2021 17th International Conference on Mobility, Sensing and Networking (MSN) | 978-1-6654-0668-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/MSN53354.2021.00074

SecQSA: Secure Sampling-Based Quantile Summary Aggregation in Wireless Sensor Networks

Aishah Aseeri
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716
aaseeri@udel.edu

Rui Zhang
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716
ruizhang@udel.edu

***Abstract*—Wireless sensor networks are widely expected to play a key role in the emerging Internet of Things (IoT)-based smart cities in which a large number of resource-constrained sensor nodes collect data about our physical environment to assist intelligent decision making. Since blindly forwarding all the sensed data to the base station may quickly deplete sensor nodes' limited energy, secure data aggregation has been considered as a key functionality in wireless sensor networks that allow the base station to acquire important statistics about the sensed data. While many secure data aggregation schemes have been proposed in the literature, most of them target simple statistics such as Sum, Count, Min/Max, and Median. 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 provide a more accurate characterization of the data distribution. How to realize secure quantile summary aggregation in wireless sensor networks remains an open challenge. In this paper, we fill this void by first evaluating the impact of a range of attacks on quantile summary aggregation using simulation and then introduce a novel secure quantile summary aggregation protocol for wireless sensor networks. Detailed simulation studies confirm the efficacy and efficiency of the proposed protocol.**

I. Introduction

Wireless sensor networks are widely expected to play a key role in emerging Internet of Things (IoT)-based smart cities in which a large number of sensor nodes continuously sense the physical environment and generate data to assist intelligent decision making [1], [2]. Since sensor nodes are typically resource-constrained with limited computation capability, memory, and energy, blindly forwarding all the sensed data to a base station may quickly deplete sensor nodes' limited energy. Data aggregation has been widely considered as a key functionality [3] for reducing data redundancy, improving energy efficiency, and prolonging the lifetime of wireless sensor networks, in which sensed data are aggregated enroute by intermediate sensor nodes, which allow a base station to acquire important statistics about the sensed data.

Secure data aggregation is necessary to safeguard the aggregation process from malicious attacks. Resource-constrained sensor nodes in unattended environments are subject to physical capture and may be compromised by attackers. Once compromised, a sensor node may carry out a wide range of attacks under the attacker's instruction. For example, a compromised node may change the intermediate aggregation results that can significantly deviate the final aggregation result at the base station. As another example, a compromised node may drop the data from its children nodes to prevent them from reaching the base station. As a result, secure data aggregation has been investigated extensively over the past to allow the base station to acquire statistics about the sensed data in the presence of attacks [4]–[14]. Unfortunately, all existing solutions target simple statistics such as Sum, Count, Min/Max, and Median.

Quantile summary aggregation allows a base station to learn a more accurate distribution of the sensed data than simple statistics functions. Specifically, a quantile summary allows one to extract the ϕ-quantile for any $0 < \phi < 1$ of all the sensor readings in the network and thus can provide a more accurate characterization of the data distribution. Given a set of n distinct data values with a total order, the ϕ-quantile is the value x with rank $r(x) = \lfloor \phi n \rfloor$ in the set, where $r(x)$ is the number of values in the set smaller than x. 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 $(\phi + \epsilon)n$ is usually sought in the literature. While several quantile summary aggregation protocols [15]–[18] have been proposed in the past, none of them were designed to withstand potential attacks. How to realize secure quantile summary aggregation in wireless sensor networks thus remains an open challenge.

In this paper, we fill this void by introducing SecQSA, a novel secure quantile summary aggregation protocol for wireless sensor networks. SecQSA is built upon the non-secure quantile summary aggregation protocol proposed by Huang *et al.* [17], which we choose because it can guarantee a constant individual node communication cost independent of network size even for the nodes close to the base station with many decedents. We observe that the key for securing quantile summary aggregation is to ensure the integrity of sample readings of the quantile summary as well as the correctness of the operation that merges multiple local quantile summaries into one. SecQSA achieves these two goals using efficient cryptographic primitives. Our contributions in this paper can be summarized as follows.

- To the best of our knowledge, we are the first to study secure quantile summary aggregation in wireless sensor

978-1-6654-0668-0/21/$31.00 ©2021 IEEE
DOI 10.1109/MSN53354.2021.00074

networks.

- We introduce SecQSA, a novel secure quantile summary aggregation protocol based on efficient cryptographic primitives.
- We confirm the efficacy and efficiency of the proposed protocol via detailed simulation studies.

The rest of this paper is structured as follows. Section II discusses the related work. Section III introduces the network and adversary models. Section IV evaluates the impact of different attacks on quantile summary aggregation. Section V introduces the design of SecQSA. Section VI reports the simulation results. Section VII finally concludes this paper.

II. Related Work

Secure data aggregation in wireless sensor networks have been studied extensively in the past. Most of the existing solutions target simple aggregation functions such as Sum, Count, Average, and Min/Max. The resilience of different aggregation functions under a single aggregator model was analyzed in [8]. Przydatek *et al.* [6] introduced a secure aggregation scheme that can support Median, Min/Max, and Average aggregation. In [4], Chan *et al.* presented a secure hierarchical additive aggregation scheme, which was subsequently improved by Frikken *et al.* with reduced communication cost [10]. A commitment-based hop-by-hop aggregation scheme was introduced in [9] which allows the base station to verify abnormal aggregate via hypothesis testing. A secure hierarchical data aggregation scheme based on synopsis diffusion was proposed in [7], [13], which can support additive aggregation functions such as Count and Sum against falsified sub-aggregate attacks. In [11], Papadopoulos *et al.* introduced a secure aggregation scheme for exact Sum aggregation. Chen and Yu presented a scheme [19] that realizes secure approximate Sum aggregation via secure Min aggregation, which was later shown to be vulnerable to a special enumeration attack [14].

