Secure Data Aggregation Using Clusters in Sensor Networks
Prakash G L, Thejaswini M, S H Manjula, K R Venugopal, L M Patnaik

Abstract—Wireless sensor network can be applied to both abominable and military environments. A primary goal in the design of wireless sensor networks is lifetime maximization, constrained by the energy capacity of batteries. One well-known method to reduce energy consumption in such networks is data aggregation. Providing efficient data aggregation while preserving data privacy is a challenging problem in wireless sensor networks research. In this paper, we present privacy-preserving data aggregation scheme for additive aggregation functions. The Cluster-based Private Data Aggregation (CPDA) leverages clustering protocol and algebraic properties of polynomials. It has the advantage of incurring less communication overhead. The goal of our work is to bridge the gap between collaborative data collection by wireless sensor networks and data privacy. We present simulation results of our schemes and compare their performance to a typical data aggregation scheme TAG, where no data protection is provided. Results show the efficacy and efficiency of our schemes.

Keywords—Aggregation, Clustering, Query Processing.

I. INTRODUCTION

A wireless sensor network (WSN) is an ad-hoc network composed of small sensor nodes deployed in large numbers to sense the physical world. Wireless sensor networks have very broad application prospects including both military and civilian usage. Sensors are usually resource-limited and power-constrained. They suffer from restricted computation, communication, and power resources. Sensors can provide ungrained raw data.

Due to these limitations data aggregation is an important consideration for sensor networks. The idea is to combine the data coming from different sources and enroute it further, after eliminating redundancy, minimizing number of transmissions and thus saving energy. This being different from traditional address centric approaches, shifts the focus to data centric approaches.

They can be used for a wide variety of monitoring and research applications, inventory maintenance, health care, military, object recognition and tracking, research and study of biological and environmental phenomenon. All the applications depend on the ability to extract data from the network. Sensor networks generate a large amount of data. For extracting information, we need to collect and query the data from sensor networks. The primary focus is on aggregates- summarized data.

Prakash G L, Thejaswini M, S H Manjula and K R Venugopal are with the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001, e-mail: glprakash@yahoo.com.

L M Patnaik is a Vice Chancellor, Defence Institute of Advanced Technology (Deemed University), Pune, India.

In many sensor network applications, the designer is usually concerned with aggregate statistics such as SUM, AVERAGE, or MAX/MIN of data readings over a certain region or period. As a result, data aggregation in WSNs has received substantial attention. Consequently, providing a reasonable guideline on building systems that perform private data aggregation is desirable. It is well-known that end-to-end data encryption is able to protect private communications between two parties (such as the data source and data sink), as long as the two parties have agreement on encryption keys. However, end-to-end encryption or link level encryption alone is not a good candidate for private data aggregation.

In this paper, we present privacy-preserving data aggregation schemes called Cluster-based Private Data Aggregation (CPDA), for additive aggregation functions in WSNs. The goal of our work is to bridge the gap between collaborative data aggregation and data privacy in wireless sensor networks. When there is no packet loss, the sensor network can obtain a precise aggregation result while guaranteeing that no private sensor reading is released to other sensors. Observe that this is a stronger result than previously proposed protocols that are able to compute approximate aggregates only (without violating privacy). Our presented schemes can be built on top of existing secure communication protocols. Therefore, both security and privacy are supported by the proposed data aggregation scheme.

The CPDA scheme, sensor nodes are formed randomly into clusters. Within each cluster, our design leverages algebraic properties of polynomials to calculate the desired aggregate value. At the same time, it guarantees that no individual node knows the data values of other nodes. The intermediate aggregate values in each cluster will be further aggregated (along an aggregation tree) on their way to the data sink. The rest of the paper is organized as follows. Section II summarizes the related work. Section III describes the problem definition for data aggregation in wireless sensor networks. Section IV provides our algorithm for private data aggregation. Section V evaluates the proposed schemes. We summarize our findings and layout future research directions in Section VI.

