Secure Top-k Query Processing in Unattended Tiered Sensor Networks

Rui Zhang, *Member, IEEE*, Jing Shi, Yanchao Zhang, *Senior Member, IEEE*, and Xiaoxia Huang, *Member, IEEE*

***Abstract*—Many future large-scale unattended sensor networks (USNs) are expected to follow a two-tier architecture with resource-poor sensor nodes at the lower tier and fewer resource-rich master nodes at the upper tier. Master nodes collect data from sensor nodes and then answer the queries from the network owner on their behalf. In hostile environments, master and sensor nodes may be compromised by the adversary and return incorrect data in response to data queries. Such application-level attacks are more harmful and difficult to detect than blind denial-of-service attacks on network communications, particularly when the query results are the basis for critical decision making. This paper presents a suite of novel schemes to enable verifiable top-k query processing in USNs, which is the first work of its kind. The proposed schemes are built upon symmetric cryptographic primitives and enable the network owner to detect any incorrect top-k query results. Detailed theoretical and simulation results confirm the high efficacy and efficiency of the proposed schemes.**

***Index Terms*—Security, top-k query, unattended tiered sensor networks (UTSNs).**

I. Introduction

UNATTENDED sensor networks (USNs) are sensor networks operating without an online data collection entity [2], [3]. USNs are ideal for remote and extreme environments such as oceans, volcanos, animal habitats, and battlefields. Instead of maintaining a costly high-speed stable communication link between the network and its external network owner, the USN relies on in-network data storage [4]–[7] for continuously produced sensed data. The network owner can access the data via an on-demand communication connection (e.g., a satellite link) or by physical means such as dispatching mobile sinks to the USN [5].

As shown in Fig. 1, many future large-scale USNs are expected to follow a two-tier architecture with resource-poor sensor nodes at the lower tier and resource-rich master nodes at the upper tier, which we refer to as *unattended tiered sensor networks* (UTSNs). This two-tier architecture is known to be indispensable for increasing network capacity and scalability, reducing system complexity, and prolonging network lifetime [4], [8]. Sensor nodes perform the sensing tasks and periodically submit sensed data to nearby master nodes for in-network storage, whereas master nodes answer ad hoc data queries issued by the network owner via an on-demand wireless link to some mater nodes. UTSNs may support various data queries, and top-k queries [9], [10] are among the most important and also the focus of this paper. A top-k query asks for data items with numeric attributes or scores [9] among the k highest, where k is an application-dependent parameter. An exemplary top-10 query is "Return the data whose temperature attribute is among the 10 highest between 2 P.M. and 3 P.M."

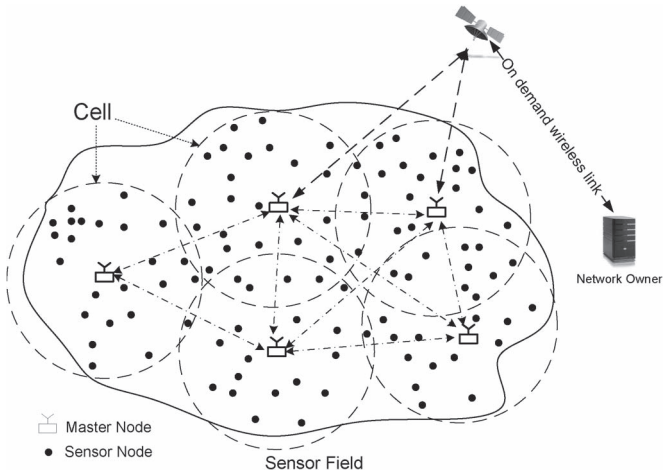


Fig. 1. Remote two-tier sensor network.

The unattended nature of UTSNs unfortunately renders top-k query processing very vulnerable to attacks in hostile environments. For example, master and/or sensor nodes in military or homeland security applications may be compromised by the adversary; those in commercial UTSNs may likewise be compromised by malicious business competitors to degrade their quality of data service.[1] The adversary may launch a number of attacks via compromised master and/or sensor nodes. For example, the adversary may instruct a compromised master node to return fake or juggled data in response to top-k queries from the network owner. Such application-level attacks are more subtle and harmful than blind denial-of-service attacks, particularly

Manuscript received August 17, 2013; revised December 14, 2013; accepted March 2, 2014. Date of publication March 14, 2014; date of current version November 6, 2014. This work was supported by the U.S. National Science Foundation under Grant CNS-0844972 (CAREER), Grant CNS-1117462, and Grant CNS-1320906. This paper was presented in part at the 29th IEEE Conference on Computer Communications, San Diego, CA, USA, March 15–19, 2010. The review of this paper was coordinated by Prof. S. Chen.

R. Zhang is with the Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822 USA (e-mail: ruizhang@hawaii.edu).

J. Shi is with the School of Public Administration, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: shi.jing@hust.edu.cn).

Y. Zhang is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706 USA (e-mail: yczhang@asu.edu).

X. Huang is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: xx.huang@siat.ac.cn).

Digital Object Identifier 10.1109/TVT.2014.2312014

[1] Similar incidents have been increasingly reported over the Internet, where companies hired botnet operators to wreck the business of their competitors.

0018-9545 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

when query results are the basis for making critical military or business decisions. As another example, compromised sensor nodes may forge sensed data with extremely large scores such that the data items generated by legitimate sensor nodes will have little chance to appear in the query result even if the master node behaves well. Moreover, compromised sensor nodes may assist master nodes to escape detection.

These given situations necessitate proactive mechanisms to enable verifiable top-k query processing, by which the network owner can verify the *authenticity* and *soundness* of top-k query responses. Authentication check is needed to detect fake data in query responses, whereas soundness verification is necessary to ensure that the returned data items are indeed those satisfying the query conditions, i.e., indeed the data items with scores among the k highest. A query result is considered correct only if it is both authentic and sound.

This paper investigates verifiable top-k queries in UTSNs with the following contributions.

- We first propose VTQ, which is a novel scheme whereby the network owner can detect any incorrect top-k query results returned by a compromised master node. VTQ relies on sensor nodes embedding some relationships among the data items they generated so that the network owner can detect any incorrect top-k query result by examining the embedded information.
- We then propose a random probing (RP) scheme to detect possible colluding attack from compromised master and sensor nodes. RP works by letting the network owner probe some randomly chosen sensor nodes for additional proofs after a top-k query result passes the verification under VTQ.
- We further propose a query conversion (QC) scheme to mitigate the impact of compromised sensor nodes forging data items with extremely high scores. The basic idea is that the network owner converts a top-k query into another such that the query result for the converted query contains the true top-k data items generated by legitimate sensor nodes with overwhelming probability.
- We also propose a lightweight scheme called RW to detect possible compromised sensor nodes framing a legitimate master node. RW relies on randomly chosen sensor nodes serving as witnesses for those submitting sensed data to the master node. In case of dispute, the network owner can detect framings by examining the testimonies from witness nodes.

All our proposed schemes are built upon symmetric cryptographic primitives and, thus, are very suitable for resource-constrained UTSNs. Their efficacy and efficiency are confirmed by detailed theoretical analysis and simulation results.

The rest of this paper is structured as follows. Section II introduces our network, query, and adversary models. Section III presents our problem formulation and the evaluation metrics. Section IV illustrates VTQ. Section V illustrates RP, QC, and RW for defending against compromised sensor nodes. All the proposed schemes are theoretically analyzed in Section VI and evaluated via detailed simulations in Section VII. Section VIII discusses the related work, and Section IX concludes this paper.

II. Network, Query, and Adversary Models

Here, we introduce our network, query, and adversary models.

A. Network Model

We assume a similar network model as in [7] and [11]–[13]. The UTSN is partitioned into many *cells*, each consisting of many sensor nodes and one master node. We assume that master and sensor nodes know their respective locations and affiliated cells. The localization requirement is fundamental in most sensor network applications and can be satisfied by many existing techniques such as in [14] and [15]. There might be sensor nodes in the overlapping area of multiple cells, in which case they are affiliated with all those cells.

Master and sensor nodes significantly differ in their resources. In particular, master nodes have abundant resources in storage, energy (e.g., a heavy-duty battery or solar panel), and computation, whereas sensor nodes are much more constrained in every regard. In addition, each master node can communicate with neighboring master nodes via relatively long-range and high-rate radios, thus forming an upper-tier multihop network.

As in [7], [11], and [12], we assume that time is divided into *epochs*. At the end of each epoch, each sensor node submits to its affiliated master node all the data (if any) it generated during that epoch. We assume that there is no stable communication link connecting the sensor network to the external network owner; hence, data must be stored at master nodes. The network owner can issue top-k queries via an on-demand wireless (e.g., satellite) link to some master node(s), which is often both costly and of a relatively low rate. As a result, the communication cost incurred by top-k queries over such on-demand wireless links should be kept as low as possible.

B. Top-k Query Basics

Data generated by sensor nodes may have multiple attributes, each corresponding to one type of sensor or one aspect of a detected event. Each data item can be scored by some scoring functions [9] and ranked based on its score. In this paper, we focus on top-k queries with a single score function. For the sake of simplicity, the following *primitive* top-k queries will be considered:

$$(\text{cell} = \mathcal{C}) \wedge (\text{epoch} = t) \wedge (\text{num} = k) \wedge (\text{query region} = \mathcal{I}_t).$$

Here, $\mathcal{C}$ and t are the interested cell ID and epoch number, respectively; k refers to the number of desired data items; and $\mathcal{I}_t$ denotes the physical *query region*. We will subsequently abuse the notation $\mathcal{I}_t$ to also denote the set of sensor node IDs in the query region. Our assumption here is that both the network owner and the master node know the mappings between sensor node IDs and their respective geographic locations. We aim to support fine-grained top-k queries, in which $\mathcal{I}_t$ may cover one or more random sensor nodes in cell $\mathcal{C}$.

