Locally Differentially Private Quantile Summary Aggregation in Wireless Sensor Networks

Aishah Aseeri[1] and Rui Zhang[2](✉)

[1] King Abdulaziz University, Jeddah 21589, Kingdom of Saudi Arabia
aaaseeri@kau.edu.sa
[2] University of Delaware, Newark, DE 19716, USA
ruizhang@udel.edu

Abstract. Privacy-preserving data aggregation has been widely recognized as a key enabling functionality in wireless sensor networks to allow the base station to learn valuable statistics of the sensed data while protecting individual sensor node's data privacy. Existing privacy-preserving data aggregation schemes all target simple statistic functions such as SUM, COUNT, and MAX/MIN. In contrast, a quantile summary allows a base station to extract the ϕ-quantile for any $0 < \phi < 1$ of all the sensor readings in the network and can thus provide a more accurate characterization of the data distribution. Unfortunately, how to realize privacy-preserving quantile summary aggregation remains an open challenge. In this paper, we introduce the design and evaluation of PrivQSA, a novel privacy-preserving quantile summary aggregation scheme for wireless sensor networks, which enables efficient quantile summary aggregation while guaranteeing ϵ-Local Differential Privacy for individual sensors. Detailed simulation studies confirm the efficacy and efficiency of the proposed protocol.

Keywords: Data aggregation · Wireless sensor network · Local differential privacy · Quantile summary

1 Introduction

Data aggregation [16] has been widely recognized as a key technique for reducing energy consumption and prolonging network lifetime by allowing sensed data to be aggregated by intermediate nodes along the route to the base station to eliminate possible redundancy. Data privacy is a key concern in many applications. For example, data generated by sensor nodes in an IoT-based smart-home system may contain a variety of sensitive information about users such as appliance usage and home occupancy. Since directly submitting such information to a base station would reveal sensitive information, there is strong need for privacy-preserving data aggregation solutions that can allow the base station to learn valuable statistic of the data generated in the network while ensuring the data privacy of individual sensor nodes.

N. T. Nguyen et al. (Eds.): ACIIDS 2022, LNAI 13757, pp. 364–376, 2022.
https://doi.org/10.1007/978-3-031-21743-2_29

Privacy-preserving data aggregation has received significant attentions over the past decade. Existing privacy-preserving data aggregation schemes [7,9,11, 14,15,18,21] all target simple statistic functions such as SUM, COUNT, and MAX/MIN. In contrast, a quantile summary allows a base station to learn a more accurate distribution of the sensed data than simple statistics functions [5,6,10,19]. Specifically, a quantile summary allows the base station to retrieve the ϕ-quantile for any $0 \leq \phi \leq 1$, which can provide a much better characterization of the distribution of data generated by a wireless sensor network. Unfortunately, how to realize privacy-preserving quantile summary aggregation remains an open challenge.

In this paper, we introduce the design and evaluation of PrivQSA, a novel privacy-preserving quantile summary aggregation scheme for wireless sensor networks. Specifically, we design PrivQSA to satisfy ϵ-Local Differential Privacy (LDP), which is a model widely considered as the gold standard for data privacy. Under PrivQSA, every sensor node randomly perturbs its set of readings to ensure ϵ-LDP. All the nodes then participate in the quantile summary aggregation to allow the base station to obtain a quantile summary of the perturbed readings. The base station then estimates a quantile summary of the original sensed data based on the perturbation mechanism followed by individual sensor nodes. Our contributions in this paper can be summarized as follows.

- We are the first to study privacy-preserving quantile summary aggregation in wireless sensor networks.
- We introduce PrivQSA, a novel privacy-preserving quantile summary aggregation scheme that can allow the base station to learn a quantile summary of the sensed data while ensuring ϵ-LDP for individual sensor nodes.
- We confirm the efficacy and efficiency of PrivQSA via both theoretical analysis and detailed simulation studies, which demonstrate significant advantages over other baseline solutions.

The rest of this paper is structured as follows. Section 2 discusses the related work. Section 3 introduces the network model and some preliminaries. Section 4 introduces the design of PrivQSA. Section 5 evaluates PrivQSA via both theoretical analysis and simulation studies. Section 6 finally concludes this paper.

2 Related Work

In this section, we review some related work in quantile summary aggregation, privacy preserving data aggregation in WSNs.