II. RELATED WORK

In typical wireless sensor networks, sensor nodes are usually resource-constrained and battery-limited. In order to save resources and energy, data must be aggregated to avoid overwhelming amounts of traffic in the network. There has been extensive work on data aggregation schemes in sensor networks, including [1], [2]. These efforts share the assumption...
that all sensors are trusted and all communications are secure. Work presented in [3, 4] investigates secure data aggregation schemes in the face of adversaries who try to tamper with nodes or steal the information.

To reduce computational overhead, Girao et al. [5] and Castelluccia et al. [6] propose using homomorphic encryption ciphers, which allow efficient aggregation of encrypted data without decryption involved in the intermediate nodes. Though these schemes are efficient to preserve data privacy in data aggregation, they do not protect the trend of private data of a node from being known by its neighboring nodes. This is because when the neighboring nodes can always overhear the sum of the private data and an xed unknown number. In contrast, the private data aggregation schemes we present in this paper ensures that no trend about private data of a sensor node is released to any other nodes.

In [7], the authors address the problem of data gathering and compression at relay nodes by using the theory of concave cost applied to single source aggregation. The authors develop an elegant algorithm that finds good trees that simultaneously maximize several concave cost functions of interest. Their model is significantly different from ours in the sense that their setting assumes information sources supply a constant amount of information. Moreover, their model does not take into account possible collaborations among nodes.

TAG [8] proposes the tag approach for sensor networks with Mote nodes. It assumes that if node A listens node B then node B also listens node A. This may not be true always. The spanning tree need to be created for energy eficient routing of messages. The root node starts the broadcast by sending a message with level 0 and its sensor Id. All nodes hearing the message increase the level eld attach their id and broadcast it again. They select the source of the message as their parent. The process continues down the tree.

The TAG offers a lot of advantages, saves energy, minimizes the number of messages transfer, use of epoch allows nodes to sleep during idle time thereby saving energy. In addition, fixed tree structures also have the long stretch problem. A stretch of two nodes u and v in a tree T on a graph G is the ratio between the distance from node u to v in T and their distance in G. Long stretch implies that packets from adjacent nodes have to be forwarded many hops away before aggregation.

This problem has been studied as Minimum Stretch Spanning Tree (MSST) and Minimum Average Stretch Spanning Tree (MST) [9].

In [10], the authors introduce the cluster based LEACH algorithm. In their model, the cluster head nodes compress data arriving from nodes that belong to the respective cluster, and send an aggregated packet to the base station. The work in [9] introduces the PEGASIS algorithm, that uses the energy X delay metric over the routing tree; their algorithms find chains of nodes instead of clusters.

In-networking processing can significantly improve the scalability and lifetime of microsensor networks. At each sensor, the local raw data is rst combined with partially processed data delivered from sensors farther away from the sink, and then the aggregated result is transmitted to the sensor closer to the sink or the sink itself for further processing. Intuitively, data is routed along a reversed multicast tree with the sink as the root.

Data aggregation happens at each non-leaf node, which summarizes the outputs based on the aggregation function (SUM, AVG, MEAN, MAX, etc.) from all sensors in the subtree rooted at itself and transmits the aggregated data to its parent. This process is termed data-centric routing [11, 12].

III. PROBLEM DEFINITION

A. Sensor Networks and the Data Aggregation Model

A sensor network is modeled as a connected graph G(V; E), where sensor nodes are represented as the set of vertices’s V and wireless links as the set of edges E. The number of sensor nodes is dened as |V| = N.

A data aggregation function is dened as y(t) = f(d1(t), d2(t), . . . , dN(t)), where di(t) is the individual sensor reading at time t for node i. Typical functions of f include sum, average, min, max and count. If di(i = 1, . . N) is given, the computation of y at a query server (data sink) is trivial. However, due to the large data traffic in sensor networks, bandwidth constraints on wireless links, and large power consumption of packet transmission1, data aggregation techniques are needed to save resources and power.

