A DISTRIBUTED ENERGY EFFICIENT CLUSTERING ALGORITHM FOR DATA AGGREGATION IN WIRELESS SENSOR NETWORKS

SEYED MOHAMMAD BAGHER MUSAVI SHIRAZI, MARYAM SABET*, AND MOHAMMAD REZA PAJOOHAN

Department of Computer and Electrical Engineering, Yazd University, Yazd, Iran.

**Corresponding author: maryam.sabet@stu.yazd.ac.ir*

(Received: 1st March 2017; Accepted: 17th Oct 2017; Published on-line: 1st June 2018)

https://doi.org/10.31436/iiumej.v19i1.825

ABSTRACT: Wireless sensor networks (WSNs) are a new generation of networks typically consisting of a large number of inexpensive nodes with wireless communications. The main purpose of these networks is to collect information from the environment for further processing. Nodes in the network have been equipped with limited battery lifetime, so energy saving is one of the major issues in WSNs. If we balance the load among cluster heads and prevent having an extra load on just a few nodes in the network, we can reach longer network lifetime. One solution to control energy consumption and balance the load among nodes is to use clustering techniques. In this paper, we propose a new distributed energy-efficient clustering algorithm for data aggregation in wireless sensor networks, called Distributed Clustering for Data Aggregation (DCDA). In our new approach, an optimal transmission tree is constructed among sensor nodes with a new greedy method. Base station (BS) is the root, cluster heads (CHs) and relay nodes are intermediate nodes, and other nodes (cluster member nodes) are the leaves of this transmission tree. DCDA balances load among CHs in intra-cluster and inter-cluster data communications using different cluster sizes. For efficient inter-cluster communications, some relay nodes will transfer data between CHs. Energy consumption, distance to the base station, and cluster heads' centric metric are three main adjustment parameters for the cluster heads election. Simulation results show that the proposed protocol leads to the reduction of individual sensor nodes' energy consumption and prolongs network lifetime, in comparison with other known methods.

ABSTRAK: Rangkaian sensor wayarles (WSN) adalah rangkaian generasi baru yang terdiri daripada nod-nod murah komunikasi wayarles. Tujuan rangkaian-rangkaian ini adalah bagi mengumpul maklumat sekeliling untuk proses seterusnya. Nod dalam rangkaian ini dilengkapi bateri kurang jangka hayat, jadi simpanan tenaga adalah satu isu besar dalam WSN. Jika beban diimbang antara induk kelompok dan lebihan beban dihalang pada setiap rangkaian iaitu hanya sebilangan kecil nod pada tiap-tiap kelompok, jangka hayat dapat dipanjangkan pada sesebuah rangkaian. Satu penyelesaian adalah dengan mengawal penggunaan tenaga dan mengimbangi beban antara nod menggunakan teknik berkelompok. Kajian ini mencadangkan kaedah baru pembahagian tenaga berkesan secara algoritma berkelompok bagi pembahagian data dalam WSN, dikenali sebagai Pembahagian Kelompok Kumpulan Data (DCDA). Melalui pendekatan baru ini, pokok transmisi optimum dibina antara nod sensor melalui kaedah baru. Stesen utama (BS) ialah akar, induk kelompok-kelompok (CHs) dan nod penyiar ialah nod perantara, dan nod-nod lain (nod-nod ahli kelompok) ialah daun bagi pokok trasmisi. DCDA mengimbangi beban CHs antara-kelompok dan dalam-kelompok komunikasi data daripada kelompok berbeza saiz. Bagi komunikasi berkesan dalam-kelompok, sebahagian nod penyampai akan memindahkan data antara CHs. Penggunaan tenaga, jarak ke stesen utama dan induk kelompok metrik sentrik adalah tiga parameter pelaras bagi pemilihan induk kelompok. Keputusan simulasi protokol yang dicadang menunjukkan pengurangan penggunaan tenaga pada nod-nod sensor individu dan memanjangkan jangka hayat rangkaian, berbanding kaedah-kaedah lain yang diketahui.

KEYWORDS: *wireless sensor networks; clustering; data aggregatio; multi-hop communication; optimal transmission tree*

1. INTRODUCTION

A wireless sensor network consists of a number of sensor nodes deployed over a geographical area to monitor physical environmental conditions such as temperature, humidity, vibrations, seismic events, and so on. Normally, a sensor node is a tiny device composed of three basic subsystems: a sensor to gather data from the physical surrounding environment, a local data processing unit and storage, and a wireless communication device for data transmission. A power source is also available to supply the energy for processing and sensing subsystems. This power source is often a limited energy budget battery. Moreover, as the nodes may be deployed in a hostile or impractical environment, recharging the battery may not be possible or convenient. Thus, the wireless sensor network should have enough lifetime to fulfill the application requirements [1-2].

There are two variations of wireless sensor network (WSN) deployment: structured and unstructured. While in Structured WSN, the sensor nodes may be deployed in a pre-planned manner, the deployment of all or some nodes in unstructured WSN is random. An unstructured WSN consists of a dense collection of sensor nodes. Once deployed, the network is left unattended to perform monitoring and reporting functions. In an unstructured WSN, network maintenance, such as managing connectivity and detecting failures, is difficult since there are so many nodes developed in the network. The advantage of a structured network is that fewer nodes can be deployed with lower network maintenance and management cost. Fewer nodes can be deployed if nodes are placed at specific locations to provide coverage whereas ad hoc deployment may leave regions uncovered [3].

WSNs are also different from other networks in other aspects: they tend to be densely deployed, nodes are susceptible to failure, and rely on broadcast communications. Network topology is dynamic, in which links are established between nodes and may break due to various reasons, such as deliberate changes to the transmission power of the nodes, node failure, or mobility. Thereupon, maintaining a fully connected topology for such networks is a challenge that requires careful application of topology control [4].

With clustering algorithms, network performance could be enhanced, particularly for large scale of WSNs. Clustering algorithms are commonly used for such networks. Using the clustering approach, the network is prolonged and the control overhead of routing is limited. Network scalability is also increased, and the total network performance is advanced [5, 6].

In this paper, we propose an energy efficient clustering algorithm for data aggregation in wireless sensor networks in a fully distributed manner. It consists of two phases, cluster setup phase and data transmission phase. Cluster setup phase has four stages. These are cluster head (CH) selection, relay node candidate specification, cluster formation, and routing tree construction stages. In cluster setup phase CHs are selected based on multiple criteria, such as each node’s residual energy, its distance to the Base station (BS) and node centric metric. Qualified nodes are selected as candidate relay nodes that could undertake transmission of data between clusters. In cluster formation stage, each usual node joins to the proper CH, considering the three assumptions related to the number of received CH-messages and distance to the CHs. A lot of energy is wasted in inter-cluster data communications, the key issue of our protocol scheme is that in the routing tree construction stage, we construct a routing tree with a wiser approach; CHs form routing tree edges, BS at root and member nodes and leaves of this tree. For effective inter-cluster communications, in relay node selection, each CH selects an eligible relay node based on residual energy and distance to that CH, which directs CH data closer to the BS. According to the energy that will be consumed, each CH communicates with its parent cluster head or relay node and then finally, in the data transmission phase, data reaches the BS through a constructed routing tree. Effective CH selection and routing tree construction leads to power savings in more sensor nodes.

The remainder of this paper is organized as follows. Section 2 provides a brief survey of some closely related works. Section 3 describes the network and energy model in our algorithm. Section 4 presents the clustering protocol in detail. Section 5 analyzes several properties of the protocol. Section 6 evaluates the performance of the algorithm and analyzes the results obtained from simulations. Finally, Section 7 gives the concluding remarks.