C. Adversary Model

We aim to support authenticity and soundness verification of top-k query results and refer the readers to the existing

rich literature (e.g., [2], [6], and [15]–[24]) for other important security issues.

We assume that the adversary has compromised some master and sensor nodes in the UTSN. Since the operations in different cells are independent from each other, the adversary will not gain more from the collaboration of compromised master/sensor nodes in different cells. Without loss of generality, our subsequent discussion thus focuses on a cell $\mathcal{C}$ consisting of a master node $\mathcal{M}$ and N sensor nodes $\{S_i\}_{i=1}^{N}$ whose IDs compose a set $\mathcal{I} = \{1, 2, \ldots, N\}$. Among them, we assume that $c \ll N$ sensor nodes are compromised.

The adversary may launch different attacks through compromised $\mathcal{M}$, sensor nodes, or both. In particular, we consider the following attacks in this paper.

- *Attack* 1: Compromised $\mathcal{M}$, with the possible assistance of compromised sensor nodes, may return incorrect query results in response to the network owner's top-k queries.
- *Attack* 2: Compromised sensor nodes may forge data items with extremely high scores such that the data items generated by legitimate sensor nodes will have little chance to appear in the query result.
- *Attack* 3: Compromised sensor nodes may frame a good master node by exploiting our verification mechanism, e.g., deviating from protocol execution, such that the network owner will falsely identify $\mathcal{M}$ as malicious.

Different from [7], [11], and [12], we do not intend to ensure data confidentiality against master nodes. Many sensor network applications do not require data confidentiality but only query-result authenticity and soundness. For example, intrusion events in a sensor network for battlefield reconnaissance are known to the adversary and, thus, need not be secret. In other words, the adversary knows that he has been detected, but he can instruct compromised master nodes to return fake and/or unsound query responses so that the network owner cannot precisely determine his itinerary. In such cases, enabling query-result authenticity and soundness verifications becomes a must. Achieving secure top-k query-processing and data confidentiality is still an open challenge.

III. Problem Statement

Here, we formulate the problem and introduce our design goals and evaluation metrics.

A. Problem Formulation

For ease of presentation, we assume that during each epoch t, each node $S_i \in \{S_i\}_{i=1}^{N}$ generates μ data items, denoted by $\mathcal{D}_i = \{D_{i,j}\}_{j=1}^{\mu}$. Our scheme can be easily adapted to support the case in which each node generates different number of data items. The master node $\mathcal{M}$ thus receives $N\mu$ data items at the end of epoch t, which are denoted by $\mathcal{D} = \bigcup_{i=1}^{N} \mathcal{D}_i$. We assume that all the data items generated in cell $\mathcal{C}$ during epoch t have mutually different scores. For example, we can break a tie between two different data items by considering their corresponding node IDs or the times when they are generated. This assumption implies that a *unique* correct response exists for any top-k query. We will denote by $s_{i,j}$ the score of $D_{i,j}$, i.e., $s_{i,j} = f(D_{i,j})$, where $f(\cdot)$ is a public scoring function [9]. In addition, we will equate $D_{i,j} \leq D_{i',j'}$ with $s_{i,j} \leq s_{i',j'}$ for any i, i', j, and j'.

Given a query $\mathcal{Q}_t = \langle \mathcal{C}, t, k, \mathcal{I}_t \rangle$ as introduced in Section II-B, we define the corresponding *candidate data set* as $\mathcal{D}_t = \bigcup_{i \in \mathcal{I}_t} \mathcal{D}_i$, which contains $\mu_t = n\mu$ *candidate data items*, where $n = |\mathcal{I}_t|$. It is possible that there are less than k candidate data items, i.e., $\mu_t < k$. This situation, however, has very little impact on our schemes. For simplicity, we hereafter assume $\mu_t \geq k$ in most descriptions and will point out the additional actions that need be taken for $\mu_t < k$ when appropriate.

Assuming that $\mathcal{M}$ returns a query response containing k data items, denoted by $\mathcal{R}_t$, the problem of interest is how the network owner can efficiently verify the compliance of $\mathcal{R}_t$ with the following conditions.

- **Authenticity**: All data items in $\mathcal{R}_t$ were generated by nodes in the query region or, equivalently, $\mathcal{R}_t \subseteq \mathcal{D}_t$.
- **Soundness**: $\mathcal{R}_t$ contains the top-k data items among all the candidates or, equivalently, $D_{i,j} > D_{i',j'}$, for all $D_{i,j} \in \mathcal{R}_t$ and $D_{i',j'} \in \mathcal{D}_t \setminus \mathcal{R}_t$.

B. Performance Metrics

The following performance metrics will be used throughout.

- P_{det}—**detection probability**: the probability that an incorrect (i.e., forged and/or unsound) top-k query result is detected.
- C_{cell}—**in-cell communication cost**: the total additional communication energy consumption in bits incurred by enabling verifiable top-k queries in cell $\mathcal{C}$ per epoch. Here, we assume the same energy consumption in transmitting and receiving every bit across each hop.
- C_{query}—**query communication cost**: the total additional information in bits transmitted between $\mathcal{M}$ and the network owner for enabling verifiable top-k queries. The route connecting $\mathcal{M}$ to the network owner may traverse multiple master nodes and the on-demand wireless link. For simplicity, we associate an energy cost of transmitting and receiving every bit with this route, which is usually much larger than that between neighboring sensor nodes.

IV. Verifiable Top-k Queries

Here, we present VTQ, which enables the network owner to verify the authenticity and soundness of any top-k query result in UTSNs against a compromised master node. For clarity, we defer the discussion of other attacks launched by compromised sensor nodes in Section V.

A. Overview

VTQ is essentially built upon the following two facts.

Fact 1: Suppose that each node S_i sorts its data items in descending order such that $D_{i,j} > D_{i,j+1}$ for all $j \in [1, \mu - 1]$. If $D_{i,j}$ is among the top k, so is $D_{i,x}$ for all $x \in [1, j)$; likewise, if $D_{i,j}$ is not among the top k, neither is $D_{i,y}$ for all $y \in (j, \mu]$.

Fact 2: Any top-k data item is larger than any non-top-k data item in the query region.

Fact 1 implies that adjacent data items generated by the same sensor node are very likely to satisfy or dissatisfy a top-k query at the same time. If node S_i has $k_i > 0$ data items among the top k, then they must be $D_{i,1}, \ldots, D_{i,k_i}$. On the other hand, Fact 2 implies that for any two nodes S_i and $S_j, i \neq j$, if $D_{i,k_i+1} > D_{j,1}$, then node S_j has no data item among the top k, i.e., $k_j = 0$.

To exploit these two facts, we let each sensor node sort their data items and exchange its highest score with its nearby nodes. Each node then chains adjacent data items with other nodes' highest scores using a cryptographic hash function. On receiving a top-k query $\mathcal{Q}_t$, we require master node $\mathcal{M}$ to return some additional information in addition to the top-k data items in the query result whereby the network owner can verify both the authenticity and soundness of the query result.

For our purpose, we assume that each S_i is preloaded with a distinct initial key $K_{i,0}$ uniquely shared with the network owner. At the end of epoch $t \geq 1$, S_i generates an epoch key by $K_{i,t} = H(K_{i,t-1})$ and erases $K_{i,t-1}$ from its memory, where $H(\cdot)$ denotes a good hash function. We also introduce an extremely small public value $\underline{\chi}$ and an extremely large public value $\overline{\chi}$, both out of the known domain of the data score. Assuming that $N = nm$, we partition each cell $\mathcal{C}$ into m virtual subcells of equal size and assume that each sensor node knows its affiliated subcell. We denote the m subcells and their respective node ID sets by $\{\mathcal{C}_y\}_{y=1}^m$ and $\{\mathcal{J}_y\}_{y=1}^m$, respectively.

In what follows, we detail the VTQ design, which consists of three phases. In the *data-submission* phase, each sensor node preprocesses its sensed data using cryptographic methods for submission. In the subsequent *query-processing* phase, $\mathcal{M}$ answers a top-k query by returning the query result and certain proofs to the network owner. In the final *verification* phase, the network owner verifies the authenticity and soundness of the query result by examining the proofs.

B. Data Submission

At the end of each epoch, sensor nodes in each subcell $\mathcal{C}_y$ exchange some information about their sensed data. Consider node S_i as an example. Node S_i broadcasts its highest score and node ID within subcell $\mathcal{C}_y$ as follows:

$$S_i \rightarrow * : i, s_{i,1}.$$

Here, we assume a suitable broadcast authentication protocol like multilevel μTESLA [16] for secure and reliable transmissions of such broadcast messages.

Node S_i waits for sufficient time to receive all the highest scores $\{s_{j,1}\}_{j \in \mathcal{J}_y \setminus \{i\}}$ from all the other nodes in $\mathcal{C}_y$. It then sorts its own data scores and the received data scores $\{s_{i,j}\}_{j=1}^{\mu} \bigcap \{s_{x,1}\}_{x \in \mathcal{J}_y}$ in descending order, resulting in a list of $\mu + n - 1$ scores, where n is the size of each subcell. Recall our assumption that all the data items generated during each epoch in cell $\mathcal{C}$ have different scores. Node S_i then replaces the scores received from other nodes with their corresponding node IDs, resulting in $\mu + 1$ lists of node IDs $\mathcal{L}_{i,1}, \ldots, \mathcal{L}_{i,\mu+1}$, separated by S_i's own scores $s_{i,1}, \ldots, s_{i,\mu}$. More specifically, for any node ID x appears in $\mathcal{L}_{i,1}, \mathcal{L}_{i,j}$ $(2 \leq j \leq \mu)$, and $\mathcal{L}_{i,\mu+1}$, we have $s_{x,1} > s_{i,1}$, $s_{i,j-1} > s_{x,1} > s_{i,j}$, and $s_{x,1} < s_{i,\mu}$, respectively. In addition, if x and y both appear in $\mathcal{L}_{i,j}$, x is on the left-hand side of y if and only if $s_{x,1} > s_{y,1}$. We call each $\mathcal{L}_{i,j}$ an *auxiliary ID list* henceforth.