In this paper, we focus on additive aggregation functions, i dN d N i t i ( ). It is worth noting that using additive aggregation functions is not too restrictive, since many other aggregation functions, including average, count, variance, standard deviation and any other moment of the measured data, can be reduced to the additive aggregation function sum.

IV. SYSTEM DESIGN

A. Requirements of Private Data Aggregation

Protecting the data privacy in many wireless sensor network applications is a major concern. The following criteria summarize the desirable characteristics of a private data aggregation scheme:

Privacy: Each nodes data should be only known to itself. Furthermore, the private data aggregation scheme should be able to handle to some extent attacks and collusion among compromised nodes. When a sensor network is under a malicious attack, it is possible that some nodes may collude to uncover the private data of other node(s). Furthermore, wireless links may be eavesdropped by attackers to reveal private data. A good private data aggregation scheme should be robust to such attacks.

Efficiency: The goal of data aggregation is to reduce the number of messages transmitted within the sensor network, thus reduce resource and power usage. Data aggregation achieves bandwidth eficiency by using in-network processing. In private data aggregation schemes, additional overhead is introduced to protect privacy. However, a good private data aggregation scheme should keep that overhead as small as possible.

Accuracy: An accurate aggregation of sensor data is desired, with the constraint that no other sensors should know the exact value of any individual sensor. Accuracy should be a criterion to estimate the performance of private data aggregation schemes.
B. Cluster-based Private Data Aggregation (CPDA)

In this section, we present private data aggregation protocols focusing on additive data aggregation called Cluster-based Private Data Aggregation (CPDA). It consists of three phases: cluster formation, calculation of the aggregate results within clusters, and cluster data aggregation.

1) Formation of Clusters: The rst step in CPDA is to construct clusters to perform intermediate aggregations. We propose a distributed protocol for this purpose. The cluster formation procedure is illustrated in Figure 1. A query server Q triggers a query by a HELLO message. Upon receiving the HELLO message, a sensor node elects itself as a cluster leader with a probability \( p_c \), which is a preselected parameter for all nodes. If a node becomes a cluster leader, it will forward the HELLO message to its neighbors; otherwise, the node waits for a certain period of time to get HELLO messages from its neighbors, then it decides to join one of the clusters by broadcasting a JOIN message. As this procedure goes on, multiple clusters are constructed.

2) Calculation within Clusters: The second step of CPDA is the intermediate aggregations within clusters. To simplify the discussion, we use a simple scenario, where a cluster contains three members: A, B, and C. a, b and c represent the private data held by nodes A, B, and C, respectively. Let A be the cluster leader of this cluster. Let B and C be cluster members. Our privacy-preserving aggregation protocol based on the additive property of polynomials. Figure 2 illustrates the message exchange among the three nodes to obtain the desired sum without releasing individual private data. First, nodes within a cluster share a common (non-private) knowledge of non-zero numbers, refer to as seeds, x, y, and z, which are distinct with each other (as shown in Figure 2(a)). Then node A calculates

\[
\begin{align*}
v_A^1 &= a + r_A^1 x + r_A^2 x^2, \\
v_B^1 &= a + r_A^1 y + r_A^2 y^2, \\
v_C^1 &= a + r_A^1 z + r_A^2 z^2
\end{align*}
\]

where \( r_A^1 \) and \( r_A^2 \) are two random numbers generated by node A, and known only to node A. Similarly, node B and C calculate \( v_A^B \), \( v_B^B \), \( v_A^C \), \( v_B^C \), \( v_C^A \), and \( v_C^B \) independently as:

Node B:

\[
\begin{align*}
v_A^B &= a + r_B^1 x + r_B^2 x^2, \\
v_B^B &= a + r_B^1 y + r_B^2 y^2, \\
v_C^B &= a + r_B^1 z + r_B^2 z^2
\end{align*}
\]

Node C:

\[
\begin{align*}
v_A^C &= a + r_C^1 x + r_C^2 x^2, \\
v_B^C &= a + r_C^1 y + r_C^2 y^2, \\
v_C^C &= a + r_C^1 z + r_C^2 z^2
\end{align*}
\]