2. LITERATURE SURVEY

Much research has been done to cope with the energy consumption constraint problem in wireless networks. Clustering techniques have mostly been the focus, since data is aggregated and fused in each CH and energy would be conserved in every sensor node. This would lead to longer lifetime of the overall wireless sensor network. However, these algorithms face some problems. One of these limitations is that the energy metric is not directly considered while selecting CHs. Also, more relevant parameters, such as distance, density of nodes, and number of hops to reach the BS, are available that are not considered in the selection of the basic nodes. Considering equal cluster sizes causes the early death of CHs near the BS and those in the denser areas. Besides, constructing an optimal routing tree based on an accurate method is a key to saving energy. Therefore, defining a protocol to overcome these shortages is needed. Some proposed algorithms are listed herein. Low energy adaptive clustering hierarchy (LEACH) [7] is a well-known distributed clustering protocol in which most nodes transmit to cluster heads, and the cluster heads aggregate and compress the data and broadcast their decisions to the base station. The remaining nodes join the closest cluster head in terms of the communication energy cost. Then the role of the cluster head is periodically rotated among the nodes to balance energy consumption, since cluster heads have the extra burden caused by a long-range transmission to a distant BS. Nodes that have been cluster heads cannot become cluster heads again. Thus, LEACH counteracts the non-uniform energy dispensation problem by rotating the CH role. However, in *LEACH*, node energy is drained quickly as in CH selection the energy parameter is not involved. A hybrid, energy-efficient, distributed clustering algorithm (HEED) [8] is also an energy-efficient protocol designed for sensor networks. It introduces a variable known as the cluster radius which defines the transmission power used for intra-cluster broadcast. The initial probability for each node to become a cluster head depends on its residual energy and other parameters like adjustment degree and intra-cluster communication cost. HEED relies on the assumption that cluster heads can communicate with each other and form a connected graph. In article [9], a cluster-based routing protocol for wireless sensor networks with a non-uniform node distribution which includes an energy-aware clustering algorithm and a cluster-based routing algorithm, is proposed. A cluster-based routing protocol (EADC) constructs clusters of equal sizes using competition range to balance the energy consumption of the cluster nodes. The criteria used to select cluster heads are the residual energy of each sensor node and some relevant parameters. Simultaneously, the routing algorithm increases forwarding tasks of the nodes in scarcely covered areas by forcing the cluster heads to choose the next hop with more remaining energy, less cluster members and not further away from the BS than the relay node. To balance the energy consumption among cluster heads a cluster-based inter-cluster routing algorithm that adjusts intra-cluster and inter-cluster energy consumption among sensor nodes is proposed. This protocol may prolong the network lifetime significantly.

Article [10] presents an energy constrained minimum dominating set based efficient clustering called (ECDS) to model the problem of optimally choosing cluster heads with energy constraints. A randomized distributed algorithm for the constrained dominating set runs in few rounds with high probability. Multiple extensions to the distributed algorithm for the energy constrained dominating set are also proposed. These extensions perform relatively well in terms of energy usage, node lifetime, and clustering time and, thus, are very suitable for wireless sensor networks. In article [11], the energy and coverage-aware distributed clustering protocol, which is an integrated protocol involving both energy and coverage in the clustering stage, is proposed. For different application coverage problems, corresponding coverage importance cost metrics are introduced. Only nodes with higher energy and smaller coverage importance are reliable to be selected as cluster heads and routers. The idea behind this approach is that nodes deployed in sparse areas, or those that can cover fewer interest points alone, are less selected as routing nodes such that they can collect data for a longer period of time. In article [12], an energy efficient clustering algorithm for data aggregation is proposed. The cluster head is selected considering the node's residual energy as well as the distance between the node and its neighbors. Furthermore, a type of data aggregation tree is constructed with the help of the clusters to reduce the amount of data transmission, which effectively extends the network's lifetime.

In order to solve the problem of minimum energy, data aggregation enhanced converge cost operations in wireless sensor networks paper [13] presents two algorithms: one based on minimum spanning tree and the other on single source shortest path spanning tree. These algorithms are shown to be energy efficient for different extreme values of data growth factor. Besides, the data latency was reduced by re-structuring the energy-efficient tree. The goal of the method proposed in article [14] is to lay the foundations to develop algorithms and techniques that simultaneously minimize the data gathering latency and balance the energy consumption among the nodes, thus maximizing the network lifetime. Following an incremental-complexity approach, several mathematical programming problems focused on different network performance metrics are proposed. In article [15], an energy-efficient distributed clustering protocol for wireless sensor networks is introduced. The proposed GESC protocol is based on a localized metric for measuring the value of a node in covering the neighborhood when rebroadcasting. The protocol achieves small and linear communication and computation complexity, respectively.

In article [16], a structure-free and energy-balanced data aggregation protocol, SFEB, is proposed. SFEB features both efficient data gathering and balanced energy consumption resulting from its two-phase aggregation process and the dynamic aggregator selection mechanism. In article [17], a cluster head weight selection method that takes service parameters for enhancing performance of the overall network has been discussed. In a clustering-based approach, one of the main concerns is the selection of appropriate cluster heads in the network and the formation of balanced clusters. Cluster heads are selected first in a weight metric based network; then a cluster formation is executed. This approach aims to conserve the energy of the sensors and balances load in the network. In paper [18], a distributed clustering and routing algorithm, jointly referred to as DFCR, is proposed. The DFCR uses a distributed runtime recovery of the sensor nodes due to sudden failure of the cluster heads. It takes care of the sensor nodes that have no CH within their communication range. The algorithm is shown to be energy efficient and fault tolerant.

In paper [19], energy-efficient clustering and power management schemes for virtual MIMO operation in a multi-hop WSN are suggested. Schemes are integrated into a comprehensive protocol, which involves clustering the WSN into several clusters. This protocol achieves energy efficiency by properly selecting the master CHs and slave CHs, adapting the antenna elements and powers in the inter-cluster communications phase, and using a cross layer MIMO-aware route selection algorithm for multi-hop operation. The conditions on the transmission powers of CHs and non-CHs that ensure the connectivity of the inter-cluster topology are established. Paper [20] focuses on designing a prediction-based approach. It exploits both spatial and temporal correlations to form highly stable clusters of nodes sensing similar values. The sink node uses only the local prediction models of cluster heads to forecast all readings in the network without direct communication.

In paper [21], clustering problem in lossy WSNs with a mobile collector is studied, where a large number of links are unreliable and retransmission is required over such links to provide successful data delivery. These retransmissions cause energy wasting in intra-cluster communications. The objective is to construct clusters in the network such that packets from cluster members can be transmitted to cluster heads over reliable links, and the mobile collector can collect data from cluster heads directly. Two distributed algorithms to construct one hop and k-hop clusters, are also introduced.

3. NETWORK AND ENERGY MODEL

In this section, we explain our assumed energy as well as network model. The network model comprises a BS and some sensor nodes that are distributed in a target area. We make a few assumptions as follows:

1. The network is shaped by N sensor nodes deployed in an M×M (m^2) square field.
2. The base station is located outside the sensing field and it doesn’t have any energy constraint so that it can increase its power to send data all around the network.
3. Nodes are deployed randomly and they are stationary after deployment. Our method determines a fix inter-cluster and intra-cluster data transmission range, since this works well on static networks.
4. Sensor nodes can be heterogeneous, which means that they can have different transmission power (which we set properly), but are equipped with an unchangeable battery. Our goal is to save energy in unchangeable sensors.
5. The nodes are equipped with power control capabilities to change their transmission power.
6. All sensor nodes are location-unaware. Therefore, no GPS is needed.
7. Each node has a unique ID to identify one to others.

