Clustering based Data Collection using Data Fusion in Wireless Sensor Networks

S.G. Santhi
Assistant Professor,
Dept of CSE,
Annamalai University,
Annamalainagar, India

R. Ramya
P.G Student,
Dept. of CSE,
Annamalai University,
Annamalainagar, India

ABSTRACT
Clustering is one of the important techniques in Wireless Sensor Networks (WSNs). In this paper introduce Double Cluster Head Model (DCHM) for secure and accurate data fusion in WSNs. Data fusion is used to reduce the traffic load and conserves energy of the sensors. In this clustering technique each cluster has two Cluster Heads (CHs) and they are assuming to be trust. After clustering each sensor nodes needs to maintain a reputation and trust table which is used to find the compromised nodes. Each CH perform the data fusion process independently and its sends the fused data to the base station. In this base station dissimilarity coefficient is computed and compared with threshold value which is preset by the users. If the dissimilarity coefficient exceeds the threshold, the CHs will be added to blacklist, and the CHs must be reelected by the sensor nodes in a cluster. And also feedback is sent from the base station to the reputation and trust system, which can helps to identify and delete the compromised sensor nodes. Through a series of extensive simulations, it can found that the DCHM performed very well in data fusion security and accuracy.

Keywords
Double Cluster Heads; clustering; data fusion; security; accuracy; threshold.

1. INTRODUCTION
WSN is a collection of sensor nodes which is used to sense the surrounding environment. Each sensor node contain the limited number of resources such as energy, storage, and computational capacity then each node can communicate with one another within a small distance. Energy (sending and receiving) is a driving constraint. Minimum Energy consumption and traffic load is an important issue in WSNs. Clustering is a standard approach for reducing energy consumption and traffic load. Group of similar sensor nodes are organized into a cluster. Each cluster has a CH which is used to collect the information from respective sensor nodes and forward to the base station.CH is selected by two circumstances. They are Energy and time. Aim of this paper is to secure and accurate data fusion. A data fusion node collects the results from multiple nodes, and it sends the fused data to base station. Data fusion is a good approach for reducing the traffic load and conserves energy of the sensors. In this paper probability based Bayesian algorithm is used for data fusion.

Security is also an important issue in WSNs. Normally, cryptography is used to provide a security problem, but cryptography is not only sufficient to protect network. To overcome this problem in this paper it is proposed to use reputation and trust table for security. Reputation creates the opinion of neighbor node and trust creates the expectation of neighbor node. Each cluster separately built and maintains this reputation and trust table because, it is very difficult to build and maintain a reputation and trust system for whole network. The reputation and trust based systems for WSNs use one of the three following methods to share information among the nodes: friends list, blacklist, and reputation table. Positive informations are shared in a friends list, negative informations are shared in a blacklist, while a reputation table is shared both positive and negative information.

After clustering DCHM is introduce for secure and accurate data fusion in WSNs. In DCHM each cluster has two CHs. One is Master Cluster Head (MCH) another one is Vice Cluster Head (VCH). MCH receives and fuse the data from its member nodes. The fuse data are sent to the VCH. The VCH transmits fused data to the base station directly.

2. RELATED WORK
In WSNs, several problems are researched related to proposed work.

2.1 Clustering for reducing the energy consumption in wireless sensor network
This section discussed about some published clustering algorithm uses for wireless sensor network. A distributed clustering algorithm for ad hoc networks [1]. They have produced a node that consumes minimum energy among its one-hop neighbors is selected as the cluster head. Based on the weight the heads get changed in the network. The change in the network conditions is reflected in the table for ready reference. Hierarchical clustering algorithm also used for energy consumption [2]. They have proposed a distributed algorithm for organizing sensors into a hierarchy of clusters with an objective of reducing the total energy spent in the system to communicate the information collected by these sensors to the information processing center. They have found that the best parameter values for these algorithms that minimize the energy spent in the network. This paper also has proposed some equation for time complexity. Multiple cluster tree routing algorithm for sensor networks used in [17].Clustering algorithm is not only used for wireless sensor network it is also used for mobile ad hoc network to reduce the energy consumption.

