Optimization of Cluster Heads for Energy Efficiency in Large-Scale Wireless Sensor Networks

Yi Gu and Qishi Wu

Department of Computer Science
University of Memphis
Memphis, TN 38152
{yigu,qishiwu}@memphis.edu

Abstract. Many complex sensor network applications require deploying a large number of inexpensive and small sensors in a vast geographical region to achieve quality through quantity. Hierarchical clustering is generally considered as an efficient and scalable way to facilitate the management and operation of such large-scale networks and minimize the total energy consumption for prolonged lifetime. Judicious selection of cluster heads for data integration and communication is critical to the success of applications based on hierarchical sensor networks organized as layered clusters. We investigate the problem of selecting nodes in a pre-deployed sensor network to be the cluster heads to minimize the total energy needed for data gathering. We rigorously derive an analytical formula to optimize the number of cluster heads in sensor networks under uniform node distribution, and propose a Distance-based Crowdedness Clustering algorithm to determine the cluster heads in sensor networks under general node distribution. The results from an extensive set of experiments on a large number of simulated sensor networks illustrate the performance superiority of the proposed solution over the clustering schemes based on *k*-means algorithm.

Keywords: sensor networks, optimization, energy consumption, cluster heads.

1 Introduction

Multiple sensor systems have been the target of active research since the early 90s due to their widespread use in many agricultural, civil, industrial, and military applications that involve environmental monitoring, target tracking, and situational assessment [10, 11]. Recent developments in Micro-Electro-Mechanical Systems (MEMS) make it now possible to deploy a large number of inexpensive and small sensors to perform complex tasks by obtaining quality through quantity. In some important applications, sensor networks are deployed for remote operations in vast unstructured geographical areas. In such deployments, wireless networks with low bandwidth are usually the only means of communication among the sensors. These sensors are typically powered by irreplaceable batteries with limited energy supply, a large portion of which has to be spent on data communication among sensors and the Base Station (BS). Therefore, minimizing the Total Energy Consumption (TEC) for sensor data gathering is critical to ensuring sustained operations of these large-scale Wireless Sensor Networks (WSNs), even though

J. Zheng et al. (Eds.): ADHOCNETS 2009, LNICST 28, pp. 33–48, 2010.

minimizing TEC does not necessarily maximize network lifetime, which also depends on the balance of residual energy across the network.

It has been well recognized that clustering provides an efficient and scalable way to design and organize large-scale WSNs for energy efficiency of data communication. In a typical hierarchical WSN that deploys a large number of homogeneous or heterogeneous sensors, clusters are formed around a set of strategically selected or randomly designated Cluster Heads (CHs). The sensors within each cluster, often referred to as Leaf Nodes (LNs), collect and send environmental measurements to their corresponding CH, which performs an appropriate form of data processing (i.e. aggregation and compression) and sends the result to a BS for integration with data from other clusters. The BS, which is located either inside or outside the sensor network region, could be wire-connected or rechargeable, and hence is often considered to have unlimited energy supply. Apparently, the number and location of CHs in hierarchical WSNs have a significant effect on the energy consumption for data communication from LNs to the BS. Many existing clustering algorithms assume that the number of clusters is predetermined and each CH is also designated *a priori* or on a random basis [9, 16]. In practice, however, such information on CHs is not always readily available during network implementation and operation, especially when the network is deployed in an unstructured environment with unpredictable disturbances and threats. As a matter of fact, determining the optimal number and location of CHs has been raised as one of the most fundamental problems in clustering-based hierarchical WSNs.

The global knowledge of sensor locations is crucial for determining the optimal number and location of CHs. Once deployed, each sensor can acquire and report its geographical location through a certain location discovery process. The Global Positioning System (GPS) is certainly a very effective but prohibitively expensive solution due to its high cost and the constrained energy supply of sensors. Many other location discovery approaches use received signal strength, neighbor node position, or arrival time difference to estimate the distance between two sensors [18]. Based on the distance estimations, sensors can compute their locations by distance or angle combining (hyperbolic trilateration, triangulation, or maximum likelihood estimation). To make such combining feasible, the computation requires at least two known reference points, which can be obtained through either a GPS or a deterministic deployment.

