Hindawi
Security and Communication Networks
Volume 2021, Article ID 6234409, 10 pages
https://doi.org/10.1155/2021/6234409

Research Article

EPCT: An Efficient Privacy-Preserving and Collusion-Resisting Top-*k* Query Processing in WSNs

Qian Zhou,[1] **Hua Dai**,[1,2] **Jianguo Zhou**,[1] **Rongqi Qi**,[1] **Geng Yang**,[1,2] **and Xun Yi**[3]

[1]*Nanjing University of Post and Telecommunication, Nanjing 210023, China*
[2]*Jiangsu Security and Intelligent Processing Lab of Big Data, Nanjing 210023, China*
[3]*Royal Melbourne Institute of Technology University, Melbourne 3001, Australia*

Correspondence should be addressed to Qian Zhou; zhouqian@njupt.edu.cn

Received 8 September 2021; Revised 28 September 2021; Accepted 7 October 2021; Published 15 November 2021

Academic Editor: Weizhi Meng

Data privacy threat arises during providing top-*k* query processing in the wireless sensor networks. This article presents an efficient privacy-preserving and collusion-resisting top-*k*(EPCT) query processing protocol. A minimized candidate encrypted dataset determination model is first designed, which is the foundation of EPCT. The model guides the idea of query processing and guarantees the correctness of the protocol. The symmetric encryption with different private key in each sensor is deployed to protect the privacy of sensory data even a few sensors in the networks have been colluding with adversaries. Based on the above model and security setting, two phases of interactions between the interested sensors and the sink are designed to implement the secure query processing protocol. The security analysis shows that the proposed protocol is capable of providing secure top-*k* queries in the manner of privacy protection and anticollusion, whereas the experimental result indicates that the protocol outperforms the existing works on communication overhead.

1. Introduction

Wireless sensor networks (WSNs), as one of the important technologies in the Internet of Things (IoT), have been widely deployed to provide practical solutions in various applications, such as environment monitoring, military target sensing, and smart home application. Meanwhile, data privacy leakage in WSNs is becoming the main obstruction, which slows down its further development. For example, in the scenario of a smart home application, videos or pictures collected by wireless IP-cameras could be eavesdropped for illegal profit. As a result, privacy protection on sensitive data is a critical issue that must be addressed in WSNs.

In WSNs, the top-*k* query is one of the critical operations in data aggregation for sensor monitoring process. The top-*k* query requests the *k* lowest or highest data items collected from IoT sensors in WSNs. For example, “collecting the 10 lowest humidity data in forest area A-Z in last 2 hours” is an example of top-*k* query, which can be performed for fire monitoring. Our aim of this work is to design a secure top-*k* query approach with privacy-preserving and collusion-resisting manners.

This article presents an efficient privacy-preserving and collusion-resisting top-*k* query processing protocol (EPCT) in WSNs. We first propose a minimized candidate encrypted dataset determination model, which is the foundation of our proposed protocol. It guides the idea of query processing and guarantees the correctness of the protocol. There are two phases of interactions between the queried sensors and the sink in EPCT. In the first phase, when the queried sensors receive a top-*k* query from the sink, they first use their own private keys to encode the maximum of the collected data in the interested time slot, respectively, and then, they submit the encrypted data to the sink. In the second phase, the sink decrypts the received ciphertext and calculates the candidate sensors; after that, it unicastly informs the candidate sensors to submit the rest candidate data. Once the sink obtains enough data from the candidate sensors, the final result of the query is determined. The security analysis and performance evaluation indicate that the proposed approach

EPCT has the ability of protecting data privacy and performing efficiently in transmission overhead.

The main contributions of this article are listed as follows:

(i) We present a minimized candidate encrypted dataset determination model, which is the foundation of our proposed scheme. It guides the idea of query processing and guarantees the correctness of the protocol.

(ii) We present a novel privacy-preserving and collusion-resisting top-k query processing protocol, which consists of two phases of secure interactions between the queried nodes and the sink. We also analyse the correctness, security, and transmission overhead of the proposed method.

(iii) We perform evaluations on the transmission overhead of the proposed protocol and the existing works. The experimental result shows the advantages of the proposed scheme in transmission overhead.

The remainder of this article is organized as follows. Section 2 discusses the related work. Section 3 introduces the network model, query model, threat model, and the problem description. Section 4 proposes the minimized candidate encrypted dataset determination model. Section 5 presents the top-k query processing protocol and the analysis of this protocol. Section 6 presents the performance evaluation of query protocols on communication cost, and Section 7 gives a conclusion of this article.

2. Related Work

Secure data queries (such as top-k query, range query, and MAX/MIN query) are critical operations for sensor monitoring and data collection in security-sensitive environment. There are a lot of works [1–24] focusing confidentiality, integrity, and completeness when performing data queries.

Kui et al. [3] utilize the pairwise-key and order-preserving symmetric encryption and the together to protect the privacy of data in top-k queries in two-tiered WSNs. Peng et al. [4] encoded both sensory data and top-k query commands, and storage nodes are designed to be able to correctly perform top-k queries over those encoded data. Li et al. [6] use pseudorandom hash function with bloom filter and partition algorithm to protect data privacy and integrity for top-k queries, respectively. Tsou et al. [7] constructed a layered authentication tree by an order-preserving symmetric encryption and used it to verify the completeness of query results. Zhang et al. [12] designed a renormalized arithmetic coding method such that storage nodes can calculate exact top-k query results without knowing real values of data, and they proposed a verification scheme to detect compromised storage nodes. Peng et al. [13] encoded top-k queries by threshold-based scheme and proposed a secure protocol that storage nodes can calculate query results over encrypted sensory data. Xingpo et al. [14] proposed secure top-k query protocol with privacy and integrity preservation by deploying the secure data preprocessing in sensor nodes. Wu and Wang [17] bound the collected sensory data with the corresponding locations to achieve secure top-k query processing on hybrid sensory data. Liu et al. [18] proposed a verifiable top-k query protocol on two-tiered mobile sensor network, which adopts the distinct symmetric data encryption and maps real nodes into virtual nodes. These methods are designed for two-tiered WSNs, which adopt resource-rich storage nodes in traditional multihop WSNs. The different network architecture makes them not suitable for addressing secure top-k queries in traditional multihop WSNs.