As a concrete example, suppose that subcell $\mathcal{C}_1$ consists of sensor nodes S_1, S_2, and S_3 with data score sets $\{1, 5, 9\}$, $\{2, 3, 4\}$ and $\{6, 7, 8\}$, respectively. During data submission, node S_1 broadcasts its highest score with node ID $\langle 1, 9\rangle$ and receives $\langle 2, 4\rangle$ and $\langle 3, 8\rangle$ from nodes S_2 and S_3, respectively. Node S_1 then sorts its own data scores $\{1, 5, 9\}$ and the received 4 and 8 in descending order, resulting in $\langle 9, 8, 5, 4, 1\rangle$. It then replaces data scores 4 and 8 with their corresponding node IDs to obtain $\langle 9, 3, 5, 2, 1\rangle$. The corresponding auxiliary ID lists are then $\mathcal{L}_{1,1} = \emptyset$, $\mathcal{L}_{1,2} = \langle 3\rangle$, $\mathcal{L}_{1,3} = \langle 2\rangle$, and $\mathcal{L}_{1,4} = \emptyset$.

Let $h_*(\cdot)$ denote a *message authentication code* (MAC) computed using the key at the subscript. Node S_i then binds adjacent data items as well as auxiliary ID lists by computing

$$\mathsf{V}_{i,j} = \begin{cases} h_{K_{i,t}}\left(\overline{\chi} \| \mathcal{L}_{i,1} \| D_{i,1}\right), & j = 1 \\ h_{K_{i,t}}\left(D_{i,j-1} \| \mathcal{L}_{i,j} \| D_{i,j}\right), & 2 \leq j \leq \mu \\ h_{K_{i,t}}\left(D_{i,\mu} \| \mathcal{L}_{i,\mu+1} \| \underline{\chi}\right), & j = \mu + 1. \end{cases} \quad (1)$$

Finally, each S_i submits all its data items to the master node $\mathcal{M}$ in the following message:

$$\begin{aligned} S_i \rightarrow \mathcal{M} : i, t, &\langle \mathcal{L}_{i,1}, D_{i,1}, \mathsf{V}_{i,1}\rangle \\ &\vdots \\ &\langle \mathcal{L}_{i,\mu}, D_{i,\mu}, \mathsf{V}_{i,\mu}\rangle \\ &\mathcal{L}_{i,\mu+1}, \mathsf{V}_{i,\mu+1}. \end{aligned} \quad (2)$$

C. Query Processing

After receiving a top-k query $\mathcal{Q}_t = \langle \mathcal{C}, t, k, \mathcal{I}_t\rangle$, the master node $\mathcal{M}$ first locates the largest k data items in the candidate data set $\mathcal{D}_t$, whereby to determine the number of top-k data items for each node S_i (denoted by k_i). It follows that $\sum_{i \in \mathcal{I}_t} k_i = k$. For convenience, we will call a data item *qualified* (or *unqualified*) if it is (or not) among the top k. Similarly, we will call a sensor node qualified (or unqualified) if it has at least one (or no) qualified data item.

For each qualified node S_i (i.e., $k_i > 0$), $\mathcal{M}$ returns the following information as a part of the query response.

- Case 1: If $k_i < \mu$, the information is

$$\begin{aligned} \mathcal{M} \rightarrow \text{network owner} : i, &\langle \mathcal{L}_{i,1}, D_{i,1}, \mathsf{V}_{i,1}\rangle \\ &\vdots \\ &\langle \mathcal{L}_{i,k_i+1}, D_{i,k_i+1}, \mathsf{V}_{i,k_i+1}\rangle \end{aligned}$$

  where $D_{i,1}, \cdots, D_{i,k_i}$ are qualified data items, and D_{i,k_i+1} is unqualified but needed for later verification.
- Case 2: If $k_i = \mu$, the information is

$$\begin{aligned} \mathcal{M} \rightarrow \text{network owner} : i, &\langle \mathcal{L}_{i,1}, D_{i,1}, \mathsf{V}_{i,1}\rangle \\ &\vdots \\ &\langle \mathcal{L}_{i,\mu}, D_{i,\mu}, \mathsf{V}_{i,\mu}\rangle, \mathcal{L}_{i,\mu+1} \end{aligned}$$

  where $D_{i,1}, \ldots, D_{i,\mu}$ are all qualified data items.

In addition, if $\mathcal{M}$ does not return any data item from one subcell, the network owner cannot differentiate whether that subcell indeed has no qualified data or $\mathcal{M}$ purposefully skipped them. In view of this situation, VTQ requires $\mathcal{M}$ to return some additional information for each subcell without qualified data. Specifically, we call a subcell unqualified if it overlaps with the query region but has no qualified data. The master node $\mathcal{M}$ is required to return the largest data item in each unqualified subcell $\mathcal{C}_y$ with nodes $\mathcal{J}_y$ as follows.

- Case 3: Assuming that node S_i generated the largest data item $D_{i,1}$ in epoch t among all the nodes in $\mathcal{J}_y \bigcap \mathcal{I}_t$, $\mathcal{M}$ need return the following information in the query response:

$$\mathcal{M} \rightarrow \text{network owner} : i, \langle \mathcal{L}_{i,1}, D_{i,1}, \mathsf{V}_{i,1} \rangle.$$

D. Verification

Upon receiving the query result from $\mathcal{M}$, the network owner first verifies its authenticity by checking the MACs. In particular, for each sensor node S_i with at least one data item returned, the network owner derives the corresponding key $K_{i,t}$. Then, for each data item $D_{i,j}$ returned, the network owner recomputes the corresponding $\mathsf{V}_{i,j}$ according to (1) and compares it with the received one. If the two match, $D_{i,j}$ is considered authentic. Since each data item is bound with adjacent data items using MACs, verifying each $\mathsf{V}_{i,j}$ also ascertains that master node $\mathcal{M}$ has not inserted any forged data items or skipped any legitimate data items. If all the verifications succeed, the network owner considers the query result authentic, as each key $K_{i,t}$ is known only to himself and S_i.

The network owner proceeds to check the soundness of the query result by examining the relationships among the returned data items and auxiliary ID lists as follows.

- First, the network owner checks if there is at least one data item returned for every subcell that overlaps with the query region.
- Second, the network owner checks if the query result is consistent with Fact 2. In particular, since the information returned for each node S_i follows one of the three cases, the network owner can easily determine k_i for S_i as well as the qualified data items, i.e., $D_{i,1}, \ldots, D_{i,k_i}$ (Case 1 or 2), and the unqualified data item D_{i,k_i+1} (Case 1 or 3), if any. He can then verify if there are indeed total k qualified data items returned. If so, he further checks if the smallest qualified data item is larger than the largest unqualified data item among all those returned.
- Finally, the network owner examines all the auxiliary ID lists $\mathcal{L}_{i,j}$ contained in the query response to see if $\mathcal{M}$ has skipped all the data items for some qualified node. In particular, for each qualified data item, e.g., $D_{i,j}$ with a nonempty auxiliary ID list $\mathcal{L}_{i,j}$, the network owner checks whether there is at least one data item returned from node S_x for all $x \in \mathcal{L}_{i,j} \cap \mathcal{I}_t$. If not, the query result is considered unsound. The underlying rationale is very simple. If $x \in \mathcal{L}_{i,j} \cap \mathcal{I}_t$, node S_x must have at least one data item scoring higher than $s_{i,j}$ according to the definition of $\mathcal{L}_{i,j}$.

If all the given verifications succeed, the network owner considers the query result both authentic and sound.

V. Defenses Against Compromised Sensor Nodes

So far, we have not considered the impact of compromised sensor nodes for the sake of clarity. Here, we discuss three attacks launched by compromised sensor nodes and propose corresponding defenses.

A. Forging Auxiliary ID List

Compromised sensor nodes may collude with $\mathcal{M}$ to *overshadow* some qualified data items by forging their auxiliary ID lists. In particular, a compromised sensor node can forge its auxiliary ID lists to cheat the network owner into believing that no other node in the same subcell has qualified data items. Consider the following example. Suppose that the network owner queries the top-2 data items generated by nodes S_1 and S_2, among which S_1 is legitimate and has generated top-2 data items, and S_2 is compromised. Node S_2 can fake its auxiliary ID lists by setting $\mathcal{L}_{2,1} = \mathcal{L}_{2,2} = \emptyset$, which means that S_1 has no data item larger than $D_{2,2}$. The master node $\mathcal{M}$ can then return the top-2 data items of S_2 and provide necessary proofs to pass the authenticity and soundness verification as in VTQ.

We propose a randomized probing (RP) scheme for the network owner to ask for additional proofs from randomly chosen sensor nodes. In particular, after the query result passes all the verifications in Section IV-D, the network owner randomly chooses $\theta \geq 1$ candidate nodes in each subcell that overlaps with the query region, from which no data item has been returned. Let d be the number of subcells that overlaps with the query region. The network owner sends θd chosen node IDs to $\mathcal{M}$, which, in turn, returns the largest data item and corresponding auxiliary ID list for each of them. More specifically, for each chosen node S_i, the master node $\mathcal{M}$ need return $\langle \mathcal{L}_{i,1}, D_{i,1}, \mathsf{V}_{i,1} \rangle$.

On receiving the θd largest data items and auxiliary ID lists, the network owner first verifies the authenticity for each of them by checking the corresponding MAC as in Section IV-D. If all the information returned is authentic, the network owner proceeds to check if each pair of returned data item and auxiliary ID list is consistent with the query result.