We investigate the problem of selecting an optimal subset of sensors, which are designated as CHs to form clusters, to minimize the TEC for data communication per round from LNs to the BS in a pre-deployed WSN under a uniform or general node distribution. We rigorously derive an analytical solution to calculate the number of CHs for minimum TEC in WSNs where nodes are uniformly distributed. In WSNs where nodes are deployed according to an unknown probability distribution, we propose a heuristic algorithm, Distance-based Crowdedness Clustering (DCC), to determine the optimal number and location of CHs. The extensive experimental results on simulated WSNs ranging from small to large scales show the performance superiority of the proposed methods compared to the clustering and optimization schemes based on classical k-means algorithm in terms of TEC for data communication per round. However, the main purpose of these performance comparisons is to recognize the necessity of such CH optimization and illustrate the efficacy of the proposed research approaches. In fact,

the proposed optimization schemes, which could be performed offline using global information prior to the actual deployment, are largely orthogonal to the main body of current research efforts on sensor network clustering, and the optimization results obtained by our approaches can be used to effectively guide the operation of and therefore greatly improve the performance of the classical clustering algorithms that require *a priori* knowledge on CHs.

The rest of the paper is organized as follows. We describe related work in Section 2. The energy consumption models are presented and the optimization problems are formulated in Section 3. In Section 4, we derive an analytical solution for WSNs under uniform distribution and develop a heuristic approach for WSNs under general distribution. Implementation details and performance evaluations are provided in Section 5. We conclude our work and discuss future efforts in Section 6.

2 Related Work

In WSNs, a CH, which may collect environmental sensing data as well in some cases, is responsible for receiving, processing, and transmitting data from the LNs in its service area to the BS, and hence it consumes much more energy than an LN. The sensor nodes in the close proximity of a CH may also run out of battery quickly due to frequent data forwarding. Therefore, designating an optimal subset of sensor nodes as CHs at appropriate locations is critical to minimizing the TEC for prolonging the lifetime of the entire network. There exist a large number of research efforts in the literature that have been devoted to solving various clustering problems with different objectives [9, 12, 23, 25, 24, 14, 17, 3, 13, 8, 7].

Several clustering algorithms have been proposed to minimize the energy consumption or satisfy the network connectivity requirement with the assumption that the number of CHs is known *a priori* [9, 16, 22, 3]. Some other researchers tackle the problem of optimizing the number and location of CHs, which is closely related to our work. Kim *et al.* derived analytical formulas to estimate the optimal number of CHs to achieve the minimum TEC of the entire network, based on the assumption of even partition (i.e. each cluster covers an equal number of sensors) and one-hop communication in both intra-cluster and inter-cluster routing [12]. In [23], Wang *et al.* calculated the optimal number of CHs for a WSN by applying a cross-layer approach from both perspectives of the power efficiency in the medium access control (MAC) layer and the coverage performance in the physical layer. In [24], Xu *et al.* determined the optimal location of CHs for minimum communication energy consumption in a two-level heterogeneous WSN. They formulated the optimization problem with the constraint that each LN is connected to at least p CHs and each CH has at most q LNs as a Mixed Integer Nonlinear Programming (MINLP) problem and achieved a global optimum based on an iterative decomposition algorithm and a randomized multi-start technique. Srinivas *et al.* studied the problem of minimizing the number of backbone nodes and referred to it as the Connected Disk Cover problem in [20], where they controlled the mobility of the backbone nodes to maintain the connectivity of the network. In [21], Tao *et al.* proposed an improved algorithm that combines the optimal number of CHs with energy adaptive CH selection algorithm based on the classical LEACH protocol.