In traditional multihop WSNs, the earlier studies [19, 25–29] proposed various top-k query schemes but without concerning any security issues. Huang et al. [30] designed a privacy-protection top-k query algorithm using a filter and a data distribution table. The algorithm adopts conic section function to protect the privacy of the sensory data. But, the algorithm is vulnerable when collusion attacks happen. It is because all sensor nodes share the same secure keys and functions. If a sensor node colludes with adversaries, these secure keys and functions will be disclosed, and the adversaries could obtain the private data of other innocent sensors. In our previous work [31], we gave the first solution providing the privacy-protecting and anticollusion top-k query processing scheme in wireless sensor networks. It adopts the bloom filter and HMAC when performing interactions between nodes and the sink to achieve secure top-k query processing. However, there is some space for transmission overhead saving because of the redundant data submission and the false positive of bloom filter. This article presents an efficient and secure top-k query processing protocol, which can address the above problems.

Additionally, some previous studies have focused on the privacy-preserving range queries [8, 11, 32] and MAX/MIN queries [15, 16] in WSNs. Because the query types are different, the ideas of these works cannot be applied to achieve the secure top-k queries in WSNs.

3. Problem Description

3.1. Network Model. The architecture adopted is shown in Figure 1. The network routing topology is structured as a tree, which is following TAG protocol [33]. Assuming that in our scenario, n sensors $S = \{s_1, s_2, \ldots, s_n\}$ are deployed and a sink. Sensor nodes are sensory devices with limited resources in energy, storage, and computation. They are in charge of collecting data items from their neighboring areas and then submitting the collected data to the sink through the tree route. The sink is a resourceful device, which executes query commands from users and returns query results to users. When receiving a query command, the sink cooperates with those queried sensors in S to process queries according to predeployed protocols. After the sink obtains the query result, it returns the result to the upper-level users.

3.2. Top-k Query Model. A top-k query is a data aggregation operation to get k highest or lowest sensory data from queried sensors. It is denoted as a triple $\text{Query}_t = (t, S, k)$

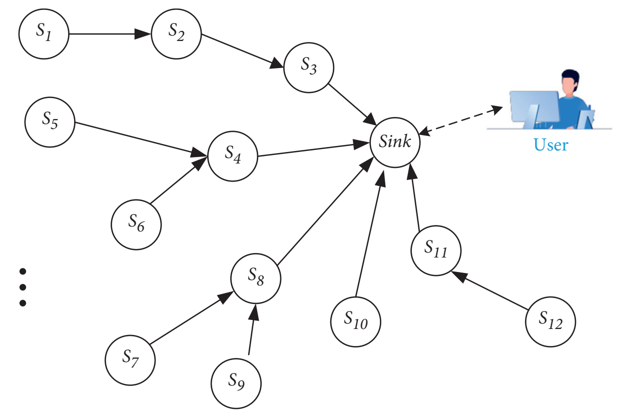

Figure 1: A example of tree routing topology.

where t is the queried time slot identity, S is the set of interested sensors, and k is the number of interested data items. For example, $(t, \{s_1, s_2, \ldots, s_{12}\}, 3)$ is a top-3 query to obtain the 3 highest or lowest data items during sensors $\{s_1, s_2, \ldots, s_{12}\}$ in the time slot t.

Each sensor $s_i \in S$ is assumed to collect N data items in a time slot, which is denoted as $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,N}\}$, and each data item collected by a sensor is assumed to have an unique score. The uniqueness of collected data items can be achieved by integrating the data collecting time and the sensor identity into the data item score calculation. It ensures the uniqueness and correctness of a top-k query result.

3.3. Threat Model. The honest-but-curious threat model [9] is adopted in this article. The sink is trustful while sensors could collude with adversaries to leak out their collected or forwarded data. But the sensors that has been attacked still perform the pre-deployed protocols and cooperate with other innocent (noncompromised) sensors to process query commands. We have to note that the innocent sensors are the majority in WSNs; otherwise, the network will be useless.

The goal of the proposed secure top-k query protocol is described as follows:

(1) A sensor only owns the data collected by itself, and the data can be shared with the sink. It has no idea of the data collected by other sensors even when they are colluding with the adversaries.

(2) Query results can only be obtained by the sink, but the adversaries have no idea of them even when there are a few compromised sensors colluding with the adversaries.

(3) The k data items obtained by the sink are the k highest or lowest data items collected by the queried sensors, which means that the query result is correct.

Because sensors have limited energy, the network lifetime is usually determined by the energy consumption of the sensors. Reference [33] shows that sensors consume most energy in data transmission. Thus, the transmission overhead of network is an important metric for performance evaluation. We will perform the evaluation on this metric in Section 6.

4. Minimized Candidate Encrypted Dataset Determination Model

Based on the idea making, the proposed protocol efficient in transmission overhead. We propose the minimized candidate encrypted dataset determination model in this session.

4.1. Minimized Candidate Sensor Set. Let $\text{Query}_t = (t, S, k)$ be a query command, and each sensor $s_i \in S$ collects N data items in a time slot, the set of collected data of all sensors in S is $D = \{d_{i,j} | s_i \in S \wedge 1 \leq j \leq N\}$.

Definition 1. For a top-k query, the query result R_t is a dataset having the k largest data items of D. $L(R_t)$ is denoted as the lower bound of R_t, which is the minimum of R_t.

Definition 2. For a sensor $s_i \in S$, the in-node-maximum of s_i is the maximum data item.

For example, if the collected data of s_i are $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,N}\}$ and $d_{i,1} > d_{i,2} > \cdots > d_{i,N}$, then $d_{i,1}$ is the in-node-maximum of s_i.