The expended radio energy for transmitting l -bit data E_t, is based on Eqn. (1) [7].

$$=E_t \begin{cases} l \times E_{elec} + l \times \varepsilon fs \times d^2 & d < d_0 \\ l \times E_{elec} + l \times \varepsilon mpf \times d^4 & d \geq d_0 \end{cases} \qquad (1)$$

where d is the transmission distance, E_{elec} is the energy being dissipated to operate the transmitter or receiver circuitry per bit, εfs and εmpf are the amplifier energy factors for free space and multi-path fading channel models, respectively. d_0 is the threshold distance that depends on the environment and the amplifier energy factors. In our work, we consider both the free space (d^2 power loss) and multi-path fading (d^4 power loss) channel models depending on the distance d between the transmitter and receiver.

The radio also expends energy to receive l-bit data given by Eqn. (2).

$$E_r = l \times E_{elec} \tag{2}$$

Because of short distance between sensors in a cluster ($d < d_0$) free space model and long distance between CHs and the BS multi-path fading model is intended. According to Fig. 1, energy consumption for receiving data from cluster members, data aggregation, and transmission of aggregated data to the next cluster head is as Eq. (3).

$$E_{CH} = N \times l \times E_{elec} + (N+1) \times l \times E_{DA} + l \times (E_{elec} + \varepsilon mpd_c^4) \tag{3}$$

where N is the number of member nodes in a cluster, E_{DA} is data aggregation energy consumed by CH and d_c is distance to the next CH. Besides, energy consumption of CHs for transmitting data to the intermediate relay nodes follows Eq. (4).

$$E_{CH} = N \times l \times E_{elec} + (N+1) \times l \times E_{DA} + l \times (E_{elec} + \varepsilon fsd_r^2) \tag{4}$$

where N is the number of non-CH nodes in cluster k, E_{DA} is the data aggregation energy and d_r is distance to the next relay node. Therefore, the energy consumption model of CHs either follows Eqn. (3), or follows Eqn. (4), according to the method of determining the next hop node.

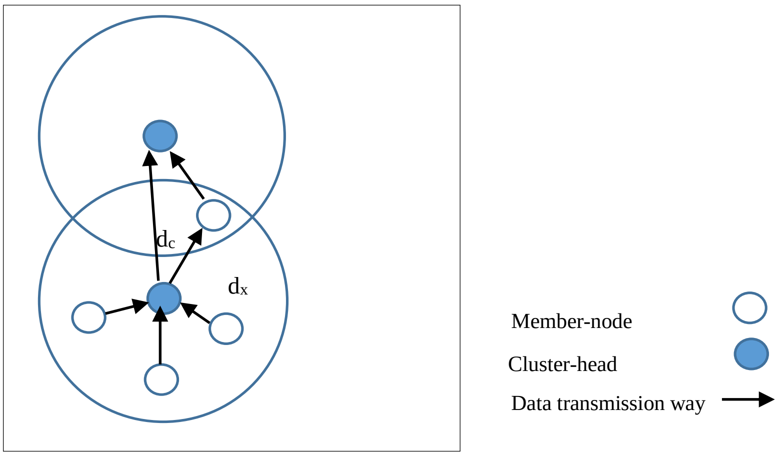

Fig. 1: Inter-cluster communication path.

4. CLUSTERING PROTOCOL

Distributed Clustering for Data Aggregation (DCDA) contains two different stages. In first stage, to have an effective cluster head selection, CHs would be selected based on domestic and local information. Cluster member nodes choose a deserving cluster head for data aggregation objectives and relay nodes characterized concurrently. Since the maximum amount of energy dissipation occurs in inter-network communications, thereupon with constructing an optimal spanning tree among sensor nodes, CHs would preserve energy in inter-network and intra-network data communications. In the data transmission stage, each CH navigates its data directly or via other CHs to the BS with multi-hop communications. According to the distance between CHs, each data could traverse through a routing tree with the help of relay nodes. By this approach, energy conserves in sensor nodes and hence leads to a longer lifetime of the sensor network. Descriptions of control messages, which are used in the proposed scheme, are illustrated in Table 1.

Table 1: Description of control messages

Message	**Description**
init-Msg	Tuple(token)
intro-Msg	Tuple(node-id,node-energy,node-distoBS)
CH-Msg	Tuple(node-id,node-energy,node-distoBS,node-ndeg)
routing-Msg	Tuple(node-id,node-energy,node-distoBS,node-ndeg,node-hopcount)
Relay-Msg	Tuple(node-id,node-energy,node-distoCH)
Join-Msg	Tuple(node-id, ch-id)
Schedule-Msg	Tuple(scheduleorder)

4.1 The Protocol Details

In the cluster setup phase of DCDA, CHs and relay nodes are specified and the routing tree is constructed among sensor nodes. In the data transmission phase, each member node sends its data to the corresponding CH. Afterward, each CH sends data directly or via an intermediate relay node to the BS. To increase the network lifetime and reduce the overhead of the algorithm, the data transfer phase is much longer than the cluster setup phase. In each round, the cluster heads are changed and a new path is created.

4.1.1 Min-Max Normalization

Min-Max Scaling or normalization means that one function linearly transforms real data values such that the minimum and the maximum of the transformed data take certain values frequently by 0 and 1. This depends on the context. For example the Eqn. (5) maps a value d of S to d' in the range $[new\_min_S, new\_max_S]$.[22]:

$$d' = \frac{[d - min_S]}{[max_S - min_S]} \times [new\_max_S - new\_min_S] + new\_min_S \tag{5}$$

where min_S is the minimum value of the attributes and max_S is the maximum value of the attributes. For normalization purposes, we want to map a value d of S to d' in the range [0,1], therefore we consider the value of new_max_S equal to one ($new\_max_S = 1$) and the value of new_min_S equal to zero ($new\_min_S = 0$) in the above Eqn. (5). So the formula is converted to the simple min-max normalization as Eqn. (6).

$$d' = \frac{[d - min_S]}{[max_S - min_S]} \tag{6}$$

Later on, to evaluate our results in an identical condition and eliminating the effects of different units, the data is scaled to a fixed range (0 – 1). So min-max normalization based on Eqn. (6) is used in all equations.

4.1.2 Cluster Setup Phase

a) CH selection stage

At the beginning, BS broadcasts an init-msg message. All nodes set their state as usual node. Each node that receives a message from BS and its residual energy is higher than threshold value ($Er_i > E_{threshold}$), takes part in the CH selection competition. CHs Candidate compute distance to the BS based on received signal power strength and broadcast intro-msg in Rc based on Eqn. (7). Competition radius (*Rc*) is a fraction of the transmission range (*Rr*). This equation shows how Rc is determined for each sensor node. The intro-msg contains information that is required for a CH selection procedure, such as residual energy and distance to the BS.

$$Rc = 1/2 \times (a \times dtoBS_i + b \times Er_i) \times Rr \tag{7}$$

where $dtoBS$ is each node's distance (Euclidean distance) to the BS and Er is the residual energy of that node. a and b are real values uniformly distributed in [0,1] such that their summation is equal to one. CHs near the BS have to collect and aggregate data from member nodes. Besides, they are located in the path of the entire network traffic directed to the BS. For load balancing between inter-cluster and intra-cluster communications, clusters closer to the BS should have a shorter radius. Moreover, when energy decreases, cluster radius has to be decreased and a smaller number of member nodes would be covered.