Consider as an example $D_{i,1}$ and $\mathcal{L}_{i,1}$ returned from node S_i in subcell $\mathcal{C}_y$ with nodes $\mathcal{J}_y$. According to VTQ query processing, $\mathcal{M}$ must have returned at least one data item from nodes among $\mathcal{J}_y \bigcap \mathcal{I}_t$, i.e., the intersection between subcell $\mathcal{C}_y$ and the query region $\mathcal{I}_t$. Without loss of generality, assume that $\mathcal{M}$ has returned data items $\mathcal{D}_y$ from nodes $\mathcal{J}_{q,y} \subseteq \mathcal{J}_y \bigcap \mathcal{I}_t$. If there was an overshadowing attack in $\mathcal{C}_y$, then $\mathcal{M}$ must have omitted data items from at least one node in $\mathcal{J}_y \bigcap \mathcal{I}_t$ with data items larger than the smallest data item among returned $\mathcal{D}_y$.

The network owner first checks if $\mathcal{J}_{q,y} \subseteq \mathcal{L}_{i,1}$, i.e., if every node with at least one data item returned has its ID in $\mathcal{L}_{i,1}$. If not, he considers that there was an overshadowing attack in $\mathcal{C}_y$. The reason is that any data item among $\mathcal{D}_y$ must be larger than $D_{i,1}$ and has its corresponding node ID embedded in $\mathcal{L}_{i,1}$ according to VTQ. Assume that $\mathcal{L}_{i,1} = \langle j_1, \ldots, j_z \rangle$, where $z = |\mathcal{L}_{i,1}|$. The network owner finds the maximum $x \in [1, z]$ such

that at least one data item is returned from node S_{j_x}. Then, for each $w \in [1, x-1]$, the network owner checks if node S_{j_w} satisfies one of the following two conditions.

- Condition 1: $j_w \notin \mathcal{I}_t$, i.e., S_{j_w} is not in the query region.
- Condition 2: At least one data item has been returned from node S_{j_w}.

If not, the network owner considers that there was an overshadowing attack, i.e., node S_{j_w}'s data items have been overshadowed. If the given verifications succeed for each of the θd returned largest data items and auxiliary ID lists, the network owner considers that there was no overshadowing attack. The efficacy of randomized probing is analyzed in Section VI-B.

B. Forging Data With Extremely High Scores

Compromised sensor nodes may also overshadow some qualified data items by forging data items with extremely high scores. In particular, compromised sensor nodes in cell $\mathcal{C}$ each submits μ fake data items with extremely high scores to $\mathcal{M}$, which are properly authenticated and chained as in VTQ. If any compromised node appears in the query region and k is small, the data from legitimate sensor nodes will have little chance to appear in the query result and, thus, be overshadowed. It is fundamentally difficult to tell if a data item is fake or legitimate without special assumptions. The only feasible solution is to tolerate such fake data items while retrieving the true top-k data items generated by legitimate sensor nodes.

Our defense is to let the network owner query more data items than needed to tolerate possible forged data items from compromised sensor nodes. By doing so, the quality of data queries will not be significantly affected as long as the query result contains the true top-k data items generated by legitimate sensor nodes. Moreover, the network owner could analyze all the returned data items offline using advanced statistical technique to detect compromised sensor nodes.

The remaining challenge is how to minimize the query overhead while, at the same time, ensuring that the query result contains the true top-k data items without knowing which sensor nodes are compromised. In what follows, we introduce QC, which is a query conversion scheme that converts an original top-k query $\mathcal{Q}_t$ into a top-k' query, such that the query result of the converted query contains the true top-k data items generated by legitimate sensor nodes in the query region with high probability. QC is built upon the following two ideas.

First, the network owner can simply increase k to k' to tolerate forged data items from compromised sensor nodes. In particular, suppose that the network owner intends to tolerate up to c compromised sensor nodes. Recall that each node generates μ data items in each epoch. Given a top-k query $\mathcal{Q}_t = \langle \mathcal{C}, t, k, \mathcal{I}_t \rangle$, a simple conversion is let $k = c\mu + k$. By doing so, the returned k' data items will certainly contain the top-k data items generated by legitimate sensor nodes as long as the number of compromised sensor nodes is smaller than c, as each compromised sensor node can forge at most μ data items. The limitation of this method is that c might be difficult to choose in practice. When μ is large, a conservative choice of c may incur significant query overhead.

Second, the network owner can exploit certain prior knowledge about the sensed data distribution to reduce query communication overhead. In particular, assuming that each data item generated by legitimate sensor nodes is equally possible to be among the top k, it is not likely for a single legitimate node to have too many qualified data items. For example, suppose that $\mu = 10$ and that the network owner queries the top-5 data items generated by ten sensor nodes. The probability that any single node generates all top-5 data items can be computed as $10 \times (1/10^5) = 10^{-4}$, which is negligible. The network owner can thus purposefully restrict that any sensor node can contribute at most $\delta < 10$ data items to the query result. By doing so, the network owner can tolerate more compromised sensor nodes for given k' because each compromised node can have at most δ forged data items in the query result, while ensuring that the query result contains the true top-k data items with sufficiently high probability.

We now detail the query conversion mechanism that incorporates the given two ideas. Given an original top-k query $\mathcal{Q}_t = \langle \mathcal{C}, t, k, \mathcal{I}_t \rangle$, the network owner converts $\mathcal{Q}_t$ into a δ-*constrained* top-k' query $\mathcal{Q}_t^c = \langle \mathcal{C}, t, k', \mathcal{I}_t, \delta \rangle$, where $\mathcal{C}, t, k$, $\mathcal{I}_t$ have the same meanings as in the original top-k query definition, and $\delta \leq \min(\mu, k)$ denotes the maximum number of data items that can be returned from any single node. Alternatively, we can view $\mathcal{Q}_t^c$ as the top-k' query over the candidate data set $\{D_{i,j} | i \in \mathcal{I}_t, 1 \leq j \leq \delta\}$, which is a subset of the original candidate data set $\mathcal{D}_t$.

The network owner sends $\mathcal{Q}_t^c$ to $\mathcal{M}$, which, in turn, returns the corresponding query result under VTQ. On receiving the query result, the network owner can verify its authenticity and soundness verification as in VTQ. The probability that the query result contains the true top-k data items, which are denoted by P_{true}, is jointly determined by the number of compromised sensor nodes in the query region and the choice of δ and k', which will be analyzed in Section VI-C.

C. Framing Legitimate Master Node

Our previous discussion focuses on detecting a compromised master node $\mathcal{M}$, which might be assisted by some compromised sensor nodes. The adversary, however, may also exploit our techniques to frame some legitimate master nodes. For example, assume that the adversary only compromises some sensor nodes in cell $\mathcal{C}$ while the master node $\mathcal{M}$ is legitimate. The compromised sensor nodes can frame $\mathcal{M}$ by sending it data authenticated using incorrect keys. Since $\mathcal{M}$ does not know the correct keys, it cannot detect such misbehavior. Consequently, the network owner will falsely identify $\mathcal{M}$ as malicious.

Our previous works [12], [25] suggest that an effective countermeasure against the framing attack is to let each sensor node and master node digitally sign every message transmitted and received. In case of dispute, the network owner can detect the misbehaving entities by analyzing related messages and signatures. This solution, however, requires public-key operations not suitable for resource-constrained sensor nodes.

Now, we introduce a symmetric-key-based solution (called RW) to defend against the framing attack, which relies on randomly chosen nodes serving as *witnesses* for sensor nodes

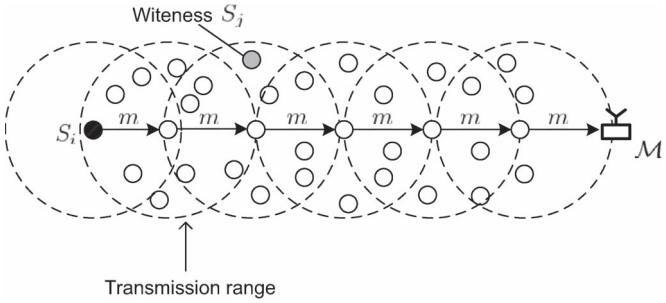


Fig. 2. Example of witness selection.

submitting sensed data to the master node. We assume that every sensor node can work in the promiscuous mode. Consider Fig. 2 as an example. Suppose that node S_i wants to submit a message msg to the master node $\mathcal{M}$ in epoch t. Each intermediate node S_j along the route that overhears the message checks if

$$h_{K_{j,t}}(i\|j\|t) \mod X \leq Y \tag{3}$$

where $X \geq Y$ are two integer-valued system parameters. If so, node S_j computes a *testimony* for msg as follows:

$$\mathsf{T}_{i,j,t} = h_{K_{j,t}}(\mathsf{msg}\|i\|t)\,. \tag{4}$$

Each node submits all the testimonies generated in epoch t to $\mathcal{M}$ at the beginning of epoch $t+1$. We can see that $\rho = Y/X$ determines the ratio of witness nodes of node S_i in epoch t among all the intermediate nodes that overheard the message msg. Since $K_{j,t}$ is only known to S_j and the network owner, the adversary cannot predict which nodes will be chosen as witnesses for msg. The adversary thus cannot compromise all the witness nodes in advance before framing a legitimate master node.

Later, if there is a dispute between node S_i and $\mathcal{M}$, the network owner can retrieve all the related testimonies to determine whether $\mathcal{M}$ is malicious or framed. Continue the previous example. Suppose that $\mathcal{M}$ later returns a top-k query result based on msg and is detected as inauthentic by the network owner. The network owner first derives the IDs of all the witnesses of node S_i during epoch t according to (3) and then requires $\mathcal{M}$ to return the original message msg as well as all the testimonies on message msg. The network owner then recomputes each testimony according to (4) using the corresponding key of each witness node S_j. If a majority of the testimonies indicate that msg is indeed the original message submitted by node S_i, the network owner considers that $\mathcal{M}$ is framed and excludes node S_i from a future query region.

It is worth noticing that the nodes far away from the master node will have more witnesses than those close to the master node for the same ratio $\rho = Y/X$ because its messages will be overhead by more intermediate nodes. The network owner may assign different values of ρ for different nodes according to their distances to the master node.

