WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2015; **15**:615–628
Published online 22 March 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/wcm.2369

RESEARCH ARTICLE

PRDA: polynomial regression-based privacy-preserving data aggregation for wireless sensor networks

Suat Ozdemir[1]*, Miao Peng[2] and Yang Xiao[2]

[1] Computer Engineering Department, Gazi University, Maltepe, Ankara, TR-06570, Turkey
[2] Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487-0290, U.S.A.

ABSTRACT

In wireless sensor networks, data aggregation protocols are used to prolong the network lifetime. However, the problem of how to perform data aggregation while preserving data privacy is challenging. This paper presents a polynomial regression-based data aggregation protocol that preserves the privacy of sensor data. In the proposed protocol, sensor nodes represent their data as polynomial functions to reduce the amount of data transmission. In order to protect data privacy, sensor nodes secretly send coefficients of the polynomial functions to data aggregators instead of their original data. Data aggregation is performed on the basis of the concealed polynomial coefficients, and the base station is able to extract a good approximation of the network data from the aggregation result. The security analysis and simulation results show that the proposed scheme is able to reduce the amount of data transmission in the network while preserving data privacy. Copyright © 2013 John Wiley & Sons, Ltd.

KEYWORDS

polynomial regression; privacy; data aggregation; wireless sensor networks

***Correspondence**

Suat Ozdemir, Computer Engineering Department, Gazi University, Maltepe, Ankara, TR-06570, Turkey.
E-mail: suatozdemir@gazi.edu.tr

1. INTRODUCTION

Recent advances in wireless communications accelerated the deployment of large-scale wireless sensor networks (WSN), which consist of spatially distributed and resource constrained sensing devices. These sensing devices, that is, sensor nodes, rely on small batteries and are usually capable of measuring physical phenomenons such as temperature, sound, vibration, and pressure. In many cases, WSNs are employed to gather data from a hostile or unattended area, which makes sensor node battery replacement too expensive or even impossible [1]. Therefore, a WSN must perform the data gathering task in an energy-efficient manner so that its lifetime is prolonged. As illustrated in Figure 1(a), additive data aggregation is an important primitive that aims to combine and summarize data packets of several sensor nodes so that overall communication bandwidth and energy consumption are reduced [2].

Wireless sensor networks are usually used for mission-critical applications such as military surveillance and health monitoring. Hence, securing these networks is another essential issue, and widespread deployment of these networks could be curtailed by the lack of adequate security [3]. Although data aggregation and security are both indispensable tools for WSNs, in terms of security, there is a significant risk of data aggregation. A sensor node that is compromised by an adversary can illegally disclose the data it collects from other nodes. If the compromised node is a data aggregator, the effect of the adversary is more severe. For example, by capturing a few number of data aggregators that are positioned close to the base station, an adversary can attack on the privacy of the data of a large portion of the network. Figure 1(b) presents such a scenario. Hence, a security aware data aggregation protocol should keep sensor data secret from data aggregators, that is, provide data privacy, while still allowing them to perform data aggregation task.

In order to preserve the privacy of the data being aggregated, this paper proposes a polynomial regression-based privacy-preserving data aggregation protocol, called PRDA. PRDA is a secure additive data aggregation protocol. The novel idea behind PRDA can be explained

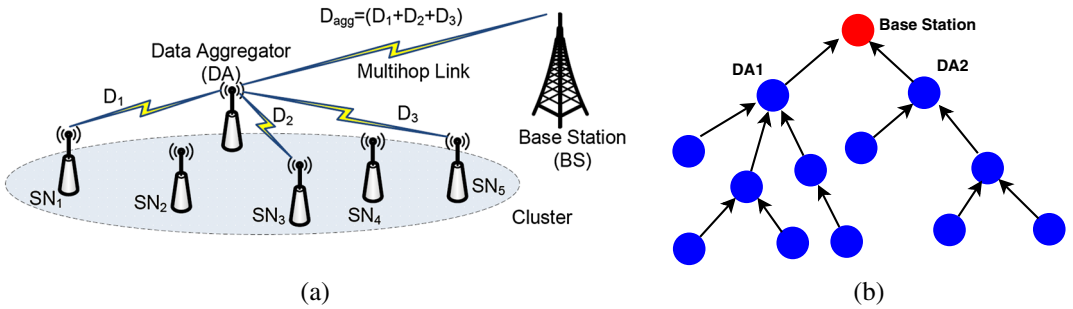

Figure 1. (a) The motivation behind additive data aggregation. Data of multiple sensor nodes can be summarized by the data aggregator node so that the amount of data transmission is reduced. (b) If DA1 or DA2 is compromised, the privacy of large portion of the sensor network is violated.

as follows. In each data aggregation session, a sensor node stores its last n sensor readings and fits these readings to an m-order polynomial curve (i.e., polynomial function) where $m < n$. Instead of sending the complete data series, the sensor node sends coefficients of its polynomial function to the data aggregator and empties its buffer. In order to make the data aggregation process privacy preserving, before sending the coefficients of its polynomial curve, the sensor node adds a random number to the coefficients, thereby concealing its data from the data aggregator. The random number is generated on the basis of the secret key shared by the sensor node and the base station. The data aggregator collects concealed coefficients from sensor nodes and performs data aggregation by adding them to each other resulting in a single concealed aggregated polynomial curve. The base station obtains the aggregated polynomial curve by subtracting the previously added random numbers from the coefficients of the concealed aggregated curve and regenerates the sum of the sensor data from the aggregated curve. Performance analysis and simulation results show that PRDA protocol is able to reduce the amount of data transmission between sensor nodes and data aggregators, and preserve data privacy during data aggregation. Simulation results also show that PRDA offers a high data accuracy depending on the m/n ratio. However, it should be noted that there is a tradeoff between data aggregation accuracy and data transmission amount, that is, energy efficiency, in PRDA protocol. Increasing the order of the polynomial functions increases not only the accuracy of aggregated data but also the energy consumption of sensor nodes because of data transmission. The degree of privacy protection is independent of this m/n ratio.

Our contribution in this paper is to propose a polynomial regression-based data aggregation protocol that preserves the privacy of the sensor data during data aggregation. The proposed protocol employs a simple yet efficient cryptographic technique to achieve privacy-preserving data aggregation. To the best of our knowledge, this is the first paper that applies polynomial regression concept to the secure data aggregation problem. The rest of the paper is organized as follows. In Section 2, the state-of-the-art in privacy-preserving data aggregation is presented. Section 3 describes the system and attack model. The details of PRDA protocol are given in Section 4. Security analysis and performance evaluation of PRDA protocol are presented in Section 5. Finally, concluding remarks are made in Section 6.

2. RELATED WORK

In WSN domain, the issue of privacy-preserving data aggregation has been very attractive for researchers [4–8]. The schemes proposed in [4] aim to bridge the gap between collaborative data aggregation and data privacy. The authors present two privacy-preserving data aggregation schemes for additive aggregation functions, namely cluster-based private data aggregation (CPDA) and slice-mix-aggregate (SMART). CPDA leverages clustering protocol and algebraic properties of polynomials so that the communication overhead is reduced. SMART employs data slicing techniques and the associative property of addition. In the proposed scheme, each sensor node slices its data into n pieces, and the pieces are then securely distributed to $n-1$ nearest sensor nodes for aggregation.

The authors of [5] propose a family of secret perturbation-based schemes that can protect sensor data confidentiality without disrupting additive data aggregation. In the proposed schemes, the base station shares a secret with each sensor node; when a sensor node has a sensory data item to report, it does not report the original data, but the sum of the original data and the secret shared with the base station. The proposed schemes provide confidentiality protection for both raw and aggregated data with lower overhead compared with existing schemes. In [6], sensor nodes form clusters to perform secret aggregation. In order to hide the individual sensor readings, the proposed scheme employs pairwise keys shared by node pairs within a cluster. The cluster aggregates are sent in clear text to be further aggregated to compute the final aggregate value. The proposed scheme ensures that the individual values as well as the identity of the contributing nodes cannot be derived by any node in the network.

