mechanism, such that the derived range counting result satisfies the specified approximation degree and privacy preservation can be optimized. Specifically, our formulation traverses all valid intermediate approximation aggregations, and calculates the minimized differential privacy budget for them, while the final result still guarantees the input approximation degree. We theoretically prove that the derived range counting result meets the specified approximation degree while providing the strongest privacy preservation.

Finally, to develop a pricing mechanism to avoid arbitrage, a sufficient and necessary condition for all arbitrage-avoiding pricing functions is proposed, with the information regarding how fast arbitrage-avoiding pricing functions can decrease with the approximation degree. Our key contributions are summarized as follows.

- We propose a privacy-preserved approximate range counting aggregation algorithm, in which an unbiased estimator with bounded variance for range counting aggregation is presented.
- An optimization problem is formulated to achieve the strongest differential privacy, while satisfying the approximation degree specified by data customers. A solution is provided accordingly.
- An arbitrage-avoiding pricing mechanism is proposed to eliminate arbitrage attacks. A set of pricing functions can be constructed based on our identified critical condition to guarantee justice of trading.
- We extensively evaluate the performance of our approach based on real-world dataset, *i.e.*, the CityPulse Smart City Datasets.

The rest of this paper is organized as follows. Section II introduces our system and adversary models, together with some preliminary knowledge on differential privacy and our problem definition. Section III presents the sampling based range counting aggregation algorithm and the optimization mechanism for providing the strongest differential privacy. The pricing mechanism is introduced in Section IV. Section V illustrates our evaluation results. Section VI discusses the related works, and Section VII concludes the paper.

II. Problem Formulation

This section first presents our system model and adversary model. Then necessary preliminary knowledge on *differential privacy* and the concept of *Arbitrage Avoiding* is introduced. At the end of this section, we present the problem definition of differentially private (α, δ)-range counting.

A. System Model

As shown in Fig. 1, there are three major entities in our system model including IoT networks, data brokers and data consumers. IoT networks consist of large scales of smart devices, which collect data generated by sensing modules or other input channels. Denote D as the global dataset collected by all smart devices in IoT networks. Instead of transferring the entire D to the base station, each smart device only sends a sample of its locally collected data to the base station. This will significantly reduce the communication cost of data transmission. Then a sample S of D is stored in the base station, which opens the data access API to data brokers.

In this paper, we consider *range counting* queries on datasets collected by smart devices, and the definition of range counting is as below.

Definition 2.1: **Range Counting.** Given range parameters l and u $(l \leq u)$ together with dataset D, the range counting of D with lower bound l and upper bound u is $\gamma(l, u, D) = |\{x | l \leq x \leq u, x \in D\}|$.

Computing exact range counting from scratch is expensive in terms of real-time communication in IoT networks. In many scenarios, an approximate range counting with acceptable accuracy suffices to meet customers' requirements. Definition 2.2 presents the notion of (α, δ)-range counting, which parameterizes range counting with accuracy parameters specified by customers.

Definition 2.2: (α, δ)**-Range Counting.** Given $0 \leq \alpha \leq 1$ and $0 \leq \delta \leq 1$, for any range parameters l and u such that $l \leq u$, the (α, δ)-range counting of dataset D, denoted as $\hat{\gamma}(l, u, D)$, satisfies that $\mathbf{Pr}[|\hat{\gamma}(l, u, D) - \gamma(l, u, D)| \leq \alpha |D|] \geq \delta$.

Data customers send (α, δ)-range counting requests denoted by $\Lambda(\alpha, \delta)$ to a data broker. The data broker may access S to response these requests. However, the sensitive information may still be inferred by adversaries with background knowledge, even if approximate aggregates, rather than the raw dataset, are released to data customers, Thus, the IoT network entrusts the protection of data privacy to the data broker. The data broker first accesses S to compute a (α', δ')-range counting, where $\alpha' \leq \alpha$ and $\delta' \geq \delta$. Then the data broker employs the standard notion of differential privacy, and adds carefully controlled noise to the (α', δ')-range counting. Subsequently, the (α', δ')-range counting and the noise jointly composite an (α, δ)-range counting. Finally, the data broker responses a customer with the composited (α, δ)-range counting, and charges the customer with price $\pi(\alpha, \delta)$.

B. Adversary Model

The adversaries in this paper are closefisted or malicious customers, who look for arbitrage opportunities against the trading pricing designed by a data broker. For example, an adversary is interested in an aggregate result with low variance. However, instead of making full payment, the adversary

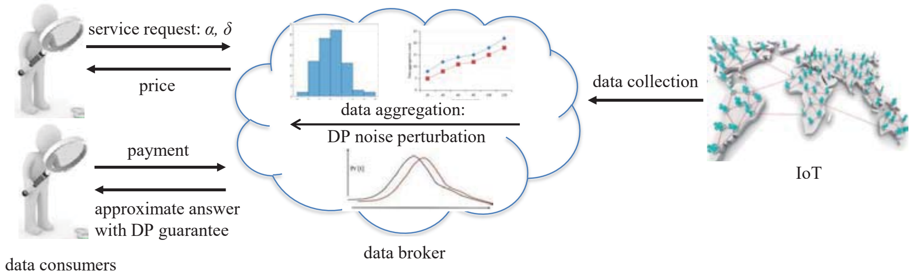
service request: α, δ
price
payment
approximate answer with DP guarantee
data consumers
data aggregation: DP noise perturbation
data broker
data collection
IoT

Fig. 1. System Model.

turns to purchase several aggregates with high variance at a cheaper price. Then the adversary can reduce the variance by averaging the returned aggregates. A benefit-concerned data broker would like to rule out such arbitrage behaviors. Thus, the pricing mechanism should provide the property of arbitrage avoiding, which is defined as follows [9]:

Definition 2.3: **Arbitrage Avoiding.** A pricing function $\pi\{\alpha, \delta\}$ is arbitrage avoiding if $\forall m \geq 1$, $\{\Lambda(\alpha_1, \delta_1), \ldots, \Lambda(\alpha_m, \delta_m)\} \mapsto \Lambda(\alpha, \delta)$ implies:

$$\pi(\alpha, \delta) \leq \sum_{j=1}^{m} \pi(\alpha_j, \delta_j), \tag{1}$$

where $\{\alpha_i, \delta_j | i, j \in \{1, \ldots, m\}, \alpha_i > \alpha, \delta_j < \delta\}$; $\mapsto$ is an operation that composites these m range counting results to a result with (α, δ)-approximation.

Definition 2.3 implies that for a desirable $\pi(\alpha, \delta)$, $\pi(\alpha, \delta) \leq \sum_{j=1}^{m} \pi(\alpha_j, \delta_j)$. Then adversaries cannot obtain aggregation service $\Lambda(\alpha, \delta)$ with a lower price through buying multiple aggregates with diverse higher variances and averaging aggregates at a cheaper price than $\Lambda(\alpha, \delta)$.

C. Differential Privacy

Differential privacy is a well accepted standard notion for protecting sensitive information in statistical aggregates. The formal definition of differential privacy is given as follows:

Definition 2.4: **ϵ-Differential Privacy.** A randomized algorithm G satisfies ϵ-Differential Privacy (ϵ-DP) if and only if for any two neighboring datasets D and D' that differ in only one item and for any possible output O of G, the following condition holds:

$$\Pr[G(\gamma(D)) = O] \leq e^{\epsilon} \cdot \Pr[G(\gamma(D')) = O].$$

The *Laplace mechanism* [10], introduced as below, is a standard approach to achieve differential privacy.

Laplace Mechanism. For a function $\gamma : \mathbb{D} \to \mathbb{R}^d$, Laplace mechanism derives the following result:

$$G(D) = \gamma(D) + \text{Lap}\left(\frac{\Delta\gamma}{\epsilon}\right)^d,$$

where

$$\Delta\gamma = \max_{D \simeq D'} \|\gamma(D) - \gamma(D')\|_1,$$

and

$$\Pr[\text{Lap}(\eta) = x] = \frac{1}{2\eta} e^{-|x|/\eta}$$

to achieve ϵ-differential privacy.