Definition 3. For a top-k query, we define Φ is a sensor set consisting of k sensors whose in-node-maximums are the k largest in-node-maximums of sensors in S, that is

$$\Phi \subseteq S \wedge |\Phi| = k \wedge \left(\forall s_i \in \Phi, s_j \in S - \Phi \longrightarrow d_{i,1} > d_{j,1}\right). \tag{1}$$

Lemma 1. $L(R_t) \geq \min(\{d_{i,1} | s_i \subseteq \Phi\})$

Proof. According to Definition 1, $L(R_t)$ is the lower bound of R_t, which is the kth largest data of D. Because $|\Phi| = k$, there are k in-node-maximums of sensors of Φ, that is, $|\{d_{i,1} | s_i \in \Phi\}| = k$. Thus, $\min(d_{i,1} | s_i \in \Phi\})$ is the kth largest data of $\{d_{i,1} | s_i \in \Phi\}$, where $\min(*)$ represents the minimum of a dataset. Because $\{d_{i,1} | s_i \in \Phi\} \subseteq D$, we have that $L(R_t) \geq \min(\{d_{i,1} | s_i \in \Phi\})$ holds. □

Lemma 2. *Φ is the candidate sensor set of a query, which means that all data in the query result R_t are contributed by sensors of Φ, that is,*

$$R_t \subseteq \bigcup_{s_i \in \Phi} D_i. \tag{2}$$

Proof. We give the proof by contradiction. We are assuming that there is at least one data of R_t, which is not contributed by a sensor of Φ. It means that $\exists x (x \in R_t \wedge x \in D_j)$, where x is collected by s_j and s_j is not in Φ, i.e., $s_j \in S - \Phi$. We are assuming that x is the lth largest data of D. Then, we can deduce two results: 1. $1 \leq l \leq k$ holds because of $x \in R_t$ and $|R_t| = k$. 2. According to the definition of Φ, for $\forall y \in \{d_{i,1} | s_i \in \Phi\}$, because x is assumed to be collected by $s_j \in S - \Phi$, we have $y > d_{j,1} \geq x$, where $d_{j,1}$ is the in-node-maximum of s_j. Additionally, there are k in-node-maximums contributed by sensors in Φ, i.e., $|d_{i,1} | s_i \in \Phi\}| = k$. Therefore, we can deduce that $l > k$ holds.

Obviously, there are contradictions between 1 and 2. As a result, we deduce that Lemma 2 holds.

Lemma 2 It indicates that all sensors in Φ are candidate sensors, which contribute the query result. In addition, Φ is also the minimized candidate sensor set, and we prove it in Lemma 3. □

Lemma 3. *Φ is the minimized candidate sensor set that contribute the query result R_t.*

Proof. To prove this lemma, we have to prove the following two observations. □

Observation 1. For $\forall d_{j,h} \in D_j$ where $\forall s_j \in S - \Phi$, $L(R_t) > d_{j,h}$ holds.

Observation 2. Any sensor deletion from Φ could incur the incompleteness of query result. If and only if the two observations hold simultaneously, then we can deduce that Φ is the minimized candidate sensor set that contribute the query result.

Proof to Observation 1. According to Definition 3, for $\forall s_i \in \Phi$ and $\forall s_j \in S - \Phi$, $d_{i,1}$ and $d_{j,1}$ are their in-node-maximums and $d_{i,1} > d_{j,1}$ holds. Because $d_{j,1}$ is the in-node-maximum of s_j, $d_{i,1} \geq d_{j,h}$ holds where $d_{j,h} \in D_j$. Thus, $d_{i,1} > d_{j,h}$ holds. In addition, because s_i could be any sensor of Φ, we can deduce that $\min(\{d_{i,1} | s_i \in \Phi\}) > d_{j,h}$. At last, Lemma 1 indicates that $L(R_t) \geq \min(\{d_{i,1} | s_i \in \Phi\})$; therefore, $L(R_t) > d_{j,h}$ holds, and the first observation is proved. □

Proof to Observation 2. To prove the second observation, we just need to prove that, for any sensor of Φ, its collected data could belong to the query result R_t. If it is true, then deleting any sensor from Φ could cause the incompleteness of R_t. We are assuming that the collected data of sensors of Φ satisfy: $\forall d_{i,j} (s_i \in \Phi \wedge 2 \leq j \leq N) \longrightarrow d_{i,j} < \min(d_{p,1} \mid s_p \in \Phi)$. Because $|\Phi| = k$, the top-k query result R_t is determined and $R_t = \{d_{p,1} | s_p \in \Phi\}$. It means that the in-node-maximums of all sensors of Φ are just the elements of R_t. It is obvious that, in such circumstance, deleting any sensor from Φ will incur the incompleteness of R_t. Therefore, the second observation is proved.

According to the proofs, the above two observations both hold. Thus, Φ is the minimized candidate sensor set that contribute the query result. □

4.2. Minimized Candidate Encrypted Dataset. To protect data privacy, each sensor owns its private key only by itself. When a query is started, sensors first encrypt the qualified data by their keys and then submit the encrypted data to sink. For sensor s_i, we are assuming its key is g_i, which is only shared by s_i and sink. The encrypted data of $d_{i,j}$ is denoted as $(d_{i,j})_{g_i}$.

Definition 4 (minimized candidate encrypted dataset). For a top-k query, the minimized candidate encrypted dataset, denoted as Γ, is contributed by sensors of Φ and consists of the minimum number of encrypted data that have the encrypted query result in it.

We are assuming that the candidate sensors are $\Phi = \{s_1, s_2, \ldots, s_k\}$ and their in-node-maximums are $\{d_{1,1}, d_{2,1}, \ldots, d_{k,1}\}$, respectively, where $d_{1,1} > d_{2,1} > \cdots > d_{k,1}$. For any sensor $s_i \in \Phi$, its collected data items are $\{d_{i,1}, d_{i,2}, \ldots, d_{i,N}\}$, where $d_{i,1} > d_{i,2} > \cdots > d_{i,N}$. Thus, the calculation of Γ is given as follows:

$$\Gamma = \bigcup_{s_i \in \Phi} \Gamma_i, \tag{3}$$

where

$$\Gamma_i = \begin{cases} \left\{ \left(d_{i,j}\right)_{g_i} | 1 \leq j \leq k - i + 1 \right\}, & N \geq k - i + 1, \\ \left\{ \left(d_{i,j}\right)_{g_i} | 1 \leq j \leq N \right\}, & N < k - i + 1. \end{cases} \tag{4}$$

We give an example to describe the minimized candidate encrypted dataset. As shown in Figure 2, we are assuming that there are 5 nodes $\{s_1, s_2, s_3, s_4, s_5\}$, and each sensor has collected 4 data items. Their in-node-maximums satisfy $d_{1,1} > d_{2,1} > \cdots > d_{5,1}$. For sensor s_i, its collected data satisfy $d_{i,1} > d_{i,2} > d_{i,3} > d_{i,4}$. According to Definition 4, the minimized candidate encrypted datasets when $k = 3$ and $k = 5$ are shown in the dotted-lined area and solid-lined area, respectively.