The difference between [6] and our work is the way that PRDA performs data aggregation. In particular, our scheme does not need key management and uses random numbers to hide polynomial coefficients of the real sensor data. This means that data size is reduced at the source, and then data are made secret using random numbers. The authors of [7] propose several efficient mechanisms for privately querying WSNs. Two network models are presented. In the first model, access to sensor readings is provided by a single organization. In the second one, any of multiple, mutually distrusting organizations can perform this operation. In [8], a novel protocol, called PriSense, is proposed for privacy-preserving data aggregation in people-centric urban sensing systems. PriSense is based on the concept of data slicing and supports both additive and non-additive aggregation functions with accurate aggregation results. Moreover, PriSense provides strong user privacy against a tunable threshold number of colliding users and aggregation servers. In [9], authors propose regression-based aggregation scheme. Although there is a similarity in terms of regression-based data aggregation, PRDA's novelty comes from the fact that it is a lightweight privacy-preserving additive data aggregation scheme. In [10], the authors propose DyDAP (a dynamic data aggregation scheme for privacy aware wireless sensor networks), which is able to dynamically handle in-network data fusion to reduce the communication load. DyDAP uses a discrete-time control loop to balance the communication load and therefore avoids network congestion and improves WSN estimation accuracy while guaranteeing anonymity and data integrity. Riggio *et al*. [11] proposed a hybrid sensor network architecture based on a sharing of tasks between mesh routers and sensor nodes. The proposed protocol is able to concurrently support multiple WSNs while ensuring both end-to-end encryption and hop-by-hop authentication. The authors of [12] proposed cross-layer protocol (CLP) to improve data quality on the basis of an integrated solution that considers a privacy management policy coupled with a secure localization protocol. CLP takes advantage of consistency between the information on nodes behavior gathered during localization phase and privacy compliance verification to evaluate nodes reputation. In [13], an integrity protecting data aggregation protocol in WSNs is proposed. The proposed protocol is based on a two-hop verification mechanism of data integrity, which does not require referring to the base station for verifying and detecting faulty aggregated readings.

Protocols in [14,15] utilize symmetric and asymmetric privacy homomorphic encryption to allow aggregation of encrypted data, respectively. However, in [14], sensor data must be encrypted with a single key to perform concealed data aggregation. Using a single symmetric key is not secure as an adversary can fake aggregated results through compromising only a sensor node. In addition, symmetric key-based privacy homomorphism is shown to be insecure for chosen plain text attacks for some specific parameter settings [16]. The scheme proposed in [15] relies on asymmetric key-based privacy homomorphism for data aggregation. To achieve privacy-preserving data aggregation, the authors of [17] employed an extension of the one-time pad encryption technique using additive operations modulo n. The protocol adds a random number to data that is being sent and then perform modulo operation.

Our protocol PRDA differs from the existing work by providing a simple and efficient mechanism for preventing privacy in data aggregation. The aforementioned schemes employ either energy consuming complex cryptographic algorithms or redundancy-based data transmissions to provide data privacy; hence, their energy consumption is expected to be higher compared with those of traditional data aggregation schemes that do not offer any security. PRDA, on the other hand, is based on simple polynomial additions and does not increase the energy consumption of the network while preserving privacy in data aggregation. Our simulation results show that PRDA outperforms two popular data aggregation algorithms (PDA and Tiny AGgregation (TAG)) in terms of energy efficiency. In addition, to the best of our knowledge, there is no existing work that employs polynomial fitting for secure data aggregation.

3. SYSTEM MODEL AND PRELIMINARIES

We consider a large sensor network with densely deployed sensor nodes that are assigned unique identification numbers, and each sensor node shares a secret key with the base station. Because of the dense deployment, sensor nodes have overlapping sensing regions, and events are detected by multiple sensor nodes, thereby requiring data aggregation to reduce the amount of data transmission. The base station has unlimited computation and communication resources, whereas sensor nodes have limited computation and communication capabilities. The network is divided into clusters, and each cluster has a dynamically selected data aggregator node [18]. The details of cluster forming and data aggregator selection processes are out of the scope of this paper.

3.1. Data collection and aggregation

In PRDA, data are periodically collected and aggregated in data aggregation sessions. The duration of data aggregation sessions is determined on the basis of the size of each sensor reading and sensor node buffer so that data losses due to buffer overflow are minimized. Each data aggregator notifies the sensor nodes in its cluster at the beginning and end of each data aggregation session. In a data aggregation session, each sensor node performs n sensor readings resulting in a data set of size n. After the data collection, the data aggregator requests sensor nodes to send their data sets to calculate the aggregated data. The data aggregator sends aggregated data to the base station over multihop paths. For the sake of convenience, we assume that hierarchical data aggregation is not allowed in the network, and each

data aggregator aggregates its cluster data only. However, it is worth to mention that PRDA protocol can be used for hierarchical data aggregation as well.

3.2. Polynomial representation of sensor data

We assume that each sensor node is preloaded with a curve-fitting algorithm (least squares regression), and data collected by sensor nodes are correlated. In order to reduce the amount of data transmission from sensor nodes to the data aggregator, each sensor node i (SN_i) fits its data set of size n to the following m-degree polynomial

$$f_i(x) = a_{i0} + a_{i1}x + a_{i2}x^2 + a_{i3}x^3 + \cdots + a_{im}x^m,$$
$$\text{where} \quad n > m \quad \text{and} \quad i = 1 \ldots n$$

Note that the preceding polynomial can be represented as follows:

$$f_i(x) = \sum_{j=0}^{m} a_{ij}x^j, \quad n > m \text{ and } i = 1 \ldots n$$

The number of sensor readings in data sets (n) and the degree of the polynomial functions (m) are system parameters and determined before the network deployment. Additive data aggregation is performed using these polynomials as follows.

$$D_{\text{agg}}(x) = \sum_s f_s(x) = \sum_j \left[\left(\sum_s a_{sj} \right) x^j \right]$$

3.3. Key management

Each SN_i in the network shares a unique secret key $K_{\text{bs},i}$ with the base station *BS*. The keys are assigned to sensor nodes before the network deployment, and *BS* has a list of sensor node IDs and their respective secret keys. As secret keys are distributed before the deployment and there is no key transmission over the air, an attacker cannot obtain a secret key unless he compromises the sensor node that posses the key. We assume that, before the network deployment, *BS* and all sensor nodes are provided a pseudo-random number generator (PRNG) [19] that takes $K_{\text{bs},i}$ as the seed. As we explain in the subsequent sections, *BS* and sensor nodes use PRNG to generate random numbers that make sensor data secret from data aggregators and outsider eavesdroppers.

3.4. Attack model

In WSNs, adversaries can physically compromise sensor nodes and obtain their secret information. Hence, an adversary can perform a wide range of attacks including denial-of-service, eavesdropping, false data injection, modification, forgery, or replay. PRDA protocol is designed to mitigate attacks against data privacy. Replay attacks can be prevented by using a simple sequence number approach. Detecting modification, false data injection, and forgery attacks require extensive monitoring mechanisms as in [20]. Denial-of-service attack that jams the wireless channel cannot be effectively prevented in WSNs. Sensor nodes add message authentication codes to data packets. Therefore, attacks targeting data integrity, denial-of-service attacks, and replay attacks are not addressed in this paper. We also assume that the adversary have no more computational capabilities than a laptop, which means that it is computationally infeasible for an adversary to perform brute force attacks on message authentication codes or pseudo-random numbers.