Then node A encrypts \( v_A^B \) and sends to B using the shared key between A and B. It also encrypts \( v_A^C \) and sends to C using the sharing key between A and C (Figure 2(b)). Similarly node B encrypts and sends \( v_B^A \) to A and \( v_B^C \) to C; node C encrypts and sends \( v_C^A \) to A and \( v_B^B \) to B. When node A receives \( v_B^A \) and \( v_B^C \), it has the knowledge of \( v_B^A = a + r_A^1 x + r_A^2 x^2, v_B^B = b + r_B^1 x + r_B^2 x^2 \) and \( v_B^C = c + r_C^1 x + r_C^2 x^2 \). Next, node A calculates assembled value \( F_A = v_A^A + v_B^B + v_C^C = (a + b + c) + r_1 + r_2 \), where \( r_1 = r_A^1 + r_B^1 + r_C^1 \) and \( r_2 = r_A^2 + r_B^2 + r_C^2 \). Similarly node B and C calculate their assembled values

\[
\begin{align*}
F_B &= v_B^B + v_B^C + v_B^A = (a + b + c) + r_3 + r_4, \\
F_C &= v_C^A + v_B^B + v_C^C = (a + b + c) + r_5 + r_6
\end{align*}
\]

respectively. Then node B and C broadcast \( F_B \) and \( F_C \) to the cluster leader A (Figure 2(c)). Then the cluster leader A can deduce the aggregate value \( (a + b + c) \). This is because \( x, y, z, F_A, F_B, F_C \) are known to A.

It is necessary to encrypt \( v_B^B, v_B^C, v_A^B, v_B^A, v_C^B, \) and \( v_C^C \). For example, if node C overears the value \( v_B^B \), then C knows \( v_B^A, v_B^C \), and \( F_A \), then C can deduce \( v_A^B = F_A v_B^A v_C^A \), and further it can obtain a if \( x, v_A^A, v_B^B, v_C^C \) are known. However, if node A encrypts \( v_B^B \) and sends it to node B, then node C cannot get \( v_B^A \). With only \( v_A^A, v_B^B, v_B^C, v_C^C \) to each other, then As data will be disclosed. To prevent such collusion, the cluster size should be large. In a cluster of size \( m \), if less than \( (m - 1) \) nodes collude, the data wont be disclosed.

3) Cluster Data Aggregation: A common technique for data aggregation is to build a routing tree. We implement CPDA on top of the TAG Tiny AGgregation [9] protocol. Each cluster leader routes the derived sum within the cluster back towards the query server through a TAG routing tree rooted at the server.

4) Discussions on Parameter Selection in CPDA: In CPDA, a larger cluster size introduces a larger computational overhead. However, a larger cluster size is preferred for the sake of improved privacy under node collusion attacks. In CPDA, we should guarantee a cluster size \( m \geq 3 \). Generally, lets dene \( m_e \) as the minimum cluster size. We should set \( m_e \geq 3 \). Next, we discuss how to ensure every cluster has a cluster size larger than \( m_e \), and how to tune parameter \( p_c \) to reduce...
together with TAG (no privacy protection). In our experiments, we did extensive simulations and collected results to compare these two schemes. A goal of these schemes is to protect data privacy. This introduces some communication overhead in cluster formation phase. If a cluster $C_i$ has a size smaller than $m_c$, the cluster leader of $C_i$ needs to broadcast a merge request to join another cluster.

V. EVALUATION

In this section we evaluate the private-preserving data aggregation schemes presented in this paper. We evaluate how our schemes perform in terms of privacy-preservation, efficiency, and aggregation accuracy. We use TAG, a typical data aggregation scheme as the baseline. Since the design of TAG does not take privacy into consideration, no data privacy protection is provided. We only use it to evaluate the efficiency and aggregation accuracy compared with our proposed schemes.