One commonly adopted way to ensure load balance and meet energy constraint of the entire network is to rotate the role of a CH and form a corresponding cluster on a random and periodical basis among all sensor nodes. The classical energy-efficient algorithm, Low-Energy Adaptive Clustering Hierarchy (LEACH) [9], employed randomized rotation of CHs to evenly distribute the energy consumption among the sensors in the network. Yang *et al.* [25] applied a sleep-wakeup scheme based on a decentralized MAC protocol to LEACH and further proposed an analytical framework for achieving the optimal probability with which a sensor becomes a CH in order to minimize network energy consumption and prolong network lifetime, assuming uniform distribution of all the sensors. A similar node election strategy that considers the amount of residual energy was also used in [19] with multihop data communication. Du *et al.* presented an energy-efficient Chessboard Clustering routing protocol for heterogeneous sensor networks [5] to balance node energy consumption and increase network lifetime, in which some sensors are initially set to be in sleeping mode and can be activated later on according to a certain procedure. Some research efforts along a different line allow sensors to adjust their locations after initial deployment. Mao *et al.* proposed two new sensor location updating algorithms, VFSec and Weighted Centroid, to jointly optimize sensing coverage and secure connectivity [15].

Nonlinear programming has been widely used to model wireless sensor data communication as network flows to minimize TEC in WSNs. Besides [24], Ergen *et al.* also formulated their problem as a nonlinear programming problem that determines the optimal location of relay nodes and the optimal energy provided to them so that the network is alive during the desired lifetime with minimum total energy [6]. In [4], Dasgupta *et al.* presented the sensor deployment problem for maximum lifetime with coverage constraints and proposed an algorithm motivated by force-directed/potential-field based approaches in robotics and graph drawing to place and assign the role of each sensor in the system to maximize network lifetime.

The probability of a sensor to be a CH is also considered in some work. Bandyopadhyay *et al.* proposed a distributed and randomized clustering algorithm to organize the sensors in a WSN into clusters, and they further extended this algorithm to generate a hierarchy of CHs and observed more energy savings as the number of levels in the hierarchy increases [2].

The main differences between our work and the aforementioned ones arise from the following aspects: (i) we investigate both uniform distribution and general random deployment for large-scale sensor networks; (ii) we propose both analytical derivation and heuristic algorithm to solve the CH optimization problems; (iii) we consider complete data aggregation and use a compression ratio to reflect the reduction in data size at CHs; (iv) we do not assume even partition of sensor networks as in [12].

3 Cost Model and Problem Formulation

3.1 Energy Consumption Model

We consider two different types of energy consumption for data transmission and receiving, respectively: a transmitter consumes energy to run both the radio electronics

and the power amplifier, while a receiver only consumes energy to drive the radio electronics. Since the inter-cluster communication distance is typically much longer than the intra-cluster communication distance, we employ (i) the free space (*fs*) fading channel model for intra-cluster wireless communication that incurs a d^2 power loss, and (ii) the multipath (*mp*) fading channel model for inter-cluster wireless communication that incurs a d^4 power loss [9, 12, 19]. In a real communication system, the transmission power could be adjusted by suitably configuring the power amplifier. Therefore, the energy dissipation in transmitting one unit of data message over a directed wireless communication link can be modeled as $E_t(i)$:

$$E_t(i) = E_{elec} + E_{amp}(d_{i,j}) = \begin{cases} E_{elec} + \varepsilon_{fs} \cdot {d_{i,j}}^2, & \text{intra-cluster} \\ E_{elec} + \varepsilon_{mp} \cdot {d_{i,j}}^4, & \text{inter-cluster} \end{cases} \quad (1)$$

where E_{elec} denotes the energy for driving the electronics, which depends on various factors including digital coding, modulation, filtering and spreading of the signals, for both transmitter electronics and receiver electronics; ε_{fs} and ε_{mp} are the coefficients for calculating the amplifier energy E_{amp}, which depends on the Euclidean distance $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ between transmitter v_i located at (x_i, y_i) and receiver v_j located at (x_j, y_j) as well as the acceptance bit-error rate. The energy consumed by a sensor v_i in receiving one unit of data packet is denoted as $E_r(i) = E_{elec}$. Note that the above transmission and receiving energy models assume a contention free MAC protocol, where interferences from simultaneous transmission can be avoided.