4. PRDA PROTOCOL

This section explains the details of PRDA protocol. Consider the scenario given in Figure 1 where there is only one cluster shown for the sake of simplicity. After the cluster is formed and the data aggregator node (*DA*) is selected, *DA* broadcasts a message to its cluster members indicating that the data aggregation session is started. The message includes a data aggregation session number d so that sensor nodes can synchronize their PRNGs for the dth data aggregation session. All sensor nodes acknowledge this message and start sensing the environment in certain intervals. Each sensor node performs n data readings in a data aggregation session. Sensor nodes store data readings in their buffer. If the buffer of a sensor node is full, the oldest data reading is overwritten. After n rounds of data reading, *DA* broadcasts another message requesting data from sensor nodes. Upon receiving the data request from *DA*, SN_i first fits its last n data readings to an m-degree polynomial ($f_i^d(x) = \sum_{j=0}^{m} a_{ij}x^j$). Then SN_i generates a random number R_i^d using the PRNG and the secret key $K_{\text{bs},i}$. In order to preserve data privacy, SN_i adds R_i^d to each coefficient of its m-degree polynomial and obtains $Concealed[f_i^d(x)]$.

$$Concealed\left[f_i^d(x)\right] = \sum_{j=0}^{m} \left(a_{ij} + R_i^d\right) x^j$$

As R_i^d can only be generated by SN_i and *BS*, addition of R_i^d to polynomial coefficients makes SN_i's data concealed from *DA*. Each node SN_i sends coefficients of $Concealed[f_i^d(x)]$ to *DA* along with its node ID. *DA* waits for a certain time to ensure that all sensor nodes that has data send their concealed polynomials and then aggregate the concealed polynomials as follows.

$$Concealed\left[D_{\text{agg}}^d(x)\right] = \sum_s Concealed\left[f_s^d(x)\right]$$
$$= \sum_j \left[\left(\sum_s \left(a_{sj} + R_s^d\right) \right) x^j \right]$$

Note that, during the data aggregation, *DA* cannot obtain the data of sensor nodes as all coefficients are added a random number, which is secret to *DA*; hence, privacy of the aggregated data is ensured. *DA* adds the list of sensor node IDs that contributed to dth data aggregation session to the concealed aggregated data and sends it to *BS*.

When *BS* receives the concealed aggregated data $Concealed[D^d_{\text{agg}}(x)]$, it first checks the list of node IDs to generate the random number R^d_i for each node SN_i in the list using the PRNG and the secret key $K_{\text{bs},i}$ that it shares with SN_i. Then *BS* reveals the aggregated data D_{agg} by subtracting all R^d_i s from the concealed aggregated data $Concealed[D^d_{\text{agg}}(x)]$ as follows:

$$\begin{aligned} D_{\text{agg}}(x) &= \sum_s f_s(x) \\ &= \sum_j \left[\left(\sum_s \left(a_{sj} + R^d_s - R^d_s \right) \right) x^j \right] \\ &= \sum_j \left[\left(\sum_s a_{sj} \right) x^j \right] \end{aligned}$$

When generating random numbers, *BS* and sensor nodes must be synchronized so that *BS* can correctly reveal the network data from $Concealed[D^d_{\text{agg}}(x)]$. In order to achieve the synchronization among *BS* and sensor nodes, each data aggregator adds the node ID list and data aggregation session number to aggregated data packet. Note that adding sensor node ID lists to aggregated data packets may increase the communication overhead significantly. Hence, PRDA uses an indexing scheme to reduce the communication overhead due to synchronization. In what follows, we explain how PRDA achieves the synchronization among *BS* and sensor nodes at the expense of a low communication overhead.

4.1. Synchronizing random numbers

Each sensor node generates a random number R^d_i for the dth data aggregation session and uses R^d_i to conceal its data. In order to *BS* to correctly obtain the aggregated data, *BS* keeps a PRNG for each sensor node SN_i that generates the same random number R^d_i for each SN_i. When *DA* broadcasts the message that initiates the data aggregation session, it appends the data aggregation session number d to the message. The data aggregation session number d is iteratively increased in every data aggregation session and used by sensor nodes to synchronize their PRNGs. In the first data aggregation session, sensor nodes run their PRNGs for the first time and so on.

However, it is expected that not every sensor node participates in all data aggregation sessions. Hence, PRNGs of sensor nodes may be out of synch. For example, assume that a sensor node SN_i did not participate in the dth data aggregation session and hence *DA* did not send SN_i's node ID to *BS*. Now, *BS*'s PRNG that it keeps for SN_i and SN_i's own PRNG are left at the $(d-1)$th random number. Now assume that *DA* initiates the $(d+1)$th data aggregation session. If SN_i is willing to participate in this data aggregation session, then it must synchronize its PRNG by running it two times and generate $(d+1)$th random number R^{d+1}_i. SN_i uses $(d+1)$th random number to conceal its data collected in $(d+1)$th data aggregation session. When *DA* sends $(d+1)$th aggregated data to the *BS*, it attaches the data aggregation session number and node ID list, which includes SN_i's node ID. As this is the $(d+1)$th aggregated data and SN_i is in the node ID list, upon receiving $Concealed[D^{d+1}_{\text{agg}}(x)]$, *BS* synchronizes the PRNG that it keeps for SN_i by running it two times and generates $(d+1)$th random number R^{d+1}_i.

4.2. Sending node IDs

As mentioned before, when sending its aggregated data, *DA* must inform *BS* about the sensor nodes that contributed to the data aggregation. Usually, to reduce the communication overhead, Bloom filters [21] are employed to succinctly represent a set of node IDs. However, Bloom filters incur minor false- positive and zero false-negative rates. In PRDA, *BS* needs to have the exact list of sensor nodes that contributed to the data aggregation (i.e., cannot tolerate any false positive) making it impossible to use Bloom filters. Hence, in PRDA, sensor node IDs must be transmitted from data aggregators to *BS*, but transmission of sensor node IDs (which is usually 2 bytes) along with aggregated data packets increases the communication overhead of the network significantly. Therefore, PRDA employs an indexing scheme where each data aggregator keeps an index table that consists of sensor node IDs and their corresponding small index numbers. The effect of using index numbers instead of sensor node IDs is evaluated in Section 5.

4.3. Example

Let us present an example to show how PRDA protocol achieves privacy-preserving data aggregation. In the example, the system parameters are selected as $n = 5$ and $m = 2$; hence, sensor nodes collect five data measurements in each data aggregation session and use 2-degree polynomial to represent these measurements. For the sake of simplicity, let us assume that the network consists of only one cluster as shown in Figure 2. Upon receiving *DA*'s message that starts zth data aggregation session, sensor nodes that need to collect data synchronize their PRNG's and generate random numbers for the zth data aggregation session. Assume that only sensor nodes SN_1, SN_2, and SN_3 need to collect data in this session. Then SN_1, SN_2, and SN_3 collect five data measurements and fit these measurements to a 2-degree polynomial. Let the random numbers and polynomials be as follows:

- SN_1's $f_1(x) = 3x^2 + 2x + 1$ and $R_1 = 4$

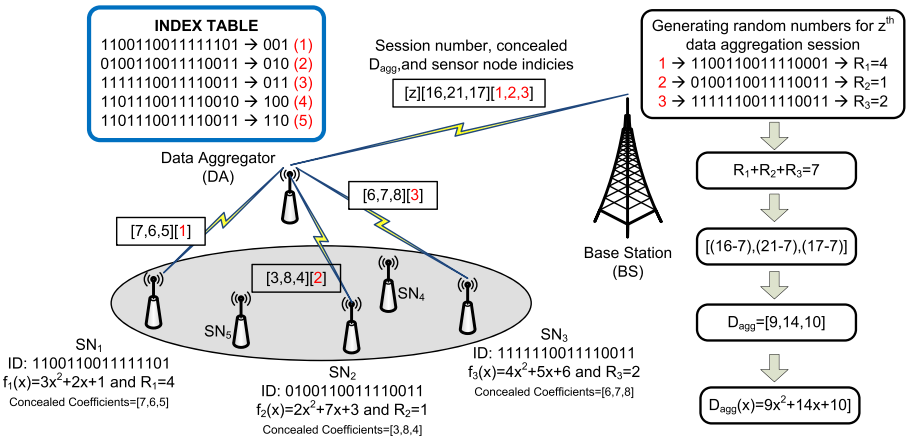

Figure 2. Example of PRDA protocol.