Lemma 4. *Γ is the minimized candidate encrypted dataset that has the encrypted query result.*

Proof. To prove this lemma, the following two observations need to be proved. □

Observation 3. For any $(d_{i,j})_{g_i} \notin \Gamma$, which is generated by s_i, $L(R_t) > d_{i,j}$ holds.

Observation 4. Any encrypted data deletion from Γ could incur the incompleteness of query result. If and only if the two observations hold simultaneously, then we can deduce that Γ is the minimized candidate encrypted dataset that has the encrypted query result.

Proof of Observation 1. For sensor s_i, it has two alternative cases, which are $s_i \notin \Phi$ or $s_i \in \Phi$. We give the proofs in such two cases:

(i) Case I: $s_i \notin \Phi$. According to Lemma 3, Φ is the minimized candidate sensor set that contribute the query result R_t. Because $s_i \notin \Phi$, we have $d_{i,j} \notin R_t$, where $d_{i,j} \in D_j$ and then $L(R_t) > d_{i,j}$ is deduced.

(ii) Case II: $s_i \in \Phi$. Because $(d_{i,j})_{g_i} \notin \Gamma$, $k - i + 2 \leq j \leq N$ is deduced according to equation (4). In the calculation of Γ, $d_{1,1} > d_{2,1} > \cdots > d_{i,1} > \cdots > d_{k,1}$ and $d_{i,1} > d_{i,2} > \cdots > d_{i,j} > \cdots > d_{i,N}$ are the given assumption. Thus, there are at least $k' = i + j - 2$ data

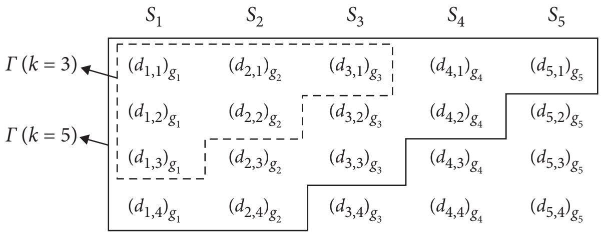

Figure 2: Minimized candidate encrypted datasets when $k = 3$ and 5.

larger than $d_{i,j}$. According to $k - i + 2 \le j \le N$ and $k\prime = i + j - 2$, then we have $k \le k\prime \le i + N - 2$. It means that there are at least k data larger than $d_{i,j}$. Definition 1 shows that the query result R_t has the k largest data, so the minimum of R_t is obviously larger than $d_{i,j}$, that is, $L(R_t) > d_{i,j}$. The deductions in two cases both lead to the same result $L(R_t) > d_{i,j}$, and the first observation is proved. □

Proof of Observation 2. To prove the second observation, we just need to prove that, for any $(d_{i,j})_{g_i} \in \Gamma$, the corresponding plaintext data $d_{i,j}$ could belong to R_t. If it is true, then deleting any encrypted data from Γ could cause the incompleteness of R_t. According to the assumptions of the calculation of Γ that the minimized candidate sensor set is $\Phi = \{s_1, s_2, \ldots, s_k\}$, where their in-node-maximums satisfy $d_{1,1} > d_{2,1} > \cdots > d_{k,1}$ and the collected data of any $s_i \in \Phi$ satisfy $d_{i,1} > d_{i,2} > \cdots > d_{i,N}$, for the data $d_{i,j}$ and $C = \{d_{1,1}, d_{2,1}, \ldots, d_{i-1,1}, d_{i,1}, d_{i,2}, \ldots, d_{i,j-1}\}$, each data in C is larger than $d_{i,j}$ and $|C| = i + j - 2$. If the following equation holds, then $d_{i,j}$ is the $(i + j - 2)$th largest data.

$$\forall d_{p,q}\Big(\left(d_{p,q}\right)_{g_p} \in \Gamma \wedge d_{p,q} \notin C\Big) \longrightarrow d_{p,q} < d_{i,j}. \qquad (5)$$

According to the calculation of Γ in equations (3) and (4), we have $j \le k - i + 1$, then $|C| \le k - 1$ holds. It means that $d_{i,j}$ is at least the kth largest data when equation (5) hold. In such scenario, $d_{i,j}$ always belongs to R_t. Therefore, we have that deleting any encrypted data from Γ could cause the incompleteness of R_t. Observation 2 is proved.

According to the above proofs, two observations both hold. Thus, Γ is the minimized candidate encrypted dataset that has the encrypted query result. Lemma 4 is proved.

Lemma 4 It indicates that Γ is the minimized candidate encrypted dataset that has the encrypted query result. It is a key to achieve efficient privacy-preserving query processing method. □

5. Top-*k* Query Processing

At first, an efficient privacy-preserving and collusion-resisting top-k (EPCT) query scheme is introduced here. Then, the correctness and security analysis, and performance of the proposed EPCT protocol will be presented.

5.1. Query Processing Protocol. The queried nodes and the sink are involved as the cooperators in this EPCT protocol. To perform the protocol, sensors and the sink are firstly settled with keys in the network deployment. Each sensor is deployed a private key, and it only shares the key with the sink. The sink owns keys of all sensors, whereas sensors have no idea of each other's keys. The protocol has two phases, shown in Figure 3. The command is broadcasted to sensors in S, before the sink receives a top-k query $\text{Query}_t = (t, S, k)$ in the first phase from the user. Once the sensor s_i gets Query_t, it transmits the encrypted in-node-maximum in the queried time slot t to the sink. As the first phase ends, the second phase begins. In the second phase, the minimized candidate sensor set is determined according to the maximum values of the queried sensors. Then, the sink transmits the second phase data request command to those candidate sensors. After each candidate sensor submits the qualified encrypted data, the sink obtains the minimized candidate encrypted dataset, and then, it will get the final query result after decryption. The processing of the top-k query Query_t is finished.

The detailed procedures of the query processing protocol are shown in Protocol 1.

Protocol 1. EPCT protocol is shown as follows:

(1) Phase 1:

(1) As a query $\text{Query}_t = (t, S, k)$ is running, the first phase starts to process. Sink broadcasts Query_t through all the networks and initials the dataset $\Gamma = \varnothing$. Then, it waits till the first phase responses from the queried nodes in the networks.