Given the privacy budget ϵ, the amount of noise is also denoted as $\text{Lap}(\epsilon)$ for abbreviation.

D. Problem Definition

The problem of computing differentially private (α, δ)-range counting is defined as follows:

Input:

1) Data set D;

2) Range parameters l and u ($l \leq u$), accuracy parameters α and δ ($0 \leq \alpha \leq 1$ and $0 \leq \delta \leq 1$).

Output:

1) Differentially private (α, δ)-range counting with the minimum privacy budget.

In addition to the above problem definition, this paper studies the design of arbitrage avoiding pricing mechanisms for trading (α, δ)-range counting aggregation.

III. (α, δ)-DIFFERENTIALLY PRIVATE RANGE COUNTING

In this section, we propose a sampling based algorithm to provide (α, δ)-differentially private range counting. Section III-A presents our RankCounting Estimator to answer (α, δ)-range counting based on samples. Then we define an optimization problem and present a solution in section III-B to achieve the optimal differential privacy under the constraint that (α, δ)-range counting is guaranteed.

A. Sampling Based (α, δ)-Range Counting

In this section, we handle the first part of our work, namely the sampling-based (α, δ)-range counting. High accuracy and low communication cost are essential to the performance of the entire system. To achieve this, we propose an estimator namely RankCounting for answering (α, δ)-range counting aggregations.

We assume the network is organized in a flat model, in which each node communicates with the base station directly. Note that algorithms on flat models can be easily extended to a general tree model. After samples are collected from underlying nodes, they will be used to answer future range counting aggregations if the required accuracy can be satisfied. Otherwise, the base station will inform the underlying nodes to collect more samples from the network.

A straightforward estimation (denoted as $\mathsf{BasicCounting}$) to the range counting is $\gamma_B(l,u,S) = \frac{|\{x|x\in \mathsf{S}, l\leq x\leq u\}|}{p}$. This estimator is unbiased and its variance is $\frac{|\{x|x\in D, l\leq x\leq u\}|(1-p)}{p}$, which may grow to $\frac{|D|(1-p)}{p}$ when a large range is queried. This in turn increases the communication cost of sample transmission since more samples should be drawn to guarantee query accuracy.

To reduce communication cost, we leverage the rank of sampled data elements to present the $\mathsf{RankCounting}$ estimator. Let S_i be the set of sample drawn by node i, and D_i denote the set of data collected by node i, $i = 1,\ldots,k$. Let $S = \cup_1^k S_i$, $D = \cup_1^k D_i$ be the global set of sample and data respectively. Let n_i be the number of data collected at node i, and n denote the total number of data collected at k nodes. Let fst and lst denote the first and last data collected at node i, $fst \leq lst$, respectively. Given the lower and upper ranges l, u, our estimator first computes $\hat{\gamma}(l,u,i)$ using S_i, namely the range counting with parameter (l,u) at node i, and then obtains the range counting at S, denoted as $\hat{\gamma}(l,u,S)$.

The RankCounting Estimator. Each node i first independently samples each of its data with a certain probability p (to be determined later). For each sampled data x, it computes $r(x,i)$, the local rank of x at node i, *i.e.*, the rank of x in D_i. Finally, node i sends all the sampled data and corresponding ranks to the base station. Similarly, if the existing samples are unable to satisfy the query accuracy requirement, more samples should be drawn and their ranks are also transferred to the base station.

Given query parameters l and u, we denote $r(l,i)$ as the smallest rank of any element in D_i whose value is no smaller than l, and $r(u,i)$ is defined as the largest rank of any element no larger than u. Note that l and u may not exist, in most cases, in the sampled data S_i, so neither the broker nor the $\mathsf{RankCounting}$ estimator is able to obtain the ranks $r(l)$ and $r(u)$. Nevertheless, the concept of $r(l,i)$ and $r(u,i)$ are only employed in the accuracy analysis of the proposed estimator, and they are not involved in estimation calculation at all.

Let $\mathfrak{p}(x,i)$ denote the predecessor of x in the sampled data S_i, *i.e.*, the largest sampled data no larger than x. Likewise, let $\mathfrak{s}(x,i)$ denote the successor of x in the sampled data from node i, *i.e.*, the smallest value larger than x. It is worth noting that $\mathfrak{p}(x,i)$ and $\mathfrak{s}(x,i)$ may not exist. To distinguish different cases with regards to the existence or non-existence of $\mathfrak{p}(l,i)$ and $\mathfrak{s}(u,i)$, we make the following denotations, together with their probabilities below:

ω_p: $\mathfrak{p}(l,i)$ exists, $\mathbf{Pr}[\omega_p] = 1-(1-p)^{r(l)}$;
$\overline{\omega_p}$: $\mathfrak{p}(l,i)$ does not exist, $\mathbf{Pr}[\overline{\omega_p}] = (1-p)^{r(l)}$;
ω_s: $\mathfrak{s}(u,i)$ exists, $\mathbf{Pr}[\omega_s] = 1-(1-p)^{n_i - r(u)}$
$\overline{\omega_s}$: $\mathfrak{s}(u,i)$ does not exist, $\mathbf{Pr}[\overline{\omega_s}] = (1-p)^{n_i - r(u)}$.

Based on the above notations, we carry out our estimator $\mathsf{RankCounting}$, and $\mathsf{RankCounting}$ estimates $\gamma(l,u,i)$ as below.

$$\hat{\gamma}(l,u,i) = \begin{cases} \gamma(\mathfrak{p}(l,i), \mathfrak{s}(u,i), i) - \frac{2}{p}; & \text{if } \omega_p, \omega_s; \\ \gamma(\mathfrak{p}(l,i), lst, i) - \frac{1}{p}; & \text{if } \omega_p, \overline{\omega_s}; \\ \gamma(fst, \mathfrak{s}(u,i), i) - \frac{1}{p}; & \text{if } \overline{\omega_p}, \omega_s; \\ \gamma(fst, lst, i) & \text{else.} \end{cases}$$

In the calculation of $\hat{\gamma}(l,u,i)$, ranks of certain samples are employed to improve estimation accuracy. Given query parameters l, u and the collected samples in S_i, $\mathsf{RankCounting}$ determines which of the four cases should be adopted. Then the corresponding calculation is carried out, and in this process terms $\gamma(\mathfrak{p}(l,i), \mathfrak{s}(u,i), i)$, $\gamma(\mathfrak{p}(l,i), lst, i)$, $\gamma(fst, \mathfrak{s}(u,i), i)$ and $\gamma(fst, lst, i)$ can be exactly calculated with the ranks of $\mathfrak{p}(l,i)$, $\mathfrak{s}(u,i)$, fst and lst. Here, $\mathfrak{p}(l,i)$ and $\mathfrak{s}(u,i)$ are in S_i, so $\mathsf{RankCounting}$ can obtain their ranks. The ranks of fst and lst are simply 1 and n_i.

Based on $\hat{\gamma}(l,u,i)$, $\mathsf{RankCounting}$ estimates $\gamma(l,u,D)$ as

$$\hat{\gamma}(l,u,S) = \sum_{i}^{k} \hat{\gamma}(l,u,i). \tag{2}$$

Next, we illustrate the high accuracy of $\mathsf{RankCounting}$ with Theorem 3.1, which shows that $\mathsf{RankCounting}$ can use $\hat{\gamma}(l,u,i)$ to accurately estimate $\gamma(l,u,i)$.

Theorem 3.1: For any l and u, $\hat{\gamma}(l,u,i)$ is an unbiased estimation of $\gamma(l,u,i)$ with variance $\text{Var}[\hat{\gamma}(l,u,i)] \leq \frac{8}{p^2}$.