- SN_2's $f_2(x) = 2x^2 + 7x + 3$ and $R_2 = 1$
- SN_3's $f_3(x) = 4x^2 + 5x + 6$ and $R_3 = 2$

When *DA* broadcasts the data collection message, sensor nodes SN_1, SN_2, and SN_3 conceal their data by adding random numbers to coefficients of the polynomials and send to *DA* as follows:

- SN_1's $Concealed[f_1(x)] = 7x^2 + 6x + 5$
- SN_2's $Concealed[f_2(x)] = 3x^2 + 8x + 4$
- SN_3's $Concealed[f_3(x)] = 6x^2 + 7x + 8$

DA aggregates data of SN_1, SN_2, and SN_3 by simply adding coefficients of the polynomials and obtains the concealed aggregated data

$$Concealed[D_{\text{agg}}] = [(7+3+6), (6+8+7), (5+4+8)] = [16, 21, 17]$$

DA appends the data aggregation session number (z) and the indices of sensor nodes IDs to the concealed aggregated data and sends it to *BS*. *BS* uses its index table that it keeps for *DA* and finds out which nodes participated in the zth data aggregation session, namely SN_1, SN_2, and SN_3. Then *BS* synchronizes PRNG's of these sensor nodes and generates their random numbers for the zth data aggregation session ($R_1 = 4$, $R_2 = 1$, $R_3 = 2$). Finally, *BS* obtains the aggregated data by subtracting the sum of the random numbers from concealed coefficients as follows:

$$D_{\text{agg}} = [(16-7), (21-7), (17-7)] \text{ where } \sum_i R_i = 7$$

$D_{\text{agg}} = [9, 14, 10]$ are the coefficients of the 2-degree aggregated polynomial $D_{\text{agg}}(x) = 9x^2 + 14x + 10$. *BS* uses this polynomial to regenerate the sum of the data sent by SN_1, SN_2, and SN_3.

4.4. Hierarchical data aggregation in PRDA

Although PRDA is explained using a single-level data aggregation, it can also be used in hierarchical data aggregation scenarios. In order to employ hierarchical data aggregation in PRDA, intermediate data aggregators should act as local base stations for the aggregation trees under themselves. However, because of security reasons, only the base station has the PRNG to generate the random numbers; hence, intermediate data aggregators are not allowed to obtain the content of the aggregated data. For hierarchical data aggregation, when a data aggregator receives the aggregated data from its sub-aggregation tree, it can add its own sensed data (or aggregated data received from a lower-level data aggregator) to the aggregated data and send the new aggregated data to the next data aggregator on the path.

The drawbacks of the hierarchical data aggregation can be stated as follows: (i) As the number of sensor nodes that contributes to aggregated data increases, the length of the node ID list appended to the aggregated data packet also increases. (ii) In order to employ hierarchical data aggregation, data of all sensor nodes should be correlated all over the network. Otherwise, sensor nodes compute the coefficients from strictly distinct data sets and the precision of aggregated data would be low.

5. PERFORMANCE ANALYSIS AND EVALUATION

In this section, we first analyze PRDA theoretically and verify the theoretical analysis by simulation results. First, we analyze the security of PRDA in terms of privacy preservation. Then we present analysis of the communication efficiency and the data aggregation accuracy of PRDA. For a complete comparison of secure data aggregation protocols, please refer to [2].

5.1. Privacy-preservation analysis

In WSNs, data privacy of a sensor node SN_i can be disclosed in the following two cases: (i) an intruder eavesdrops the data transmitted by SN_i or (ii) the data aggregator of SN_i is compromised. We analyze the privacy-preservation performance of PRDA for these two cases.

Lemma 5.1. *An intruder cannot compromise sensor node SN_i's data privacy unless it compromises SN_i and obtains the secret key of SN_i.*

Proof. PRDA protocol ensures data privacy via the random numbers generated by PRNGs that are possessed by sensor nodes and *BS*. Each sensor node SN_i seeds its PRNG with the secret key that it shares with *BS* and uses a different random number R_i^d for each data aggregation session d. Hence, in order to eavesdrop the data transmitted by SN_i, an intruder must have the data aggregation session number and the secret key shared by SN_i and *BS*. As secret keys are distributed to sensor nodes before the network deployment, the intruder cannot obtain the secret key without compromising SN_i. Therefore, SN_i's data privacy is ensured. □

Lemma 5.2. *A compromised data aggregator cannot violate the privacy of data that it collects from its cluster sensor nodes.*

Proof. In PRDA, each sensor node SN_i sends coefficients of its polynomial to a data aggregator instead of its real data. Suppose the data aggregator of SN_i is compromised. To conceal its data from compromised data aggregators and eavesdroppers, SN_i adds a random number R_i^d to its polynomial coefficients. Random numbers are changed in every data aggregation session, and they are generated using the secret key between SN_i and *BS*. In order to access the data transmitted by SN_i, the compromised data aggregator must have the secret key shared by SN_i and *BS*. Moreover, secret keys are distributed to sensor nodes offline before the network deployment. Hence, the compromised data aggregator cannot obtain the secret key, and therefore, the privacy of the aggregated data is protected. □

5.2. Communication efficiency analysis

As opposed to traditional data aggregation algorithms in which sensor nodes send their original data to data aggregator for aggregation, in PRDA sensor nodes send coefficients of their polynomials. Therefore, in PRDA, reduction in data transmission occurs before data aggregation process. Moreover, traditional data aggregation algorithms cannot provide data privacy at data aggregators as they need to decrypt the sensor data. Hence, in this section, we show how PRDA reduces the amount of data transmission without sacrificing data privacy. To analyze the communication efficiency of PRDA, let us first define a metric called *data traffic*.

Data traffic: Data traffic metric represents the number of data transmissions per cluster during a data aggregation session.

As a traditional data aggregation protocol cannot provide data privacy at data aggregators, we compare the data traffic generated by PRDA protocol in a data aggregation session with a protocol without data aggregation. In order to achieve a fair comparison, we assume that network architecture is cluster-based in both cases. The cluster head in PRDA is also called data aggregator, and it is responsible for data aggregation and communication with base station. In protocol without data aggregation, the cluster head is only responsible for communication with the base station.

Suppose that the number of nodes in each cluster is k. Thus, there is one cluster head and $(k-1)$ regular sensor nodes in each cluster. In PRDA protocol, each sensor node performs n data measurements in each data aggregation session, and the number of coefficients of the regression polynomial is m. Let us assume that each coefficient is s bytes. The data traffic in each data aggregation session generated by communication between regular sensor nodes and the data aggregator is $(k-1)\times(m\times s)$.

For the data flows between the data aggregator and the base station, we can use index table to further decrease the communication overhead of the network. As described earlier, each cluster has $(k-1)$ regular sensor nodes. Therefore, each data aggregator keeps an index table containing $(k-1)$ index numbers in PRDA. The minimum number of bits to represent $(k-1)$ numbers is $\lceil \lg(k) \rceil$, so each sensor node ID can be denoted by $\lceil \lg(k) \rceil/8$ bytes. Furthermore, data aggregator aggregates all the polynomials generated by regular sensor nodes and forms an m-degree polynomial. Therefore, the data traffic generated by the data aggregator is given by $(k-1)\times\lceil \lg(k) \rceil/8 + m\times s$. Moreover, we assume that the average number of hops from each cluster to the base station is t. Therefore, the total traffic during each data aggregation session in PRDA is given by

$$(k-1)\times(m\times s) + ((k-1)\times\lceil \lg(k) \rceil/8 + m\times s)\times t$$

In protocol without data aggregation, the regular nodes send their encrypted data to the cluster head, then the cluster head relays these messages to the base station. The average number of hops from each cluster to the base station is t. Thus, the average number of hops from each sensor node to the base station is $t+1$. The encrypted data consists of two parts, namely encrypted sensing time and encrypted sensed data. Assuming that each encrypted part occupies s bytes, the size of each encrypted data pair is $2s$.

As each sensor node performs n data readings in each data aggregation session in PRDA, the total traffic in the protocol without data aggregation during the same data aggregation session is given by

$$2\times s\times n\times(k-1)\times(t+1)$$

Hence, the ratio of average traffic between PRDA and the protocol without data aggregation can be expressed as follows:

$$R = \frac{(k-1)\times(m\times s)+((k-1)\times\lceil \lg(k)\rceil/8+m\times s)\times t}{2\times s\times n\times(k-1)\times(t+1)}$$

For evaluation purpose, let us assume that the number of sensor nodes in each cluster is 64 ($k = 64$) and the size of each coefficient is 2 bytes ($s = 2$), and the average number of hops between a data aggregator and the base station is 12 ($t = 12$). Thus, we have

$$R = \frac{63\times(m\times 2)+(63\times 6/8+m\times 2)\times 12}{2\times 2\times n\times 63\times(12+1)} = \frac{150\times m+567}{3276\times n}$$

Table I. The ratio of average traffic between PRDA and protocol without data aggregation.