2.2 Reputation and trust table for wireless sensor network
Reputation and trust table is one of the main techniques in wireless sensor network. Reputation techniques implemented in data aggregation [4]. This paper has presented a novel Reliable Data Aggregation and Transmission (RDAT) protocol that introduces functional reputation concept. RDAT is more reliability and security when compared with normal reputation. This functional reputation overcomes the cryptographic concept. Reputation is also used for high

integrity sensor network [5]. Cryptography cannot provide data authentication needed for countering misbehavior from internal adversaries and faulty nodes. At this time reputation is produced for secure aggregation.

2.3 Secure data fusion for wireless sensor network

Data fusion in wireless sensor networks has not been widely researched, though there are several protocols recently proposed in the literature. In this section we discussed about some published techniques related to the data fusion. Secure data aggregation used for wireless sensor network [6].Number of challenges remains in the area of secure data aggregation. In this paper we introduced tree based data approaches for aggregation. It is more reliability compare with normal data aggregation. Aggregation is an important technique to achieve power efficiency in wireless sensor network [7]. One of the main challenges to data aggregation is how to secure aggregated data from release during aggregating process as well as obtain accurate aggregated results, and also they discussed about number of protocols for secure data aggregation. This paper introduced a brief discussion of wireless sensor network, data aggregation, various approaches of data aggregation in WSN, Security needs to data aggregation, overview of various security protocols and their comparison.

3. OUTLINE OF THE WORK

This paper deals with secure and accurate data fusion in WSNs based on DCHM. For this data fusion the probability based Bayesian method is developed. In this paper first discuss about the cluster formation. Group of similar sensor nodes are organized into a cluster which is based on the distance or proximity and some logical organizing. Clustering has advantages for reducing useful energy consumption. Next select the CH. CH is used to collect the information from the number of sensor nodes and forward to the base station. In a cluster, the CH is obstructing by several attacks such as message dropping, information falsifying and so on. This problem can overcome by DCHM. In DCHM each cluster has two CH. After the CHs selection, each sensor node maintains the reputation and trust table which is sent to CHs. Based on the reputation and trust table CHs detect the weighted outlier's data by using Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. And also data fusion process has done by CHs. Each CHs detect and fuse the data independently and uploaded to the base station. At the base station dissimilarity coefficient is computed based on both outlier results and fusion results. This coefficient is compared with threshold value which is preset by users. If the coefficient is lesser than or equal to the threshold, which means that both the CHs are credible, only the reputation and trust system should be updated. Otherwise reelect the CHs.

4. PROPOSED APPROACH

The general overview of the proposed approach is illustrated in figure 1. This approach uses the secure and accurate data fusion in WSNs. This fusion is based on combining clustering techniques, reputation & trust system and data fusion algorithms.

4.1 Clustering techniques

Group of similar sensor nodes are organized into a cluster. Each cluster has a number of member nodes and coordinator, referred to as a CH, and a. Clustering has been shown to improve network lifetime, a primary metric for evaluating the performance of a sensor network. Clustering techniques were primarily proposed for energy efficiency and expand the network lifetime. A clustering approach can identify the parameter(s) used for partitioning the network and whether clustering is performed in a centralized or a distributed fashion [8]. Typically a cluster member maintains complete state information about all other members within its cluster. Clustering scheme also used to create a layered hierarchy [9].

4.2 Reputation and Trust system

Reputation and trust are two very useful tools that are used to help decision making in different fields of business and deal. Reputation**:** observation that an agent creates about another agent's goal and rules, through direct and indirect inspection of its past measures. Trust**:** a subjective anticipation an agent has about another's future performance with respect to a specific achievement. The important goals of reputation and trust-based systems for wireless communication networks as follows:(i) Provide information that allows nodes to differentiate between reliable and unreliable nodes in the network, (ii) Support the nodes in the network to cooperate with each other and become reliable, (iii) Reduce the unreliable nodes to participate in the network actions, (iv) To be able to manage with any kind of visible misconduct, and (v) To reduce the break reason by any insider attacks.

Reputation and trust-based system for wireless communication networks must have three essential properties as follows:

(i) The system must have prolonged entities that motivate outlook for future relations, (ii) The system must be able to capture and distribute feedbacks about current relations with its mechanism and such information must also be available in future, and (iii) The system must use feedback to guide trust conclusions. Categories and issues of reputation & trust system are discussed in [10].