Quantile summary aggregation in wireless sensor networks has been discussed in a number of articles throughout literature. Greenwald et. al. [5] studied quantile summary aggregation in wireless sensor networks. Shrivastava *et al.* [19]proposed a quantile digest summary structure to facilitate quantile aggregation. In [6], the authors designed a distributed algorithm to realize an ϵ-approximate quantile summary of all the sensor nodes data. Later, Huang *et al.* [10] introduced an improvement to the forementioned algorithm with the aim to reduce the maximum per node communication cost. Also, several efficient gossip distributed algorithms were proposed in [8] to compute the exact and approximate quantiles. However, none of the discussed quantile aggregation schemes accounts for privacy restrictions and therefore cannot be adapted to our proposed problem.

Privacy-preserving data aggregation in sensor networks has received a lot of attention over the past two decades [2–4,9,11,18,20]. Generally speaking, existing solutions for privacy preserving data aggregation can be classified into two categories. The first category uses encryption techniques such as homomorphic encryption [2,3,11,18,20,22], secure multiparty computation [12], and modulo addition-based encryption [1]. All these solutions target simple statistic functions such as SUM, COUNT, and MAX/MIN and cannot be adopted to support quantile summary aggregation. The second category uses random perturbation [4,9,17], in which each sensor node randomly perturbs its data according to a suitable probability distribution before participating in data aggregation, and the base station can still infer valuable statistics from the perturbed data. To the best of our knowledge, there is no prior work tackling privacy-preserving quantile summary aggregation.

3 Problem Formulation

In this section, we first introduce the network model and a background on quantile summary. We then provide the definition of Local Differential Privacy.

3.1 Network Model

We consider a wireless sensor network model consisting of a base station and n sensor nodes that form an aggregation tree. Let $R = \{1, \ldots, d\}$ be the range of possible readings. We assume that every sensor node i has a set of m readings $V_i = \{v_{i,1}, \ldots, v_{i,m}\}$, where every reading $v_{i,j} \in R$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$. The set of all the sensed data generated in the sensor network is then $V = \bigcup_{i=1}^{n} V_i$. The base station aims to obtain a quantile summary of V.

3.2 Quantile Summary

A quantitle summary is a subset of readings along with their (estimated) global ranks which can support *value-to-rank* query over any $v \in R$ as well as ϕ-quantile queries for any $0 < \phi < 1$. Specifically, given a set of N distinct data values with a total order, the ϕ-quantile is the value v with rank $r(v) = \lfloor \phi N \rfloor$ in the set, where $r(v)$ is the number of values in the set smaller than v. Since a quantile summary that can provide the exact quantiles must contain the all N values in the worst case, an ϵ'-approximate ϕ-quantile is a value with rank between $(\phi - \epsilon')N$ and $(\phi + \epsilon')N$.

3.3 Local Differential Privacy (LDP)

Local Differential Privacy is a strong privacy notion widely considered as the gold standard for data privacy, which ensures that an adversary cannot differentiate two inputs based on the output he observe beyond certain predefined threshold. We give the definition of ϵ-Local Differential Privacy below.

Definition 1. *(ϵ-Local Differential Privacy). A randomized mechanism $\mathcal{M}$ satisfies ϵ-local differential privacy if and only if*

$$\frac{Pr[\mathcal{M}(x) = y]}{Pr[\mathcal{M}(x') = y]} \leq e^{\epsilon}$$

for any two inputs $x, x' \in X$ and any output $y \subseteq Range(\mathcal{M})$, where X is the domain of the input, $Range(\mathcal{M})$ is the domain of the output, and ϵ is commonly referred to as the privacy budget.

3.4 Design Goals

We seek to design a privacy-preserving quantile summary aggregation scheme with the following goals in mind.

- *Local Differential Privacy.* The scheme should satisfy ϵ-LDP for individual sensor nodes.
- *High accuracy.* The quantile summary obtained by the base station should be able to answer value-to-rank queries with high accuracy.
- *Communication efficiency.* The scheme should incur low communication overhead.

4 PrivQSA: Quantile Summary Aggregation with LDP

In this section, we first give an overview of PrivQSA and then detail its design.

4.1 Overview

We design PrivQSA by exploring the inherent connection between a quantile summary and a histogram. Specifically, a quantile summary can be viewed as an equi-depth histogram in which every bucket has the same number of values, and all the buckets in a standard histogram have the same width but different number of values. In addition, a quantile summary can be converted into a standard histogram under moderate assumptions, and vice versa. Based on this observation, we first let every sensor node randomly perturb its set of readings to generate a set of perturbed readings to ensure ϵ-LDP. All the sensor nodes then participate in quantile summary aggregation to allow the base station to receive a quantile summary, i.e., an equi-depth histogram, of the perturbed readings. The base station can then convert the quantile summary of the perturbed readings into a an equi-width histogram whereby to estimate an equi-width histogram of the original readings. Finally, the base station can convert the estimated equi-width histogram of the original readings into a quantile summary of the original readings whereby to answer any value-to-rank and percentile queries. In what follows, we detail the design of PrivQSA.