R	$m = 10$ (%)	$m = 20$ (%)	$m = 30$ (%)
$n = 40$	1.6	2.72	3.86
$n = 80$	0.79	1.36	1.93
$n = 120$	0.526	0.907	1.29

On the basis of the aforementioned parameter values for k, t, and s, Table I presents the ratio of average traffic between PRDA and the protocol without data aggregation under a different number of sensor node measurements in a data aggregation session (n) and different polynomial degrees (m). As seen from Table I, the ratio R is proportional to the degree of the polynomial and inversely proportional to the number of sensor node measurements in a data aggregation session.

We also analyzed the impact of cluster size on the ratio (R) of average traffic between PRDA and the protocol without data aggregation. Using the same k, t, and s values, Figure 3(a–c) presents the ratio R when the cluster size is 16, 32, and 64, respectively. On the basis of the comparison of ratios illustrated in Figure 3, we can observe that larger cluster size leads to lower traffic ratio for same size of data set and same degree of regression polynomial. Figure 3 shows that increasing the cluster size also increases the data aggregation efficiency, thereby reducing the communication overhead. It is also worth to mention that for $k = 16$, we use 4-bit index values and for $k = 64$, 6-bit index values. The results show that 50% increment in the size of the index table values does not affect the communication efficiency of PRDA.

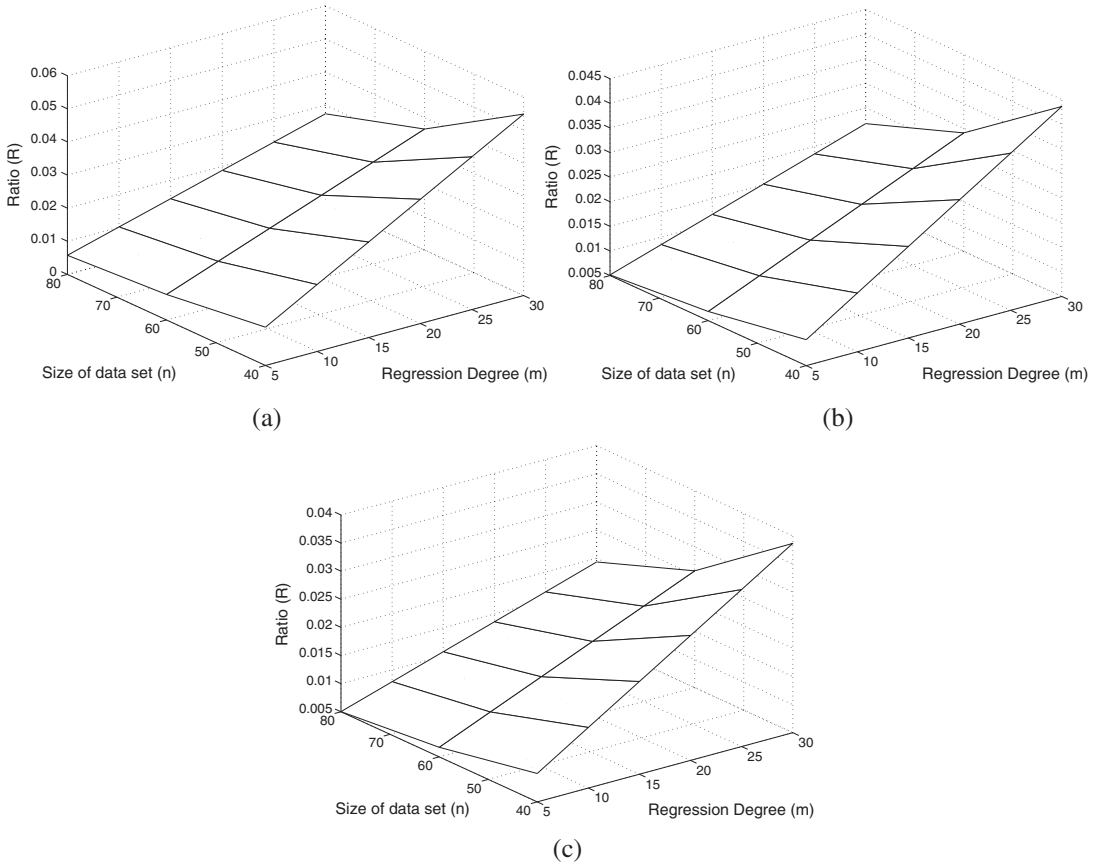

Figure 3. The ratio of average traffic between PRDA and the protocol without data aggregation: (a) $k = 16$, (b) $k = 32$, and (c) $k = 64$.

5.3. Data accuracy analysis

In this section, we theoretically analyze the data accuracy of PRDA. Let us first define *mean square error* metric to evaluate the accuracy of the data aggregated by the data aggregator.

Mean square error (MSE): We assume that there are k sensor nodes in each cluster, and n is the size of data set read by a sensor node in a given data aggregation session (i.e., the number of sensor measurements). Moreover, let $f_j(x_i)$ and y_{ij} denote the value of polynomial regression function and the sensing data of jth sensor node in the cluster at the time x_i, respectively. As data aggregation is performed in clusters, the MSE of a data aggregation result is defined as follows:

$$MSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}[f(x_i) - y_i]^2}$$

where $f(x_i) = \sum_{j=1}^{k} f_j(x_i)/k$ and $y_i = \sum_{j=1}^{k} y_{ij}/k$.

It should be obvious from the preceding equation that when the result of the data aggregation function approaches to the real sum of the data measured by the sensor nodes in the cluster, the MSE value reduces. Hence, the data aggregation accuracy of PRDA increases as MSE values decrease.

There are two major factors affecting the data aggregation performance of PRDA: MSE and variance of sensor node measurements. In our analysis, we assume that each cluster contains 20 regular sensor nodes and a data aggregator. Previously defined metric MSE must be used to measure the data accuracy performance. Because of polynomial regression, variance of sensor node measurements plays an important role on the data aggregation accuracy of PRDA as well. Hence, we also analyze the impact of maximum variance of sensor node measurements. Remembering that y_{ij} denotes the value of sensing data of jth sensor node at the time of x_i, the maximum variance (MV) is defined as follows:

$$MV = \max\{|y_{ij} - y_{ik}|\} \text{ for all } i \in (0, n] \text{ and any } j, k$$

We have showed that the degree of regression polynomial is inversely proportional to communication efficiency