(2) For each node $s_i \in S$, s_i encrypts its in-node-maximum $d_{i,1}$ by using its private key g_i, after s_i gets the Query_t. Then, s_i generates the encrypted data $(d_{i,1})_{g_i}$, submitting the message as follows to the sink.

$$s_i \longrightarrow \text{sink: } \langle t, i\ d(s_i), \left(d_{i,1}\right)_{g_i} \rangle \qquad (6)$$

(2) Phase 2:

(1) As the submitted message from a queried sensor $s_i \in S$ arrives, $\langle t, i\ d(s_i), (d_{i,1})_{g_i} \rangle$, the sink decrypts $(d_{i,1})_{g_i}$ with the shared private key g_i and gets the plaintext in-node-maximum of s_i. s_i obtains all the decrypted in-node-maximums of the nodes in S, $\{d_{i,1} | s_i \in S\}$, before it determines the top-k data. If the determined top-k data are $\{d_{1,1}, d_{2,1}, \ldots, d_{k,1}\}$ where $d_{1,1} > d_{2,1} > \cdots > d_{k,1}$ and the corresponding sensor list according to the decent sequence of data are $\Phi = \{s_1, s_2, \ldots, s_k\}$. According to Lemma 3, Φ are the set of minimized candidate sensors. Then, the sink appends $\{d_{1,1}, d_{2,1}, \ldots, d_{k,1}\}$ into Γ and transmits the following messages to the k-1 candidate nodes in $\Phi - \{s_k\}$ in unicast mode.

$$\text{sink} \longrightarrow s_i\text{: } \langle t, (k - i)_{g_i} \rangle, \quad \forall s_i \in \Phi - \{s_k\}. \qquad (7)$$

(2) For each candidate node $s_i \in \Phi - \{s_k\}$, as the message $\langle t, (k - i)_{g_i} \rangle$ arrives, s_i decrypts the ciphertext and gets the plaintext number $k - i$.

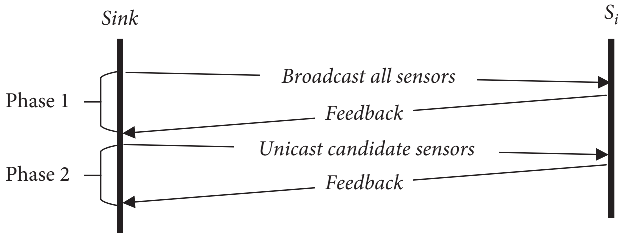

Figure 3: EPCT protocol query process.

Then, s_i encrypts $k-i$ collected data items and sends them to the sink, e.g.,

$$s_i \longrightarrow \text{sink: } \langle t, i\ d(s_i), LR_i \rangle, \tag{8}$$

where

$$LR_i = \begin{cases} \left\{ \left(d_{i,j}\right)_{g_i} \mid 2 \le j \le k-i+1 \right\}, & N \ge k-i+1, \\ \left\{ \left(d_{i,j}\right)_{g_i} \mid 2 \le j \le N \right\}, & N < k-i+1. \end{cases} \tag{9}$$

(3) The sink obtains the message $\langle t, i, LR_i \rangle$ transmitted by the candidate node $s_i \in \Phi - \{s_k\}$ in the second phase, before the ciphertext of the message is decrypted. The plaintext data after decryption are denoted as $\text{Dec}(LR_i, g_i)$ and appended into Γ. After all messages submitted from the candidate nodes are processed, the minimized candidate encrypted dataset Γ is determined, where

$$\Gamma = \bigcup_{s_i \in \Phi} \left(\left\{ d_{i,1} \right\} \cup \text{Dec}\left(LR_i, g_i\right) \right). \tag{10}$$

(4) The sink gets the top-k data of Γ, which is the exact query result R_t.

$$R_t \subseteq \Gamma \wedge \left|R_t\right| = k \wedge \left(\forall x \in R_t, y \in \left(\Gamma - R_t\right) \longrightarrow x > y\right). \tag{11}$$

As presented in Protocol 1, the query command Query_t arrives from the user in the first phase, before the sink broadcasts it through the whole network. As a queried sensor knows Query_t, it encodes the in-node-maximum before transmitting the encrypted data to the sink, where the received ciphertext is decrypted to obtain the in-node-maximums of the queried sensors in the second phase. Afterwards, the sink uses the in-node-maximums to determine the candidate sensor set $\Phi - \{s_k\}$, and then, it unicasts each candidate sensor in $\Phi - \{s_k\}$ to start the second phase. Once a candidate sensor receives the unicast message, it submits the rest data in ciphertext according to the request to the sink. As the sink obtains all the needed data from candidate nodes, the query result is determined in the end.

5.2. Protocol Analysis

5.2.1. Correctness Analysis. In the proposed EPCT protocol, when a user starts a query command Query_t, the sink will know the minimized candidate encrypted dataset Γ after interactions of the sink and sensors within two phases. Γ is consisting of the coded data items of query result. According to Lemma 4, for any $(d_{i,j})_{g_i} \notin \Gamma$, $(d_{i,j})_{g_i}$ does not belong to the query result R_t, definitely. Additionally, Γ is the minimized candidate encrypted dataset that contains the encrypted query result. Any encrypted data deletion from Γ could incur the incompleteness of query result. As Γ received by the sink, it can get the query result by obtaining the top-k data from Γ. Therefore, our proposed scheme is capable of guaranteeing the correctness of top-k query result.

5.2.2. Security Analysis. The security analysis is conducted here for the privacy of the collected data and the query results. With the cooperation of the sink and the sensors in EPCT in these two phases, each node is deployed with a private key, which is only shared with the sink. The collected data of sensors only exists in data submission from sensors to the sink. When a top-k query is started, two phases of query processing are performed. In the first phase, each sensor performs a symmetric encryption to encrypt its in-node-maximum and then transmits it to the sink. Secondly, candidate nodes are unicastly informed by the sink. They encrypted a fixed number of collected data according to the request and then sends the enciphered date to the sink node. Clearly, the data collected and transmitted through the network are all in the form of ciphertext. Every node in WSN owns a unique private key, so it can only get access to the data it collected. However, it fails to know the data collected by other sensors because of the computational infeasibility of symmetric encryption. Even a few nodes probably are attacked and colluded with adversaries, they can only snoop the collected data of those colluded sensors, but they have no idea of the collected data of innocent sensors. Besides, due to the query result is decrypted and computed in the sink and sensor nodes only process the encrypted data for the query, it is hard for the attackers to know the plaintext query result even if a few compromised sensors are colluded with them. Therefore, this proposed EPCT is a privacy-preserving and anticollusion top-k query processing protocol, which can protect the privacy of collected data of sensors even a few compromised sensors are in collusion with the adversaries, which can protect the privacy of collected data from adversaries even a few compromised sensors are in collusion with the adversaries.