Proof. Denote $\Psi = \hat{\gamma}(l,u,i) - \gamma(l,u,i)$, then Ψ could be formulated under different cases as follows,

$$\Psi = \begin{cases} \gamma(\mathfrak{p}(l,i), \mathfrak{s}(u,i), i) - \gamma(l,u,i) - \frac{2}{p}; & \text{if } \omega_p, \omega_s; \\ \gamma(\mathfrak{p}(l,i), lst, i) - \gamma(l,u,i) - \frac{1}{p}; & \text{if } \omega_p, \overline{\omega_s}; \\ \gamma(fst, \mathfrak{s}(u,i), i) - \gamma(l,u,i) - \frac{1}{p}; & \text{if } \overline{\omega_p}, \omega_s; \\ \gamma(fst, lst, i) - \gamma(l,u,i). & \text{else} \end{cases}$$

The proof first shows that $\mathbf{E}(\Psi) = 0$ and then illustrates the bounded variance of $\hat{\gamma}(l,u,i)$. We observe that $\gamma(\mathfrak{p}(l,i), \mathfrak{s}(u,i), i)$, $\gamma(\mathfrak{p}(l,i), lst, i)$, $\gamma(fst, \mathfrak{s}(u,i), i)$ and $\gamma(fst, lst, i)$ are all no smaller than $\gamma(l,u,i)$, since they contain additional data elements compared to the query range (l,u). The number of additional data elements can be studied in four cases according to the existence or non-existence of $\mathfrak{p}(l,i)$ and $\mathfrak{s}(u,i)$. $\gamma(\mathfrak{p}(l,i), \mathfrak{s}(u,i), i) - \gamma(l,u,i)$ represents the number of additional data elements located in range intervals $(\mathfrak{p}(l,i), l)$ and $(u, \mathfrak{s}(u,i))$, when both $\mathfrak{p}(l,i)$ and $\mathfrak{s}(u,i)$ exist. $\gamma(\mathfrak{p}(l,i), lst, i) - \gamma(l,u,i)$ represents the number of additional data elements located in $(\mathfrak{p}(l,i), l)$ and (u, lst), when only $\mathfrak{p}(l,i)$ exists. Likewise, $\gamma(fst, \mathfrak{s}(u,i), i) - \gamma(l,u,i)$ represents the number of additional data elements located in (fst, l) and $(u, \mathfrak{s}(u,i))$, when only $\mathfrak{s}(u,i)$ exists. $\gamma(l,u,i) - \gamma(fst, lst, i)$ represents the number of additional data elements located in (fst, l) and (u, lst), when neither $\mathfrak{p}(l,i)$ or $\mathfrak{s}(u,i)$ exists.

We introduce term $\mathsf{C}[i]_s^e = |\{x | x \in D_i, s \leq x \leq e\}|$ to denote the number of data elements from D_i located in the range interval (s,e). For example, $\mathsf{C}[i]_l^{\mathfrak{p}(l,i)}$ and $\mathsf{C}[i]_u^{\mathfrak{s}(u,i)}$ represent the number of data elements located in $(\mathfrak{p}(l,i), l)$

and $(u, \mathfrak{s}(u, i))$, when $\mathfrak{p}(l, i)$ and $\mathfrak{s}(u, i)$ exist, respectively. For a given j satisfying $1 \leq j \leq r(l)$, the probability of $\mathsf{C}[i]^l_{\mathsf{p}(l)} = j$ is $p(1-p)^{j-1}$, when $\mathsf{C}[i]^l_{\mathsf{p}(l)}$ exists. Similarly, for a given j satisfying $1 \leq j \leq n_i - r(u)$, the probability of $\mathsf{C}[i]^{\mathsf{s}(u)}_u = j$ is $p(1-p)^{j-1}$, when $\mathsf{C}[i]^{\mathsf{s}(u)}_u$ exists. According to the definition of C, we have $\mathsf{C}[i]^l_{fst} = r(l, i)$ and $\mathsf{C}[i]^{lst}_u = n_i - r(u, i)$. Note that for given query parameters l and u, $\mathsf{C}[i]^l_{\mathfrak{p}(l,i)}$, $\mathsf{C}[i]^{\mathfrak{s}(u,i)}_u$, $\mathsf{C}[i]^l_{fst}$ and $\mathsf{C}[i]^{lst}_u$ can be regarded as random variables. $\mathsf{C}[i]^{\mathfrak{p}(l,i)}_l$ is independent of $\mathsf{C}[i]^{\mathfrak{s}(u,i)}_u$ and $\mathsf{C}[i]^{lst}_u$, while $\mathsf{C}[i]^{\mathfrak{s}(u,i)}_u$ is independent of $\mathsf{C}[i]^l_{\mathfrak{p}(l,i)}$ and $\mathsf{C}[i]^l_{fst}$.

The notation of Ψ could be formulated using C as follows:

$$\Psi = \begin{cases} \mathsf{C}[i]^l_{\mathfrak{p}(l,i)} + \mathsf{C}[i]^{\mathfrak{s}(u,i)}_u - \frac{2}{p}; & \text{if } \omega_p, \omega_s; \\ \mathsf{C}[i]^l_{\mathfrak{p}(l,i)} + \mathsf{C}[i]^{lst}_u - \frac{1}{p}; & \text{if } \omega_p, \overline{\omega_s}; \\ \mathsf{C}[i]^l_{fst} + \mathsf{C}[i]^{\mathfrak{s}(u,i)}_u - \frac{1}{p}; & \text{if } \overline{\omega_p}, \omega_s; \\ \mathsf{C}[i]^l_{fst} + \mathsf{C}[i]^{lst}_u. & \text{else} \end{cases}$$

In the following analysis, we omit the identifier i of node in $r(\cdot, i)$, $\mathsf{C}[i]^{(\cdot)}_{(\cdot)}$, $\mathfrak{p}(\cdot, i)$ and $\mathfrak{s}(\cdot, i)$, and use $r(\cdot)$, $\mathsf{C}^{(\cdot)}_{(\cdot)}$, $\mathfrak{p}(\cdot)$ and $\mathfrak{s}(\cdot)$ respectively for abbreviation when focusing on node i. The expectation of Ψ can be computed as shown below:

$$\begin{aligned}
\mathbf{E}(\Psi) = & \sum_{\substack{1 \leq m \leq r(l) \\ 1 \leq n \leq n_i - r(u)}} (m + n - \frac{2}{p}) \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \mathsf{C}^{\mathfrak{s}(u)}_u = n] \\
& + \sum_{1 \leq m \leq r(l)} (m + n_i - r(u) - \frac{1}{p}) \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \overline{\omega_s}] \\
& + \sum_{1 \leq n \leq n_i - r(u)} (r(l) + n - \frac{1}{p}) \mathbf{Pr}[\mathsf{C}^{\mathfrak{s}(u)}_u = m, \overline{\omega_p}] \\
& + (r(l) + n_i - r(u)) \mathbf{Pr}[\overline{\omega_p}, \overline{\omega_s}] \\
= & \mathbf{Pr}[\omega_s] \mathbf{E}[\mathsf{C}^l_{\mathfrak{p}(l)} - \frac{1}{p}] + \mathbf{Pr}[\omega_p] \mathbf{E}[\mathsf{C}^{\mathfrak{s}(u)}_u - \frac{1}{p}] \\
& + \mathbf{Pr}[\overline{\omega_s}] \mathbf{E}[\mathsf{C}^l_{\mathfrak{p}(l)} - \frac{1}{p}] + \mathbf{Pr}[\omega_p](n_i - r(u)) \mathbf{Pr}[\overline{\omega_s}] \\
& + \mathbf{Pr}[\omega_s] r(l) \mathbf{Pr}[\overline{\omega_p}] + \mathbf{Pr}[\overline{\omega_p}] \mathbf{E}[\mathsf{C}^{\mathfrak{s}(u)}_u - \frac{1}{p}] \\
& + (r(l) + n_i - r(u)) \mathbf{Pr}[\overline{\omega_s}] \mathbf{Pr}[\overline{\omega_p}] \\
= & \mathbf{E}[\mathsf{C}^l_{\mathfrak{p}(l)}] - \frac{1}{p} \mathbf{Pr}[\omega_p] + \mathbf{E}[\mathsf{C}^{\mathfrak{s}(u)}_u] - \frac{1}{p} \mathbf{Pr}[\omega_s] \\
& + r(l) \mathbf{Pr}[\overline{\omega_p}] + (n_i - r(u)) \mathbf{Pr}[\overline{\omega_s}] \\
= & \sum_{j=1}^{r(l)} jp(1-p)^{j-1} - \frac{1}{p}(1 - (1-p)^{r(l)}) \\
& + \sum_{j=1}^{n_i - r(u)} jp(1-p)^{j-1} - \frac{1}{p}(1 - (1-p)^{n_i - r(u)}) \\
& + (1-p)^{r(l)} r(l) + (1-p)^{n_i - r(u)} (n_i - r(u)) \\
= & 0
\end{aligned}$$