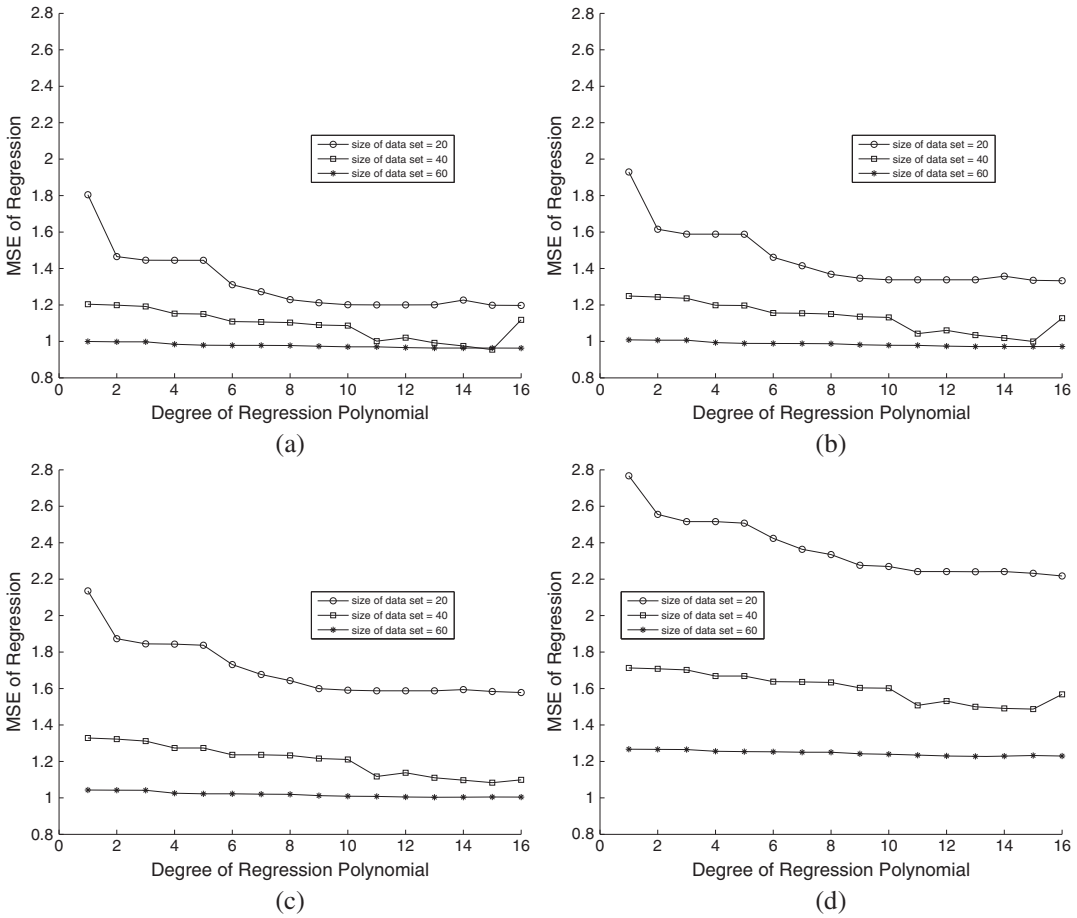

Figure 4. The MSE of regression versus the degree of regression polynomial. (a) $MV = 2$, (b) $MV = 4$, (c) $MV = 8$, and (d) $MV = 16$.

of PRDA, and we need low-degree regression polynomials to reduce the data transmission amount of sensor nodes. However, the analysis results show that in Figure 4(a–d), when the degree of regression polynomial decreases, the data accuracy of PRDA decreases as well. This is because the MSE values show a decreasing trend as the degree of regression polynomial increases. Hence, there is a tradeoff between the communication efficiency and data accuracy of PRDA protocol. Furthermore, we can also observe that the size of data set (n) in an aggregation session has a positive impact on the data accuracy of PRDA. For each fixed degree of regression polynomial, if the size of data set is larger, the data accuracy of PRDA is better. However, the MSE is not very sensitive to the size of data set because the variance of MSE is relatively stable as the size of data set varies from 20 to 60 in Figure 4. Finally, through comparing the analysis results illustrated in Figure 4, we can see that when maximum variance (MV) increases, MSE values become larger and the data accuracy level decreases.

5.4. Simulation results

The energy model proposed in [22] shows that energy consumption (i.e., lifetime) of a sensor node is proportional to the amount of data transmissions (i.e., communication) by the node. With this model, to evaluate the communication efficiency, PRDA is simulated using TinyOS 2.0 Simulator (TOSSIM) [23] in a scenario where a cluster-based sensor network is deployed to monitor the temperature of a terrain. Four hundred sensor nodes are placed in uniformly distributed random locations within a square area where the base station is located on one corner. Because of poor radio conditions of WSNs, a retransmission mechanism is implemented, and the retransmission limit is set to 5. The default bit error rate is 5%, which can be accepted as a poor radio condition. The medium access scheme is chosen as carrier sense multiple access by using the default TinyOS 2.0 CC2420 stack, which has 4 bits per symbol and 64K symbols per second, for 256 kb/s. Each simulation is run 20 times, and results are averaged. The confidence interval is 95%. We used the total number of bytes of all packets transmitted during a data aggregation session as the performance metric.

In our simulations, we used two benchmark protocols to compare the efficiency of PRDA. The first benchmark is a highly cited traditional data aggregation scheme, called TAG [24], and the second one is a recently proposed privacy protecting data aggregation scheme, called PDA [4].

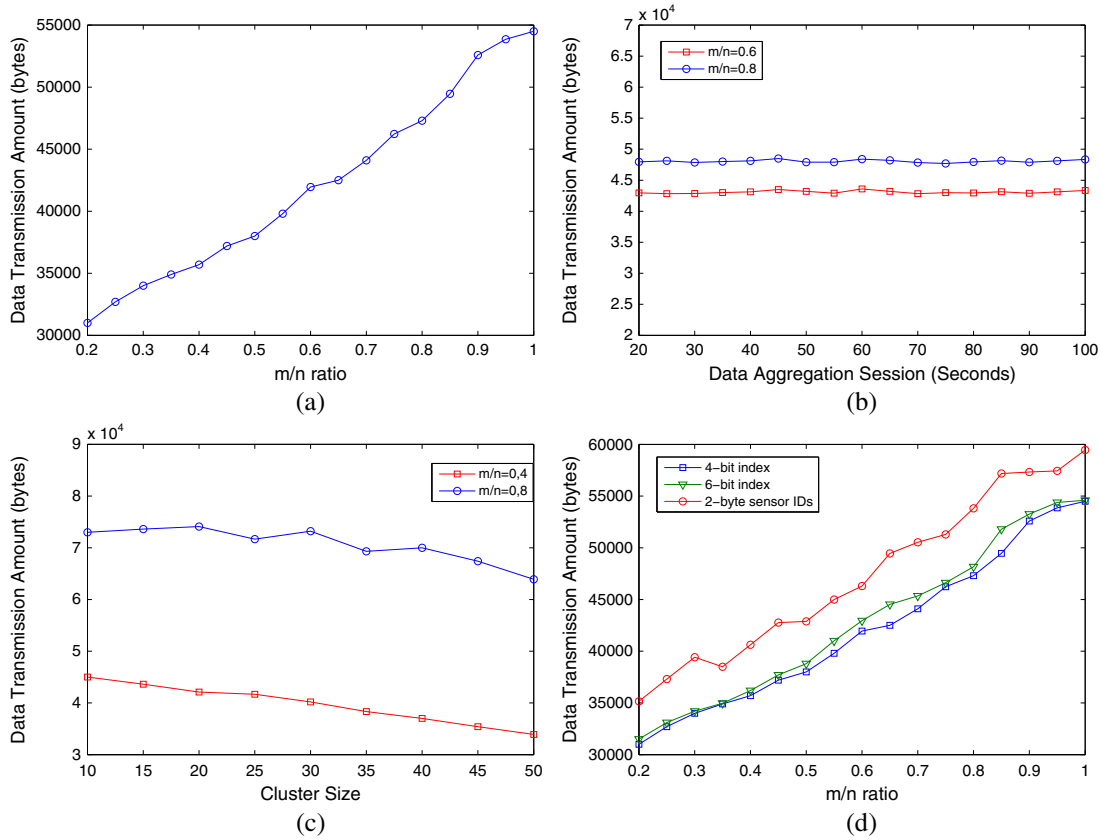

Figure 5. (a) The communication efficiency of PRDA under different (m/n) values. (b) The communication efficiency of PRDA for different data aggregation session durations. (c) The communication efficiency of PRDA for different cluster sizes. (d) The effect of using index numbers instead of sensor node IDs.

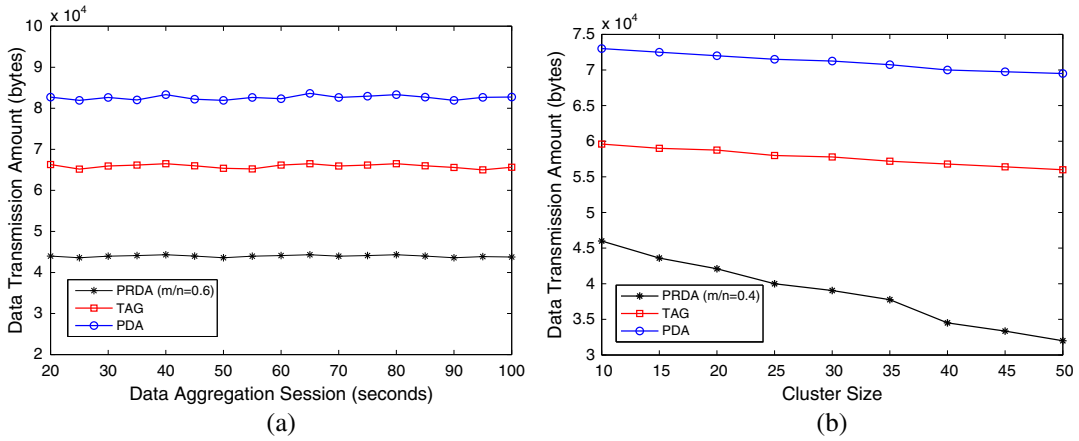

Figure 6. (a) Communication efficiency comparison of PRDA, TAG, and PDA for different data aggregation session durations. (b) Communication efficiency comparison of PRDA, TAG, and PDA for different cluster sizes.