Each CH candidate that receives an intro-msg from others, computes the average energy of neighbor nodes as in Eqn. (9), a node –centric metric with Eqn. (10) and $funCH_i$ value with Eqn. (8) and compares these values. If this new node has a maximum value, then it becomes a CH and promulgates a CH-msg message in *Rc*.

$$funCH_i = a \times \frac{Er_i}{Er_{av}} + b \times \frac{1}{dtoBS_i} + c \times \frac{1}{cent_i} \tag{8}$$

$$Er_{av} = \frac{1}{m} \times \sum E_{nei} \tag{9}$$

$$cent_i = \frac{1}{m} \times \sum dtos_i \tag{10}$$

where Er_i denotes the residual energy of node S_i, and E_{nei} is the residual energy of neighbor node. $dtoBS_i$ is used to measure node's distance to the BS. m is the number of adjacent nodes. $dtos_i$ is each adjacent node's distance to nearby node. a, b and c are real values uniformly distributed in [0,1] such that their sum is equal to one. A node with more energy, less distance to the BS, and fewer member nodes' average distance to the CH would be a better choice to undertake the CH task. By selecting CHs closer to the cluster members, less intra-cluster transmission energy would be consumed. In the proposed scheme, just CHs take the responsibility of data aggregation in the network.

b) Relay Node Candidate

Each node that receives a CH-msg from at least two CH and its residual energy is more than the threshold value ($Er_i > Er_{th}$) changes its state to relay node candidate and determines its distance to each CH based on received signal strength. After distance estimation, relay node candidates compute d with Eqn. (11) and send this value to CHs in relay-msg message.

$$d = d_x + d_y \tag{11}$$

where d_x and d_y are the distance of the relay node candidate to each of two CHs used in relay node selection decision.

CHs that received a relay-msg in routing tree construction stage decide whether to send data directly or via relay nodes to the upper CH to finally reach the BS. d_x and d_y are illustrated in Fig. 2. The pseudo-code segment 1 gives the details of the CH selection stage.

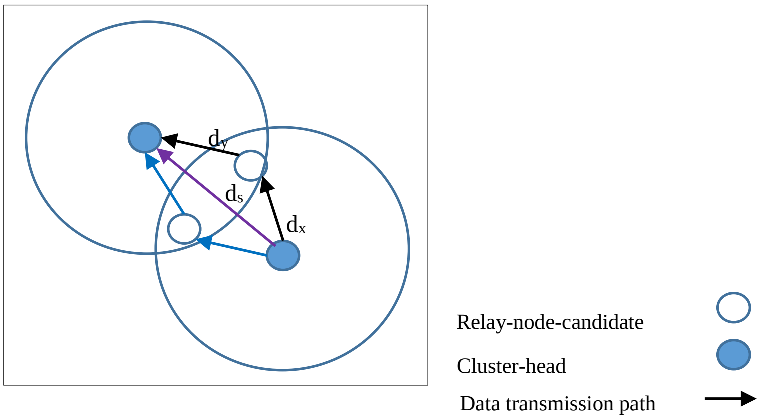

Fig. 2: Three inter-cluster communication paths.

```
Begin (CH selection algorithm)
  If $Er_i > E_{threshold}$ then
 $state \leftarrow candidate\ CH$
    Calculate Rc and broadcast intro-msg in Rc
  End
  If $state = candidate\ CH$ then
    Calculate $funCH$ and compare with neighbors
    If $Max\ value\ of\ funCH_{neigh} < funCH_i$ do
        $state \leftarrow CH$
      Broadcast CH-msg in Rc
    End
  End
  If receive CH-msg from at least two CHs && $Er_i > Er_{th}$ then
    $state \leftarrow candidate\ relay\ node$
    Calculate $d$ and send relay-msg to CHs
  End
End
```

The pseudo-code segment 1

c) Cluster Formation Stage

According to CH-msg information each member node has to choose a proper CH from the eligible CH set to join it. To do this, some assumptions are considered as follow.

1. If a member node receives a CH-msg from just one CH, it joins that cluster by broadcasting a join-msg and sets its next hop to be that specific CH.
2. If a member node receives CH-msg from two or more CHs with equal distance, it joins the CH with more residual energy and less distance to the BS in a low density area. Eqn. (12) indicates the level of competence for each CH.

$$funJ_i = a \times ErCH_i + b \times 1/CHdtoBS_i + c \times 1/den_i \qquad (12)$$

where $ErCH_i$ is the residual energy of CH, $CHdtoBS_i$ is distance of CH to the BS and den_i is density of cluster or number of CH's neighbors. a, b and c are real values uniformly distributed in [0,1] such that their summation is equal to one.

3. If a member node receives a CH-msg from CHs with different distances, it computes $funJN_i$ with Eqn. (13) and joins a cluster with higher $funJN_i$ value. This means a CH with less distance from a nearby member node and closer to the BS is more qualified to become a next hop node. Also, residual energy and proximity degree are other CH selection criteria.

$$funJN_i = a \times ErCH_i + b \times \frac{1}{CHdtoBS_i} + c \times \frac{1}{dtoCH_i} + d \times \frac{1}{den_i} \tag{13}$$

where $ErCH_i$ is the residual energy of the CH, $CHdtoBS_i$ is a distance of CH to the BS, $dtoCH_i$ is distance of member node to the CH and den_i is the density of the cluster. a, b, c and d are real values uniformly distributed in [0,1] such that their sum is equal to one. The pseudo-code segment 2 gives the details of this stage.

```
Begin (cluster formation algorithm)
  If $state! = CH$ && has not sent Join-Msg then
    If receive CH-msg from one CH
      $next\ hop \leftarrow selected\ CH$
      Send join-msg to selected CH
    End
    If receive CH-msg from at least two CHs
      If CHs have equal distances
        Calculate $funJ_i$ and find the maximum
        $next\ hop \leftarrow CH\ with\ maximum\ funJ$
        Send join-msg to selected CH
      End
      If CHs have unequal distances
        Calculate $funJN_i$ and find the maximum
        $next\ hop \leftarrow CH\ with\ maximum\ funJN_i$
        Send join-msg to selected CH
      End
    End
  End
End
```

The pseudo-code segment 2

d) Routing Tree Construction Stage

In transmission of data through the routing tree in the network, because of the long distance between CHs, the maximum amount of energy is wasted in inter-cluster data communications. Upper level nodes near the BS whose distance to the BS is less than *Rr* ($CHdtoBS_i < Rr$) and hence, could communicate with the BS without any interface, set their next hop to be the BS and others have to specify proper relay nodes. Finally, the routing tree is constructed among sensor nodes.

e) Routing Details

At the beginning, upper level nodes broadcast a routing-msg in *Rr* and set a hop-count parameter equal to 1. Single hop nodes receive this message and perform a routing tree construction procedure. Each CH that receives messages selects its next CH candidate based on Eq. (14) and increases hop-count value by one ($hopcount = hopcount + 1$), then broadcasts a routing-msg for lower level CHs and $funNCH$ is computed in the same way. This will be continued until all CHs in the network have been reached.

$$funNCH_i = \ a \times \ ErCH_i \ + \ b \times \frac{1}{CHdtoBS_i} + \ c \times \frac{1}{hopcount_i} + d \times \frac{1}{den_i} \tag{14}$$

Where $ErCH_i$ is the residual energy of CH, $CHdtoBS_i$ is a distance of CH to the BS, $hopcount$ is the number of hops to the BS and den_i is density of the cluster. a, b, c and d are real values uniformly distributed in [0,1] such that their sum is equal to one. According to $funNCH$, each CH selects next CH candidate, which has higher residual energy, less distance to the BS and lower number of nodes in the path to the BS (less data transmission) in a low density area. Fig. 3 shows some hypothetical inter-cluster paths.