5.2.3. Communication Cost Analysis. In WSNs, sensors have limited energy resource, and the energy are mainly consumed by communication. During the top-k query procedures, the communication cost of the network is mainly caused by transmission overhead of sensors. The parameters used in sensor networks are introduced in Table 1.

We are assuming that the transmission overhead of phase 1 and phase 2 are C_1 and C_2, respectively. According to the proposed EPCT protocol, all sensors participate in phase 1, whereas only the candidate sensors participate in phase 2. Then, we obtain

Table 1: Parameters description.

Para	Description
$l_{i\,d}$	The space size of a sensor ID
l_t	The space size of a time-slot
l_c	The space size of a coded data item
l_q	The space size of a query command
L	The average path length from sensors to the sink

$$\begin{aligned} C_1 &= n \cdot l_q + n \cdot \left(l_{i\,d} + l_t + l_c\right) \cdot L, \\ C_2 &= (k-1) \cdot \left(l_t + l_c\right) \cdot L + \sum_{i=1}^{k-1} \left(l_{i\,d} + l_t + i \cdot l_c\right) \cdot L \\ &= (k-1) \cdot l_{i\,d} \cdot L + (k-1) \cdot \left(l_{i\,d} + l_t\right) \cdot L + \frac{k \cdot (k-1)}{2} \cdot l_c \cdot L \\ &= (k-1) \cdot \left(l_{i\,d} + 2l_t + l_c\right) \cdot L + \frac{k \cdot (k-1)}{2} \cdot l_c \cdot L \end{aligned} \tag{12}$$

The total communication overhead of the whole network is computed as follows:

$$\begin{aligned} C_{\text{total}} = C_1 + C_2 &= n \cdot l_q + (n + k - 1) \cdot \left(l_{i\,d} + l_t + l_c\right) \cdot L \\ &\quad + (k-1) \cdot l_t \cdot L + \frac{k \cdot (k-1)}{2} \cdot l_c \cdot L. \end{aligned} \tag{13}$$

6. Performance Evaluation

Based on the improved simulator of [34], we implement three protocols, EPCT, PCTQ [31], and a naive protocol (Naive). For Naive scheme, each node queried firstly encodes its k highest data items and then submits them to the sink. After the sink gets all the ciphertext from sensors, it decrypts them to obtain the final query result. The performance is evaluated by the communication overhead in WSNs.

This experiment is conducted on a PC with an AMD R5-3600 (6 cores 12 threads 4.2 Ghz) CPU and 32 GB RAM, running 64-bit win 10 professional OS and Java JDK 1.8. In the simulation, we generate 10 networks with random topologies, and each network is distinguished by different network IDs. In each network, sensors are randomly distributed in area covering a $200 \times 200\,\text{m}^2$, and the communicating radius of a sensor is 6 m. The collected data of sensors are randomly generated in each time slot. The network communication cost C_{total} is measured by computing the average result of these 10 networks. The default settings of other parameters are shown in Table 2.

(1) C_{total} versus Network ID. Figure 4 presents that the transmission overhead of these methods are distributed uniformly in different networks. Naive has much higher cost compared with PCTQ and EPCT. Statistically, the communication overhead of EPCT is averagely 89.06% and 43.23% lower than that of Naive and PCTQ, respectively.

Table 2: Default settings of parameters.

Parameter	n	k	$l_{i\,d}$	l_q	l_t	l_c
Value	500	10	4 byte	8 byte	4 byte	16 byte

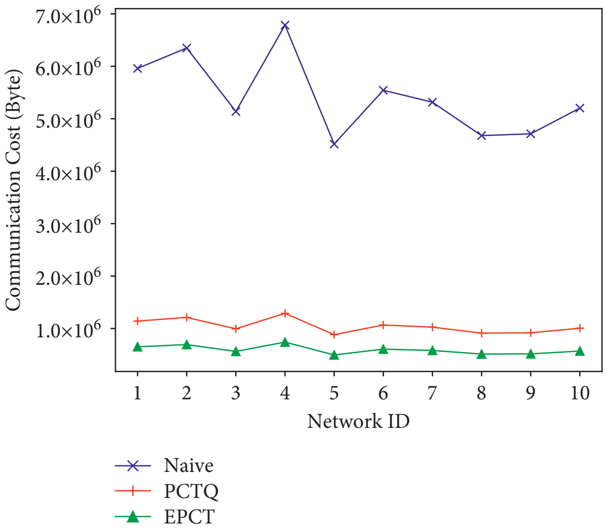

Figure 4: C_{total} vs. Network ID.

(2) C_{total} versus l_c. Figure 5 shows that the communication overhead of EPCT, PCTQ, and Naive increases as the space size of an encrypted data item l_c increases. The reason is that the transmission overhead of three approaches are all in proportion to the space size of an encrypted data item. The growth rates of communication overhead in EPCT and PCTQ are smaller than that in Naive. Statistically, EPCT reduces about 89.14% and 38.32% transmission overhead than Naive and PCTQ, respectively.

(3) C_{total} versus n. Figure 6 presents that the communication overhead of three schemes grows as the number of sensors n increases. The reason is that the more sensors are queried, the more data are transmitted in the network, i.e., the higher communication costs. Moreover, the curves in Figure 6 tell that the growth rate of transmission overhead in Naive is significantly higher than that in PCTQ and EPCT. Statistically, EPCT saves about 89.51% and 42.00% communication overhead than Naive and PCTQ, respectively.

(4) C_{total} versus k. As shown in Figure 7, the transmission overhead of three methods all increases as the number of requested data items k increases. It is that when k increases, more data items are requested in all three protocols. The growth rates of communication cost in PCTQ and EPCT are both lower than that in Naive. Specifically, EPCT saves about 93.44% and 44.57% on average than Naive and PCTQ in communication cost.