VI. Performance Analysis

Here, we analyze the efficacy and overhead of the proposed schemes.

A. Analysis of VTQ

We first have the following theorem regarding the detection capability of VTQ against a compromised master node.

Theorem 1: Assuming that none of the sensor nodes are compromised, VTQ can detect any incorrect top-k query result returned by a compromised master node.

Proof: Consider a queried node S_i that has k_i qualified data items $\{D_{i,j}\}_{j=1}^{k_i}$. Since the adjacent data items are bound with MACs for which $\mathcal{M}$ does not have the corresponding key $K_{i,t}$, $\mathcal{M}$ cannot insert forged data items into or omit legitimate data items from $\{D_{i,j}\}_{j=1}^{k_i}$ without being detected during the authenticity check.

Now, assume that the master node has returned authentic but an unsound query result, from which the network owner derives an unsound top-k query result containing k data items with the lowest score s' among them. Let s denote the lowest score among the k data items in the correct query result. If $s' > s$, there must be less than k data items with a score no lower than s' in the query region; hence, it is impossible for $\mathcal{M}$ to find k authentic data items with the lowest score s', leading to a contradiction. On the other hand, if $s' < s$, the master node $\mathcal{M}$ should have deleted at least one data item in the query region with a score higher than s'. Suppose that $\mathcal{M}$ has deleted $D_{i,j}$ with $s_{i,j} > s'$ and that node S_i is in subcell $\mathcal{C}_z$. There are two cases.

- If $\mathcal{M}$ returned no data item from node S_i, then $\mathcal{M}$ must have returned at least one data item generated by some other node in the same subcell with a score lower than s', e.g., $D_{x,y}$ with $s_{x,y} < s'$. Since $s_{i,j} > s'$, we have $s_{i,j} > s_{x,y}$ and node ID i must have been embedded into one auxiliary ID list among $\mathcal{I}_{x,1}, \ldots, \mathcal{I}_{x,y}$ and returned to the network owner, from which the network owner knows that $\mathcal{M}$ omitted some valid data item from node S_i.
- If $\mathcal{M}$ has returned some data items generated by node S_i, e.g., $D_{i,y}$, it must have returned one $D_{i,y}$ with $s_{i,y} < s'$ to pass the soundness check, which means that it must also return $D_{i,1}, \ldots, D_{i,y}$ to pass the authenticity check. Since $s_{i,j} > s' > s_{i,y}$, we have $j < y$ and $D_{i,j}$ must have been returned, leading to a contradiction.

Therefore, the network owner can detect any unsound query result as well. ■

Assume that each node ID is of l_{id} bits, each score is of l_{score} bits, $h_*(\cdot)$ is of l_{mac} bits, and the average number of hops between a sensor node and $\mathcal{M}$ is L. We then have the following theorem regarding the in-cell communication costs of VTQ.

Theorem 2: The in-cell communication cost of VTQ is given by

$$C_{\mathrm{cell}} = Nn(l_{\mathrm{id}} + l_{\mathrm{score}} + l_{\mathrm{mac}}) + N(\mu+1)Ll_{\mathrm{mac}} + N(n-1)Ll_{\mathrm{id}} \tag{5}$$

where n is the number of nodes in each subcell.

Proof: The in-cell communication cost of VTQ consists of two parts: C_{score}, which is the cost incurred by exchanging highest scores within each subcell, and C_{data}, which is the cost incurred by transmitting data items and embedded auxiliary node ID lists to $\mathcal{M}$. Note that we do not consider the cost for transmitting the epoch number, original data items, and

corresponding node IDs because they have to be submitted even without VTQ.

Under VTQ, each node needs to broadcast its node ID and highest score within its subcell. Assume that μTESLA [16] is used for broadcast authentication. Each broadcasted message is of $l_{rmid} + l_{\text{score}} + l_{\text{mac}}$ bits. Assume that the simplest broadcasting mechanism is used, in which each node rebroadcasts the message it received once. C_{score} is then given by

$$C_{\text{score}} = Nn(l_{\text{id}} + l_{\text{score}} + l_{\text{mac}}). \tag{6}$$

Since each node ID appears in $n-1$ auxiliary ID lists, the total number of node IDs in all auxiliary ID lists is thus $N(n-1)$. In addition, each node needs to transmit $\mu+1$ MACs to $\mathcal{M}$ [cf. (1)]. We thus have

$$C_{\text{data}} = N(\mu+1)Ll_{\text{mac}} + N(n-1)Ll_{\text{id}}. \tag{7}$$

It follows that

$$\begin{aligned} C_{\text{cell}} &= C_{\text{score}} + C_{\text{data}} \\ &= Nn(l_{\text{id}} + l_{\text{score}} + l_{\text{mac}}) \\ &\quad + N(\mu+1)Ll_{\text{mac}} + N(n-1)Ll_{\text{id}}. \end{aligned}$$ ■

We have only been able to derive an upper bound for the query communication cost of VTQ for a special case.

Theorem 3: Assuming that the query region comprises g subcells and that each candidate data item is equally likely to be among the top k. The expected query communication cost under VTQ is bounded by

$$\begin{aligned} C_{\text{query}} \le\ & kl_{\text{mac}} + n(1-p_o)(l_{\text{data}} + l_{\text{mac}}) \\ & + g(1-\alpha)\beta(\beta-1)l_{\text{id}} + g\alpha(l_{\text{id}} + l_{\text{data}} + l_{\text{mac}}) \end{aligned} \tag{8}$$

where

$$p_o = \frac{\binom{(gn-1)\mu}{k}}{\binom{gn\mu}{k}}, \quad \alpha = \frac{\binom{(g-1)n\mu}{k}}{\binom{gn\mu}{k}}$$

and $\beta = (n(1-p_o)/1-\alpha)$.

Proof: The query communication cost of VTQ consists of three parts: 1) the communication cost incurred by transmitting data items and indexes, which is denoted by C_1; 2) the communication cost incurred by transmitting embedded auxiliary ID lists, which is denoted by C_2; and 3) the communication cost incurred by transmitting data item and MAC for unqualified subcell, which is denoted by C_3.

We first analyze C_1. Since there are total $gn\mu$ data items generated in $\mathcal{I}_t$ during epoch t, the probability of a node S_i having no top-k data item is given by

$$p_o = \frac{\binom{(gn-1)\mu}{k}}{\binom{gn\mu}{k}}. \tag{9}$$

There are thus gnp_o qualified nodes and $gn(1-p_o)$ unqualified nodes on average. For each of the top-k data items, one MAC needs to be transmitted. For each qualified node, at most one additional data item and one index need to be transmitted. We thus have

$$C_1 \le kl_{\text{mac}} + gn(1-p_o)(l_{\text{data}} + l_{\text{mac}}) \tag{10}$$

where p_o is given in (9).

We now analyze C_2. Similar to the analysis of p_o, the probability that a subcell has no top-k data item is given by

$$\alpha = \frac{\binom{(g-1)n\mu}{k}}{\binom{gn\mu}{k}}. \tag{11}$$

The expected number of subcells with at least one top-k data items is thus $g(1-\alpha)$. On average, each such subcell has $\beta = (n(1-p_o)/1-\alpha)$ qualified nodes, each of which has its ID embedded in at most $\beta-1$ auxiliary ID lists. We thus have

$$C_2 \le ng(1-\alpha)\beta(\beta-1)l_{\text{id}}. \tag{12}$$

We now derive C_3. The expected number of subcells with no top-k data item is $g\alpha$. For each of them, the master node needs to return one node ID, one data item, and one MAC. We thus have

$$C_3 = g\alpha(l_{\text{id}} + l_{\text{data}} + l_{\text{mac}}) \tag{13}$$

where α is given in (11).

Combining (10), (12), and (13), we have

$$\begin{aligned} C_{\text{query}} &= C_1 + C_2 + C_3 \\ &\le kl_{\text{mac}} + gn(1-p_o)(l_{\text{data}} + l_{\text{mac}}) \\ &\quad + ng(1-\alpha)\beta(\beta-1)l_{\text{id}} + g\alpha(l_{\text{id}} + l_{\text{data}} + l_{\text{mac}}) \end{aligned}$$

where p_o is given in (9), α is given in (11), and $\beta = (n(1-p_o)/1-\alpha)$. ■

We have not been able to find a closed-form solution for more general cases, which we will evaluate using simulations in Section VII.

B. Analysis of RP

We have the following theorem regarding the detection probability of RP against the overshadowing attack.

Theorem 4: Assume that c out of N sensor nodes are compromised. The detection probability of RP against an overshadowing attack is bounded by

$$P_{\text{det}} > 1 - \left(\frac{c}{N}\right)^{\theta}. \tag{14}$$

Proof: Since $c \ll N$, we can view each probed sensor node as being compromised with probability $p_c = c/N$. Assume that the adversary launched overshadowing attacks in $e \ge 1$ subcells. Consider one such subcell $\mathcal{C}_y$ as an example. Since the network owner probes θ randomly chosen nodes in each subcell, he cannot detect the overshadowing attack in $\mathcal{C}_y$ if all the probed nodes are compromised, which happens with probability $(c/N)^{\theta}$. He cannot detect any overshadowing attack in the query region if all θe probed nodes are compromised. We thus have

$$P_{\text{det}} = 1 - \left(\frac{c}{N}\right)^{\theta e}.$$

■

We now estimate the communication cost incurred by RP. Consider a probed node S_i as an example, from which one data item $D_{i,1}$, one MAC $\mathsf{V}_{i,1}$, and one auxiliary ID list $\mathcal{L}_{i,1}$ need to be returned. Since S_i is randomly chosen, the expected number of IDs in $\mathcal{L}_{i,1}$ is $(n-1)/2$, i.e., about half of the nodes have highest scores higher than $s_{i,1}$. We thus have

$$C_{\mathsf{RP}} = \theta d \left(l_{\text{data}} + l_{\text{mac}} + \frac{(n-1) l_{\text{id}}}{2} \right) \tag{15}$$

where d is the number of subcells that overlap with the query region.