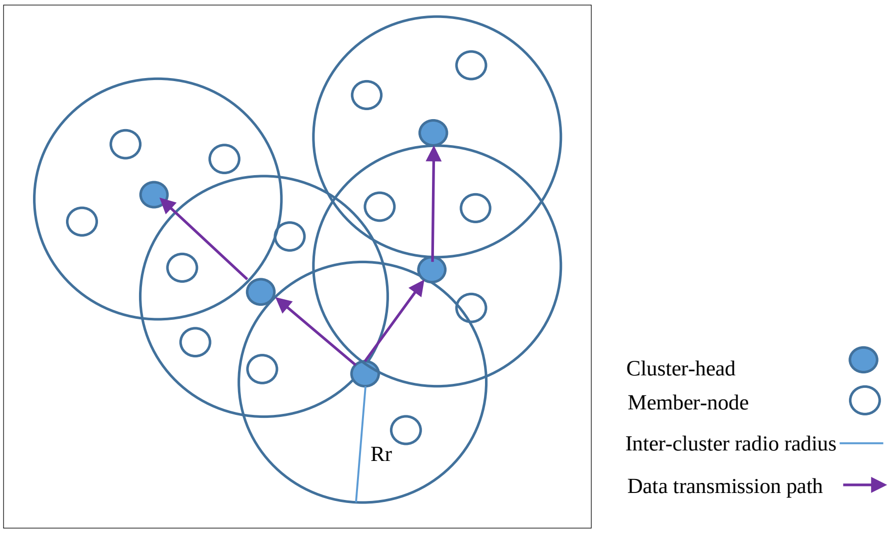

Fig. 3: Some assumptive paths between CHs.

f) Relay Node Selection

After determining the next CHs candidate, each CH decides to choose a relay node as intermediate or next CH candidate to be the final next hop node. For next hop selection, CHs use Eq. (15). At first, each CH specifies the most appropriate node as a relay and then compares energy expended for transmitting l-bit data to the next hop.

$$funR_i = a \times \ \frac{Er_r}{E_{max}} + b \times \ \frac{1}{d}$$
(15)

Where Er_r is the residual energy of relay node candidate and E_{max} is the maximum initial energy of nodes in the network. Also d is defined in Eq. (11) which is the summation of distances to the CHs. So a node with more residual energy with less d value is more competent for undertaking the relay node task. After the relay nodes competition, CHs specify the next hop by comparing the result of Eco_r and Eco_{CH} values defined in Eqn. (16) and Eqn. (17), respectively.

$$\begin{aligned} Eco_r &= l \times (E_{elec} + \varepsilon fsd_x^2) + l \times E_{elec} + l \times \left(E_{elec} + \varepsilon fsd_y^2\right) + l \times E_{elec} \\ &= 4 \times l \times E_{elec} + l \times \varepsilon fs \times (d_x^2 + d_y^2) \end{aligned} \tag{16}$$

$$Eco_{CH} = l \times (E_{elec} + \varepsilon mpd_s^4) + l \times E_{elec} \tag{17}$$

$$= 2 \times l \times E_{elec} + l \times \varepsilon mp \times d_s^4$$

In Fig. 2, d_x, d_y and d_s are illustrated. After data transmission, according to results obtained from Eqn. (16) and Eqn. (17), the one leading to lower energy consumption would be selected as a next hop. The pseudo-code segment 3 gives the details of this stage.

4.1.3 Data Transmission Phase

a) Intra-Cluster Communication

To avoid collision in inter-network transmissions, adjacent clusters use different transmission channels. Each CH based on join-msg messages received from member nodes determines a corresponding transmission time slot with the TDMA aid and declares it with schedule-msg to all cluster members. Hence, each member node transfers its data to the CH during the determined time slot. This method allows sensor nodes to sleep during other time slots and save more energy.

b) Inter-Cluster Communication

CHs aggregate received data from cluster members and then forward it to the BS along the constructed routing tree. In addition, they may undertake a forwarding task of other CHs data and conduct them through the path to the BS. This leads to reduction of data packets. As a result, network lifetime increases and more energy is saved in each sensor node. Fig. 4 illustrates one data transmission route among elected CHs and relay nodes through the routing tree.

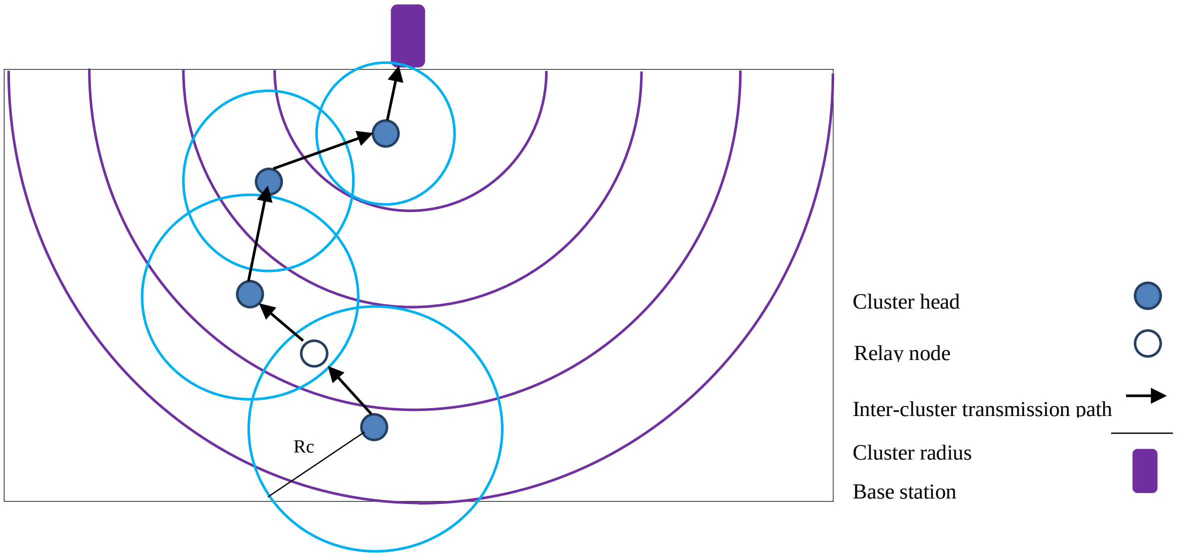

Fig. 4: A data transmission path through a routing tree.

4.1.4 Protocol Analysis

Theorem 1: *There is at most one cluster head in every cluster radius Rc.*

Proof: As previously mentioned with Eqn. (8), each CH candidate's result value is compared to the value calculated for its neighbors. If there is any adjacent node with more value in the cluster radio radius (*Rc*) it would desist the competition and set its state to be a cluster member node. Otherwise, it broadcasts the CH-msg within cluster radius (*Rc*). As soon as each node receives CH-msg from neighbors, it gives up the competition and changes its state to a member node. So in every cluster radius, just an exclusive node with maximum value will undertake a CH task and other nodes would join clusters as member nodes. Therefore, there is no more than one cluster head within each cluster radius *Rc*.