4.2 Detailed Design

PrivQSA consists of the following six steps.

Perturbation at Individual Sensor Nodes. Each sensor node i randomly perturbs its set of readings $V_i = \{v_{i,1}, \ldots, v_{i,m}\}$ into a set of n perturbed readings S'_i via the exponential mechanism to ensure ϵ-LDP. The exponential mechanism [13] is a classical technique to provide differential privacy via outcome randomization. The key idea is to associate every pair of input x and candidate outcome o with a real-value quality score $q(x, o)$, where a higher quality score indicates higher utility of the outcome. Given the output space O, a score function $q(\cdot, \cdot)$, and the privacy budget ϵ, the exponential mechanism randomly selects an outcome $o \in O$ with probability proportional to $\exp(\epsilon q(x, o))$.

Here the input is a set $V_i \subseteq R$ of m readings, and the outcome $\tilde{V}_i$ of the exponential mechanism is also a subset of R with m elements. For every pair of possible input set V_i and output set $\tilde{V}_i$, we define the quality score function as

$$q(V_i, \tilde{V}_i) = \frac{|V_i \bigcap \tilde{V}_i|}{m} , \tag{1}$$

which is the ratio of their common elements. For example, if $V_i = \tilde{V}_i$, then the quality score is one. Under the quality score function $q(\cdot, \cdot)$, each node i then randomly chooses an m-element set $\tilde{V}_i \subset R$ with probability proportional to $\exp(\frac{\epsilon |V_i \bigcap \tilde{V}_i|}{m})$.

Data Augmentation. Since all existing quantile summary aggregation schemes including Huang *et al.*'s protocol [10] requires every data value to be distinct, every sensor node augments its perturbed readings by its node ID. Let $\tilde{V}_i = \{\tilde{v}_{i,1}, \ldots, \tilde{v}_{i,m}\}$ be node i's set of perturbed readings. Each node i augments each perturbed reading $\tilde{v}_{i,j}$ as $\hat{v}_{i,j} = \tilde{v}_{i,j} || i$ for all $1 \leq j \leq m$, where $||$ denotes concatenation, and node ID i is encoded by $\gamma = \lceil \log_2 n \rceil$ bits. Doing so can ensure that every reading generated in the network is unique. We hereafter denote by $\hat{V}_i$ the set of perturbed and augmented readings of node i for all $1 \leq i \leq n$.

Quantile Summary Aggregation. Every node first generates a local quantile summary of $\hat{V}_i$. Specifically, each node i randomly samples each perturbed reading $\hat{v}_{i,j} \in \hat{V}_i$ independently with probability h to obtain a subset of perturbed readings $S_i \subseteq \hat{V}_i$, where $h \in (0, 1]$ is a system parameter. Node i's local quantile summary of $\hat{V}_i$ is then

$$Q_i = \{(\hat{v}_{i,j}, j) | \hat{v}_{i,j} \in S_i\} , \tag{2}$$

where j is the perturbed reading $\hat{v}_{i,j}$'s local rank within $\hat{V}_i$. The set $\hat{V}_i$ is commonly referred to as the *ground set* of local quantile summary Q_i

All the sensor nodes then participate in quantile summary aggregation according to Huang *et al.*'s protocol [10]. During the aggregation process, intermediate nodes may merge multiple local quantile summaries into one via random resampling to reduce the maximum per node communication cost. We refer readers to [10] for details of the merging process.

The base station performs value-to-rank query on every possible perturbed value to learn the distribution of the perturbed readings. Assume that the base station receive n' local quantile summaries $Q'_1, \ldots, Q'_{n'}$ at the end of the aggregation process, where each Q'_i corresponds to a ground set $\hat{V}'_i$ and $n' \leq n$ due to possible merging by intermediate sensor nodes. It is easy to see that $\bigcup_{i=1}^{n} \hat{V}_i = \bigcup_{i=1}^{n'} \hat{V}'_i$. For every possible augmented and perturbed value $\hat{v} = v||i$ where $v \in R$ and $i \in \{1, \ldots, n\}$, the base station estimates its global rank within $\bigcup_{i=1}^{n} \hat{V}_i$ as

$$\hat{r}(\hat{v}) = \sum_{i=1}^{n'} \hat{r}(\hat{v}, \hat{V}'_i) \ , \tag{3}$$

where

$$\hat{r}(\hat{v}, \hat{V}'_i) = \begin{cases} r(\mathsf{pred}(\hat{v}|Q'_i), \hat{V}'_i) + 1/h, & \text{if } \mathsf{pred}(\hat{v}|Q'_i) \text{ exists;} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathsf{pred}(\hat{v}|Q_i)$ denotes the predecessor of value $\hat{v}$ in Q'_i. It has been shown in [10] that $\hat{r}(\hat{v})$ is an unbiased estimator of $r(\hat{v}, \bigcup_{i=1}^{n'} \hat{V}'_i) = r(\hat{v}, \bigcup_{i=1}^{n} \hat{V}_i)$.