C. Analysis of QC

The following theorem is about the effectiveness of QC.

Theorem 5: Assume that $\mathcal{I}_t = \mathcal{I}$ and that c out of N sensor nodes are compromised, each of which generates up to μ data items with extremely large values. If the network owner converts a top-k query $\mathcal{Q}_t = \langle \mathcal{C}, t, k, \mathcal{I}_t \rangle$ into a δ-constrained top-k' query, the probability that the query result of $\mathcal{Q}_t^c$ contains the true top-k data items generated by legitimate sensor nodes is given by

$$P_{\text{true}} = \begin{cases} 0, & \text{if } \delta c + k > k' \\ \frac{P_1}{P_2}, & \text{otherwise} \end{cases} \tag{16}$$

where

$$P_1 = \sum_{\substack{0 \le x_j \le \delta, \forall\, j \in [1, N-c] \\ \sum_{j=1}^{N-c} x_j = k}} \prod_{j=1}^{N-c} \Pr[k_j = x_j]$$

$$P_2 = \sum_{\sum_{j=1}^{N-c} x_j = k} \prod_{j=1}^{N-c} \Pr[k_j = x_j]$$

$$\Pr[k_j = x] = \binom{\mu}{x} p^x (1-p)^{\mu - x}$$

$$p = \frac{k}{(N-c)\mu}.$$

Proof: First, we have $P_c = 0$ if $\delta c + k > k'$, since k' is not large enough to tolerate all the forged data items from compromised nodes in $\mathcal{I}$.

Now, consider the case $\delta c + k \le k'$. Without loss of generality, denote by $i_1, \ldots, i_{N-c}$ the IDs of legitimate sensor nodes. In addition, denote by k_j the number of true top-k data items generated by node S_{i_j}. The query result of $\mathcal{Q}_t^c$ contains the true top-k data items from the legitimate sensor nodes if $k_j \le \delta$, for all $j \in [1, N-c]$. Assume that each data item is equally likely to be among the top k. Since there are total $(N-c)\mu$ data items generated by legitimate sensor nodes, the probability of any data item being among the true top k is given by

$$p = \frac{k}{(N-c)\mu}. \tag{17}$$

When p is small, whether each data item being among the true top k can be viewed as an independent event. We can thus approximate k_j as a binomial random variable with a probability density function given by

$$\Pr[k_j = x] = \begin{cases} \binom{\mu}{x} p^x (1-p)^{\mu - x}, & \text{if } 0 \le x \le \mu \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Denote by $\mathbf{E}_1$ the event that $k_j \le \delta$ for all $j \in [1, N-c]$ and $\mathbf{E}_2$ the event that $\sum_{j=1}^{N-c} k_j = k$. We have

$$P_c = \Pr[\mathbf{E}_1 | \mathbf{E}_2] = \frac{\Pr[\mathbf{E}_1, \mathbf{E}_2]}{\Pr[\mathbf{E}_2]}. \tag{19}$$

We then have

$$\begin{aligned} \Pr[\mathbf{E}_1, \mathbf{E}_2] &= \Pr\left[k_1 \le \delta, \ldots, k_{N-c} \le \delta, \sum_{j=1}^{N-c} k_j = k \right] \\ &= \sum_{\substack{0 \le x_j \le \delta, \forall\, j \in [1, N-c] \\ \sum_{j=1}^{N-c} x_j = k}} \Pr\left[k_j = s_j, \forall\, j \in [1, N-c] \right] \\ &= \sum_{\substack{0 \le x_j \le \delta, \forall\, j \in [1, N-c] \\ \sum_{j=1}^{N-c} x_j = k}} \prod_{j=1}^{N-c} \Pr[k_j = x_j] \end{aligned} \tag{20}$$

where $\Pr[k_j = x_j]$ is given in (18).

Similarly, we have

$$\begin{aligned} \Pr[\mathbf{E}_2] &= \Pr\left[\sum_{j=1}^{N-c} k_j = k \right] \\ &= \sum_{\sum_{j=1}^{N-c} x_j = k} \Pr\left[k_j = x_j, \forall\, j \in [1, N-c] \right] \\ &= \sum_{\sum_{j=1}^{N-c} x_j = k} \prod_{j=1}^{N-c} \Pr[k_j = x_j] \end{aligned} \tag{21}$$

where $\Pr[k_j = x_j]$ is given in (18).

Substituting (20) and (21) into (19), we can then obtain (16) and prove the theory. ■

VII. Simulation Results

Here, we evaluate the performance of the proposed schemes using simulations.

We assume a cell of 1000×1000 m^2 with 400 sensor nodes randomly distributed and a master node at the center. Each sensor node has a transmission range of 100 m, leading to an average distance to the master node of $L = 3.7$ hops. We partition the cell into 25 subcells, each containing 16 sensor nodes. We also assume error-free and collision-free packet transmissions. For our purpose, the simulation code is written in C++, and each data point represents an average of 100 simulation runs, each with a different random seed. Table I summarizes the default setting used in our simulation if not mentioned otherwise.

TABLE I
DEFAULT SIMULATION PARAMETERS

Para.	Val.	Para.	Val.	Para.	Val.	Para.	Val.
N	400	k	10	n	100	μ	10
l_{data}	160	l_{score}	16	l_{mac}	160	l_{id}	16
c	10	s_c	2				

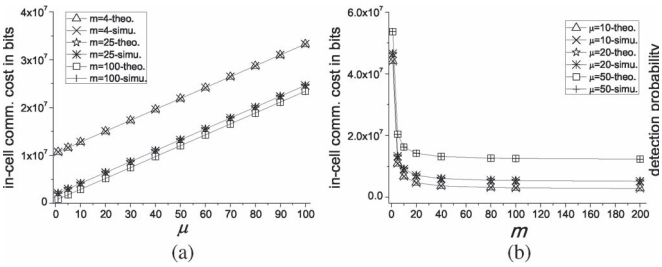


Fig. 3. Impact of μ and m on the in-cell communication cost of VTQ. (a) C_{cell} versus μ. (b) C_{cell} versus m.

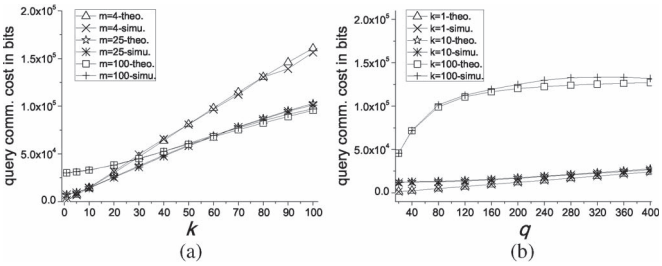


Fig. 4. Impact of k and q on query communication cost of VTQ. (a) C_{query} versus k. (b) C_{query} versus q.

A. Performance of VTQ

Since VTQ can detect any fake or unsound top-k query result returned by a compromised master node given that none of the sensor nodes are compromised, we here focus on the in-cell and query communication costs incurred by VTQ.

Fig. 3(a) shows the theoretical and simulation results of the in-cell communication cost of VTQ varying with μ, which is the number of data items generated by each node per epoch, where $m = 4, 25$ and 100, respectively. We can see that the theoretical results match the simulation results very well. Moreover, the in-cell communication cost linearly increases as μ increases. The reason is that the communication costs incurred by exchanging highest scores among each subcell is independent of μ while one MAC needs to be transmitted for each data item, resulting in a linear relationship between the in-cell communication cost and μ.

Fig. 3(b) shows the theoretical and simulation results of the in-cell communication cost of VTQ varying with m, which is the number of subcells. We can see that the theoretical results match the simulation results very well. Moreover, the in-cell communication cost rapidly decreases as the number of subcells increases. This is anticipated because the communication cost incurred by exchanging the highest scores among each subcell is proportional to the size of the subcell and, thus, inversely proportional to m [cf. (6)]. Therefore, a small m would incur significant in-cell communication cost.

Fig. 4(a) shows the theoretical and simulation results of the query communication cost of VTQ varying with k, which is the number of data items queried. We can see that C_{query} increases as k increases. The reason is that the more data items queried, the more information (e.g., additional data items and MACs) is needed to prove the authenticity and soundness of the query result, which can be easily understood. Moreover, when k is small, the larger m is, the higher the query communication cost. This is because when k is small, there will be many unqualified subcells, for each of which some information needs to be returned, leading to higher query communication cost. On the other hand, when k is large, the smaller m is, the higher the query communication cost. The reason is that as m increases, the number of unqualified subcells decreases, and the average number of IDs in each auxiliary ID list increases. Therefore, more node IDs are embedded into the data items returned, leading to higher communication cost. In general, small m may lead to higher query communication cost when k is small, so does large m when k is large.

Fig. 4(b) shows the theoretical bound and simulation results of the query communication cost of VTQ varying with the number of nodes queried, which is denoted by q. We can see that the query communication cost increases as the number of nodes in the query region increases. The reason is that for fixed k, the larger the query region is, the more candidate subcells, the more unqualified subcells, and the higher the query communication cost, and vice versa. In addition, we can also see that when k is relatively large, e.g., $k = 100$, the query communication cost rapidly increases as q increases from 20 to 100 and then slowly as q further increases. The reason is that the number of qualified nodes increases as q increases before q exceeds k. For each additional qualified node, one additional data item needs to be returned under VTQ, leading to a rapid increase in query communication cost. After q exceeds k, the number of unqualified subcells slowly increases as q further increases, leading to a slow increase in query communication cost.