4.3 Data fusion

A data fusion node collects the results from multiple nodes and it fuses the results with its own based on a probability. Send the fused data to another node or base station. Advantages of this process are reduces the traffic load and conserves energy of the sensors. Figure 1 illustrates step by step process of data fusion in WSNs. In this process eight steps and three modules are involved. In DCHM each module is performed in a particular position of the network and act as a particular role. In clusters module, each sensor node participate in the processes of clustering, CHs election, initializing and updating the reputation and trust system. This module is the basement of whole framework and also in a cluster informations are collected by the sensor nodes which are transmitted to the CHs. CHs module act as a bridge between clusters module and base station module. CHs module is performed by the two CHs and has two tasks, i.e., outlier detection and credible data fusion. These results are to be checked by the base station module. The fusion results and the lists of outliers are uploaded into the base station which is used to compute the dissimilarity coefficient (a). That coefficient is compared with threshold value (th) which is preset by users. Note that, it is understood that the base station is believable all the time. The three modules are incorporated together to maintain the network and can't be divided from each other.

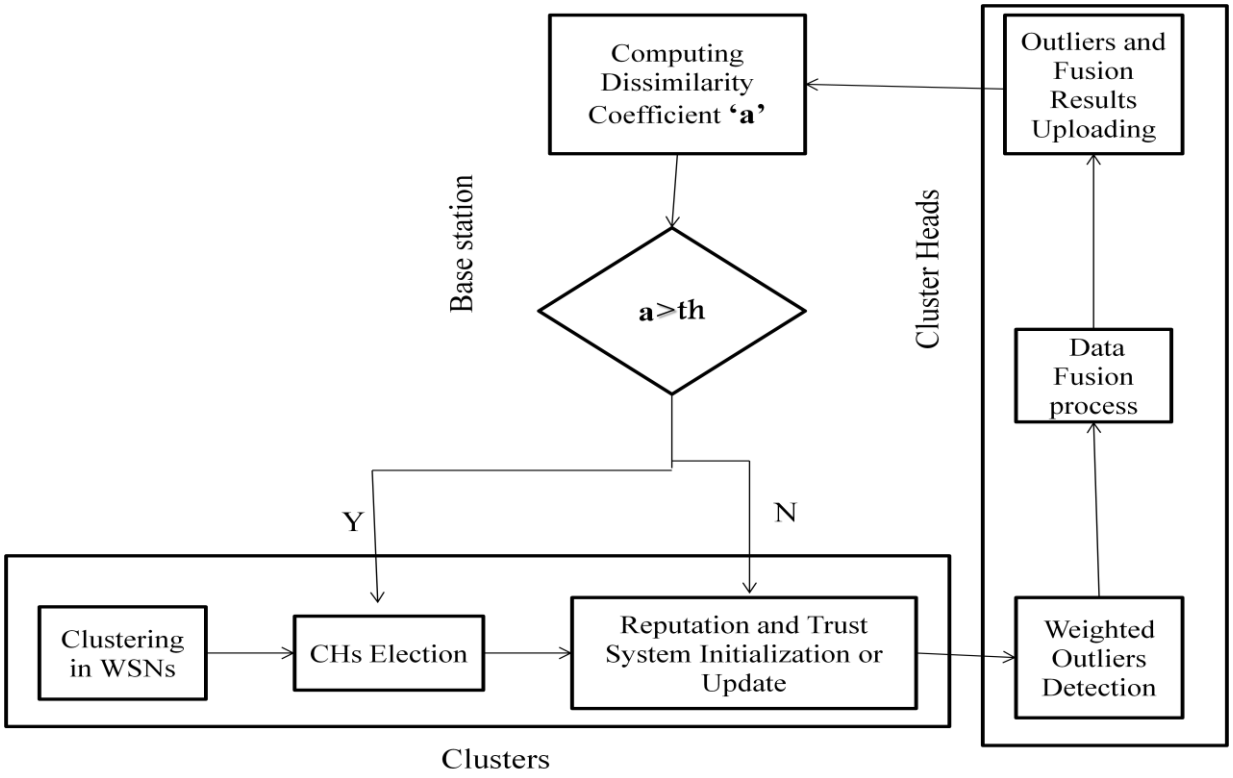

Figure 1: Architecture of Data fusion

5. MODULES

In the proposed system is divided into 3 modules.