Next, the base station computes the global rank of each possible value $v \in R$ by removing the augmented node ID from the perturbed readings. In particular, for each pair of perturbed value and estimated rank $(\hat{v}, \hat{r}(\hat{v}))$, the base station updates its value as

$$\tilde{v} = \hat{v} \mod 2^{\gamma} \tag{5}$$

and records the pair $\langle \tilde{v}, \hat{r}(\hat{v}) \rangle$.

After removing the augmented IDs from all perturbed readings, the base station obtains one or more estimated global ranks for each possible perturbed value $\tilde{v} \in R$. Without loss of generality, let $r^-(\tilde{v})$ and $r^+(\tilde{v})$ be the lowest and highest estimated global ranks, respectively, of value $\tilde{v}$ for all $\tilde{v} \in R$. If value $\tilde{v}$ has only a unique estimated global rank, then $r^-(\tilde{v}) = r^+(\tilde{v})$.

Estimating Histogram of Perturbed Readings. The base station then constructs a histogram of the perturbed readings from the received quantile summaries by estimating the frequency of each perturbed value $\tilde{v} \in R$. We formulate the histogram construction as an optimization problem. In particular, let $f_{\tilde{v}}$ be the frequency of perturbed value $\tilde{v}$ for all $\tilde{v} \in R$. It follows that value 1 is ranked from the 1st to the f_1th, and value $\tilde{v}$ is ranked from $(\sum_{i=1}^{\tilde{v}-1} f_i + 1)$th to $(\sum_{i=1}^{\tilde{v}} f_i)$th for all $1 \leq \tilde{v} \leq d$. We can then formulate the estimation of $f_1, \ldots, f_d$

as the following optimization problem

$$\begin{aligned}
\min_{(f_1,\ldots,f_d)\in\mathbb{N}^d} \quad & \sum_{\tilde{v}\in R}(\sum_{i=1}^{\tilde{v}-1} f_i + 1 - r^-(\tilde{v}))^2 + (\sum_{i=1}^{\tilde{v}} f_i - r^+(\tilde{v}))^2, \\
\text{such that} \quad & \sum_{\tilde{v}=1}^{d} f_{\tilde{v}} = nm, \\
& \sum_{i=1}^{\tilde{v}-1} f_i + 1 \leq r^-(\tilde{v}), \forall \tilde{v}\in R, \\
& \sum_{i=1}^{\tilde{v}} f_i \geq r^+(\tilde{v}), \forall \tilde{v}\in R,
\end{aligned} \tag{6}$$

where we seek to minimize the total square errors between the two boundaries and the corresponding lowest and highest estimated global ranks. In the above optimization problem, the first constraint indicates that the sum of all the values' frequencies should be nm, the second and third constraints guarantee that the lowest and highest estimated global ranks of a value $\tilde{v}$, i.e., $r^-(\tilde{v})$ and $r^+(\tilde{v})$, should fall in the range $[\sum_{i=1}^{\tilde{v}-1} f_i + 1, \sum_{i=1}^{\tilde{v}} f_i]$.

The solution of the above optimization problem is given by

$$\begin{cases}
f_1 = \frac{r^+(1)+r^-(2)-1}{2}, \\
f_i = \frac{r^+(i)+r^-(i+1)-r^+(i-1)-r^-(i)}{2}, \quad \forall 2 \leq i \leq d-1, \\
f_d = nm - \frac{r^+(d-1)+r^-(d)-1}{2}.
\end{cases} \tag{7}$$

Estimating Histogram of Original Readings. Given the estimated histogram of perturbed readings $f_1, \ldots, f_d$ obtained above, the base station further estimates the histogram of original readings based on the perturbation mechanism used by individual sensor nodes.