There are very limited efforts in developing secure aggregation schemes to support Median and Percentile aggregation. The techniques presented in [4], [6] can be used for verifying the correctness of an alleged ϕ-percentile via secure Count aggregation by counting the number of readings that are smaller than the alleged ϕ-percentile. Roy *et al.* [5] extended the secure Count aggregation scheme [4] to realize secure Median aggregation by recursively constructing an increasingly accurate histogram. However, these solutions require the base station to know the percentile of interest, i.e., ϕ, in advance and incurs a communication cost proportional to the number of percentile queries.

Quantile summary [20] aggregation in wireless sensor networks has been studied. In [15], a quantile digest summary structure was introduced to realize quantile aggregation. Greenwald *et al.* [16] introduced a distributed algorithm to compute an ϵ-approximate quantile summary of sensor data, which was later improved by Huang *et al.* [17] to reduce the maximum per node communication cost. More recently, several efficient gossip algorithms were introduced in [18] to compute exact and approximate quantiles in a fully distributed fashion. Unfortunately, none of the above quantile aggregation schemes have any security provisions. None of these works consider possible attacks, and they cannot be applied to our problem.

III. Network and Adversary Models

In this section, we introduce our system and adversary models.

A. Network Model

We consider a multi-hop wireless sensor network consisting of a base station and s sensor nodes. Every sensor node senses the environment and periodically generates readings at fixed frequency. We assume that every sensor node i has a set of n readings denoted by D_i and every reading is in the range $R = \{1, \ldots, v_{\max}\}$ it should be float numbers. It follows that the total number of readings in the network is sn. As in [17], we assume that all the readings in the sensor network are distinct. While this assumption may seem restrictive, it can be easily accommodated by imposing a total order among the readings by taking node ID and the time at which a reading is generated to break the tie.

The base station aims to obtain a quantile summary of all the readings generated in the network over a certain period. A quantitle summary is a subset of readings along with their (estimated) global ranks which can support *value-to-rank* queries. Specifically, for any value $v \in R$, the value-to-rank query returns an estimated global rank $\hat{r}(v)$. The ϕ-quantile of all the readings $\bigcup_{i=1}^{s} D_i$ is then the value x with rank $r(x) = \lfloor \phi sn \rfloor$ for any $0 < \phi < 1$.

We assume that the aggregation is performed over an aggregation tree, which is the directed tree rooted at the base station formed by the unique path from every sensor node to the base station. During network initialization, the base station learns the topologies of the network as well as the aggregation tree. We also assume that each sensor node i shares a secret key K_i with the base station. We also assume that any two nodes i and j can establish a shared key $K_{i,j}$ using existing techniques such as [21], [22].

B. Adversary Model

The attacker aims to mislead the base station into accepting a modified distribution of an aggregated summary without being detected in order to significantly shift any quantile query result from its original position. We assume that the base station is equipped with adequate computation and energy resources and is safeguarded from any malicious attacks. In contrast, sensor nodes are constrained in computation and communication resources which make them susceptible to compromising. Once a sensor node is compromised, all the information stored at the sensor node such as cryptographic keys is revealed to the attacker. The attacker can then instruct compromised sensor nodes to carry out a wide range of attacks.

Since the aggregated summary consists of a subset sampled readings and their ranks, we consider the following two attacks in this paper.

- A compromised node may forge its own readings, their ranks, or both.
- A compromised node may deviate from protocol operations, which includes dropping other nodes' readings, replacing other nodes' readings with its own, modifying other nodes' readings or their ranks.

IV. Attacks on Quantile Summary Aggregation

In this section, we first briefly review the sampling based quantile summary protocol proposed by Huang *et. al.* [17], which serves as the basis for SecQSA. We then evaluate the impact of a range of attacks on the Huang's protocol via simulation studies.

A. Review of Huang's Protocol [17]

Huang's protocol [17] is designed based on random sampling. Let $G_1, \ldots, G_k$ be a family of sets of data values, where $G_i \bigcap G_j = \emptyset$ for all $1 \leq i < j \leq k$. If we independently sample each value in G_i with probability q to obtain a subset $S_i \subseteq G_i$ for all $i = 1, \ldots, k$. Denote by $r(v, G_i)$ its local rank within the set G_i for each sampled value $v \in S_i$. Given any value x, we can estimate its local rank $\hat{r}(x, G_i)$ within G_i for all $1 \leq i \leq k$. Let $p(x|S_i)$ be the predecessor of value x in S_i. It has been shown that

$$\hat{r}(x, G_i) = \begin{cases} r(p(x|S_i), G_i) + 1/p, & \text{if } p(x|S_i) \text{ exists;} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

is an unbiased estimator of $r(x, G_i)$. The global rank of value x within $G = \bigcup_{i=1}^{k} G_i$ can then be estimated as