Theorem 2: *The cluster head set generated by DCDA can cover all nodes.*

Proof: In DCDA each node would be a CH, a cluster member node or a relay. According to CH selection function a node with the most competent would be specified as a CH and broadcast CH-msg in cluster radius. Each node that receives this message becomes a member node and would be covered with a CH. Relay nodes are those that receive a CH-msg from at least two CHs so they would belong to a cluster. Any node that has not received the CH-msg, forms a cluster alone and becomes a single CH. Thus, the CH set generated by DCDA can cover all nodes in the network and avoid the generation of isolated points [23].

Theorem 3: *The overhead complexity of control messages in the network is O(N).*

Proof: At the beginning of each round, all nodes broadcast intro-msg. Thus, at most, N intro-msg messages are exchanged in the whole network. In each round, each CH broadcasts a CH-msg, routing-msg, and a Schedule-msg. Besides, any relay node candidate would broadcast a relay-msg in the network. Each member node joins a CH with a Join-msg announcement. For example, if the number of clusters is c, the total number of Join-msg is equal to $N - c$, the maximum number of relay-msg is equal to $N - c$ and the number of each other message is c. Thus, the maximum number of control messages exchanged in the whole network follows Eqn. (18) and the minimum Eqn. (19).

$$N + (N - c) + (N - c) + c + c + c = 3N + c \tag{18}$$
$$N + (N - c) + c + c + c = 2N + 2c \tag{19}$$

Therefore, the overhead complexity of control messages in the network is O(N).

Theorem 4: *DCDA can setup the clustering topology in O(1) time.*

Proof: DCDA performs a distributed clustering strategy. Thus, the time complexity of the entire network is equal to that of a single node O(1). Therefore, the time complexity is constant and not related to the network size.

5. SIMULATIONS

In this section, we analyze the performance of a DCDA algorithm via simulations using NS2.We assume that the BS is fixed and out of the monitoring region. 100 nodes are randomly disseminated in a (200×200) m^2 network field with uniform distribution. The simulation parameters are summarized in Table 2. We performed 100 simulations independently and the statistics are averaged over these 100 runs. Fig. **5** shows the network topology of the scenarios.

We compare our algorithm with HEED (a hybrid energy-efficient distributed clustering algorithm) and LEACH (low energy adaptive clustering hierarchy), two known clustering algorithms. These are two main protocols proposed for clustering techniques in WSNs they perform clustering in a distributed manner in which each node independently selects its state as we proposed. However, our algorithm overcame the shortage of these methods. It considers much more relevant parameters and different aspects when developing a routing tree to enhance energy efficiency in data transmission.

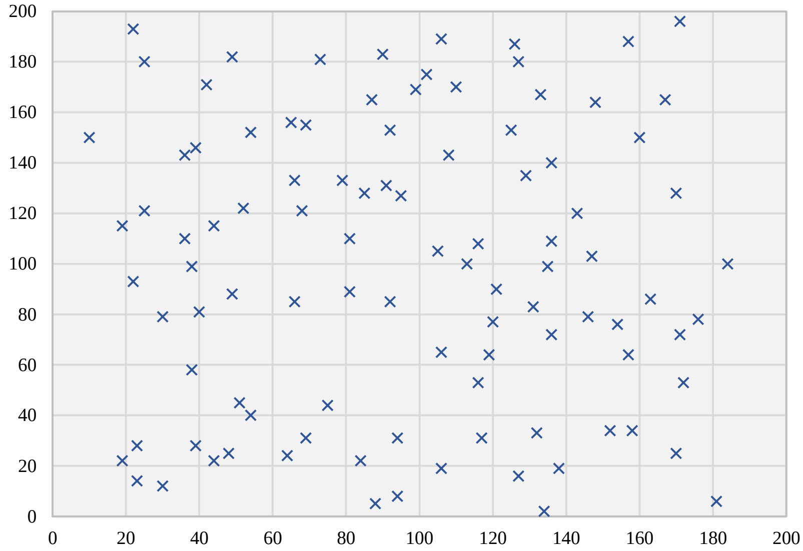

Fig. 5: Network topology.

As previously mentioned, LEACH [7] forms clusters using a distributed algorithm. It contains several rounds and each round is divided into two phases that are the establishment of cluster and data communication. In the setup stage, CHs are specified randomly and clusters are formed. In steady state phase, cluster members transmit sensed data to CH. After data aggregation each CH sends data directly to BS. Due to data fusion and communication with BS, cluster heads consume more energy. Besides, HEED [8] is also a distributed clustering protocol for ad-hoc sensors, and it uses a hybrid criterion for cluster head selection that considers the residual energy of the node and a secondary parameter, such as the node's proximity to its neighbors or the node's neighbor degree. Nodes join clusters such that communication cost is minimized. It assumes quasi-stationary networks where nodes are location-unaware and have equal significance. It exploits the availability of multiple transmission power levels at sensor nodes.

Table 2:Parameters of simulations

Parameter	Value
Number of nodes	100
Sensor field	200 m × 200 m
BS location	(250,100)
Data packet size	500 bytes
E_{elec}	50 nJ/bit
εfs	10 $pJ/(bit\ m^2)$
εmp	0.0013 pJ/(bit m4)
Initial energy of nodes	1–4 J
E_{sen}	0.2 J/bit
E_{com}	5 nJ/(bit signal)

5.1 Cluster Head Distribution

To evaluate our algorithm in different cases, first we set $Rr = 160$. Additionally, we consider equal weight for each part used in the equations. For example, in Eqn. (7) $a = b = 0.5$. Fig. **6** shows the number of cluster heads in 30 randomly selected rounds for LEACH, HEED and DCDA. The number of CHs varies wildly in LEACH since in LEACH, cluster heads are determined by a probability with random number generation.

Therefore, the number of cluster heads in LEACH varies and is not controlled. In HEED and DCDA, their competition radius ensures that there is only one cluster head within any cluster radius and good distribution of CHs over the network is ensured in most rounds. Compared with HEED and LEACH, DCDA can obtain a better distribution of CHs in the network because it considers residual energy, distance to the BS, and each member node distance to the CH when selecting CHs. Furthermore, the mentioned parameter density and hop-count are also utilized when constructing the inter-cluster routing tree.

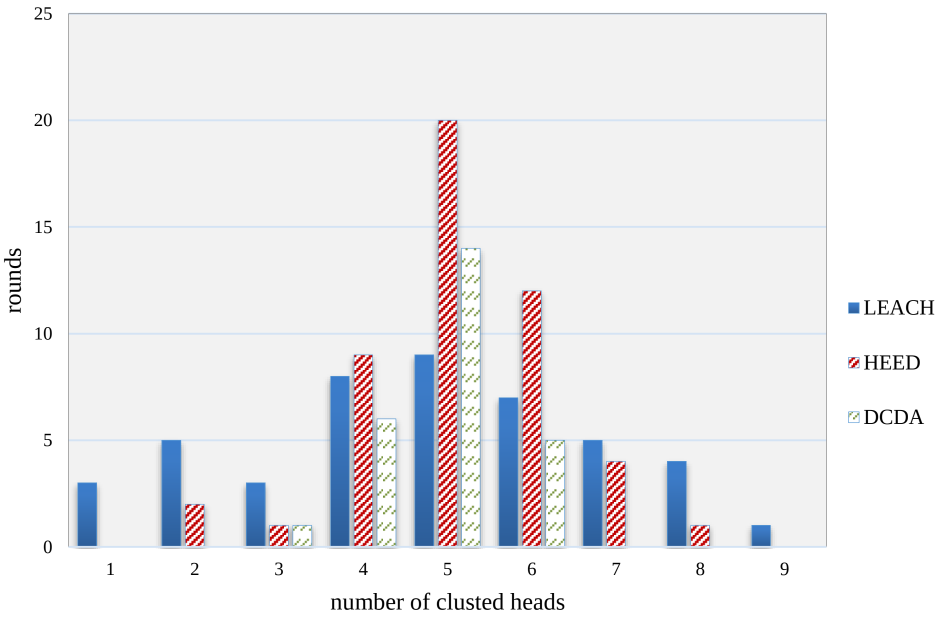

Fig. 6: Distribution of the number of cluster heads in three protocols.