Denote by g_v be the frequency of value v among the original readings $\bigcup_{i=1}^{n} V_i$ for all $v \in R$. Consider any value $v \in R$ and sensor node i's original reading set V_i. Under the random perturbation mechanism, if $v \in V_i$, the probability that v shows up in the perturbed set $\tilde{V}_i$ is given by

$$\begin{aligned}
Pr[v\in\tilde{V}_i|v\in V_i] &= Pr[v\in V_i\bigcap\tilde{V}_i|v\in V_i] \\
&= \sum_{k=1}^{m} Pr[v\in V_i\bigcap\tilde{V}_i|v\in V_i, |V_i\bigcap\tilde{V}_i| = k]\cdot Pr[|V_i\bigcap\tilde{V}_i| = k] \,.
\end{aligned} \tag{8}$$

Moreover, we have $Pr[v\in V_i\bigcap\tilde{V}_i|v\in V_i, |V_i\bigcap\tilde{V}_i| = k] = \frac{k}{m}$ and $Pr[|V_i\bigcap\tilde{V}_i| = k] = \frac{c_k\exp(\frac{\epsilon k}{m})}{\sum_{j=0}^{m} c_j\cdot\exp(\frac{\epsilon j}{m})}$, where $c_k = \binom{m}{k}\cdot\binom{d-m}{m-k}$ is the number of m-element subsets of R that shared k common elements with V_i for all $k = 0, \dots, m$. It follows that

$$Pr[v \in \tilde{V}_i | v \in V_i] = \sum_{k=1}^{m} \frac{k}{m} \cdot \frac{c_k \exp(\frac{\epsilon k}{m})}{\sum_{j=0}^{m} c_j \cdot \exp(\frac{\epsilon j}{m})} \ . \tag{9}$$

Now let us analyze the probability that v shows up in the perturbed set $\tilde{V}_i$ given that $v \notin V_i$. Following the similar analysis, we can derive

$$\begin{aligned} Pr[v \in \tilde{V}_i | v \notin V_i] &= Pr[v \in \tilde{V}_i \setminus V_i | v \notin V_i] \\ &= \sum_{k=0}^{m-1} Pr[v \in \tilde{V}_i \setminus V_i | v \notin V_i, |V_i \bigcap \tilde{V}_i| = k] \cdot Pr[|V_i \bigcap \tilde{V}_i| = k] \\ &= \sum_{k=0}^{m-1} \frac{m-k}{d-m} \cdot \frac{c_k \exp(\frac{\epsilon k}{m})}{\sum_{j=0}^{m} c_j \cdot \exp(\frac{\epsilon j}{m})} \ . \end{aligned} \tag{10}$$

Assume that value v appears in g_v sensor nodes' original reading sets. It follows that $n - g_v$ sets do not contain value v. The expected number of perturbed sets that include value v can be estimated as

$$E[f_v] = g_v \cdot Pr[v \in \tilde{V}_i | v \in V_i] + (n - g_v) \cdot Pr[v \in \tilde{V}_i | v \notin V_i] \ . \tag{11}$$

We therefore estimate the number of copies of v in $\bigcup_{i=1}^{n} V_i$ as

$$\hat{g}_v = \frac{f_v - n \cdot Pr[v \in \tilde{V}_i | v \notin V_i]}{Pr[v \in \tilde{V}_i | v \in V_i] - Pr[v \in \tilde{V}_i | v \notin V_i]} \ , \tag{12}$$

where $Pr[v \in \tilde{V}_i | v \in V_i]$ and $Pr[v \in \tilde{V}_i | v \notin V_i]$ are given in Eqs. (9) and (10), respectively.

Final Quantile Summary Construction. Given the estimated histogram of the original readings, the base station then constructs a final quantile summary of $\bigcup_{i=1}^{n} V_i$, which is equivalent to estimating the rank for every value $v \in R$ and answering ϕ-quantile query for all $0 < \phi < 1$.

To answer a value-to-rank query over value $v \in R$, the base station can simply return the median rank of value v as

$$r(v) = \lfloor \frac{r^-(v) + r^+(v)}{2} \rfloor \ , \tag{13}$$

where $r^-(v) = \sum_{i=0}^{v-1} \hat{g}_i + 1$ and $r^+(v) = \sum_{i=0}^{v} \hat{g}_i$. Moreover, to answer a ϕ-quantile query where $0 < \phi < 1$, the base station returns value v such that $r^-(v) \leq nm\phi \leq r^+(v)$.

5 Performance Evaluation

5.1 Theoretical Analysis

We first have the following theorem regarding the privacy guarantee of PrivQSA.