Module 1: Cluster

Module 2: Cluster heads

Module 3: Base station

These modules are explained as follows.

5.1 Cluster

Three steps are involved in this cluster module (i) clustering in WSN, (ii) CHs Election, (iii) Constructing reputation & trust table and uploading.

5.1.1 Clustering in WSN and CHs Election

Clustering is one of the most challenging issues in wireless sensor networks. In WSN each sensor nodes should be known about the information of location which is used to form the cluster. Clustering is formed based on the highest residual energy and communication cost. Among the group of nodes, two nodes are elected as a cluster head using centroid method. The minimum distance between the cluster nodes and the centroid point is elected as a cluster heads. Clustering can extend the network lifetime and reduce the network traffic. In these two CHs one is act as a master CH (MCH) another one is act as a vice CH (VCH).MCH receives and fuse the data from its member nodes and it sends to the VCH. The VCH transmits fuse data to the base station directly.

5.1.2 Reputation and trust table

Reputation-based frameworks where nodes maintain reputation of other nodes and use it to evaluate their trustworthiness which are used to provide scalable, find diverse node in WSN. Trust can simply be defined as the expectation of one node about the actions of other nodes. It is used by the first node to make a choice, when an action must be taken before the actions of other nodes. Reputation is defined as the opinion of one node about another node. Note that reputation is not a physical quantity but it is a faith; it can only be used to statistically expect the future behavior of other nodes. Trust is a function of insecurity. The level of trust can be measured by a continuous real number, referred to as the trust value.

Trust and reputation-based systems are constructed in one of the following three ways:

1. All the nodes in the networks are initially assumed to be reliable. Every node trusts other nodes in the network. The reputations of the nodes reduce with every bad behavior.
2. Every node is considered to be unreliable in the system bootstrapping stage, and the nodes do not trust each other initially. The reputations of nodes with such system increase with every good behavior.
3. Every node in the network is considered to be neither reliable nor unreliable. All nodes start with a neutral reputation value to. With every good or bad behavior, the reputation value is increased or decreased respectively.

In the network reputation systems can be categorized into two groups: (i) centralized and (ii) distributed. In centralized systems, one central node maintains the reputations of all nodes in the network. This central node can be a source of security, susceptibility and performance block in the system. In distributed systems, each node maintains reputation information of all the nodes about which it is involved. In such systems, maintaining consistency in reputation values maintained in different nodes may be a major challenge. In a distributed system each node may maintain reputation of the nodes that are within its communication range, or may

maintain reputation information of all the nodes in the network. In sensor network applications, every node maintains reputation information only for its neighbors. This reduces the memory overhead for reputation information maintenance.

5.2 Cluster Heads

Weighted outlier detection and data fusion process are performed in this module. These steps are explained as follows.

5.2.1 Weighted outlier detection

Outlier: A data object that deviates significantly from the normal objects as if it were generated by a different mechanism. Example: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky likes that. Outliers are different from the noise data because Noise is random error or variance in a measured variable and it should be removed before outlier detection. Outliers are violates the mechanism that generates the normal data. Applications of the outliers are (i) Credit card fraud detection, (ii) Telecom fraud detection, (iii) Customer segmentation, (iv) Medical analysis.

5.2.1.1 Clustering based detection methods

Outlier detection methods can be categorized into two ways (i) Based on whether user-labeled examples of outliers can be obtained: Supervised, semi-supervised vs. unsupervised methods (ii) Based on assumptions about normal data and outliers: Statistical, proximity-based, and clustering-based methods. Here we can implement clustering-based methods because it is sufficient for our process. Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.

5.2.2 Data fusion process

In this paper introduces a new Bayesian fusion algorithm to merge more than one trust component (data trust and communication trust) to gather the overall trust between nodes. Bayesian fusion algorithms are the most suitable tools to join trust component.