5.2 Energy Consumption

Fig. **7** illustrates the energy consumption of nodes in LEACH, HEED and DCDA. It shows that in LEACH nodes, the residual energy decreases at a higher rate because, in LEACH, a node with very low energy may be selected as a CH in several rounds which may drain its energy more quickly. In addition, CHs would be selected randomly with a certain probability in a cyclic way through which the energy load of the entire network may not be distributed efficiently among nodes. Each CH sends data directly to the BS so it may lead to earlier death of the CHs far from the BS. In HEED due to considering a hybrid of the node’s residual energy and a secondary parameter, such as node proximity to its neighbors or node degree, more energy is saved in each sensor node. However, HEED doesn’t take into account network structure, which causes an imbalance in energy consumption among the clusters. DCDA could greatly prolong the network lifetime and reduce energy consumption of nodes within the clusters by considering proximity to the CHs when the CHs are selected. Moreover, as the great amount of energy is consumed in inter-cluster communication, an energy efficient routing tree is established using effective parameters like energy, adjacent degree, distance to the BS and number of hops to the BS. One more solution considered for energy dispensation among sensor nodes is to determine a set of eligible relay nodes that forward CH data to the CHs closer to the BS.

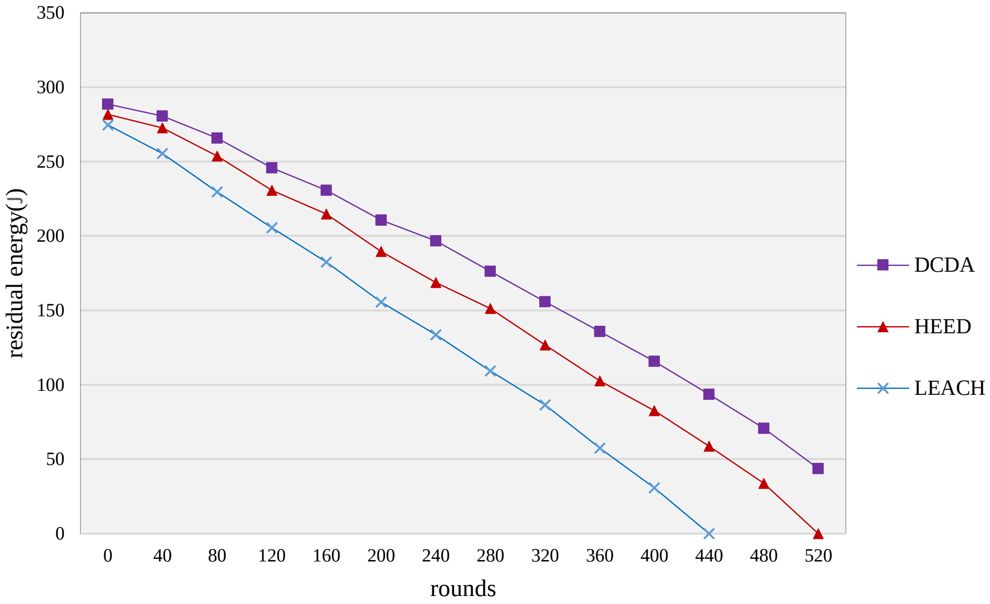

Fig. 7: Residual energy of nodes in different rounds.

5.3 Network Lifetime

According to the definitions given in [24], network lifetime could be defined in different metrics, the time from the deployment of the network to the death of the first node (FND), the time when a certain percent of nodes remain alive (PNA), and the time when all the nodes are dead in the network (LND). Here we define the network lifetime as the time from the deployment of the network to the death of its first node (First Node Dies, FND).

We ran LEACH, HEED and the proposed method several times. The main aim of DCDA is to prolong network lifetime with effective CH selection and an optimal routing tree construction among sensor nodes. Thus, residual energy, distance to the BS, and CH centric metrics are used to choose a set of qualified CHs. Next hop nodes through the routing tree are selected with regard to various peripheral and internal features of each node. As shown in Fig. 8, DCDA improves the lifetime of nodes compared with other algorithms. The improvements are approximately 61% and 18% compared to LEACH and HEED, respectively.

Fig. **9** illustrates the number of alive nodes in some rounds, which is a downward trend. HEED improves LEACH as it uses a variable known as the cluster radius that defines the transmission power to be used for intra-cluster broadcast. Also, in the CH selection stage, tentative CHs would be determined with regard to the residual energy and final cluster heads are selected according to the intra-cluster communication cost. In HEED, clusters do not overlap with each other, so the number of alive nodes is greater in higher rounds. DCDA could effectively balance the energy consumption among cluster heads, since relay nodes aid CHs in data transmission between clusters and hence balance energy consumption among sensors in the routing tree construction stage. DCDA forms clusters of different sizes to balance energy consumption among CHs. Since nodes closer to the BS take more responsibility for the forwarding task from all around the network to the BS, the cluster size should be decreased compared with those far from the BS. It could help to balance load among CHs in intra-cluster and inter-cluster data communications. So compared to HEED, DCDA saves more energy in sensor nodes and hence the number of alive nodes is greater in higher rounds.

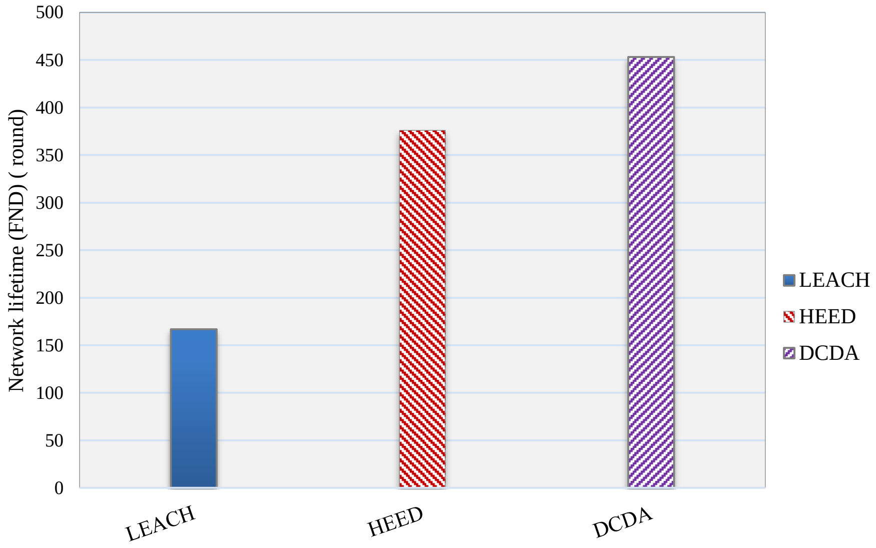

Fig. 8: Network lifetime.