So we have $\mathbf{E}[\hat{\gamma}(l, u, i)] = \gamma(l, u, i)$, and the RankCounting estimator produces an unbiased estimation to $\gamma(l, u, i)$. Given query parameters l and u, the exact query result $\gamma(l, u, i)$ is a constant, and it indicates that $\mathbf{Var}[\hat{\gamma}(l, u, i)] = \mathbf{Var}[\Psi]$ since $\Psi = \hat{\gamma}(l, u, i) - \gamma(l, u, i)$. Next, we investigate the variance of $\hat{\gamma}(l, u, i)$ by calculating the variance of Ψ.

$$\begin{aligned}
& \mathbf{Var}[\hat{\gamma}(l, u, i)] = \mathbf{Var}[\Psi] = \mathbf{E}[\Psi^2] - \mathbf{E}[\Psi]^2 = \mathbf{E}[\Psi^2] \\
= & \sum_{\substack{1 \leq m \leq r(l) \\ 1 \leq n \leq n_i - r(u)}} (m + n - \frac{2}{p})^2 \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \mathsf{C}^{\mathfrak{s}(u)}_u = n] \\
+ & \sum_{1 \leq m \leq r(l)} (m + n_i - r(u) - \frac{1}{p})^2 \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \overline{\omega_s}] \\
+ & \sum_{1 \leq n \leq n_i - r(u)} (r(l) + n - \frac{1}{p})^2 \mathbf{Pr}[\mathsf{C}^{\mathfrak{s}(u)}_u = m, \overline{\omega_p}] \\
+ & (r(l) + n_i - r(u))^2 \mathbf{Pr}[\overline{\omega_p}, \overline{\omega_s}] \\
< & \sum_{\substack{1 \leq m \leq r(l) \\ 1 \leq n \leq n_i - r(u)}} (2m^2 + 2n^2 + \frac{4}{p^2}) \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \mathsf{C}^{\mathfrak{s}(u)}_u = n] \\
- & \sum_{\substack{1 \leq m \leq r(l) \\ 1 \leq n \leq n_i - r(u)}} \frac{2(m+n)}{p} \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \mathsf{C}^{\mathfrak{s}(u)}_u = n] \\
+ & \sum_{1 \leq m \leq r(l)} (2m^2 + 2(n_i - r(u))^2 + \frac{1}{p^2}) \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \overline{\omega_s}] \\
- & \sum_{1 \leq m \leq r(l)} \frac{2(m + n_i - r(u))}{p} \mathbf{Pr}[\mathsf{C}^l_{\mathfrak{p}(l)} = m, \overline{\omega_s}] \\
+ & \sum_{1 \leq n \leq n_i - r(u)} (2r^2(l) + 2n^2 + \frac{1}{p^2}) \mathbf{Pr}[\mathsf{C}^{\mathfrak{s}(u)}_u = m, \overline{\omega_p}] \\
- & \sum_{1 \leq n \leq n_i - r(u)} \frac{2(r(l) + n)}{p} \mathbf{Pr}[\mathsf{C}^{\mathfrak{s}(u)}_u = m, \overline{\omega_p}] \\
+ & (2r^2(l) + (n_i - r(u))^2) \mathbf{Pr}[\overline{\omega_p}, \overline{\omega_s}] \\
< & 2\mathbf{E}[(\mathsf{C}^l_{\mathfrak{p}(l)})^2] + 2\mathbf{E}[(\mathsf{C}^{\mathfrak{s}(u)}_u)^2] + 2\mathbf{Pr}[\overline{\omega_p}] r^2(l) \\
+ & 2\mathbf{Pr}[\overline{\omega_s}](n_i - r(u))^2 + \frac{4}{p^2} - \frac{2}{p}(\mathbf{E}[\mathsf{C}^l_{\mathfrak{p}(l)}] + \mathbf{E}[\mathsf{C}^{\mathfrak{s}(u)}_u]) \\
= & \frac{4}{p^2} + 2\sum_{j=1}^{r(l)} j^2 p(1-p)^{j-1} + 2\sum_{j=1}^{n_i - r(u)} j^2 p(1-p)^{j-1} \\
+ & 2(1-p)^{r(l)} r(l)^2 + 2(1-p)^{n_i - r(u)} (n_i - r(u))^2 \\
- & \frac{2}{p}(\sum_{j=1}^{r(l)} jp(1-p)^{j-1} + \sum_{j=1}^{n_i - r(u)} jp(1-p)^{j-1}) \leq \frac{8}{p^2}
\end{aligned}$$

□

Compared with BasicCounting whose variance is bounded by $\frac{|D|(1-p)}{p}$, our estimator does provide advantage to improve the communication cost. The total number of samples drawn in the system is expected to be $|S| = |D|p$. In general cases, only a small fraction of the raw data will be sampled, so we have $1 - p > 0.5$. If $|S| = |D|p > 16k$ and $\frac{8}{p^2} < \frac{|D|(1-p)}{p}$, it indicates that the proposed estimator provides smaller variance and incurs smaller communication cost. In contrast, if

$|S| = |D|p \leq 8k$, it means that the average number of samples transferred by each node is no larger than 16. In this case, a node could pack the samples into an ordinary heartbeat message to the broker, and no more communication cost is incurred either.

Since the global range counting aggregation is the sum of local range counting aggregations, which is produced independently, we conclude that $\hat{\gamma}(l,u,S)$ is an unbiased estimator of $\gamma(l,u,D)$ with bounded variance in Theorem 3.2.

Theorem 3.2: For any l and u, $\hat{\gamma}(l,u,S) = \sum_i^k \hat{\gamma}(l,u,i)$ is an unbiased estimation of $\gamma(l,u,D)$ with variance $\text{Var}[\hat{\gamma}(l,u,i)] \leq \frac{8k}{p^2}$, where k is the number of nodes.

Therefore, by setting $p = \sqrt{8k}/\alpha n$, the variance will be $(\alpha n)^2$. According to Chebyshev's inequality, this means that $\hat{\gamma}(l,u,S)$ approximates $\gamma(l,u,D)$ within an additive error of αn with constant probability. We can make this constant probability arbitrarily close to 1 by enlarging p by appropriate constant factors. The total communication overhead in this case is $\sqrt{8k}/\alpha$, since this is the expected number of samples to be transferred. Furthermore, according to Chebyshev's inequality, we can extend the above approximation, which is within additive error of αn with constant probability, to the case with a certain probability guarantee (say δ):

Theorem 3.3: Given query parameters l and u, the number of nodes k, for any $0 < \alpha < 1$ and $0 < \delta < 1$, if the sampling probability p in the RankCounting estimator satisfies that $p \geq \frac{\sqrt{2k}}{\alpha n}\frac{2}{\sqrt{1-\delta}}$, then $\hat{\gamma}(l,u,S)$ is an (α,δ)-range counting.