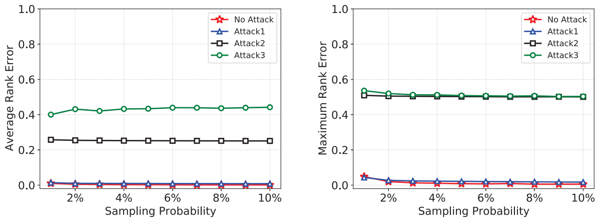
$$\hat{r}(x) = \sum_{i=1}^{k} \hat{r}(x, G_i) \ .$$

Under Huang's protocol [17], every node i first samples each reading of its own independently to generate a local quantile summary. All the nodes then participate in quantile summary aggregations in which local quantile summaries are forwarded and merged with others into one along the way before reaching the base station. A key advantage of Huang's scheme [17] over prior solutions [15], [16] is that it can guarantee an individual node communication cost of $O(1/\epsilon)$ even for those nodes close to the base station and have many decedents by carefully designed merging conditions. We refer readers to [17] for details of Huang's scheme.

B. Impact of Attacks

We now evaluate the impact of several attacks on Huang's protocol [17], which will guide the design of SecQSA.

Several possible attacks can be launched by a compromised sensor node. First, a compromised sensor node can arbitrarily forge its own readings and their local ranks, which is fundamentally difficult to detect without any special assumption. Moreover, since a quantile summary consists of a subset of sample values with their local ranks, a compromised sensor node can also modify the readings of its decedent nodes and corresponding ranks. In addition, Huang's protocol [17] requires that every reading is sampled independently during merging operations, but a compromised node may not follow by discarding all the readings from one or more of its decedent nodes. Due to symmetry, we only consider the case in which the attacker intends to inflate the estimated rank of any value and consider the following three attacks.



(a) Average rank error

(b) Maximum rank error

Fig. 1. Comparison of ARE and MRE under different attacks.

- *Attack 1*: Modify its own sampled values to the minimum and their ranks to the maximum.
- *Attack 2*: Modify children nodes' sampled values to the minimum and their ranks to the maximum.
- *Attack 3*: Modify its own sampled values to the minimum and their ranks to the maximum and drop all the children nodes' value from the quantile summary.

We use the following two metrics to evaluate the impact of the above three attacks on 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 $v \in \{1, \ldots, v_{\max}\}$. The normalized average rank error (ARE) and maximum rank error (MRE) are defined as

$$\mathsf{ARE} = \frac{\sum_{v=1}^{v_{\max}} |\hat{r}(v) - r(v)|}{v_{\max}^2} \ , \quad (2)$$

and

$$\mathsf{MRE} = \frac{\max_{v=\{1,\ldots,v_{\max}\}}(|\hat{r}(v) - r(v)|)}{v_{\max}} \ . \quad (3)$$

We simulate a wireless sensor network consisting of $s = 62$ sensor nodes which form an aggregation tree of height 6 where each sensor node has two children nodes. We assume that each node has $n = 1000$ readings. Every point in the following figures is the average of 100 runs each with a distinct random seed for the sampling process.

Figs. 1(a) and 1(b) compare the ARE and MRE under the three types of attack as well as in the absence of attack with the sampling probability varying from 0.01 to 0.1. As we can see, both ARE and MRE decreases as the sampling probability increases in the absence of attack. This is expected as the higher the sampling probability, the more readings are included in the final quantile summary received by the base station, the more accurate the value-to-rank query results, the lower ARE and MRE, and vice versa. In addition, the ARE and MRE under Attack 1 are very close to those under no attack. The reason is that a single compromised node forging its own readings and local ranks has very limited impact on the accuracy of final quantile summary. In contrast, the ARE and

MRE under Attack 2 and Attack 3 are significantly higher than those under Attack 1. In particular, we can see from Fig. 1(a) that the AREs under Attack 2 and Attack 3 are 0.24 and 0.42, respectively. Similarly, the MREs under Attack 2 and Attack 3 are both around 0.5. These results clearly demonstrate the severe impact of Attacks 2 and 3 on the final quantile summary.

V. SecQSA: Secure Quantile Summary Aggregation

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

A. Overview

We find that the key to secure quantile summary aggregation is to ensure the integrity of the readings and their ranks during merging operations. Specifically, SecQSA is designed to achieve the following goals.

1) *Integrity of sample readings*: every reading in the final quantile summary must be generated by a sensor node and has not been altered during the aggregation process.
2) *Integrity of local ranks*: as readings being aggregated into different quantile summaries through the process, their local ranks within quantile summaries must be correctly computed according to [17].
3) *Compliance of uniform sampling*: when multiple quantile summaries are merged, every reading should be sampled independently according to [17].

We do not intend to defend against a compromised node forging its own readings and their local ranks, which has very limited impact on the aggregation results as shown in Section IV-B.

SecQSA is designed to meet the above goals using efficient cryptographic primitives. Under SecQSA, sensor nodes send, receive, and merge local quantile summaries in a secure fashion. Specifically, a quantile summary Q is associated with a ground set G and represented by

$$Q = \langle ID, O, q \rangle,$$

where ID is the node that generates the quantile summary, O is a set of *sample objects*, and q is the sampling probability. Every sample object $o \in O$ corresponds to one reading drawn from the ground set G and has the form

$$o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle.$$

where v is the reading, and σ_{init} and $\sigma_{\mathsf{current}}$ carry the necessary verification information about v in the quantile summary. More specifically, the first component σ_{init} carries the initial rank of v and has the form

$$\sigma_{\mathsf{init}} = \langle ID_i, r(v, D_i), H_{K_i}(v||r(v, D_i)) \rangle,$$

where ID_i is the ID of the node that generates reading v, $r(v, D_i)$ is the initial local rank of v in node i's set D_i, K_i is the secret key node i shared with the base station, and $H_*(\cdot)$ denotes a message authentication code keyed with the subscript. The second component $\sigma_{\mathsf{current}}$ has the form

$$\sigma_{\mathsf{current}} = \langle ID_j, r(v, G) \rangle,$$

where ID_j is the ID of the node which merges value v into the current quantile summary Q, and $r(v, G)$ is the local rank of v in the current ground set G.