TAG does not provide any security mechanism, and it is a lightweight data aggregation protocol. PDA provides two aggregation functions, namely CPDA and SMART. We used CPDA because it uses algebraic properties of polynomials.

Initially, we evaluated the communication efficiency of PRDA under different (m/n) ratios. The simulation results verified the theoretical analysis results obtained previously. Figure 5(a) shows that as the ratio of m/n reduces, the communication efficiency of PRDA increases. This is because the lower values of m reduces the amount of data transmitted from sensor nodes to data aggregators. Then we evaluated the communication efficiency for different data aggregation session durations by keeping the number of data measurements (n) the same. As Figure 5(b) shows, the length of data aggregation sessions does not have any notable impact on the communication efficiency. Next, we measured the effect of cluster size on the communication efficiency. As seen in Figure 5(c), increasing the cluster size increases the communication efficiency more than 20%. However, large clusters may incur congestion at data aggregators and require high management overhead [25].

We also evaluated the effect of using index numbers instead of sensor node IDs. The results are depicted in Figure 5(d). In the simulation, initially 2-byte original sensor IDs were used without any indexing scheme. Then 6-bit and 4-bit index values are used instead of original sensor node IDs. Results show that, in comparison with real sensor node IDs, using 4-bit index values increases the communication efficiency 15% on average. It should be noted that, however, using 4-bit index values limits the cluster size 16 sensor nodes. Hence, if larger cluster sizes are needed, index size must be increased. For example, by using 6-bit index values, each cluster can include up to 64 sensor nodes. However, increasing index values from 4 to 6 bits has a negligible impact on the communication efficiency.

PRDA's communication efficiency was also compared with those of TAG [24] and PDA [4]. The results are presented in Figure 6(a, b). The figure clearly show that the communication efficiency of PRDA is better than those of TAG and PDA. These results can be explained by analyzing the protocols. In PDA, there is an extensive communication overhead due to encrypted data exchange. On the other hand, PRDA uses a simple XOR-based security mechanism and does not have any security-related overhead. In TAG, data are collected at the data aggregator and aggregated, so reduction in data transmission occurs between data aggregator and the base station. However, in PRDA, sensor nodes send coefficients of their data sets to data aggregators; hence, reduction in data transmission occurs between sensor nodes and the base station. As a result of these fundamental differences, PRDA outperforms TAG and PDA in terms of communication efficiency.

5.5. Data aggregation accuracy

We evaluated the data aggregation accuracy of PRDA for different (m/n) ratios, cluster sizes, and data aggregation session durations as well. The data aggregation accuracy of the network is defined as [1 − Error in the aggregation result], and the results are presented in Figure 7(a, b). The error in the aggregation result is the difference between the average of data sensed by sensor nodes and the average of the aggregated data computed by the data aggregator. Again, the simulation results are in parallel with the results of the theoretical analysis. Figure 7(a) shows that the ratio of m/n positively affects the data accuracy. The reason is that, as the value of m increases, the polynomial representation of sensor data approaches to the original sensor data. Figure 7(b) indicates that the cluster size has a negative effect on data accuracy. This is because congestion occurs at the data aggregators of large clusters. As a result of the congestion, data aggregators cannot aggregate the data of all sensor nodes, thereby reducing the data accuracy.

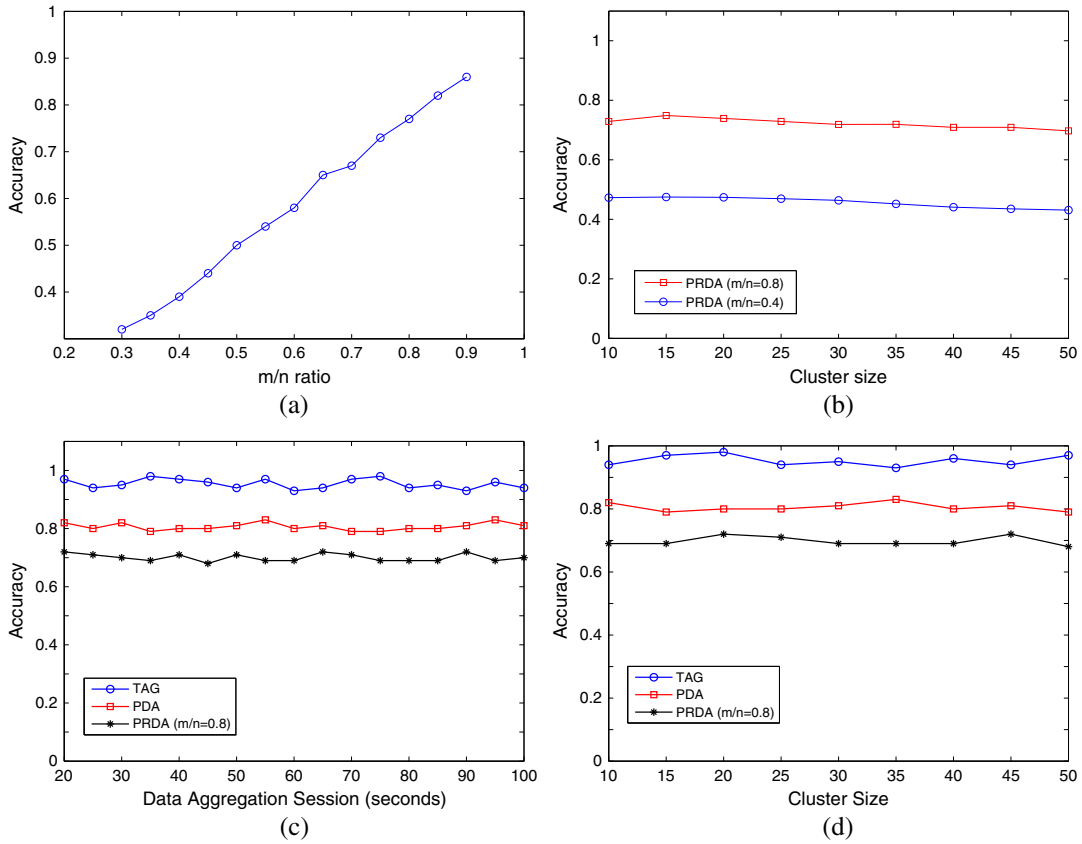

Figure 7. (a) Data accuracy of PRDA under different (m/n) values. (b) Data accuracy of PRDA for different cluster sizes. (c) Data aggregation accuracy comparison of PRDA, TAG, and PDA for different data aggregation session durations. (d) Data aggregation accuracy comparison of PRDA, TAG, and PDA for different cluster sizes.

Figure 7(c, d) compares the data aggregation accuracy of PRDA with TAG and PDA protocols. The following observations can be made from Figure 7(c, d): (i) The duration of data aggregation session does not have any impact on the aggregation accuracy of PRDA. (ii) Both TAG and PDA achieve higher data aggregation accuracy than PRDA regardless of the cluster size and data aggregation session duration. This is because PRDA uses polynomial regression that reduces the precision of the sensor node data before the data aggregation. (iii) The traditional data aggregation protocol TAG has the highest data aggregation accuracy. This is because TAG uses simple aggregation functions such MAX, MIN, or AVERAGE. It should be noted that if the correlation of sensor data increases, data aggregation accuracy of PRDA significantly increases. Therefore, PRDA should be employed in applications in which sensor data are correlated.