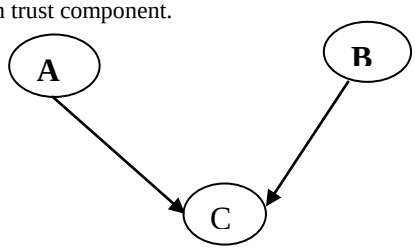

Figure 2: Bayesian future structures

The Bayesian fusion structure is illustrated in Figure 2. Here 'A' represents the communication trust, 'B' represents the data trust and C represents the total trust. Using Bayes' theorem, the probability of the total trust, given the data trust and communication trust, can be obtainable, as shown in (1).

$$P(C/A,B)=\frac{P(A/C,B)*P(C/B)}{P(A/B)} \quad (1)$$

A node is always communicating but not reporting the data, B and A are independent. Because of that independence, the probability function $P(A/C,B)$ in (1) can be presented, as in (2),

$$P(A/C,B)=P(A/C) \quad (2)$$

By substituting equation (2) in equation (1), the probability of the total trust will be as given in (3),

$$P(C/A,B)=\frac{P(A/C)*P(C/B)}{P(A/B)} \quad (3)$$

Applying Bayes' theorem, $P(A/C)$ can be calculated as in (4),

$$P(A/C)=\frac{P(C/A)*P(A)}{P(C)} \quad (4)$$

By substituting (4) in (3), the result is given in (5),

$$P(C/A,B)=\frac{P(C/A)*P(C/B)*P(A)}{P(A/B)*P(C)} \quad (5)$$

Equation (5), after ignoring the normalizing factor and the other constants, it can be seen that the probability of the combined trust C is mainly equal to the multiplication of the probabilities of both trust components, B and A, as shown in (6),

$$P(C/A,B)=P(C/A)*P(C/B) \quad (6)$$

In this way data fusion performs by both CHs. After perform the outliers and fusion process both results are uploaded to the base station module.

5.3 Base station module

The main function of Base Station Module is computing dissimilarity coefficient, denoted by *"a"*, and comparing with a threshold *"th"*, which is preset by users. We define *"a"* based on both two data fusion results F_1 , F_2 and the outlier lists $O_1 O_2$ as follows:

$$a=\varepsilon_1*\frac{2*|F_1-F_2|}{|F_1+F_2|}+\varepsilon_2*(1-\frac{|O_1\cap O_2|}{|O_1\cup O_2|}) \quad (7)$$

Where ε_1 and ε_2 are the weights preset by the users and the sum of them equals to 1. $|O_1\cap O_2|$ is the number of the common outliers. $|O_1\cup O_2|$ is the number of all the outliers . The threshold is another parameter preset by the users. In simulation part, ε_1 and ε_2 both are set to be 0.5 and *th* is set to be 0.2. If $a \le th$, which means that both the cluster heads are credible, only the reputation and trust system should be updated.

6. SIMULATION RESULTS

In the simulation, the ns-2 simulator implements, here 50 nodes are created and each sensor node begins with only 0.01 J of energy. Simulation results are discussed in bellow.

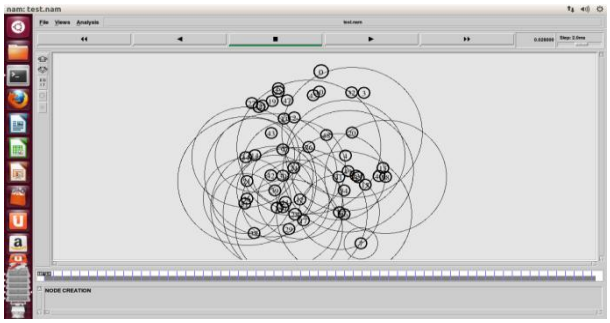

Figure 3: Node Creation

In figure 3, 0 to 49 nodes are created. In figure 3.1 is showed cluster formation. Here four clusters are formed based on the distance or proximity. Each cluster differentiates from the different colors.