As a reading v moves through the aggregation process, the first component σ_{init} remains unchanged and will allow the base station to verify the integrity of the reading and compliance of random sampling of any intermediate node. In contrast, the second component $\sigma_{\mathsf{current}}$ will be updated whenever reading v is merged into a new quantile summary.

In what follows, we detail how quantitle summaries are generated by individual sensor nodes and merged through the aggregation process.

B. Initialization

To initiate a quantile summary aggregation process, the base station broadcasts a command with a random seed d using a proper broadcast authentication protocol such as μ-Telsa [23].

On receiving the command, each sensor node i first generates a local quantile summary Q_i with respect to its set of readings D_i. Let $H(\cdot)$ be a cryptographic hash function that maps any input to an integer in the range $\{0, \ldots, \lambda-1\}$. Node i samples every reading $v \in D_i$ with probability q_{init}, where q_{init} is a system parameter that determines the accuracy of the quantile summary and communication overhead. Specifically, every reading v is selected to be included in the local quantile summary Q_i if

$$H(ID_i||r(v, D_i)||d) \leq q_{\mathsf{init}}\lambda \ . \tag{4}$$

It is easy to see that each reading is sampled independently with probability q_{init}. We subsequently denote by $S_i \subseteq D_i$ the subset of readings included in Q_i.

For each selected sample reading $v \in S_i$, node i constructs a sample object as $o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$, where

$$\sigma_{\mathsf{init}} = \sigma_{\mathsf{current}} = \langle ID_i, r(v, D_i), H_{K_i}(v||r(v, D_i)) \rangle \ .$$

C. Secure Quantile Summary Aggregation

All the nodes then participate in the quantile summary aggregation based on the aggregation tree. Specifically, every leaf node i of the aggregation tree sends its local quantitle summary Q_i to its parent node, say j, as

$$Q_i = \langle ID_i, O_i, q_{\mathsf{init}}, H_{K_{i,j}}(\mathsf{info}) \rangle \ ,$$

where $O_i = \{o | v \in S_i\}$ is the set of sample objects and $\mathsf{info} = ID_i||O_i||q_{\mathsf{init}}$ is the concatenation of all the prior information.

On receiving a local quantile summary Q_i from one of its children nodes, node j first verifies its integrity by checking $H_{K_{i,j}}(\mathsf{info})$ using the shared key $K_{i,j}$. If succeed, node j checks if local quantile summary Q_i exhibits any inconsistency. Specifically, node j checks if the reading in every sample object is in the range R. Without loss of generality, suppose that $O_i = \langle o_1, \ldots, o_x \rangle$, where $o_x = \langle v_i, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$ and

$v_1 < \cdots < v_x$. Node j checks if $r(v_1, D_i) < \cdots < r(v_x, D_i)$. If so, node j considers quantile summary Q_i valid.

Node j then processes Q_i in one of the two possible ways. In the first case, node j directly forwards Q_i to its parent node, say k, by sending

$$j \to k : \langle ID_i, O_i, q, H_{K_{j,k}}(\mathsf{info}) \rangle \ ,$$

which allows node k to verify its integrity. In the second case, node j merges Q_i with one or more other local quantile summaries to produce a single quantile summary if the conditions specified in [17] are met. In what follows, we use an example to illustrate how multiple local quantile summaries are merged at an intermediate node.

Suppose that node j intends to merge l local quantile summaries $Q_1, \ldots, Q_l$ into one local quantile summary Q. Each local quantile summary

$$Q_x = \langle ID_x, O_x, q_x \rangle \ ,$$

is sampled from a ground set G_x with sampling probability q_x independently for all $x = 1, \ldots, l$. The resulting quantile summary $Q_j = \langle ID_j, O_j, q \rangle$ corresponds to the ground set $G = \bigcup_{x=1}^{l} G_x$ where every reading in G is sampled independently with probability q.

The merging operation involves four steps. First, node j resamples every reading in $Q_1, \ldots, Q_l$ to obtain the set of readings to be included in Q. Specifically, for each quantile summary Q_x, $1 \leq x \leq l$, node j samples every sample unit $o \in O_x$ independently with probability q/q_x. In particular, each sample object $o \in O_x$ is selected if

$$H(ID_j || r(v, G_x) || d) \leq \frac{q\lambda}{q_x} \ . \tag{5}$$

It follows that each reading v in the ground set G_x is selected to be in Q with probability

$$\begin{aligned} \mathsf{Pr}(v \in Q) &= \mathsf{Pr}(o \in Q | v \in Q_x) \mathsf{Pr}(v \in Q_x) \\ &= \frac{q}{q_x} \cdot q_x \\ &= q \ . \end{aligned}$$