A. Privacy-preservation Analysis of CPDA

In the CPDA scheme, private data may be disclosed to neighbors only when the sensor nodes exchange messages within the cluster. Given a cluster of size $m$, a node needs to send $m!\prod_{i=1}^{m-1} (m-i)$ encrypted messages to other $m-1$ members within the cluster. Only if a node knows all $m$ keys of a given member, can it crack the private data of the member. Otherwise, the private data cannot be disclosed. Consequently, the communication overhead is $m\prod_{i=1}^{m-1} (m-i)$. We consider networks with 600 sensor nodes. These nodes are randomly deployed over a 400meters X 400meters area. The transmission range of a sensor node is 50 meters and data rate is 1 Mbps.

At the beginning of each simulation, a query is delivered from the query server to the sensor nodes. Similar to TAG, the query species an epoch duration $E$, which is the amount of time for the data aggregation procedure to finish. Upon receiving such a query, a parent node on the aggregation tree subdivides the epoch such that its children are required to deliver their data (protected data in CPDA, or unprotected data in TAG) in this parent-dened time interval.

B. Communication Overhead

CPDA uses data-hiding techniques and encrypted communication to protect data privacy. This introduces some communication overhead. In order to investigate bandwidth efficiency of these schemes, we implemented CPDA in ns2 on top of the data aggregation component of TAG. We did extensive simulations and collected results to compare these two schemes together with TAG (no privacy protection). In our experiments, simulation results can be explained by analyzing the number of exchanged messages in each scheme. In TAG, each node needs to send 2 messages for data aggregation: one Hello message to form an aggregation tree, and one message for data aggregation. In our implementation of CPDA, a cluster leader sends roughly 4 messages and cluster members sends 3 messages for private data aggregation. Accordingly, $4p_c + 3(1 - p_c)$ is the average number of messages sent by a node in CPDA. Thus, the message overhead in CPDA is less than twice as that in TAG.

Now let us further study the effect of $p_c$ on the communication overhead in CPDA. Figure 4 shows the result with $p_c = 0.1, 0.2, 0.3$ respectively. As we can see, the larger the $p_c$ value, the larger the communication overhead. It is very interesting to notice that when $p_c = 0.1$, communication overhead is much lower than TAG. This is because when $p_c$ is too small, many nodes cannot be covered due to insufficient number of cluster leaders.

In ideal situations when there is no data loss in the network, CPDA should get 100 percent accurate aggregation results. However, in wireless sensor networks, due to collisions over wireless channels and processing delays, messages may get lost or delayed. Therefore, the aggregation accuracy is affected. We define the accuracy metric as the ratio between the collected sum by the data aggregation scheme used and the real sum of all individual sensor nodes. A higher accuracy value means the collected sum using the specific aggregation scheme is more accurate. An accuracy value of 1.0 represents the ideal situation.

Figure 3 shows the communication overhead of TAG, CPDA with $p_c = 0.3$ under different epoch durations. We use the total number of bytes of all packets communicated during the aggregation as the metric. Each point in the figure is the average result of 50 runs of the simulation. In each run, one randomly generated sensor network topology is used. The vertical line of each data point represents the 95 percent confidence interval of the data collected.

In this section we evaluate the private-preserving data aggregation schemes presented in this paper. We evaluate how our schemes perform in terms of privacy-preservation, efficiency, and aggregation accuracy. We use TAG, a typical data aggregation scheme as the baseline. Since the design of TAG does not take privacy into consideration, no data privacy protection is provided. We only use it to evaluate the efficiency and aggregation accuracy compared with our proposed schemes.

Communication overhead

CPDA uses data-hiding techniques and encrypted communication to protect data privacy. This introduces some communication overhead. In order to investigate bandwidth efficiency of these schemes, we implemented CPDA in ns2 on top of the data aggregation component of TAG. We did extensive simulations and collected results to compare these two schemes together with TAG (no privacy protection). In our experiments, we consider networks with 600 sensor nodes. These nodes are randomly deployed over a 400meters X 400meters area. The transmission range of a sensor node is 50 meters and data rate is 1 Mbps.