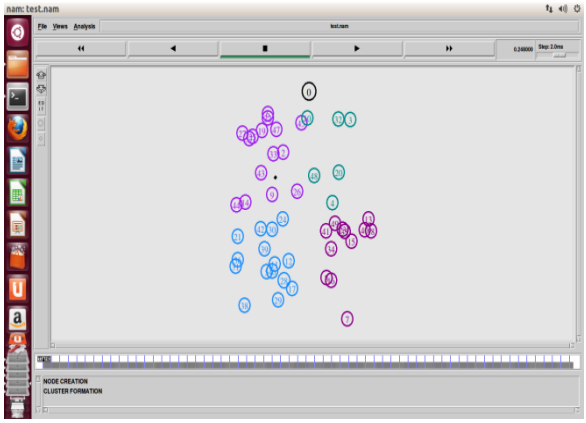

Figure 3.1: Cluster Formation

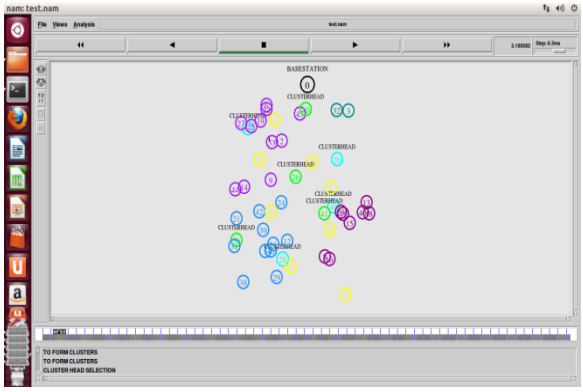

Figure 3.2: CHs Selection

In figure 3.2 is described the CHs selection. In this result blue color node represents MCH as well as green color node represents VCH.Each cluster has two CHs.

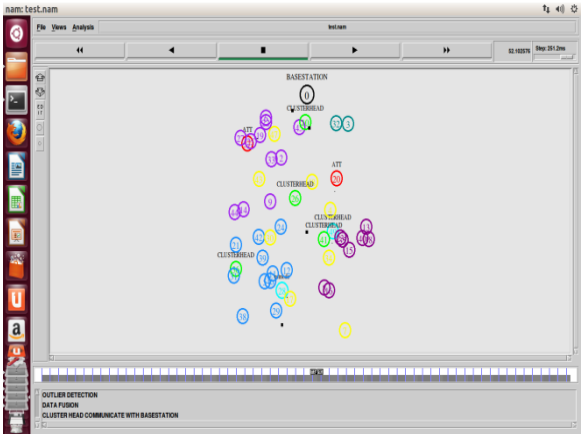

Figure 3.3: Weighted Outliers Detection

Weighted outliers detection explained in figure 3.3, here some nodes are in yellow color which is produced number of outliers as well as some nodes are in red color which CH have high dissimilarity. In this base station process dissimilarity coefficient is compared with threshold value. In our simulation part dissimilarity is high so CHs reelection process is done; it is shown in figure 3.4

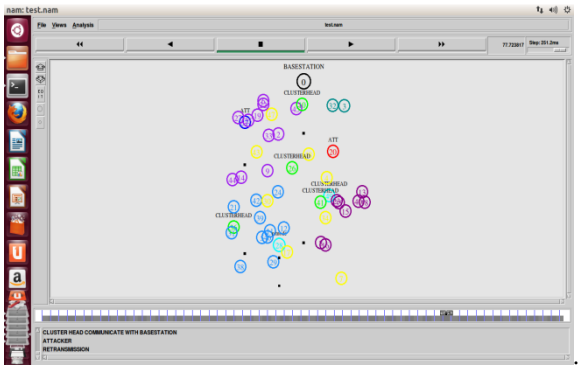

Figure 3.4: Base station Process

7. CONCLUSION

In this proposed work is discussed the problem of secure and accurate data fusion in WSNs. As well as discussed how to avoid selecting the compromised sensor nodes as cluster heads, detect the outliers and update the reputation and trust systems. Outliers are discussed in section 5.2. In this paper DCHM used for secure and accurate data fusions as well as Simulation results are discussed clearly in section 6. Accuracy of data fusion results is also improved because the compromised sensor nodes are detected and eliminated in the networks. Simulation proved this model performs well in improving the security and accuracy of data fusion. Main objective of this future work is to retrieve the collective information from the sink. And use DSR (Destination Source Routing) protocol for reducing the data loss.

8. REFERENCES

[1] Amit Kumar, Dhirendra Srivastav, "Simulator for Energy Efficient Clustering in Mobile Ad Hoc Networks." CS & IT-CSCP 2012.

[2] Seema Bandyopadhyay and Edward J. Coyle, "Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks." IEEE INFOCOM 2003.

[3] Bakhta Meroufel, Ghalem Belalem, "Clustering-based data in ad-hoc networks." Proceedings ICWIT 2012.

[4] Bhoopendra Singh M Tech, Mr. Gaurav Dubey, "Report on Reputation Based Data Aggregation for Wireless Network." International Journal of Computational Engineering Research, Volume 03, Issue 6.