Second, node j computes the local rank of every reading in the quantile summary Q. Let $O'_x \subseteq O_x$ be the subset of sample objects in Q_x selected to be included in Q for all $1 \leq x \leq l$. Consider a sample unit $o \in O'_j$ as an example where $o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$ and $\sigma_{\mathsf{current}} = \langle ID_z, r(v, G_x), H_{K_z}(v || r(v, G_x)) \rangle$. It follows that v is ranked $r(v, G_x)$ within the ground set G_x. Its local rank within the new ground set $G = \bigcup_{y=1}^{l} G_x$ can then be estimated as

$$r(v, G) = r(v, G_x) + \sum_{y=1, y \neq x}^{l} r(v, G_y) \ ,$$

where

$$r(v, G_y) = \begin{cases} r(p(v|O_y), G_y) + 1/q_y, & \text{if } p(v|O_y) \text{ exists;} \\ 0 & \text{otherwise,} \end{cases}$$

and $p(v|O_y)$ is the predecessor of value v in O_y. It has been shown that $r(v, G)$ is an unbiased estimator of v's local rank within G [17].

Third, node j updates each sample object in Q. Specifically, for each $o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$ selected, node j updates $\sigma_{\mathsf{current}}$ to

$$\sigma_{\mathsf{current}} = \langle ID_j, r(v, G) \rangle \ .$$

Next, node j, its children nodes, and its parent node k execute a protocol whereby node j's children nodes verify and endorse the new local rank of each sample object in G. Among the l local quantile summaries $Q_1, \ldots, Q_l$, there is at most one local quantile summary generated by node j itself. Without loss of generality, suppose that local quantile summary Q_j is generated by node j itself and that each quantile summary Q_y is received from children node y for all $y = 1, \ldots, l$ and $y \neq j$.

Node j first broadcasts the quantile summary as

$$j \to * : \langle Q, H_{K_{j,k}}(Q) \rangle \ ,$$

where $Q = \langle ID_j, O, q \rangle$ and O is the set of sample objects. This message will be received by both node j's parent node k and all the children nodes as they are all in node j's transmission range. On receiving the message, node k verifies its integrity using the shared key $K_{j,k}$.

Node j then seeks its children nodes' endorsement for the new quantile summary Q. Since every children node y knows Q_y it sent earlier and also overheard the quantile summary Q, it knows the subset of sample object $O'_y \subseteq O_y$ being included in Q. Each node y first verifies whether node j faithfully perform random sampling for O_y according to Eq. (5). Moreover, node y also checks whether node j correctly computes the new local rank $r(v, G)$ of each sample object $o \in O$. Specifically, for each sample object $o \in O$ where $o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$, node j broadcasts the following message

$$j \to * : v, r(v, G_1), q_1, r(v, G_2), q_2, \ldots, r(v, G_l), q_l \ ,$$

Without loss of generality, consider sample object $o \in O'_y \subseteq O_y$ and Q_y was sent by child node y. Node y first verifies whether

$$r(v, G) = r(v, G_y) + \sum_{z=1, z \neq y}^{l} r(v, G_z) \ .$$

If so, node y sends its endorsement to node j as

$$x \to j : H_{K_{y,k}}(Q) \ ,$$

where $K_{y,k}$ is the shared key between node y and j's parent node k. Similarly, every other children node z ($z = 1, \ldots, l$, $z \neq y$, and $z \neq j$) which sent Q_z finds $p(v|O_z)$, i.e., the predecessor of v in O_z, and verifies whether

$$r(v, G_z) = r(p(v|O_z), G_z) + 1/q_z \ .$$

If so, node z sends its endorsement to node j as

$$y \to j : H_{K_{z,k}}(Q) \ ,$$

On receiving the endorsement from each of its children, node j sends an aggregated endorsement of Q to its parent node k

$$j \to k : \bigoplus_{y=1, y \neq j}^{l} H_{K_{y,k}}(Q) \ .$$

Since the parent node j has previously verified the integrity of Q using $H_{K_{j,k}}(Q)$, it further verifies the aggregated endorsement $\bigoplus_{y=1, y \neq j}^{l} H_{K_{y,k}}(Q)$ using the keys shared with each children node y ($y = 1, \ldots, l$). If all the verifications succeed, node k accepts Q as a valid quantile summary.

D. Final Verification at the Base Station

At the end of the aggregation process, the base station receives one or multiple quantile summaries from its children nodes. For every quantile summary it receives, the base station verifies the quantitile summary in the following steps.

First, for each sample object $o = \langle v, \sigma_{\mathsf{init}}, \sigma_{\mathsf{current}} \rangle$ where $\sigma_{\mathsf{init}} = \langle ID_i, r(v, D_i), H_{K_i}(v||r(v, D_i)) \rangle$, the base station first verifies v's integrity by recomputing $H_{K_i}(v||r(v, D_i))$ using the shared key K_i.

Second, the base station verifies if every node that performed merging operations have faithfully followed random sampling. Since the base station knows the number of readings each node has and the aggregation tree structure and the random sampling performed at each node is based on each reading's initial rank, the ID of the node that performs sampling, and the seed d, the base station can emulate the entire aggregation process to predict (1) the subset of readings sampled in each initial local quantile summary, (2) the number of quantile summaries received at each intermediate node and their corresponding sizes, (3) which nodes should have performed merging operations, and (4) the subset of readings that should have been selected in each merged quantile summary. Specifically, for each node i and every possible local rank $1, \ldots, n$, the base station verifies if (1) for every initial rank that is supposed to survive the entire aggregation process, the corresponding reading is indeed included in the final quantile summary, and (2) if there is any reading in the final quantile summary received should have been dropped by any intermediate node.