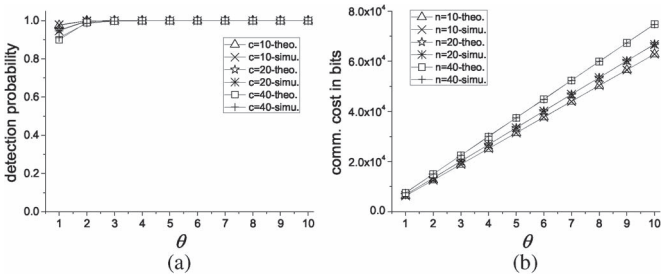


Fig. 5. Impact of θ on the detection probability and communication cost of RP. (a) P_{det} versus θ. (b) C_{RP} versus θ.

B. Performance of RP

Fig. 5(a) shows the theoretical and simulation results of the detection probability of RP against the overshadowing attack varying with θ, which is the number of nodes probed in each subcell. We can see that the theoretical results match the simulation result very well. Moreover, the more nodes probed in each candidate subcell, the higher the detection probability against overshadowing attack. The reason is that the overshadowing attack cannot be detected only if all the probed nodes are

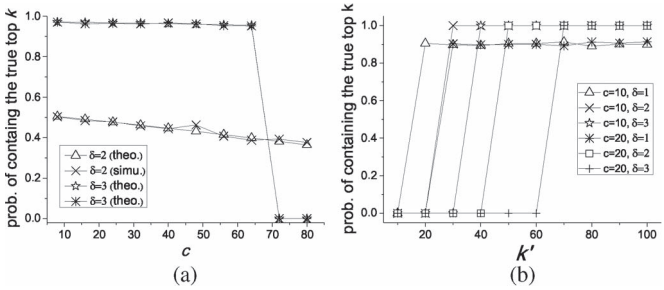


Fig. 6. Impact of c and k' on QC, where $k = 100$ and $k' = 300$. (a) P_{true} versus c. (b) C_{QC} versus m.

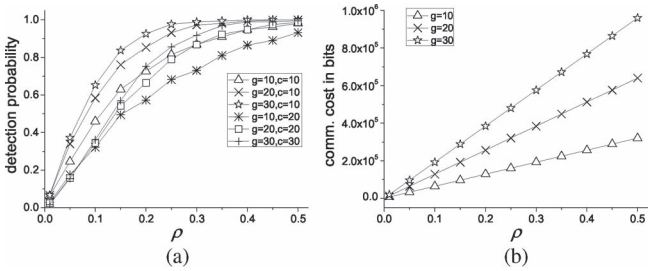


Fig. 7. Impact of witness ratio ρ on the framing detection probability and communication cost of RW. (a) Detection probability. (b) Communication cost.

compromised and that the probability that at least one probed node is not compromised increases as θ increases. We can see that even 10% of the sensor nodes are compromised, the detection probability is higher than 0.98 when $\theta = 2$ and close to one as θ further increases. It is thus unnecessary to choose a large θ in practice.

Fig. 5(b) shows the theoretical and simulation results of the additional communication cost incurred by RP varying with the number of candidate subcells. It is easy to see that the communication cost linearly increases as the number of candidate subcells increases, which is anticipated. This also implies that for a fixed query region, the communication cost incurred by RP increases as the total number of subcells increases as there will be more candidate subcells.

C. Performance of QC

Fig. 6(a) shows the theoretical results and simulation results of P_{true}, the probability of the query result containing true top k varying with c, the number of compromised sensor nodes, where $k = 100$ and $k' = 300$, respectively. We can see that P_{true} first decreases slowly as c increases and then drops to zero after c exceeds 65. The reason can be explained as follows. When c is smaller than the threshold $(k' - k)/\delta$ [cf. (16)], the query result contains the true top-k data items if none of the legitimate sensor nodes have more than δ qualified data items. As the number of compromised nodes increases, the number of legitimate nodes decreases, and the probability of at least one legitimate sensor node has more than δ increases, as the same number of qualified data items are allocated among fewer legitimate nodes. Once c exceeds $(k' - k)/\delta$, the query result can no longer tolerate all the $c\delta$ forged data items, and P_{true} thus drops to zero. Moreover, we can see that the choice of δ affects P_{true}. In particular, when $\delta = 2$, P_{true} is about 0.5 even if none of the sensor nodes are compromised. The reason is that it is very likely that a legitimate sensor node can have more than two qualified data items. On the other hand, when $\delta = 3$, P_{true} is higher than 0.95 when the number of compromised sensor nodes is smaller than $(k' - k)/\delta$, but drops to zero as c exceeds 65.

Fig. 6(b) shows P_{true} varying with the k', the number of compromised sensor nodes, where $k = 10$. We can see that the probability remains zero before k' exceeds the threshold $k + c\delta$, as k' is not large enough to tolerate all the $c\delta$ forged data items. After k' exceeds the threshold, the probability significantly increases and remains constant as k' further increases. The probability is not one because it is still possible that one sensor node has more than δ qualified data items.

In general, smaller δ leads to lower P_{true} but could tolerate more compromised sensor nodes.

D. Performance of WT

To simulate the performance of WT, we assume the worst case in which the sensor node that launches the framing attack is one hop away from the master node and, thus, has the least number of witnesses on average for fixed witness ratio $\rho = Y/X$.

Fig. 7(a) shows the detection probability against a framing attack varying with witness ratio ρ. We can see that the detection probability increases as the witness ratio increases. This is anticipated since the higher the ratio ρ is, the more witnesses are selected for each message transmission. The network owner can detect the framing attack as long as the number of legitimate witnesses is larger than that of compromised witnesses. Moreover, the higher the node density, the more neighbors each node has, the more witnesses, and vice versa. In practice, the ratio ρ should be chosen according to the node density, i.e., the higher the node density, the lower the ratio.

Fig. 7(b) shows the communication cost incurred by WT varying with ρ. We can see that the communication cost linearly increases as witness ratio ρ increases. The reason is that each witness node needs to transmit one testimony. Hence, the higher the ratio ρ is, the more witnesses for each message transmission, the higher the communication cost, and vice versa. Since each testimony $\mathsf{T}_{i,j,t}$ is essentially a MAC, which is much shorter than a data item, the communication cost incurred by transmitting testimonies is relatively small in comparison with that incurred by data submissions.

E. Discussion

We summarize the evaluation results as follows.

- VTQ can detect any fake and/or unsound top-k query result returned by a compromised master node provided that none of the sensor nodes are compromised. The in-cell and query communication costs of VTQ can be adjusted by choosing proper m, which is the number of subcells. Small m leads to high in-cell communication cost and low query communication cost when k is small, whereas large

m leads to low in-cell communication cost and high query communication cost when k is large.

- RP can detect an unsound top-k query result returned by colluding compromised master and sensor nodes with very high probability and incurs low communication cost.
- QC can tolerate forged data items from compromised sensor nodes by increasing the number of data items queried while limiting the number of qualified data items that can be returned from each candidate node.
- WT can detect possible framing attacks against a legitimate master node with high probability and incurs low communication cost.

In practice, all four schemes should be deployed together to enable verifiable top-k query processing in UTSNs. Built upon symmetric cryptographic primitives, our schemes are very suitable and practical for resource-constrained sensor networks.

VIII. Related Work

Here, we discuss some work most germane to our work.

Top-k queries are a common and important type of queries in sensor networks. Tremendous efforts have been devoted to realizing efficient top-k query processing in sensor networks (see, for example, [9], [10], and [26]–[29]). These works nevertheless do not take security issues into account.

Verifiable data queries in UTSNs have received attention only recently. In [7], and [30], Sheng and Li proposed a novel scheme to enable verifiable privacy-preserving 1-D range queries in UTSNs, which is subsequently improved by Shi *et al.* in [11]. Secure multidimensional range queries are later addressed in [12], [13], [25], and [31]. None of these schemes can be applied to top-k queries. While verifiable top-k queries against a compromised master node was tackled in [1], the impact of and defense against compromised sensor nodes were untouched.

Secure top-k queries can be viewed as a special instance of secure aggregation. In [32], Nath *et al.* proposed a set of secure aggregation schemes for wide-area sensing, including top-k queries. Their schemes rely on public-key cryptographic operation, i.e., RSA encryption, and are thus unsuitable for resource-constrained sensor networks.

Our work is also loosely related to secure data outsourcing [33], in which a data owner outsources its data to a third-party service provider answering the data queries on behalf of the data owner. Significant effort has been devoted to ensuring query integrity, i.e., that a query result was indeed generated from the outsourced data and contains all the data satisfying the query (the soundness requirement). Many techniques were proposed to realize a wide range of data queries, such as relational query [34]–[36], location-based range queries [37], [38], shortest path queries [39], and moving kNN queries [39]. None of these schemes consider top-k queries and, thus, are not applicable to our scenario.

IX. Conclusion

In this paper, we have presented a suite of novel schemes to secure top-k queries in UTSNs against a wide range of attacks from compromised master and/or sensor nodes. The proposed schemes enable the network owner to verify the authenticity and soundness of any top-k query results. Detailed analysis and simulation results confirm the high efficacy and efficiency of the proposed schemes. In the future, we intend to investigate the verifiability of other types of data queries in UTSNs.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and helpful advice.

References

[1] R. Zhang, J. Shi, Y. Liu, and Y. Zhang, "Verifiable fine-grained top-k queries in tiered sensor networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[2] R. D. Pietro, L. V. Mancini, C. Soriente, A. Spognardi, and G. Tsudik, "Catch me (if you can): Data survival in unattended sensor networks," in *Proc. IEEE PerCom*, Hong Kong, Mar. 2008, pp. 185–194.

[3] D. Ma, C. Soriente, and G. Tsudik, "New adversary and new threats: Security in unattended sensor networks," *IEEE Netw.*, vol. 23, no. 2, pp. 43–48, Mar. 2009.

[4] P. Desnoyers, D. Ganesan, and P. Shenoy, "TSAR: A two tier sensor storage architecture using interval skip graphs," in *Proc. ACM SenSys*, San Diego, CA, USA, Nov. 2005, pp. 39–50.

[5] B. Sheng, Q. Li, and W. Mao, "Data storage placement in sensor networks," in *Proc. ACM MobiHoc*, Florence, Italy, May 2006, pp. 344–355.