Theorem 1. *PrivQSA satisfies ϵ-LDP.*

Proof. Let V_i and V_j be two arbitrary sets of readings such that $|V_i| = |V_j| = m$. Let $\mathcal{M}$ denote the randomized perturbation mechanism used by each individual sensor node. Consider any possible output $\tilde{V}$ of $\mathcal{M}$. Since $0 \leq |\tilde{V} \cap V_i| \leq m$ and $0 \leq |\tilde{V} \cap V_j| \leq m$ for any V_i and V_j, we have

$$\frac{\mathsf{Pr}[\mathcal{M}(V_i) = \tilde{V}]}{\mathsf{Pr}[\mathcal{M}(V_j) = \tilde{V}]} = \frac{\exp(\epsilon \cdot \frac{|\tilde{V} \cap V_i|}{m})}{\exp(\epsilon \cdot \frac{|\tilde{V} \cap V_j|}{m})} \leq \frac{\exp(\epsilon \cdot \frac{m}{m})}{\exp(\epsilon \cdot \frac{0}{m})} \leq \exp(\epsilon). \tag{14}$$

The theorem is thus proved. □

5.2 Simulation Settings

We simulate a wireless sensor network consisting of $n = 1022$ sensor nodes. We assume that each node has $m = 10$ readings and every original reading is in the range $R = \{1, \ldots, 100\}$. Table 1 summarizes our default settings unless mentioned otherwise.

Table 1. Default simulation settings

Para.	Val.	Description.
ϵ	50	The privacy budget
h	0.5	The sampling probability
n	1022	The number of sensor nodes
m	10	The size of user value set
d	100	The maximum number in the range of user values

Since there is no prior solution for private quantile summary aggregation, we compare PrivQSA with the following two baseline schemes.

- *Baseline 1*: Every node randomly perturbs its reading set as in PrivQSA and independently samples its perturbed readings with probability h. It then submits only the sampled perturbed readings without rank information to the base station. The base station estimates the original distribution and the final quantile summary using the same method as in the last two steps of PrivQSA. Baseline 1 satisfies ϵ-LDP and does not involve any quantile summary aggregation.
- *Baseline 2*: Every node participates in quantile summary aggregation according to Huang *et al.* [10] without any privacy guarantee.

We use two metrics to measure the accuracy of the final quantile summary at the base station. Let $r(v)$ and $\hat{r}(v)$ be the true rank and estimated rank of a value v, respectively, for all $v \in \{1, \dots, d\}$. Also let $r_{\max} = nm$ be the maximum global rank in the network which is the total number of readings in the network. The normalized average rank error (ARE) is defined as

$$\mathsf{ARE} = \frac{\sum_{v=1}^{d} |\hat{r}(v) - r(v)|}{r_{\max} d} , \tag{15}$$

and the maximum rank error (MRE) is defined as

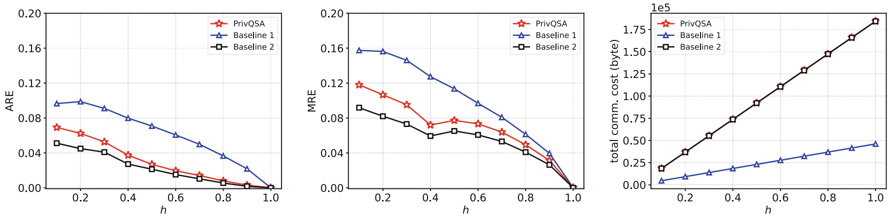
$$\mathsf{MRE} = \frac{\max_{v=\{1,\dots,d\}}(|\hat{r}(v) - r(v)|)}{r_{\max}} . \tag{16}$$

In addition, we also use total communication cost to compare PrivQSA and the two baseline solutions.

5.3 Simulation Results

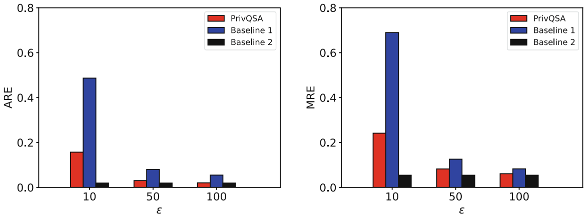
Impact of Sampling Probability *h*. Figs. 1a to 1c compare the ARE, MRE, and total communication cost of PrivQSA and the two baseline solutions, respectively, with sampling probability varying from 0.1 to 1.0. We can see from Fig. 1a and 1b that both the ARE and MRE decrease as the sampling probability h increases under all three schemes. This is expected as the more readings we sample, the more accurate the value-to-rank query results, and vice versa. Moreover, we can see that the ARE and MRE of Baseline 2 is the lowest among the three because it does not involve any random perturbation. PrivQSA comes in the second place with a small difference compared to Baseline 2 which is the cost of providing ϵ-LDP. Finally, Baseline 1 incurs the largest rank errors because its does not make use of any rank information in estimating the original distribution. On the other hand, Fig. 1c shows that the total communication cost increases as the sampling probability increases under all three schemes, which is anticipated. Moreover, we can see that PrivQSA and Baseline 2 have the same communication cost, which is larger than Baseline 1's communication cost. This is because under both PrivQSA and Baseline 2 every sensor node needs to send the rank information besides the sampled values whereas under Baseline 1 only sampled readings need to be sent.