VI. Simulation Results

In this section, we evaluate the performance of SecQSA via simulation.

A. Simulation Setting

We again consider a wireless sensor network consisting of $s = 62$ sensor nodes which form an aggregation tree of height 6 where each sensor node has two children nodes. Every point in the following graphs is the average of 100 runs, each with a distinct random seed for the sampling process. We adopt the SHA-256 for the message authentication code which results in a 32 bytes code. Also, we assume that each reading is of 16 bits, and each local rank is of 32 bits.

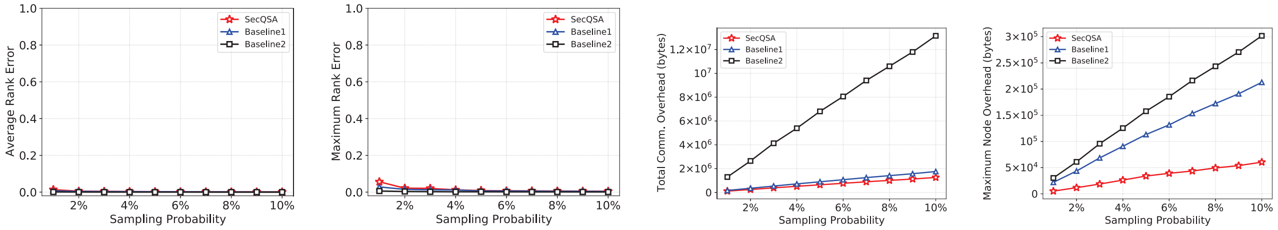
Since there is no prior solution for secure quantile summary aggregation, we compare the proposed protocol with two baseline schemes.

- *Baseline 1*: every node independently samples its readings with probability q and then submits the sampled readings along with their associated ranks and a MAC to the base station. The base station verifies the integrity of each reading and answers value-to-rank queries according to Eq. (1).
- *Baseline 2*: every node independently samples its readings with probability q and then submits the sampled readings along with a MAC to the base station with no ranking information. On receiving all the sample readings, the base station broadcasts all the readings to all the sensor nodes. Finally, all the nodes participate in multiple parallel secure SUM aggregations according to [11] to allow the base station to obtain the global rank of each reading, whereby to answer value-to-rank queries according to Eq. (1).

Besides the ARE and MRE, we also use *total communication overhead* and *maximum per node communication overhead* to evaluate the performance of proposed scheme and the two baseline schemes.

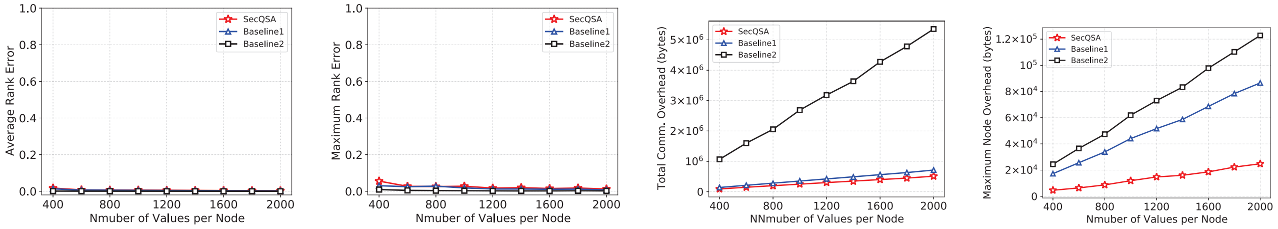
B. Simulation Results

Figs. 2(a) to 2(d) compare the ARE, MRE, total communication overhead, and maximum per node communication overhead of SecQSA and the two baseline solutions, respectively, with sampling probability varying from 0.01 to 0.1. We can see from Fig. 2(a) and 2(b) that both ARE and MRE decrease as the sampling probability q increases under all three schemes. This is expected because the higher the sampling probability, the more readings we sample, the more accurate the rank estimation, and vice versa. Moreover, we can see from Fig. 2(a) that the AREs of all three schemes are close to zero if the sampling probability exceeds 0.02. Similar trend can be observed for the MRE in Fig. 2(b) for the same reason. We can see that the accuracy of SecQSA is comparable to the two baseline schemes. On the other hand, Fig. 2(c) shows the total communication overhead under SecQSA and the other two baseline schemes. Generally speaking, the total communication overhead increases as the sampling probability increases, which is expected. Moreover, we can see that Baseline 2 has the largest communication overhead among the three because multiple parallel secure SUM aggregations need to be performed to obtain the global rank for every reading in the quantile summary. In contrast, neither Baseline 1 nor SecQSA involve such operations. Moreover, SecQSA incurs less total communication overhead than Baseline 1 because of the merging operations in SecQSA. Finally, Fig. 2(d) plots the maximum per node communication overhead for each of the three schemes. We can see that the maximum per node overhead increases under all three schemes as the sampling probability increases, which is anticipated. Most importantly, SecQSA beats both Baseline 1 and Baseline 2 with significant margins because of its merging operations.