[5] Saurabh Ganeriwal,Mani B. Srivastava, "Reputation-based Framework for High Integrity Sensor Networks." SASN'04, October 25, 2004.

[6] Sanjeev SETIA, Sankardas ROY, "Secure Data Aggregation in Wireless Sensor Networks." Proceedings ICWIT 2012.

[7] Jyoti Rajput, Naveen Garg, "A Survey on Secure Data Aggregation in Wireless Sensor Network." International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[8] Ossama Younis, Marwan Krunz, "Node clustering in wireless sensor networks: Recent development and deployement changes." IEEE Network May 2006.

[9] Suman Banerjee, Samir Khuller, "A Clustering Scheme for Hierarchical Control in Multi-hop WirelessNetworks." Algorithmica, Volume 20,1998.

[10] Jaydip Sen, "A Survey on Reputation and Trust-Based Systems for WirelessCommunication Networks." ACSC 2008.

[11] Ossama Younis, Marwan Krunz, "Node clustering in wireless sensor networks: Recent development and deployement changes." IEEE Network, May 2006.

[12] Hani Alzaid, Ernest Foo, Juan Gonzalez Nieto, "Secure Data Aggregation in Wireless SensorNetwork: a survey." ACSC 2008.

[13] Saurabh Ganeriwal, Laura K. Balzano, "Reputation-based Framework for High Integrity Sensor Networks." ACM Transactions on Sensor Networks, March 2007.

[14] S.G.Santhi, Dr.K.Venkatachalapathy, "energy consumption based rejoin procedure for cluster tree in 802.15.4 sensor networks", International Journal of Scientific & Engineering Research (IJSER), Volume 4, November 2013.

[15] Ramesh Rajgopalan, "Data Aggregation techniques in sensor networks: survey."Computer science commons 2006.

[16] Selvadurai Selvakennedy, "An Energy-Efficient Clustering Algorithm for Multihop Data Gathering in Wireless Sensor Networks." Journal of Computers, April 2006.

[17] Audun Jøsang, "The Beta Reputation System." 15th Bled Electronic Commerce Conference, June 2002.

[18] S.G.Santhi, K.Venkatachalapathy, "Ant based Multiple Cluster Tree Routing for 802.15.4 Sensor Networks." International Journal of Computer Applications (0975 – 888), Volume 48, June 2012.

[19] Liu Xiang, Jun Luo, "Compressed Data Aggregation: Energy Efficient and High Fidelity Data Collection." Proceedings of the 8th IEEE SECON, 2011.

[20] S.G.Santhi, K. Chitralakshmi, "Mobility Based Tree Construction for ZigBee Wireless Networks", International Journal of Computer Science & Engineering Technology (IJCSET).January 2014.

[21] Robert Hummel, Sameera Poduri, Franz Hover, Urbashi Mitra, Guarav Sukhatme, "Mission Design for Compressive Sensing with Mobile Robots", Submitted 2011.

[22] David L. Donoho," Compressed Sensing" IEEE transactions on information theory, Volume 52, April 2006.

9. AUTHOR'S PROFILE

S.G.SANTHI obtained her Bachelor's degree in Computer Science & Engineering from Mookambigai College of Engineering, Bharathidasan University, Tamilnadu, India in 1992. Then she obtained her Master's degree in Computer Science & Engineering from Annamalai University, Tamilnadu, India in 2005.She is currently working as an Assistant Professor in the Department of Computer Science & Engineering, Faculty of Engineering & Technology, Annamalai University. She is having 15 years and 9 months of experience in teaching. She has published more than 15 Research papers in International & National Conferences. Her field of interest includes Wireless Sensor Networks. She is a life member in various professional bodies like ISTE, CSI. Etc.,

R.RAMYA obtained her Bachelor's degree in Information Technology from Annamalai University, Tamilnadu, India in 2013. Now she is doing her Master's degree in Computer Science & Engineering from Annamalai University, Tamilnadu, India. She has presented a paper in international conference, organized by mahabarathi engineering college, chinnasalem, Tamilnadu, India in 2015. She also presented a paper in COMPSEM organized by the Computer Science & Engineering from Annamalai University, Tamilnadu, India in 2014.