Proof. The RankCounting estimator provides an unbiased estimation $\hat{\gamma}(l,u,S)$ and its variance is no larger than $\frac{8k}{p^2}$. Combined with the Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbf{Pr}[|\hat{\gamma}(l,u,S) - \gamma(l,u,D)| \leq \alpha n] \\
&= \mathbf{Pr}[|\hat{\gamma}(l,u,S) - \mathbf{E}[\hat{\gamma}(l,u,D)]| \leq \alpha n] \\
&\geq 1 - \frac{\mathbf{Var}[\hat{\gamma}(l,u,S)]}{(\alpha n)^2} = 1 - \frac{\frac{8k}{p^2}}{(\alpha n)^2} \geq 1 - \frac{\frac{8k}{(\frac{\sqrt{2k}}{\alpha n}\frac{2}{\sqrt{1-\delta}})^2}}{(\alpha n)^2} \\
&= \delta.
\end{aligned}$$

To this end, we can see that $\hat{\gamma}(l,u,S)$ satisfies the requirement of (α,δ)-Range counting. □

B. Differentially Private Approximate Range Counting

To keep sensitive data in D private, we adopt a two-phase approach to response to (α,δ)-range counting in a private manner, and we formulate an optimization problem aiming at achieving the optimal differential privacy, while taking accuracy requirements as constraints. Given query parameters l, u, α and δ, a data broker firstly chooses a pair of (α',δ'), and computes (α',δ')-range counting $\hat{\gamma}(l,u,S)$. Then the Laplacian mechanism is employed with privacy budget ϵ, and finally $\gamma^*(l,u,S) = \hat{\gamma}(l,u,S) + \text{Lap}(\epsilon)$ is returned to the querying data customer. Note that $\gamma^*(l,u,S)$ should satisfy the (α,δ)-range counting accuracy. We can get $\alpha' < \alpha$ and $\delta' > \delta$, since $\hat{\gamma}(l,u,S)$ must be more accurate than $\gamma^*(l,u,S)$, otherwise, it will violate the accuracy requirement after adding Laplacian noise $\text{Lap}(\epsilon)$. For each pair of (α',δ'), the data broker should include as many samples as possible in the computation of $\hat{\gamma}(l,u,S)$. Then in the next step a larger search space of ϵ will be obtained under the final (α,δ) accuracy requirement. For a given pair of (α',δ'), the data broker calculates the smallest ϵ making $\gamma^*(l,u,S)$ a (α,δ)-range counting. After traversing all pairs of (α',δ'), the data broker is able to achieve the optimal differential privacy, *e.g.*, the smallest ϵ.

A problem in the above process is how to calculate the accuracy of $\gamma^*(l,u,S) = \hat{\gamma}(l,u,S) + \text{Lap}(\epsilon)$, given $\hat{\gamma}(l,u,S)$ is an (α',δ')-range counting and Laplacian noise is determined by ϵ. This problem is equivalent to calculate the probability $\mathbf{Pr}[|\gamma^*(l,u,S) - \gamma(l,u,D)| \leq \alpha n]$. It is known that $\mathbf{Pr}[|\hat{\gamma}(l,u,S) - \gamma(l,u,D)| \leq \alpha' n] \geq \delta'$. For a specified ϵ, suppose $\mathbf{Pr}[|\text{Lap}(\epsilon)| \leq (\alpha - \alpha')n] \geq \tau$, then we have

$$\begin{aligned}
&\mathbf{Pr}[|\gamma^*(l,u,S) - \gamma(l,u,D)| \leq \alpha n] \\
&\geq \mathbf{Pr}[|\hat{\gamma}(l,u,S) - \gamma(l,u,D)| + |\text{Lap}(\epsilon)| \leq \alpha n] \\
&\geq \mathbf{Pr}[|\hat{\gamma}(l,u,S) - \gamma(l,u,D)| \leq \alpha' n]\mathbf{Pr}[|\text{Lap}(\epsilon)| \leq (\alpha - \alpha')n] \\
&= \delta\tau.
\end{aligned}$$

From the above inequality, it is seen that if $\mathbf{Pr}[|\text{Lap}(\epsilon)| \leq (\alpha - \alpha')n] = \tau \leq \frac{\delta}{\delta'}$, then we will derive that $\mathbf{Pr}[|\gamma^*(l,u,S) - \gamma(l,u,D)| \leq \alpha n] \geq \delta$ holds. It is worth mentioning that random variables $X = \hat{\gamma}(l,u,S) - \gamma(l,u,D)$ and $Y = \text{Lap}(\epsilon)$ are independent, and if $\tau < \frac{\delta}{\delta'}$, then we cannot guarantee that $\mathbf{Pr}[|\gamma^*(l,u,S) - \gamma(l,u,D)| \leq \alpha n] \geq \delta$ holds.

After injecting a Laplacian noise scaled at $\text{Lap}(\epsilon)$, not only the inaccuracy is composited by sampling and Laplacian noise, but also the final privacy degree. In other words, the differential privacy budget achieved by our two-phase approach denoted as ϵ' is determined, in a collaborative manner, by sampling probability p and privacy budget ϵ. Lemma 3.4 generalized from [11] shows the relationship of ϵ', p and ϵ.

Lemma 3.4: If function $\phi(\cdot)$ is ϵ-differentially private, and function $S(\cdot)$ returns independent random samples with probability $0 \leq p \leq 1$, then $\phi(S(\cdot))$ is ϵ'-differential private, where $\epsilon' = \ln(1 - p + pe^{\epsilon})$.

Proof. Let D and D' be any pair of neighboring datasets, assuming $D = D' \cup \{i\}$. Let o be any output of $\phi(S(\cdot))$.

$$\begin{aligned}
&\mathbf{Pr}[\phi(S(D)) = o] \\
&= \sum_{Z \subseteq S(D')} p\mathbf{Pr}[S(D') = Z]\mathbf{Pr}[\phi(Z \cup \{i\}) = o] \\
&+ \sum_{Z \subseteq S(D')} (1-p)\mathbf{Pr}[S(D') = Z]\mathbf{Pr}[\phi(Z) = o] \\
&\leq (pe^{\epsilon} + 1 - p) \sum_{Z \subseteq S(D')} \mathbf{Pr}[S(D') = Z]\mathbf{Pr}[\phi(Z) = o] \\
&= (pe^{\epsilon} + 1 - p)\mathbf{Pr}[\phi(S(D')) = o]
\end{aligned}$$

Similarly, we can get that

$$\mathbf{Pr}[\phi(S(D')) = o] \leq (pe^{\epsilon} + 1 - p)\mathbf{Pr}[\phi(S(D)) = o],$$

and $\phi(S(\cdot))$ is ϵ'-differential private, with $\epsilon' = \ln(1 - p + p \cdot e^{\epsilon})$. □

Now we can formulate the optimization problem of achieving the optimal differential privacy (with the smallest privacy budget) under the (α, δ)-range counting accuracy requirement. In the following, we present the formulation:

$$\begin{array}{lrcl} \min & \multicolumn{3}{l}{\epsilon' = \ln(1 + p(e^{\epsilon} - 1))} \\ \text{s.t.} & \frac{\sqrt{2k}}{\alpha' n}\frac{2}{\sqrt{1-\delta'}} & \leq & p \\ & \alpha' & \leq & \alpha \\ & \delta & \leq & \delta' \\ & \Pr[|\text{Lap}(\epsilon)| \leq (\alpha - \alpha')n] & \geq & \frac{\delta}{\delta'} \\ & \epsilon & \geq & 0 \end{array} \qquad (3)$$

Given α, δ and p as inputs, the optimization problem in (3) can be constructed, taking the optimized ϵ (together with intermediate α', δ') as output, which is used to conduct an (α, δ)-range counting. The search space of (3) contains all the feasible solutions for (α, δ)-range counting conducted by the two-phase approach. Next, we show how to compute the optimal solution to (3) and how to conduct a differentially private (α, δ)-range counting based on the optimal solution.