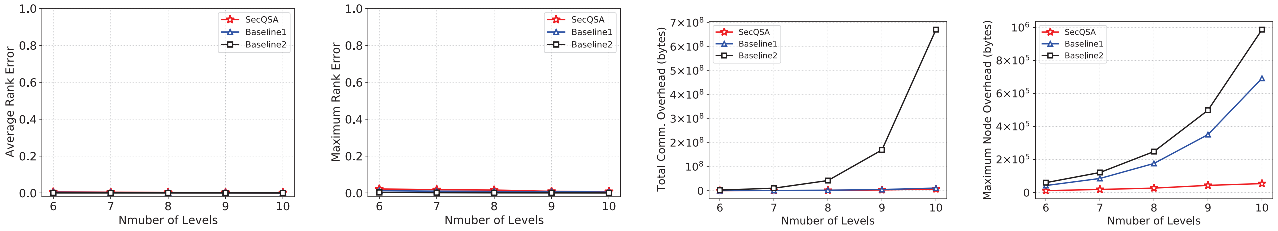
(a) Average rank error (b) Maximum rank error (c) Total comm. overhead (d) Maximum per node overhead

Fig. 2. Comparison of SecQSA and two baselines with sampling probability varying from 0.01 to 0.1.



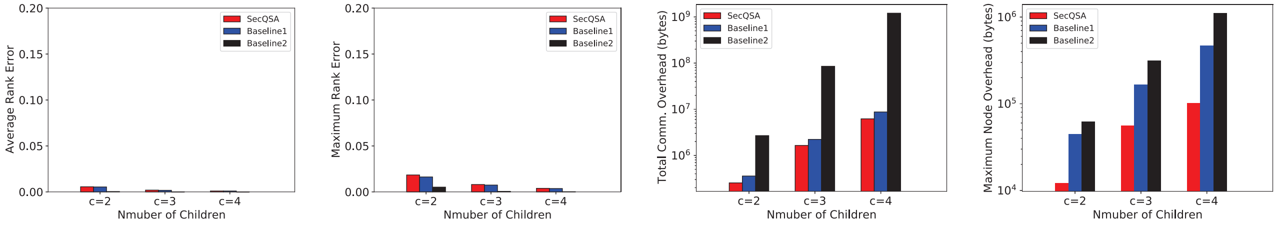
(a) Average rank error (b) Maximum rank error (c) Total comm. overhead (d) Maximum per node overhead

Fig. 3. Comparison of SecQSA and two baselines with the number of values per node varying from 400 to 2000.



(a) Average rank error (b) Maximum rank error (c) Total comm. overhead (d) Maximum per node overhead

Fig. 4. Comparison of SecQSA and the baselines with the height of the aggregation tree varying from 6 to 10.



(a) Average rank error (b) Maximum rank error (c) Total comm. overhead (d) Maximum per node overhead

Fig. 5. Comparison of SecQSA and the baselines with the number of children per node varying from 2 to 4.

Figs. 3(a) to 3(d) compare the ARE, MRE, total communication overhead, and maximum per node communication overhead under SecQSA, Baseline 1 and Baseline 2 with the number of readings per node varying from 400 to 2000. We can see from Figs. 3(a) and 3(b) that both ARE and MRE decrease as the number of readings per node increases under all three schemes. The reason is that the more readings each node has, the more readings are included in the final quantile summary, and the higher accuracy of the final quantile summary. Also, we can see from Fig. 3(a) that the ARE is almost the same for SecQSA and the two baselines for the similar reason. Similar trend can be observed for MRE in Fig. 3(b), which is of no surprise. We can also see from Figs. 3(c) and 3(d) that both the total communication overhead and maximum per node overhead produced by the three schemes increase as the number of reading per node increases. In addition, Fig. 3(c) shows that Baseline 2 incurs the highest total communication overhead among the three and SecQSA incurs the lowest total communication overhead due to its merging operations. In contrast, Fig. 3(d) shows that SecQSA incurs the lowest maximum per node communication overhead among the three and outperforms the other two by large margins.

Figs. 4(a) to 4(d) compare the ARE, MRE, total communication overhead, and maximum per node communication overhead under SecQSA, Baseline 1, and Baseline 2 with the hight of the aggregation tree varying from 6 to 10. As we can see from Figs. 4(a) and 4(b), both ARE and MRE decrease as the hight of the aggregation tree increases under all three schemes. This is expected as the higher the aggregation tree, the more sensor nodes, the more the sampled readings, the more accurate of the final quantile summary, and vice versa. We can also see from Fig. 4(a) that the ARE under SecQSA is similar to that of the other two baselines. Moreover, the MREs produced by the three schemes in Fig. 4(b) are pretty close as expected. Moreover, we can see from Figs. 4(c) and 4(d) that the total communication overhead and maximum per node overhead produced by the three schemes both increase as the height of the aggregation tree increases. The reason is that, the more levels of the aggregation tree, the more sampled readings to be collected, and the higher communication overhead. Once again, we can see that SecQSA incurs the lowest communication overhead among the three because of its merging operations.

Figs. 5(a) to 5(d) show the ARE, MRE, total communication overhead, and maximum per node communication overhead produced by SecQSA, Baseline 1 and Baseline 2 with the number of children per node varying from 2 to 4. As we can see, Figs. 5(a) and 5(b) show that both ARE and MRE decrease as the number of children increases in the network tree for all three schemes. This is anticipated as the more children nodes each non-leaf node has, the more sensor nodes in the network, the more readings collected in the final quantile summary, and the more accurate it is. Moreover, we can see from Fig. 5(a) that the ARE for SecQSA and each of the two baselines are almost matching and close to zero as we discussed before. Fig. 5(b) plots the MREs produced by the three schemes which appears to be also quite matching as expected. In addition, Figs. 5(c) and 5(d) show that the total communication overhead and maximum per node overhead produced by the three schemes both increase as the number of children per node increases. Finally, Figs. 5(c) and 5(d) once again confirm that SecQSA incurs the lowest communication overhead among the three because of its merging operations.