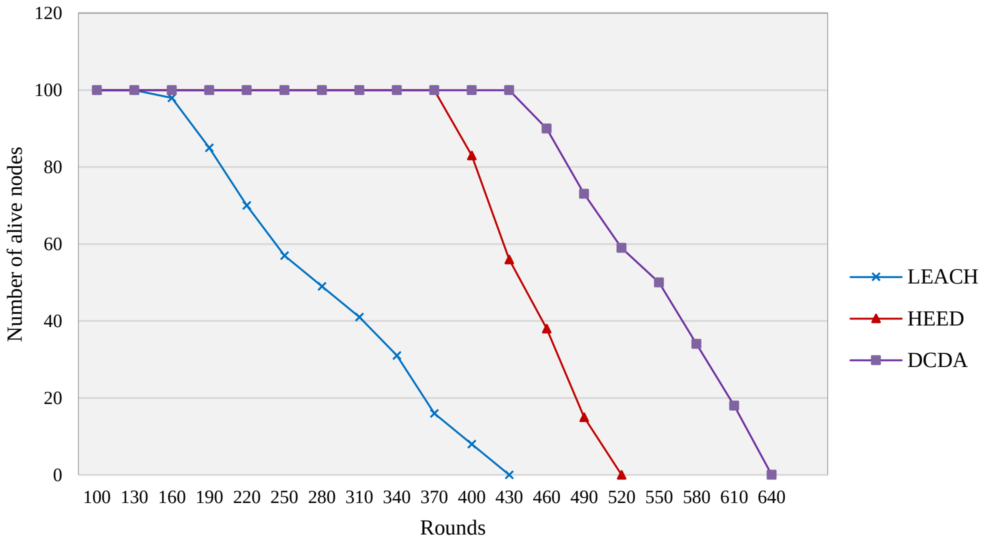

Fig. 9: Number of alive nodes in each round for three protocols.

6. CONCLUSION REMARKS

In this paper we proposed a distributed energy efficient clustering algorithm for data aggregation in wireless sensor networks (DCDA). The main objective is to save energy and balance load among sensor nodes. CDCA consists of two stages: clustering and data transmission. In the clustering stage, we form clusters with unequal sizes in a way that CHs closer to the BS have a smaller cluster range. This could prevent premature death of CHs close to the BS, due to forwarding a large amount of data from all around the network. Eligible CHs are specified based on some effective parameters, such as distance to the BS, remaining energy of sensors and nodes centric metric. Besides, some relay nodes are selected to transfer data between clusters when direct connection between CHs is more costly. Common nodes are associated with high energy CHs, in a low density area and less total distance to the BS. In next hop and relay nodes selection, DCDA utilizes more criteria such as number of hops to the BS and total cost for transmitting data to the next CH. Finally, in the data transmission stage, data is transmitted through an optimal routing tree where the BS is located at the root, a CH at each edge, and cluster member nodes in leaves. The simulation results show the proposed DCDA algorithm provides more energy efficiency and hence longer network lifetime when compared to the other discussed algorithms.

REFERENCES

[1] Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. (2002) Wireless sensor networks: a survey. Computer networks, 38:393-422.

[2] Anastasi G, Conti M, Di Francesco M, Passarella A. (2009) Energy conservation in wireless sensor networks: A survey. Ad Hoc Networks, 7:537-568.

[3] Yick J, Mukherjee B, Ghosal D. (2008) Wireless sensor network survey. Computer networks, 52:2292-2330.

[4] Aziz AA, Sekercioglu Y. A, Fitzpatrick P, Ivanovich M. (2013) A survey on distributed topology control techniques for extending the lifetime of battery powered wireless sensor networks. IEEE Communications Surveys & Tutorials, 15:121-144.

[5] Zhou W. (2011) Energy efficient clustering algorithm based on neighbors for wireless sensor networks. Journal of Shanghai University (English Edition), 15:150-153.

[6] Liu Y, Xiong N, Zhao Y, Vasilakos A. V, Gao J, Jia Y. (2010) Multi-layer clustering routing algorithm for wireless vehicular sensor networks. IET communications, 4:810-816.

[7] Heinzelman W. R, Chandrakasan A, Balakrishnan H, (2000) Energy-efficient communication protocol for wireless microsensor networks. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences.

[8] Younis O, Fahmy S. (2004) HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing, 3:366-379.

[9] Yu J, Qi Y, Wang G, Gu X. (2012) A cluster-based routing protocol for wireless sensor networks with nonuniform node distribution. AEU-International Journal of Electronics and Communications, 66:54-61.

[10] Albath J, Thakur M, Madria S. (2013) Energy constraint clustering algorithms for wireless sensor networks. Ad Hoc Networks, 11:2512-2525.

[11] Gu X, Yu J, Yu D, Wang G, Lv Y. (2014) ECDC: An energy and coverage-aware distributed clustering protocol for wireless sensor networks. Computers & Electrical Engineering, 40:384-398.

[12] Sha C, Wang R, Huang H, Sun L. (2010) Energy efficient clustering algorithm for data aggregation in wireless sensor networks. The Journal of China Universities of Posts and Telecommunications, 17:104-122.

[13] Upadhyayula S, Gupta S. K. (2007) Spanning tree based algorithms for low latency and energy efficient data aggregation enhanced convergecast (dac) in wireless sensor networks. Ad Hoc Networks, 5:626-648.

[14] Monaco U, Cuomo F, Melodia T, Ricciato F, Borghini M. (2006) Understanding optimal data gathering in the energy and latency domains of a wireless sensor network. Computer Networks, 50:3564-3584.

[15] Dimokas N, Katsaros D, Manolopoulos Y. (2010) Energy-efficient distributed clustering in wireless sensor networks. Journal of parallel and Distributed Computing, 70:371-383.

[16] Chao C. M, Hsiao T. Y. (2009) Design of structure-free and energy-balanced data aggregation in wireless sensor networks. 11th IEEE International Conference on High Performance Computing and Communications HPCC'09.

[17] Mahajan S, Malhotra J, Sharma S. (2014) An energy balanced QoS based cluster head selection strategy for WSN. Egyptian Informatics Journal, 15:189-199.

[18] Azharuddin M, Kuila P, Jana P. K. (2015) Energy efficient fault tolerant clustering and routing algorithms for wireless sensor networks. Computers & Electrical Engineering, 41:177-190.

[19] Krunz M, Siam M. Z, Nguyen D. N. (2013) Clustering and power management for virtual MIMO communications in wireless sensor networks. Ad Hoc Networks, 11:1571-1587.

[20] Ashouri M, Yousefi H, Basiri J, Hemmatyar A. M. A, Movaghar A. (2015) PDC: Prediction-based Data-aware Clustering in wireless sensor networks. Journal of Parallel and Distributed Computing, 81:24-35.

[21] Gong D, Yang Y, Pan Z. (2013) Energy-efficient clustering in lossy wireless sensor networks. Journal of Parallel and Distributed Computing, 73:1323-1336.

[22] Jain Y. K, Bhandare S. K. (2011) Min max normalization based data perturbation method for privacy protection. International Journal of Computer and Communication Technology, 2:45-50.

[23] Xinlian Z, Min W, Jianbo X. (2009) BPEC: an energy-aware distributed clustering algorithm in WSNs. J. Comput. Res. Dev, 46:723-730.

[24] Katiyar V, Chand N, Soni S. (2010) Clustering algorithms for heterogeneous wireless sensor network: A survey. International Journal of applied Engineering research Dindigul, 1:273-274.