At the beginning of each simulation, a query is delivered from the query server to the sensor nodes. Similar to TAG, the query species an epoch duration $E$, which is the amount of time for the data aggregation procedure to finish. Upon receiving such a query, a parent node on the aggregation tree subdivides the epoch such that its children are required to deliver their data (protected data in CPDA, or unprotected data in TAG) in this parent-dened time interval.

C. Accuracy

In ideal situations when there is no data loss in the network, CPDA should get 100 percent accurate aggregation results. However, in wireless sensor networks, due to collisions over wireless channels and processing delays, messages may get lost or delayed. Therefore, the aggregation accuracy is affected. We define the accuracy metric as the ratio between the collected sum by the data aggregation scheme used and the real sum of all individual sensor nodes. A higher accuracy value means the collected sum using the specific aggregation scheme is more accurate. An accuracy value of 1.0 represents the ideal situation.

In ideal situations when there is no data loss in the network, CPDA should get 100 percent accurate aggregation results. However, in wireless sensor networks, due to collisions over wireless channels and processing delays, messages may get lost or delayed. Therefore, the aggregation accuracy is affected. We define the accuracy metric as the ratio between the collected sum by the data aggregation scheme used and the real sum of all individual sensor nodes. A higher accuracy value means the collected sum using the specific aggregation scheme is more accurate. An accuracy value of 1.0 represents the ideal situation.
Fig. 3: Comparison of TAG and CPDA ($p_c = 0.3$).

Fig. 4: Communication overhead of CPDA with respect to $p_c$. The second observation is that TAG has better accuracy than CPDA. That is because without the communication overhead introduced by privacy-preservation, there will be less data collisions.

VI. CONCLUDING REMARKS

Providing efficient data aggregation while preserving data privacy is a challenging problem in wireless sensor networks. Many civilian applications require privacy, without which individual parties are reluctant to participate in data collection. In this paper, we propose aggregation scheme CPDA, focusing on additive data aggregation functions. We compare the performance of our presented schemes to a typical data aggregation scheme TAG. Simulation results and theoretical analysis show the efficacy of our scheme. Our future work includes designing private-preserving data aggregation schemes for general aggregation functions.

REFERENCES


Prakash G L is an Assistant Professor, Department of Computer Science and Engineering of Alpha College of Engineering, Bangalore, India. He received his B.E and M.E degrees in Computer Science and Engineering from Bangalore University, Bangalore. He is presently pursuing his Ph.D programme in the area of Wireless Sensor Networks in Bangalore University.

Thejaswini M is a Faculty with the Department of Computer Science and Engineering of Sri Jagadguru Mallikarjuna Murugharajendra Institute of Technology, Chitradurga, India. She received her B.E in Information Science and M.E degrees in Computer Science and Engineering from Visveswaraiyah Technological University and Bangalore University respectively. Her research interests include Wireless Sensor Networks.

Manjula S H is an Assistant Professor, Department of Computer Science and Engineering of University Visvesvaraya College of Engineering, Bangalore, India. She received her B.E and M.E degrees in Computer Science and Engineering from Bangalore University, Bangalore. She is presently pursuing her Ph.D programme in the area of Wireless Sensor Networks.

K R Venugopal is currently the Principal and Dean, Faculty of Engineering, University Visvesvaraya College of Engineering, Bangalore University. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D. in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 27 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ etc. He has been serving as the Professor and Chairman, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. During his three decades of service at UVCE he has over 200 research papers to his credit. His research interests include computer networks, parallel and distributed systems, digital signal processing and data mining.

L M Patnaik is a Vice Chancellor, Defence Institute of Advanced Technology (Deemed University), Pune, India. During the past 35 years of his service at the Indian Institute of Science, Bangalore. He has over 500 research publications in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to high performance computing and soft computing. His areas of research interest have been parallel and distributed computing, mobile computing, CAD for VLSI circuits, soft computing, and computational neuroscience.