[6] M. Shao, S. Zhu, W. Zhang, and G. Cao, "pDCS: Security and privacy support for data-centric sensor networks," in *Proc. IEEE INFOCOM*, Anchorage, AK, USA, May 2007, pp. 1298–1306.

[7] B. Sheng and Q. Li, "Verifiable privacy-preserving range query in sensor networks," in *Proc. IEEE INFOCOM*, Phoenix, AZ, USA, Apr. 2008, pp. 46–50.

[8] O. Gnawali, K.-Y. Jang, J. Paek, M. Vieira, R. Govindan, B. Greenstein, A. Joki, D. Estrin, and E. Kohler, "The Tenet architecture for tiered sensor networks," in *Proc. ACM SenSys*, Boulder, CO, USA, Oct. 2006, pp. 153–166.

[9] G. Das, D. Gunopulos, N. Koudas, and D. Tsirogiannis, "Answering top-k queries using views," in *Proc. VLDB*, Sep. 2006, pp. 451–462.

[10] M. Ye, X. Liu, W.-C. Lee, and D. L. Lee, "Probabilistic top-k query processing in distributed sensor networks," in *Proc. IEEE ICDE*, Long Beach, CA, USA, Mar. 2010, pp. 585–588.

[11] J. Shi, R. Zhang, and Y. Zhang, "Secure range queries in tiered sensor networks," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 945–953.

[12] R. Zhang, J. Shi, and Y. Zhang, "Secure multidimensional range queries in sensor networks," in *Proc. ACM MobiHoc*, New Orleans, LA, USA, May 2009, pp. 197–206.

[13] F. Chen and A. Liu, "SafeQ: Secure and efficient query processing in sensor networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[14] X. Cheng, A. Thaeler, G. Xue, and D. Chen, "TPS: A time-based positioning scheme for outdoor wireless sensor networks," in *Proc. IEEE INFOCOM*, Hong Kong, Mar. 2004, pp. 2685–2696.

[15] D. Liu, P. Ning, A. Liu, C. Wang, and W. Du, "Attack-resistant location estimation in wireless sensor networks," *ACM Trans. Inf. Syst. Security*, vol. 11, no. 4, pp. 1–39, Jul. 2008.

[16] D. Liu and P. Ning, "Multilevel μ TESLA: Broadcast authentication for distributed sensor networks," *ACM Trans. Embedded Comput. Syst.*, vol. 3, no. 4, pp. 800–836, Nov. 2004.

[17] N. Subramanian, C. Yang, and W. Zhang, "Securing distributed data storage and retrieval in sensor networks," in *Proc. IEEE PerCom*, White Plains, NY, USA, Mar. 2007, pp. 191–200.

[18] Y. Jian, S. Chen, Z. Zhang, and L. Zhang, "A novel scheme for protecting receiver's location privacy in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3769–3779, Oct. 2008.

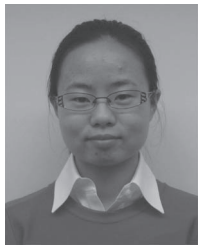
[19] F. Liu, X. Cheng, L. Ma, and K. Xing, "SBK: A self-configuring framework for bootstrapping keys in sensor networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 7, pp. 858–868, Jul. 2008.

[20] Q. Wang, K. Ren, W. Lou, and Y. Zhang, "Dependable and secure sensor data storage with dynamic integrity assurance," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 954–962.

[21] R. Lu, X. Lin, H. Zhu, and X. Shen, "TESP2: Timed efficient source privacy preservation scheme for wireless sensor networks," in *Proc. IEEE ICC*, May 2010, pp. 1–6.
[22] R. Zhang, Y. Zhang, and K. Ren, "DP^2 AC: Distributed privacy-preserving access control in sensor networks," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 1251–1259.
[23] H. Zhu, S. Du, M. Li, and Z. Gao, "Fairness-aware and privacy-preserving friend matching protocol in mobile social networks," *IEEE Trans. Emerging Topics Comput.*, vol. 1, no. 1, pp. 192–200, Jun. 2013.
[24] H. Zhu, S. Du, Z. Gao, M. Dong, and Z. Cao, "A probabilistic misbehavior detection scheme toward efficient trust establishment in delay-tolerant networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 22–32, Jan. 2014.
[25] R. Zhang, J. Shi, Y. Zhang, and J. Sun, "Secure cooperative data storage and query processing in unattended tiered sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 433–441, Feb. 2012.
[26] A. S. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang, "A sampling-based approach to optimizing top-k queries in sensor networks," in *Proc. ICDE*, Atlanta, GA, USA, Apr. 2006, pp. 68–78.
[27] M. Wu, J. Xu, X. Tang, and W.-C. Lee, "Top-k monitoring in wireless sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 962–976, Jul. 2007.
[28] B. Malhotra, M. A. Nascimento, and I. Nikolaidis, "Exact top-k queries in wireless sensor networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 10, pp. 1513–1525, Oct. 2011.
[29] B. Chen, W. Liang, R. Zhou, and J. X. Yu, "Energy-efficient top-k query processing in wireless sensor networks," in *Proc. CIKM*, Toronto, ON, Canada, Oct. 2010, pp. 329–338.
[30] B. Sheng and Q. Li, "Verifiable privacy-preserving sensor network storage for range query," *IEEE Trans. Mobile Comput.*, vol. 10, no. 9, pp. 1312–1326, Sep. 2011.
[31] Y. Yi, R. Li, F. Chen, A. X. Liu, and Y. Lin, "A digital watermarking approach to secure and precise range query processing in sensor networks," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1950–1958.
[32] S. Nath, H. Yu, and H. Chan, "Secure outsoured aggregation via one-way hash chains," in *Proc. ACM SIGMOD*, Providence, RI, USA, Jun. 2009, pp. 31–44.
[33] H. Hacigümüs, S. Mehrotra, and B. Iyer, "Providing database as a service," in *Proc. IEEE ICDE*, Aalborg, Denmark, Feb. 2002, pp. 1950–1958.
[34] M. Narasimha and G. Tsudik, "Authentication of outsourced databases using signature aggregation and chaining," in *Proc. DASFAA*, Singapore, Apr. 2006, pp. 420–436.
[35] H. Pang and K.-L. Tan, "Verifying completeness of relational query answers from online servers," *ACM Trans. Inf. Syst. Security*, vol. 11, no. 2, pp. 1–50, Mar. 2008.
[36] H. Pang, J. Zhang, and K. Mouratidis, "Scalable verification for outsourced dynamic databases," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 802–813, Aug. 2009.
[37] Y. Yang, S. Papadopoulos, D. Papadias, and G. Kollios, "Spatial outsourcing for location-based services," in *Proc. IEEE ICDE*, Cancún, México, Apr. 2008, pp. 1082–1091.
[38] W.-S. Ku, L. Hu, C. Shahabi, and H. Wang, "Query integrity assurance of location-based services accessing outsourced spatial databases," in *Proc. Int. Symp. Adv. Spatial Temporal Databases*, Aalborg, Denmark, Jul. 2009, pp. 80–97.
[39] M. Yiu, Y. Lin, and K. Mouratidis, "Efficient verification of shortest path search via authenticated hints," in *Proc. IEEE ICDE*, Long Beach, CA, USA, Mar. 2010, pp. 237–248.

Rui Zhang (M'13) received the B.E. degree in communication engineering and the M.E. degree in communication and information systems from Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively, and the Ph.D. degree in electrical engineering from the Arizona State University, Tempe, AZ, USA, in 2013.

From 2005 to 2007, he was a Software Engineer with the UTStarcom Shenzhen R&D Center, Shenzhen, China. Since July 2013, he has been an Assistant Professor with the Department of Electrical Engineering, University of Hawaii, Honolulu, HI, USA. His primary research interests include network and distributed system security, wireless networking, and mobile computing.



Jing Shi received the B.E. degree in communication engineering and the M.E. degree in communication and information systems from Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical and computer engineering from New Jersey Institute of Technology, Newark, NJ, USA, in 2010.

She is currently a Lecturer with the School of Public Administration, Huazhong University of Science and Technology. Her research interests include network and distributed system security, wireless networking, and mobile computing.

Yanchao Zhang (SM'11) received the B.E. degree in computer science and technology from Nanjing University of Posts and Telecommunications, Nanjing, China, in 1999; the M.E. degree in computer science and technology from Beijing University of Posts and Telecommunications, Beijing, China, in 2002; and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2006.

From 2006 to 2010, he was an Assistant Professor of electrical and computer engineering with the New Jersey Institute of Technology, Newark, NJ, USA. He is currently as an Associate Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA. His primary research interests include network and distributed system security, wireless networking, and mobile computing.

Dr. Zhang is an Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and a Feature Editor of the IEEE WIRELESS COMMUNICATIONS. He was a Guest Editor of the IEEE WIRELESS COMMUNICATIONS Special Issue on Security and Privacy in Emerging Wireless Networks in 2010 and a Technical Program Committee Cochair of the Communication and Information System Security Symposium, IEEE GLOBECOM 2010. He received the National Science Foundation CAREER Award in 2009.

Xiaoxia Huang (M'11) received the B.E. and M.E. degrees in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 2000 and 2002, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2007.

She is currently an Associate Researcher with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. She is the Deputy Director of the Center for Real-time Monitoring and Communications Technology. She has published over 20 papers in refereed professional journals and conferences and served as a reviewer for many refereed journals and conferences. Her research interests include cognitive radio networks, wireless sensor networks, wireless communications, and mobile computing.

Dr. Huang served as a Technical Program Committee Member of the IEEE Wireless Communications and Networking Conference (WCNC 2011), the IEEE International Conference on Communications (ICC 2011), the IEEE Global Communications Conference (GLOBECOM 2011), the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine 2010), and the International Conference on Embedded and Ubiquitous Computing (EUC 2010).