For a fixed α', to avoid repeated sampling in continuous queries, a data broker uses all the existing samples (collected with p) at the base station to compute an (α', δ')-range counting. Then, δ' can be obtained by setting $\frac{\sqrt{2k}}{\alpha' n}\frac{2}{\sqrt{1-\delta'}} = p$. According to (3), α, δ, p, k and n are all constants, and the data broker can obtain a minimum ϵ for the fixed α' with the following constraint:

$$\mathbf{Pr}[|\text{Lap}(\epsilon)| \leq (\alpha - \alpha')n] = 1 - e^{-\frac{(\alpha-\alpha')n\epsilon}{\Delta\hat{\gamma}}} \leq \frac{\delta}{\delta'}.$$

Then we know $\epsilon \geq \frac{\Delta\hat{\gamma}}{(\alpha-\alpha')n} \ln \frac{\delta'}{\delta'-\delta}$, and the optimal differential privacy is achieved by setting $\epsilon = \frac{\Delta\hat{\gamma}}{(\alpha-\alpha')n} \ln \frac{\delta'}{\delta'-\delta}$. $\Delta\hat{\gamma}$ is the sensitivity of $\hat{\gamma}(l, u, i)$. In the worst case, $\Delta\hat{\gamma}$ grows to n_i with an extremely small probability, and adopting $\Delta\hat{\gamma} = n_i$ will totally destroy the aggregation utility. A fair solution is to use the expectation of $\Delta\hat{\gamma}$, which is $\frac{1}{q}$ in the general cases. By traversing α' in $[0, \alpha]$, an optimal solution consisting of α', δ' and ϵ can be found. Although the searching range of α' is continuous, we can approximate it to a discrete domain with arbitrarily small intervals.

Given an optimal solution consisting of α', δ' and ϵ, the data broker carries out the following two steps. First, an (α', δ')-range counting is computed as $x' = \sum_1^k \hat{\gamma}(l, u, i)$. Second, Laplacian noise Lap(ϵ) is added on x' and the final result is $x'' = \sum_1^k \hat{\gamma}(l, u, i) + \text{Lap}(\epsilon)$, which is an ϵ'-differentially private (α, δ)-range counting.

IV. Pricing Mechanism

This section discusses the existence of arbitrage attacks for pricing mechanisms, and how the pricing mechanism should be designed to thwart such attacks.

Initiatively, $\pi(\alpha, \delta)$ should monotonically decrease with the approximation degree α and increase with δ, to achieve a negative correlation with the variance. However, a carelessly designed $\pi(\alpha, \delta)$ could still be vulnerable under the negative correlation.

Example 4.1: A data consumer wants to buy a range counting aggregation service $\Lambda(\alpha, \delta)$ with low aggregation variance. Therefore, she will intuitively specify a small value of α and a large value of δ, and a higher price will be charged. However, she may also circumvent to pay the full price, and turns to buy multiple cheaper services of the same range counting with higher variances, denoted as $\{\Lambda(\alpha_i, \delta_i) | i \in \{1, \ldots, m\}, \alpha_i > \alpha, \delta_i < \delta\}$. We use $V(\alpha_i, \delta_i)$ to indicate the variance derived from the parameter pair (α_i, δ_i), respectively. Afterwards, the data consumer estimates the result according to Formula (4).

$$\begin{aligned} &\{\Lambda(\alpha_1, \delta_1), \ldots, \Lambda(\alpha_m, \delta_m)\} \\ &\longmapsto \gamma(.) = \frac{1}{m}\sum_1^m \gamma_i(.) \\ &\longmapsto V(.) = \frac{1}{m^2}\sum_{i=1}^m V(\alpha_i, \delta_i), \end{aligned} \qquad (4)$$

In other words, the data consumer obtains a final range counting aggregation by computing the average of m noisy answers, with an accumulated variance $\frac{1}{m^2}\sum_{i=1}^m V(\alpha_i, \delta_i)$ potentially lower than $V(\alpha, \delta)$. If the pricing function $\pi(\alpha, \delta)$ is arbitrage avoiding, the following conditional statement must hold:

$$\frac{1}{m^2}\left\{\sum_{i=1}^m V(\alpha_i, \delta_i)\right\} \leq V(\alpha, \delta) \Rightarrow$$
$$\sum_{i,j=1}^m \pi(\alpha_i, \delta_i) \geq \pi(\alpha, \delta).$$

With the above conditional statement, we first use Lemma 4.1 to show the equivalence property of arbitrage-avoiding pricing function $\pi(\alpha, \delta)$ according to the variance.

Lemma 4.1: For an arbitrary-avoiding pricing function $\pi(\alpha, \delta)$, there must be another function $\psi(V(\alpha, \delta)) = \pi(\alpha, \delta)$, when $\pi(\alpha, \delta)$ is arbitrage free.

Proof. Assume there are two arbitrary sets of parameters α, δ and α', δ' with $V(\alpha, \delta) = V(\alpha', \delta')$. Then we must have $\pi(\alpha, \delta) = \pi(\alpha', \delta')$, *i.e.*, $\psi(V(\alpha, \delta)) = \psi(V(\alpha', \delta'))$, indicating the price is uniquely determined by the variance.

Otherwise, we have $\pi(\alpha, \delta) \neq \pi(\alpha', \delta')$, which means $\pi(\alpha, \delta)$ is not uniquely determined by $V(\alpha, \delta)$. We prove the lemma by contradiction. Assume $\pi(\alpha, \delta) > \pi(\alpha', \delta')$, then an arbitrage attack exists, as $V(\alpha', \delta') = V(\alpha, \delta)$, and $\pi(\alpha', \delta') < \pi(\alpha, \delta)$. This contradicts the assumption that $\pi(\alpha, \delta)$ is arbitrary-free. The proof is finished. □

With the above equivalent expression, the following theorem shows how to address an arbitrage attack.

Theorem 4.2: Any pricing function $\pi(\alpha, \delta)$ is arbitrage-avoiding if and only if the following properties hold:

1) $\pi(\alpha, \delta) = \psi(V(\alpha, \delta))$
2) $\forall \alpha = \alpha_0, \delta = \delta_0, \Delta\delta \geq 0$, the relative difference of $\pi(\cdot)$ and $V(\cdot)$ follows $\frac{\pi(\alpha_0,\delta_0+\Delta\delta)-\pi(\alpha_0,\delta_0)}{\pi(\alpha_0,\delta_0+\Delta\delta)} \geq \frac{V(\alpha_0,\delta_0)-V(\alpha_0,\delta_0+\Delta\delta)}{V(\alpha_0,\delta_0)}$

3) $\forall \alpha = \alpha_0, \delta = \delta_0, \Delta\alpha \geq 0$, the relative difference of $\pi(\cdot)$ and $V(\cdot)$ follows $\frac{\pi(\alpha_0,\delta_0)-\pi(\alpha_0+\Delta\alpha,\delta_0)}{\pi(\alpha_0,\delta_0)} \leq \frac{V(\alpha_0+\Delta\alpha,\delta_0)-V(\alpha_0,\delta_0)}{V(\alpha_0+\Delta\alpha,\delta_0)}$.

Proof.

Part 1: Sufficiency.

We first prove the sufficiency of the properties, where the three listed properties can guarantee the pricing function to be arbitrage-avoiding.

Assume there are two groups of parameters: $\{(\alpha, \delta)\}$ and $\{(\alpha_1, \delta_1), (\alpha_2, \delta_2), \cdots, (\alpha_m, \delta_m)\}$. Data consumers could either request for range counting once with parameters $\{(\alpha, \delta)\}$, or request multiple times with $(\alpha_2, \delta_2), \cdots, (\alpha_m, \delta_m)\}$. In the latter strategy, data consumers will apply the average value of all results as the final conclusion, with the variance equals $Var(\frac{1}{m}\sum_{i=1}^{m} R_i)$, where R_i indicates the result for the ith query.