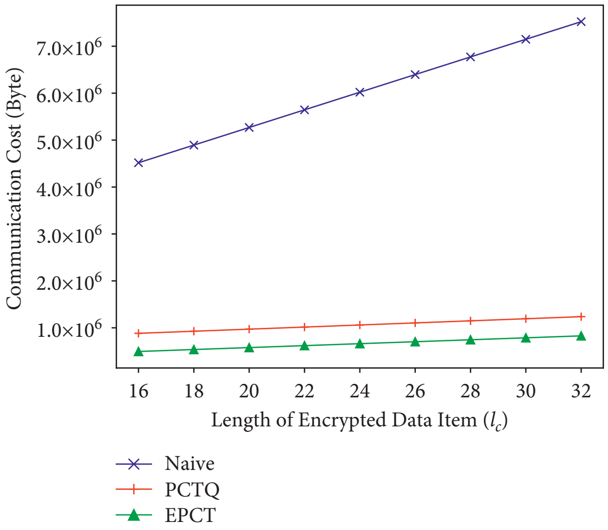

FIGURE 5: C_{total} vs. l_c.

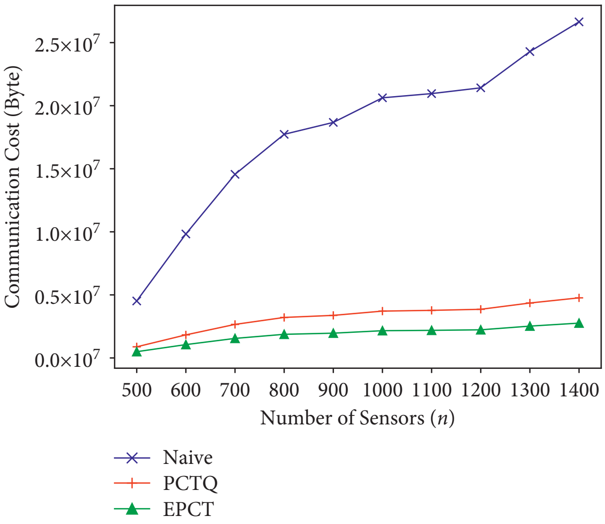

FIGURE 6: C_{total} vs. n.

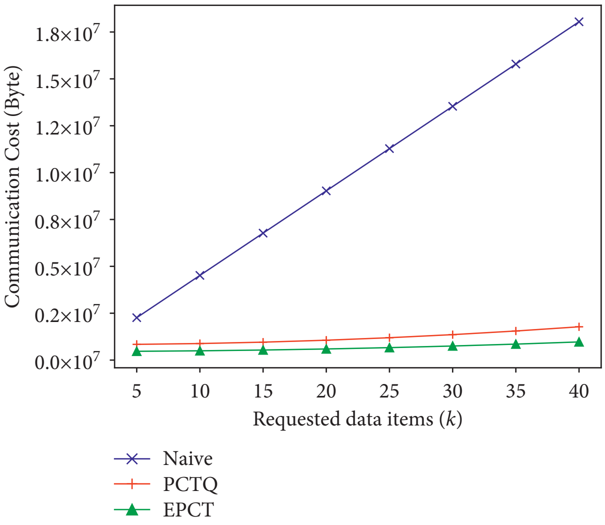

FIGURE 7: C_{total} vs. k.

According to the results of Figures 4–7, the transmission overhead of EPCT is the lowest in three protocols, whereas the overhead of Naive is much higher than the others. Because in EPCT and PCTQ, transmission only caused by candidate sensors need to, whereas in Naive scheme, all sensors are participated in transmission. Specifically, there are $k \cdot (k+1)/2$, at least k^2, and $n \cdot k$ encrypted data items are submitted from sensors to the sink in EPCT, PCTQ, and Naive, respectively. As a result, according to the above evaluations, compared with the PCTQ and Naive protocol, it has been shown that the proposed EPCT has less network communication cost and more efficient.

7. Conclusion

Data privacy threat arises during providing top-k query processing in the wireless sensor networks. To address this issue, we proposed a novel and efficient top-k query processing approach, which is capable of privacy protection and anticollusion. We fist present a minimized candidate encrypted dataset determination model, which is the foundation of the protocol. The model guides the idea of query processing and guarantees the correctness of the protocol. The symmetric encryption with different private keys in each node is employed for data privacy and even to prevent the attackers from colluding with a few nodes. Based on the above model and security setting, two phases of secure interactions between queried nodes and the sink are designed to implement the query processing protocol. The security analysis shows that our scheme is capable of providing privacy-protecting and collusion-resisting top-k queries, whereas the experimental result indicates that our approach is efficient by evaluating the network communication.

Data Availability

The data generated randomly in WSN and used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was sponsored by the Natural Science Foundation of China (Grant nos. 61902199, 61872197, 61972209, and 61872193), the Postdoctoral Science Foundation of China (Grant no. 2019M651919), and the Natural Research Foundation of Nanjing University of Posts and Telecommunications (Grant nos. NY219142 and NY217119).

References

[1] Z. H. Liang Liu and L. Wang, "Energy-efficient and privacy-preserving spatial range aggregation query processing in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 15, 2019.

[2] P. Li, Y. Liu, X. Gao, H. Li, and P. Gong, "Energy-efficient time and energy resource allocation in non-selfish symbiotic cognitive relaying sensor network with privacy preserving for smart city," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, pp. 1687–1499, 2021.

[3] X. Kui, J. Feng, X. Zhou et al., "Securing top-k query processing in two-tiered sensor networks," *Connection Science*, vol. 33, no. 1, pp. 62–80, 2021.

[4] H. Peng, X. Zhang, H. Chen, Y. Wu, Y. Wu, and J. Zeng, "Enable privacy preservation and result verification for top-k query in two-tiered sensor networks," *IEEE Trustcom/BigDataSE/ISPA*, vol. 1, pp. 555–562, 2015.

[5] X. Liao and J. Li, "Privacy-preserving and secure top-k query in two-tier wireless sensor network," in *Proceedings of the 2012 IEEE Global Communications Conference (GLOBECOM)*, pp. 335–341, Anaheim, CA, USA, December 2012.

[6] R. Li, A. X. Liu, S. Xiao, H. Xu, B. Bruhadeshwar, and A. L. Wang, "Privacy and integrity preserving top- k query processing for two-tiered sensor networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2334–2346, 2017.

[7] Y. T. Tsou, Y. L. Hu, Y. Huang, and S. Y. Kuo, "PCTopk: privacy-and correctness-preserving functional top-k query on un-trusted data storage in two-tiered sensor networks," in *Proceedings of the 2014 IEEE 33rd International Symposium on Reliable Distributed Systems*, pp. 191–200, Nara, Japan, October 2014.