VII. Conclusion

In this paper, we have initiated the study of secure quantile summary aggregation in wireless sensor networks. After examining the impact of different attacks on quantile summary aggregation via simulation, we introduced the design and evaluation of SecQSA, the first secure quantile summary aggregation protocol for wireless sensor networks. Built upon efficient cryptographic primitives, SecQSA can ensure the integrity of sampled readings and their ranks in the final quantile summary. Detailed simulation results have confirmed significant advantages of SecQSA over alternative solutions.

Acknowledgement

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] N. Khalil, M. R. Abid, D. Benhaddou, and M. Gerndt, “Wireless sensors networks for internet of things,” in *IEEE ISSNIP’14*, 2014, pp. 1–6.

[2] Y.-W. Kuo, C.-L. Li, J.-H. Jhang, and S. Lin, “Design of a wireless sensor network-based iot platform for wide area and heterogeneous applications,” *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5187–5197, 2018.

[3] R. Rajagopalan and P. K. Varshney, “Data-aggregation techniques in sensor networks: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 8, no. 4, pp. 48–63, 2006.

[4] H. Chan, A. Perrig, and D. Song, “Secure hierarchical in-network aggregation in sensor networks,” in *ACM CCS’06*, 2006, pp. 278–287.

[5] S. Roy, M. Conti, S. Setia, and S. Jajodia, “Secure median computation in wireless sensor networks,” *Ad Hoc Networks*, vol. 7, no. 8, pp. 1448–1462, 2009.

[6] B. Przydatek, D. Song, and A. Perrig, “Sia: Secure information aggregation in sensor networks,” in *ACM SenSys’03*, 2003, pp. 255–265.

[7] S. Roy, S. Setia, and S. Jajodia, “Attack-resilient hierarchical data aggregation in sensor networks,” in *ACM SASN’04*, 2006, pp. 71–82.

[8] D. Wagner, “Resilient aggregation in sensor networks,” in *ACM SASN’04*, 2004, pp. 78–87.

[9] Y. Yang, X. Wang, S. Zhu, and G. Cao, “Sdap: A secure hop-by-hop data aggregation protocol for sensor networks,” *ACM TISSEC*, vol. 11, no. 4, pp. 1–43, 2008.

[10] K. B. Frikken and J. A. Dougherty IV, “An efficient integrity-preserving scheme for hierarchical sensor aggregation,” in *ACM WiSec’08*, 2008, pp. 68–76.

[11] S. Papadopoulos, A. Kiayias, and D. Papadias, “Secure and efficient in-network processing of exact sum queries,” in *IEEE ICDE’11*, 2011, pp. 517–528.

[12] H. Yu, “Secure and highly-available aggregation queries in large-scale sensor networks via set sampling,” *Distributed Computing*, vol. 23, no. 5-6, pp. 373–394, 2011.

[13] S. Roy, M. Conti, S. Setia, and S. Jajodia, “Secure data aggregation in wireless sensor networks: Filtering out the attacker’s impact,” *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 4, pp. 681–694, 2014.

[14] A. Aseeri and R. Zhang, “Secure data aggregation in wireless sensor networks: Enumeration attack and countermeasure,” in *IEEE ICC’19*, 2019, pp. 1–7.

[15] N. Shrivastava, C. Buragohain, D. Agrawal, and S. Suri, “Medians and beyond: new aggregation techniques for sensor networks,” in *ACM SenSys’04*, 2004, pp. 239–249.

[16] M. B. Greenwald and S. Khanna, “Power-conserving computation of order-statistics over sensor networks,” in *ACM PODS’04*, 2004, pp. 275–285.

[17] Z. Huang, L. Wang, K. Yi, and Y. Liu, “Sampling based algorithms for quantile computation in sensor networks,” in *ACM SIGMOD’11*, 2011, pp. 745–756.

[18] B. Haeupler, J. Mohapatra, and H.-H. Su, “Optimal gossip algorithms for exact and approximate quantile computations,” in *ACM PODS’18*, 2018, pp. 179–188.

[19] B. Chen and H. Yu, “Secure aggregation with malicious node revocation in sensor networks,” in *IEEE ICDCS’11*, 2011, pp. 581–592.

[20] M. Greenwald and S. Khanna, “Space-efficient online computation of quantile summaries,” in *ACM SIGMOD’01*, Santa Barbara, CA, 2001, pp. 58–66.

[21] D. Liu and P. Ning, “Establishing pairwise keys in distributed sensor networks,” in *ACM CCS*, Washington, DC, October 2003, pp. 52–61.

[22] W. Zhang, M. Tran, S. Zhu, and G. Cao, “A compromise-resilient scheme for pairwise key establishment in dynamic sensor networks,” in *ACM MobiHoc*, Montreal, Canada, September 2007, pp. 90–99.

[23] A. Perrig, R. Szewczyk, V. Wen, D. Culler, and J. D. Tygar, “SPINS: Security protocols for sensor networks,” in *MobiCom*, Rome, Italy, July 2001, pp. 189–199.