To prove pricing function $\pi(\alpha, \delta)$ is arbitrage-avoiding, we need to prove that $\pi(\alpha, \delta) \leq \sum_{i=1}^{m} \pi(\alpha_i, \delta_i)$ when $Var(\frac{1}{m}\sum R_i) \leq V(\alpha, \delta)$.

Firstly, we have

$$Var(\frac{1}{m}\sum R_i) = \frac{1}{m^2}\sum_{i=1}^{m} V(\alpha_i, \delta_i) = \frac{1}{m^2}\sum_{i=1}^{m} k_i \cdot V(\alpha, \delta). \tag{5}$$

We also have

$$\sum_{i=1}^{m} \pi(\alpha_i, \delta_i) = \sum_{i=1}^{m} l_i \cdot \pi(\alpha, \delta). \tag{6}$$

For an arbitrary set (α_i, δ_i), we introduce an intermediate set (α, δ_i), and estimate its relative change on variance and price compared to (α, δ). According to Property 2), we can accumulate the difference and

$$\frac{\pi(\alpha, \delta_i)}{\pi(\alpha, \delta)} \geq \frac{V(\alpha, \delta)}{V(\alpha, \delta_i)}. \tag{7}$$

We can further estimate the change according to Property 3) by accumulating the difference, and achieve the following inequality:

$$\frac{\pi(\alpha_i, \delta_i)}{\pi(\alpha, \delta_i)} \geq \frac{V(\alpha, \delta_i)}{V(\alpha_i, \delta_i)}. \tag{8}$$

Then we can combine the two inequalities, and the following conclusion holds:

$$\frac{\pi(\alpha_i, \delta_i)}{\pi(\alpha, \delta)} \geq \frac{V(\alpha, \delta)}{V(\alpha_i, \delta_i)}, \tag{9}$$

which means $l_i \geq \frac{1}{k_i}$.

As $\frac{1}{m^2}\sum_{i=1}^{m} k_i \cdot V(\alpha, \delta) \leq V(\alpha, \delta)$, we have $\sum_{i=1}^{m} k_i \leq m^2$.

Assume $\overline{k} = \frac{\sum_{i=1}^{m} k_i}{m}$. Then we have $\overline{k} \leq m$. The conclusion can be further extended:

$$\overline{k} \geq \frac{1}{m}. \tag{10}$$

The result in Inequality (10) can be combined with the following fact:

$$\sum_{i=1}^{m} \frac{1}{k_i} \geq \sum_{i=1}^{m} \frac{1}{\overline{k}}, \tag{11}$$

and guarantees

$$\sum_{i=1}^{m} \frac{1}{k_i} \geq 1. \tag{12}$$

Finally, we combine the results:

$$\sum_{i=1}^{m} l_i \geq \sum_{i=1}^{m} \frac{1}{k_i} \geq 1. \tag{13}$$

This is the same as the conclusion that

$$\sum_{i=1}^{m} \pi(\alpha_i, \delta_i) \geq \pi(\alpha, \delta), \tag{14}$$

and the proof of sufficiency is completed.

Part 2: Necessity.

Now we prove the necessity of the properties. Assume pricing function $\pi(\alpha, \delta)$ is arbitrage-avoiding. We achieve the conclusion by contradiction.

The necessity of the first property is proved in Lemma 4.1.

As for the second property, assume there is a pair of α' and δ', where

$$\frac{\pi(\alpha', \delta' + \Delta\delta) - \pi(\alpha', \delta')}{\pi(\alpha', \delta' + \Delta\delta)} < \frac{V(\alpha', \delta') - V(\alpha', \delta' + \Delta\delta)}{V(\alpha', \delta')} \tag{15}$$

for some $\Delta\delta > 0$. Assume $V(\alpha', \delta') = k \cdot V(\alpha', \delta + \Delta\delta)$, and $\pi(\alpha', \delta') = l \cdot \pi(\alpha', \delta + \Delta\delta)$. Then there must be a pair of integers m_1 and m_2, such that data consumers can either buy range counting queries m_1 times with $V(\alpha', \delta')$, or m_2 times with $V(\alpha', \delta + \Delta\delta)$, while their general variances are identical. In this case, the total cost for the first strategy is smaller than that of the second one, as $l \leq \frac{1}{k}$ according to Inequality (15). Then it contradicts the fact that $\pi(\alpha, \delta)$ is arbitrage-avoiding, as we can achieve the same query result and variance with a lower cost.

The proof of the third property is similar with that of Property 2), thus omitted here.

Generally, any arbitrage-avoiding pricing function $\pi(\alpha, \delta)$ must guarantee all the properties, and the necessity is proved. □

V. Experiments

This section presents the evaluation results of the proposed method. Both the results for approximate range counting aggregation, and the tradeoff between privacy preservation and utility are investigated.

Datasets. The evaluation employs a real-world dataset, the pollution records in CityPulse Smart City Datasets [12]. The data are contributed by sensors with exact locations along the rode sides. Each record contains 5 air quality indexes: *ozone*, *particulate_matter*, *carbon_monoxide*, *sulfur_dioxide*, and *nitrogen_dioxide*. The dataset has 17568 records collected from 0:05am, 8/1/2014 to 0:00am, 10/1/2014.

A. *Effectiveness of the Sampling-Based Algorithm*

We evaluate the performance of the proposed sampling-based algorithm for answering approximate range counting queries.

The first evaluation validates the impact of sampling probability p on the accuracy of the sampling algorithm. In this evaluation, the maximum relative error of the sampling algorithm is calculated, while p increases from 0.0173 to 0.4048. Fig. 2 shows that the maximum relative error is high with the maximum value 27% when the sampling probability is less than 0.12. Furthermore, with less data being preserved for querying, the accuracy oscillates considerably. Querying accuracy is very high if there are more than 15% data which are preserved in the samples; and accuracies remain stable. From Fig. 2, we further conclude that the proposed algorithm has high aggregation accuracy, *i.e.*, the maximum relative error can be bounded to 3% if there are more than 5% data which are preserved in the samples.

The second evaluation studies the impact of α and δ on the accuracy of the achieved results. The accuracy is computed while α and δ increase from 0.08 to 0.8. Fig. 3 shows that the maximum relative error remains stable when δ is larger than 0.3. Furthermore, the maximum relative error is less than 0.019 when δ is larger than 0.3. When a small δ is given, *e.g.*, $\delta < 0.3$, the maximum relative error shocks significantly.

The third evaluation investigates the impact of data size on the sampling probability. In this evaluation, α and δ are set to be 0.055 and 0.5, respectively. The sampling probability is computed while the size of data increases from 10% to 100% of the original dataset. As shown in Fig. 4, our algorithm is suitable for big data range counting aggregation, as it largely reduces computation and communication overhead. When data size is very large, the sampling probability can converge to a stable state with less data collected. Therefore, our sampling probability can appropriately balance aggregation accuracy and overhead.

B. *Privacy-Utility Tradeoff*

We also evaluate the performance of the differential privacy mechanism for realizing privacy-utility tradeoff. The first evaluation investigates the impact of privacy budget ϵ on the accuracy of the range counting aggregation. The accuracy is computed while α and δ increase from 0.08 to 0.8. Furthermore, ϵ increases from 0.01 to 8, and $p = 0.4$. As shown in Fig. 5, with the decreasing of the privacy requirements, aggregation accuracy can be improved, indicating less privacy concerns lead to better utilities. Moreover, if a strong privacy guarantee is required with $\epsilon = 0.1$, our algorithm can still bound the relative error under 8% for all these 5 datasets, showing the high stability of the proposed framework.