6. CONCLUSION

This paper presents a polynomial regression-based data aggregation protocol that preserves the privacy of sensor data. Sensor nodes use regression polynomials to represent their data and secretly send coefficients of these polynomial functions to data aggregators instead of their complete data sets. The data aggregation is performed on the basis of these secret coefficients, and the base station is able to extract a good approximation of the network data from the aggregation result. The extensive performance analysis and simulation results show that the proposed scheme significantly reduces the amount of data transmission in the network while preserving data privacy.

ACKNOWLEDGEMENTS

Dr. Ozdemir's work is supported by the Gazi University Scientific Research Project Fund No. 06/2011-09. Dr. Xiao's work is supported in part by the US National Science Foundation under grant numbers CNS-0737325, CNS-0716211, CCF-0829827, and CNS-1059265.

REFERENCES

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. *IEEE Communications Magazine* 2002; **40**(8): 102–114.
2. Ozdemir S, Xiao Y. Secure data aggregation in wireless sensor networks: a comprehensive overview. *Computer Networks* 2009; **53**(12): 2022–2037.
3. Shi E, Perrig A. Designing secure sensor networks. *Wireless Communications Magazine* 2004; **11**(6): 38–43.
4. He WB, Liu X, Nguyen H, Nahrstedt K, Abdelzaher T. PDA: privacy-preserving data aggregation in wireless sensor networks, In *Proceedings of IEEE INFOCOM*, Anchorage, Alaska, USA, 2007; 2045–2053.
5. Feng T, Wang C, Zhang W, Ruan L. Confidentiality protection for distributed sensor data aggregation, In *Proceedings of IEEE INFOCOM*, Phoenix, AZ, USA, 2008; 56–60.
6. Conti M, Zhang L, Roy S, Di P, R, Jajodia S, Mancini LV. Privacy-preserving robust data aggregation in wireless sensor networks. *Security and Communication Networks (Wiley)* 2009; **2**: 195–213.
7. Carbunar B, Yu Y, Shi W, Pearce M, Vasudevan V. Query privacy in wireless sensor networks. *ACM Transactions on Sensor Networks* 2010; **6**(2): Article No. 14.
8. Shi J, Zhang R, Liu Y, Zhang Y. Prisense: privacy-preserving data aggregation in people-centric urban sensing systems, In *Proceedings of IEEE INFOCOM*, San Diego, CA, USA, 2010; 1–9.
9. Banerjee T, Chowdhury K, Agrawal DP. Distributed data aggregation in sensor networks by regression based compression, In *Proceedings of IEEE Mobile Adhoc and Sensor Systems Conference*, Washington DC, USA, 2005; 283-290.
10. Sicaria S, Griecob LA, Boggiab G, Porisinia A. DyDAP: a dynamic data aggregation scheme for privacy aware wireless sensor networks. *Journal of Systems and Software* 2012; **85**(1): 152–166.
11. Riggio R, Rasheed T, Sicari S. Performance evaluation of an hybrid mesh and sensor network, In *Proceedings of GLOBECOM*, Houston, TX, USA, December 2011; 1–6.
12. Porisinia A, Sicari S. Improving data quality using a cross layer protocol in wireless sensor networks. *Computer Networks* 2012; **56**(17): 3655–3665.
13. Bagaaa M, Challalb Y, Ouadjaouta A, Laslaa N, Badachea N. Efficient data aggregation with in-network integrity control for WSN. *Journal of Parallel and Distributed Computing* 2012; **72**(10): 1157–1170.
14. Westhoff D, Girao J, Acharya M. Concealed data aggregation for reverse multicast traffic in sensor networks: encryption, key distribution and routing adaptation. *IEEE Transactions on Mobile Computing* 2006; **5**(10): 1417–1431.
15. Ozdemir S. Concealed data aggregation in heterogeneous sensor networks using privacy homomorphism, In *Proceedings of ICPS'07 : IEEE International Conference on Pervasive Services*, Istanbul, Turkey, 2007; 165–168.
16. Wagner D. Cryptanalysis of an algebraic privacy homomorphis, In *Proceedings of Sixth Information Security Conference*, Bristol, United Kingdom, 2003; 234–239.
17. Castelluccia C, Mykletun E, Tsudik G. Efficient aggregation of encrypted data in wireless sensor networks, In *Proceedings of Conference on Mobile and Ubiquitous Systems: Networking and Services*, San Diego, CA, USA, 2005; 109–117.
18. Boyinbode O, Le H, Mbogho A, Takizawa M, Poliah R. A survey on clustering algorithms for wireless sensor networks, In *Proceedings of 13th International Conference on Network-Based Information Systems*, Takayama, Gifu, Japan, 2010; 358–364.
19. Seetharam D, Rhee S. An efficient pseudo random number generator for low-power sensor networks, In *Proceedings of 29th Annual IEEE International Conference on Local Computer Networks*, Tampa, Florida, USA, 2004; 560–562.
20. Ozdemir S, Cam H. Integration of false data detection with data aggregation and confidential transmission in wireless sensor networks. *IEEE/ACM Transactions on Networking* 2010; **18**(3): 736–749.
21. Kirsch A, Mitzenmacher M. Less hashing, same performance: building a better bloom filter. *Lecture Notes in Computer Science* 2006; **4168**: 456–467.
22. Heinzelman WR, Chandrakasan A, Balakrishnan H. Energy efficient communication protocol for wireless microsensor networks, In *Proceedings of HICSS 2000*, 2000; 1–10.
23. TinyOS Simulator. Available from: http://www.tinyos.net [accessed on March 2013].
24. Madden S, Franklin MJ, Hellerstein JM, Hong W. TAG: a Tiny AGgregation service for ad-hoc sensor networks. *ACM SIGOPS Operating Systems Review* 2002; **36**: 131–146.
25. Ozdemir S. Secure load balancing via hierarchical data aggregation in heterogeneous sensor networks. *Journal of Information Science and Engineering (JISE)* 2009; **25**(6): 1691–1705.

AUTHORS' BIOGRAPHIES

Suat Ozdemir has been with the Computer Engineering Department at Gazi University, Ankara, Turkey, since 2007. He received his MSc degree in Computer Science from Syracuse University and PhD degree in Computer Science from Arizona State University. Dr Ozdemir's research areas mainly include sensor networks, wireless networks, network security, and data mining. Dr Ozdemir is a member of IEEE and currently serving as editor/TPC member/reviewer for various leading IEEE and ACM journals and conferences.

Miao Peng received his BS degree in Applied Mathematics from Dalian University of Technology, Dalian, China, in 2004 and his MS degree in Mathematical Statistics from Jilin University, Changchun, China, in 2007. He is currently working toward his PhD degree in Computer Science at The University of Alabama, Tuscaloosa. He is currently a Research Assistant with The University of Alabama. His research interests include wireless sensor networks, wireless network security, and energy-efficient wireless networks. In particular, he is interested in mathematical modeling in wireless and sensor networks.

Yang Xiao worked in the industry as a medium access control architect involving the IEEE 802.11 standard enhancement work before he joined the Department of Computer Science at The University of Memphis in 2002. He is currently with the Department of Computer Science at The University of Alabama. He was a voting member of IEEE 802.11 Working Group from 2001 to 2004. He is an IEEE Senior Member. He serves as a panelist for the US National Science Foundation (NSF), Canada Foundation for Innovation (CFI)'s Telecommunications expert committee, and the American Institute of Biological Sciences (AIBS), as well as a referee/reviewer for many national and international funding agencies. His research areas are security and communications/networks. He has published more than 180 refereed journal papers and over 200 refereed conference papers and book chapters related to these research areas. Dr Xiao's research has been supported by the US National Science Foundation (NSF), US Army Research, The Global Environment for Network Innovations (GENI), Fleet Industrial Supply Center–San Diego (FISCSD), FIATECH, and The University of Alabama's Research Grants Committee. He currently serves as Editor-in-Chief for *International Journal of Security and Networks (IJSN)* and *International Journal of Sensor Networks (IJSNet)*. He was the founding Editor-in-Chief for *International Journal of Telemedicine and Applications (IJTA) (2007–2009)*.