Impact of Privacy Budget ϵ. Figures 2a and 2b compare the ARE and MRE under PrivQSA and Baseline 2 with the privacy budget ϵ varying from 10 to 100, where those under Baseline 2 are plotted for reference only. We can see from both figures that both ARE and MRE decrease as the privacy budget ϵ increases both under PrivQSA and Baseline 1. This is because the larger the ϵ, the closer the perturbed reading set to the original reading set, the more accurate the estimated original distribution, the more accurate the value-to-rank query results, and vice versa. In addition, PrivQSA achieves significantly lower ARE and MRE than Baseline 1 due to the rank information included in the quantile summary aggregation.



(a) Average rank error (b) Max. rank error (c) Total comm. cost

Fig. 1. Sampling probability h varying from 0.1 to 1.0.

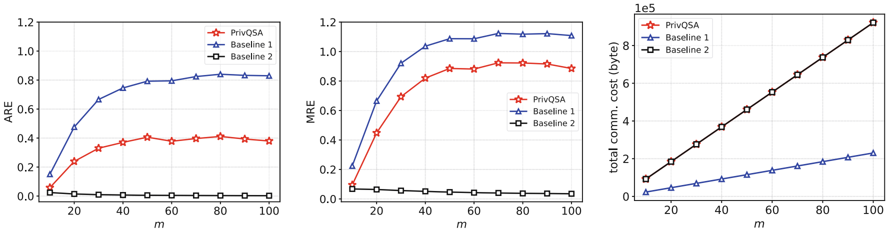


(a) Average rank error (b) Max. rank error

Fig. 2. Privacy budget ϵ varying from 10 to 100.

Impact of $\boldsymbol{m}$**.** Figures 3a to 3c compare the ARE, MRE, and total communication cost of the three schemes with m varying from 10 to 100. As we can see from Figs. 3a and 3b, both ARE and MRE decrease as the number of reading per node increases under Baseline 2. The reason is that the more readings each node has, the more sampled readings, the more accurate the value-to-rank query results, and vice versa. In contrast, both ARE and MRE increase as the number of values per node increases under PrivQSA and Baseline 1. This is because perturbing a larger set of readings with the same privacy budget results in more noise added to the perturbed reading set and thus larger rank errors. Again, Baseline 2 has the lowest ARE and MRE, which is followed by PrivQSA and Baseline 1 for the same reasons mentioned earlier. Moreover, Fig. 3c shows that the total communication cost under all three schemes increase as m increases, which is expected. Moreover, Baseline 1 incurs the lowest communication cost, and PrivQSA and Baseline 2 incur the same communication cost.

Impact of $\boldsymbol{d}$**.** Figures 4a and 4b compare the ARE and MRE of all three schemes with the size of reading domain d varying from 100 to 500. As we can see, Baseline 2 shows a slight increase in both ARE and MRE as d increases. This is expected as the larger the domain range, the more values that need to have their ranks estimated, the higher the ARE and MRE under a fixed sampling probability. For the same reason, we can see that PrivQSA and Baseline 1 incur higher ARE and MRE which also increase faster in comparison with Basline 2

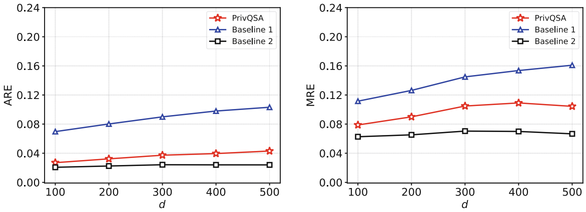


(a) Average rank error

(b) Max. rank error

(c) Total comm. cost

Fig. 3. Number of readings per node varying from 10 to 100.



(a) Average rank error

(b) Max. rank error

Fig. 4. Size of reading domain d varying from 100 to 500.

as the size of domain range increases. The reason is that the larger the value domain, the fewer common elements between the original reading set and the perturbed reading set after perturbation, and the larger the rank estimation errors, and vice versa.

6 Conclusion

In this paper, we have introduced the design of PrivQSA, the first locally differentially private quantile summary aggregation protocol for wireless sensor networks, which can guarantee ϵ-LDP for individual sensor node's readings. We have confirmed the significant advantages of PrivQSA over baseline solutions via detailed simulation studies.