The second evaluation investigates the impact of sampling probability on privacy preservation under different privacy budgets. Fig. 6 shows that querying accuracy is low when $p < 0.15$. With the increasing of sampling probability from 0.0173 to 0.25, querying accuracy is improved as more samples are collected. The above observations indicate that the global sensitivity of $\hat{\gamma}(b_l, b_u, S)$ satisfies $GS(\hat{\gamma}(b_l, b_u, S)) \propto 1/p$, and a larger p means smaller volume of differential privacy noise.

VI. Literature Review

Approximate Range Counting Aggregation. Sampling-based algorithms have been proposed for approximate data aggregation in many areas [13] [14] [15], such as data stream, tradition database systems, P2P networks, etc. However, these works are not designed for range counting queries for big data in IoT, indicating that no guarantees on performance are provided for this kind of queries.

The sampling-based algorithms have also been applied for long-term queries via continuous data collection. Considering the high correlation in temporal and spatial dimensions, the work in [16] proposes a distributed approximate aggregation algorithm which can considerably reduce aggregation overhead. The work in [17] proposes some algorithms to realize the tradeoff between aggregation overhead and aggregation accuracy in order to prolong network lifetime through allocating tolerable error bounds to each sensor node in a network. Some indexing structures are proposed in [18] to conduct spacial online sampling and data aggregation on big temporal-spatial and spatial datasets. As a sampling-based algorithm, the structures proposed in [18] have good performance for dynamic datasets. However, these works mainly focus on the reduction of long-term bandwidth consumption. They have no promise on the performance of the one sample multiple queries discussed in this paper. Furthermore, they neglect the privacy issues underlying the collected data.

Differentially Private Data Aggregation. There are many state-of-the-art paradigms investigating differentially private data aggregation [19]. Especially, the work in [20] studies the problem of hierarchical decomposition with differential privacy guarantee based on spatial decomposition trees. The paradigm proposed in [20] can efficiently answer differentially private range counting by eliminating the dependence of querying sensitivity on the maximum height of the decomposition tree. The work in [21] investigates the problem of releasing the degree distribution of a graph under differential privacy. To reduce Laplace noise volume, the work in [21] transfers high-dimensional graph data to low dimensional data based on a graph projection approach. However, these studies mainly focus on reducing the scale of noise introduced by privacy preservation, while they fail to demonstrate an unbiased estimator and cannot be applied directly for data in distributed environment.

Trading Private Aggregation. For the trade of data analyzing results, current research mainly focuses on arbitrage avoiding in querying-based pricing mechanisms [22] [23]. The work in [9] proposes some arbitrage avoiding pricing mechanisms for arbitrary formats of queries. However, these conclusions cannot be extended to the range counting aggregation, and no guarantees on privacy and utility tradeoff are given.

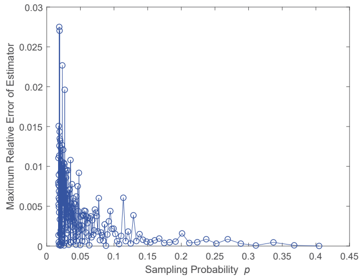

Fig. 2. Querying accuracy affected by sampling probability p.

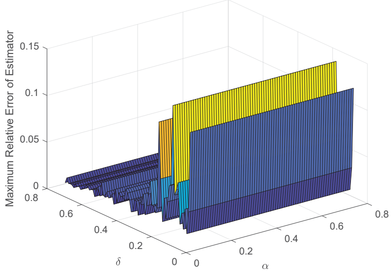

Fig. 3. Querying accuracy affected by α and δ.

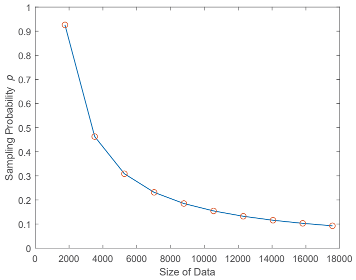

Fig. 4. Sampling probability and data size relationship.

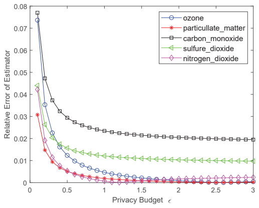

Fig. 5. Querying accuracy affected by ϵ with $p = 0.4$.

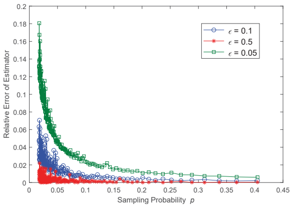

Fig. 6. Querying accuracy affected by p.

VII. Conclusion

This paper investigates the problem of trading approximate range counting for big IoT data. The objective is to derive query answers with optimal differential privacy guarantee, while the answers satisfy the specified approximation degree. A two-phase sampling-based approach with bounded variance is proposed, including an unbiased estimator based on sampled data and a perturbation mechanism designed for optimal differential privacy. The paper further studies the pricing mechanism for trading counting results, facing the arbitrage attacks from cunning consumers. As a result, the critical condition to establish a proper pricing function is established. All the proposed methods are validated towards real-world datasets.

Acknowledgment

This work is partly supported by the National Science Foundation (NSF) under grant NOs. 1252292, 1741277, 1829674, and 1704287.

References

[1] "Data brokers: A call for transparency and accountability: A report of the federal trade commission," https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014, 2014.

[2] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "The design of an acquisitional query processor for sensor networks," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '03, 2003.

[3] ——, "Tag: A tiny aggregation service for ad-hoc sensor networks," *ACM SIGOPS Operating Systems Review*, vol. 36, 2002.

[4] J. Li and S. Cheng, "(ε, δ)-approximate aggregation algorithms in dynamic sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 3, pp. 385–396, 2012.

[5] J. Li, S. Cheng, Z. Cai, J. Yu, C. Wang, and Y. Li, "Approximate holistic aggregation in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 2, p. 11, 2017.

[6] Z. He, Z. Cai, S. Cheng, and X. Wang, "Approximate aggregation for tracking quantiles and range countings in wireless sensor networks," *Theoretical Computer Science*, vol. 607, pp. 381–390, 2015.

[7] "Invasion of the data snatchers:," https://www.aclu.org/blog/speakeasy/invasion-data-snatchers-big-data-and-internet-things-means-surveillance-everything, accessed: 2010-09-30.

[8] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, 2006, pp. 1–12.

[9] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 757–768, 2014.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[12] "Citypulse smart city dataset:," http://iot.ee.surrey.ac.uk:8080/datasets.html, accessed: 2014.

[13] H. Harb, A. Makhoul, D. Laiymani, and A. Jaber, "A distance-based data aggregation technique for periodic sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 13, no. 4, p. 32, 2017.

[14] S. Cheng, J. Li, and Z. Cai, "O (ε)-approximation to physical world by sensor networks," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 3084–3092.

[15] S. Cheng, Z. Cai, and J. Li, "Approximate sensory data collection: A survey," *Sensors*, vol. 17, no. 3, p. 564, 2017.

[16] A. Boulis, S. Ganeriwal, and M. B. Srivastava, "Aggregation in sensor networks: an energy–accuracy trade-off," *Ad hoc networks*, vol. 1, no. 2-3, pp. 317–331, 2003.

[17] X. Tang and J. Xu, "Extending network lifetime for precision-constrained data aggregation in wireless sensor networks," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, April 2006, pp. 1–12.

[18] L. Wang, R. Christensen, F. Li, and K. Yi, "Spatial online sampling and aggregation," *Proceedings of the VLDB Endowment*, vol. 9, no. 3, pp. 84–95, 2015.

[19] R. Chen, B. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: a case study on the montreal transportation system," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 213–221.

[20] J. Zhang, X. Xiao, and X. Xie, "Privtree: A differentially private algorithm for hierarchical decompositions," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 155–170.

[21] W.-Y. Day, N. Li, and M. Lyu, "Publishing graph degree distribution with node differential privacy," in *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016, pp. 123–138.

[22] S. Deep and P. Koutris, "Qirana: A framework for scalable query pricing," in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 699–713.

[23] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen, "Unlocking the value of privacy: Trading aggregate statistics over private correlated data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2031–2040.