[8] H. Dai, Q. Ye, X. Yi, R. He, G. Yang, and J. Pan, "VP2RQ: efficient verifiable privacy-preserving range query processing in two-tiered wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 12, 2016.

[9] L. Dong, X. Chen, J. Zhu, H. Chen, K. Wang, and C. Li, "A secure collusion-aware and probability-aware range query processing in tiered sensor networks," in *Proceedings of the 2015 IEEE 34th Symposium on Reliable Distributed Systems (SRDS)*, pp. 110–119, Montreal, Canada, October 2015.

[10] Y.-T. Tsou, C.-S. Lu, and S.-Y. Kuo, "SER: secure and efficient retrieval for anonymous range query in wireless sensor networks," *Computer Communications*, vol. 108, pp. 1–16, 2017.

[11] J. Zeng, L. Dong, Y. Wu, H. Chen, C. Li, and S. Wang, "Privacy-preserving and multi-dimensional range query in two-tiered wireless sensor networks," in *Proceedings of the GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–7, Singapore, December 2017.

[12] X. Zhang, H. Peng, L. Dong, H. Chen, and H. Sun, "SET: secure and efficient top-k query in two-tiered wireless sensor networks," in *Proceedings of the Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pp. 495–510, Beijing, China, August 2017.

[13] H. Peng, B. Liu, J. Liu, D. Li, and L. Yun, "Dp2T: preserving data privacy for top-K query in wireless sensor networks," in *Proceedings of the 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pp. 885–888, Beijing, China, November 2018.

[14] M. Xingpo, L. Junbin, M. Wenpeng, L. Yin, L. Ran, and K. Xiaoyan, "A secure top-k query processing protocol for two-tiered wireless sensor networks," *Journal of Computer Research and Development*, vol. 55, p. 2490, 2018.

[15] H. Dai, M. Wang, X. Yi, G. Yang, and J. Bao, "Secure max/min queries in two-tiered wireless sensor networks," *IEEE Access*, vol. 5, pp. 14478–14489, 2017.

[16] H. Dai, T. Wei, Y. Huang, J. Xu, and G. Yang, "Random secure comparator selection based privacy-preserving MAX/MIN query processing in two-tiered sensor networks," *Journal of Sensors*, vol. 2016, Article ID 6301404, 13 pages, 2016.

[17] H. Wu and L. Wang, "Efficient and secure top-k query processing on hybrid sensed data," *Mobile Information Systems*, vol. 2016, Article ID 1685054, 10 pages, 2016.

[18] F. Liu, X. Ma, J. Liang, and M. Lin, "Verifiable top-k query processing in tiered mobile sensor networks," *International Journal of Distributed Sensor Networks*, vol. 11, pp. 437678–437678, 2015.

[19] J. Tang and Z. Zhou, "A priority-aware multidimensional top-k query processing in wireless sensor networks," *Procedia Computer Science*, vol. 129, pp. 149–158, 2018.

[20] J. Shiraishi, H. Yomo, and K. Huang, "Content-based wake-up for top-k query in wireless sensor networks," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 1, pp. 362–377, 2021.

[21] A. F. Baig and S. S. Eskeland, "Privacy, and usability in continuous authentication: a survey," *Sensors*, vol. 21, 2021.

[22] Q. Xie, K. Li, X. Tan, L. Han, and W. T. Bin Hu, "A secure and privacy-preserving authentication protocol for wireless sensor networks in smart city," *EURASIP Journal on Wireless Communications and Networking*, vol. 12, 2021.

[23] K. A. Shah and D. Jinwala, "Privacy preserving secure expansive aggregation with malicious node identification in linear wireless sensor networks," *Frontiers of Computer Science*, vol. 15, 2021.

[24] S. Hu, L. Liu, L. Fang, F. Zhou, and R. Ye, "A novel energy-efficient and privacy-preserving data aggregation for WSNs," *IEEE Access*, vol. 8, pp. 802–813, 2020.

[25] J. Zheng, B. Song, Y. Wang, and H. Wang, "Adaptive filter updating for energy-efficient top-k queries in wireless sensor networks using Gaussian process regression," *International Journal of Distributed Sensor Networks*, vol. 11, 2015.

[26] G. Li, X. Gao, M. Liao, and B. Han, "An iterative algorithm to process the top-k query for the wireless sensor networks," *International Journal of Embedded Systems*, vol. 7, no. 1, pp. 26–33, 2015.

[27] Z. Chen, M. He, W. Liang, and K. Chen, "Trust-aware and low energy consumption security topology protocol of wireless sensor network," *Journal of Sensors*, 2015.

[28] J. Tang, Z. Wang, Y. Sun, C. Du, and Z. Zhou, "Top-k queries in wireless sensor networks leveraging hierarchical grid

index," in *Proceedings of the 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 381–386, Birmingham, United Kingdom, July 2014.

[29] C. Zhu, L. T. Yang, L. Shu, V. C. Leung, T. Hara, and S. Nishio, "Insights of top-k query in duty-cycled wireless sensor networks," *IEEE Transactions on Industrial Electronics*, vol. 62, pp. 1317–1328, 2014.

[30] H. Haiping, F. Juan, W. Ruchuan, and Q. XiaoLin, "An exact top-k query algorithm with privacy protection in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 10, 2014.

[31] J. Zhou, H. Dai, J. Zhu, R. Qi, G. Yang, and J. Xu, "A privacy-preserving and collusion-resisting top-K query processing in WSNs," in *Proceedings of the 2020 16th International Conference on Mobility, Sensing and Networking (MSN)*, pp. 677–682, IEEE, Yokyo, Japan, December 2020.

[32] L. Wang, M. Zhao, J. Chen et al., "A novel privacy-and integrity-preserving approach for multidimensional data range queries in two-tiered wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 15, 2019.

[33] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "Tag: a tiny aggregation service for ad-hoc sensor networks," *ACM SIGOPS - Operating Systems Review*, vol. 36, no. SI, pp. 131–146, 2002.

[34] A. Coman, M. A. Nascimento, and J. Sander, "A framework for spatio-temporal query processing over wireless sensor networks," in *Proceedings of the 1st International Workshop on Data Management for Sensor Networks: in conjunction with VLDB 2004*, pp. 104–110, Toronto, Canada, August 2004.