Acknowledgements. The authors would like to thank the anonymous reviewers for their constructive comments and helpful advice. This work was supported in part by the US National Science Foundation under grants CNS-1651954 (CAREER), CNS-1933047, and CNS-1718078.

References

1. Ács, G., Castelluccia, C.: I have a DREAM! (DiffeRentially privatE smArt Metering). In: Filler, T., Pevný, T., Craver, S., Ker, A. (eds.) IH 2011. LNCS, vol.

6958, pp. 118–132. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24178-9_9

2. Castelluccia, C., Chan, A.C., Mykletun, E., Tsudik, G.: Efficient and provably secure aggregation of encrypted data in wireless sensor networks. ACM Trans. Sens. Netw. (TOSN) **5**(3), 1–36 (2009)
3. Castelluccia, C., Mykletun, E., Tsudik, G.: Efficient aggregation of encrypted data in wireless sensor networks. In: IEEE Mobiquitous 2005, pp. 109–117. IEEE (2005)
4. Dwork, C.: Differential privacy. p. 1–12. ICALP 2006, Springer-Verlag, Berlin (2006)
5. Greenwald, M., Khanna, S.: Space-efficient online computation of quantile summaries. In: ACM SIGMOD 2001, p. 58–66. Santa Barbara, CA (2001)
6. Greenwald, M.B., Khanna, S.: Power-conserving computation of order-statistics over sensor networks. In: ACM PODS, pp. 275–285 (2004)
7. Groat, M.M., Hey, W., Forrest, S.: Kipda: *k*-indistinguishable privacy-preserving data aggregation in wireless sensor networks. In: IEEE INFOCOM, pp. 2024–2032. IEEE (2011)
8. Haeupler, B., Mohapatra, J., Su, H.H.: Optimal gossip algorithms for exact and approximate quantile computations. In: ACM PODC, pp. 179–188 (2018)
9. He, W., Liu, X., Nguyen, H., Nahrstedt, K., Abdelzaher, T.: PDA: privacy-preserving data aggregation in wireless sensor networks. In: IEEE INFOCOM, pp. 2045–2053. IEEE (2007)
10. Huang, Z., Wang, L., Yi, K., Liu, Y.: Sampling based algorithms for quantile computation in sensor networks. In: ACM SIGMOD, pp. 745–756 (2011)
11. Li, Q., Cao, G.: Efficient and privacy-preserving data aggregation in mobile sensing. In: 2012 20th IEEE ICNP, pp. 1–10. IEEE (2012)
12. Lindell, Y., Pinkas, B.: Secure multiparty computation for privacy-preserving data mining. J. Priv. Confid. **1**(1) (2009)
13. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: IEEE FOCS (2007)
14. Naranjo, J.A., Casado, L.G., Jelasity, M.: Asynchronous privacy-preserving iterative computation on peer-to-peer networks. Computing **94**(8–10), 763–782 (2012)
15. Ozdemir, S., Xiao, Y.: Secure data aggregation in wireless sensor networks: a comprehensive overview. Comput. Netw. **53**(12), 2022–2037 (2009)
16. Rajagopalan, R., Varshney, P.K.: Data-aggregation techniques in sensor networks: a survey. IEEE Commun. Surv. Tutorials **8**(4), 48–63 (2006)
17. Sun, J., Zhang, R., Zhang, Y.: PriStream: privacy-preserving distributed stream monitoring of thresholded PERCENTILE statistics. In: IEEE INFOCOM, pp. 1–9 (2016)
18. Shi, J., Zhang, R., Liu, Y., Zhang, Y.: Prisense: privacy-preserving data aggregation in people-centric urban sensing systems. In: IEEE INFOCOM, pp. 1–9. (2010)
19. Shrivastava, N., Buragohain, C., Agrawal, D., Suri, S.: Medians and beyond: new aggregation techniques for sensor networks. In: SenSys, pp. 239–249 (2004)
20. Westhoff, D., Girao, J., Acharya, M.: Concealed data aggregation for reverse multicast traffic in sensor networks: Encryption, key distribution, and routing adaptation. IEEE Trans. Mob. Comput. **5**(10), 1417–1431 (2006)
21. Xue, M., Papadimitriou, P., Raïssi, C., Kalnis, P., Pung, H.K.: Distributed privacy preserving data collection. In: Yu, J.X., Kim, M.H., Unland, R. (eds.) DASFAA 2011. LNCS, vol. 6587, pp. 93–107. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20149-3_9
22. Zhang, K., Han, Q., Cai, Z., Yin, G.: Rippas: a ring-based privacy-preserving aggregation scheme in wireless sensor networks. Sensors **17**(2), 300 (2017)