A CH, which also collects environment sensing data, receives data messages from LNs within the cluster and sends all the data to the BS after performing a certain type of data processing (such as data aggregation and data compression). We use a constant E_p to represent the energy spent in processing each unit of received or sensed data. We assume that the CH performs complete data aggregation, i.e. an input of two k-bit messages produces an output of one k-bit message after aggregation. Furthermore, we use a parameter α, $0 < \alpha \leq 1$, to denote the data compression ratio: an input of k bits results in an output of $\alpha \cdot k$ bits after compression.

3.2 Problem Formulation

The problem of determining the optimal number and location of CHs for minimum TEC in sensor networks is formulated as follows. We consider a WSN where n sensor nodes have been deployed in a bounded $L \times L$ (m^2) square region and a single BS is located at (x_{BS}, y_{BS}), somewhere inside or outside the network region. The location of each sensor v_i, $i \in 0, 1, \ldots, n-1$, is denoted as (x_i, y_i). We assume a one-hop communication model for both intra-cluster (from LNs to their associated CHs) and inter-cluster (from CHs to the BS) communication: the transmission energy within each cluster is calculated using the *fs* model; while from CHs to the BS, the *mp* model is used. We consider two different sensor deployment scenarios: (i) uniform node distribution and (ii) general node distribution. The CH optimization problem is to strategically designate an appropriate subset of sensor nodes in the network as CHs, each of which forms a cluster with its neighbor nodes, such that the TEC for the transmission of each unit of data message from all LNs to CHs and to the BS per round is minimized. Here, the "round" is defined

as a time period during which every sensor in the network sends one unit of data message to the BS through its associated CH. The optimization objective is to determine the optimal number and location of CHs in the given sensor network for minimum TEC.

We consider the following general conditions or assumptions in our problem formulation:

- All sensors are pre-deployed and have constrained energy supply;
- The BS is also pre-deployed and has unlimited energy supply;
- The network is static, i.e. neither the sensors nor the BS has mobility once deployed;
- The total number of sensors is known;
- Each CH forms exactly one cluster, and besides data processing, also performs the same task of environmental sensing and data collection as a regular sensor node;
- There exists a contention free MAC protocol for wireless communication.

We consider the energy consumption for data transmission of each LN, and for data receiving, processing, and transmission of each CH. Since the energy cost for environment sensing is generally much less than communication and processing tasks, we do not consider sensing energy cost here. Obviously, the TEC depends on the network distribution, the number and location of CHs, and the compression ratio α at CHs.

4 Optimizing Number and Location of CHs

4.1 Analytical Derivation for Uniform Distribution

When CHs perform data processing or are responsible for certain intra-cluster management duties, uniform distribution of sensors among the network is usually a goal for setting up the network, which can also leverage data delay [1]. In LEACH and other similar clustering algorithms, the expected number k of CHs per round is considered as a prefixed system parameter [9, 16, 22, 3]. Here, we shall rigorously derive an analytical formula for calculating the optimal value of k to achieve the minimum TEC of data transfer from LNs to the BS through their corresponding CHs. The optimal k value determined by our approach can be used to guide the execution of the clustering algorithms that require such information.

We first calculate the expected distances from an LN to its CH and from a CH to the BS using an approach similar to the one used in [25]. Since the CHs are uniformly distributed in an $L \times L$ (m^2) sensor network region, the expected square area covered by each cluster with the CH deployed at (x_{CH}, y_{CH}) can be calculated as: $\sqrt{\frac{L^2}{k}} \times \sqrt{\frac{L^2}{k}}$ based on Voronoi tessellation. Furthermore, the LNs are also uniformly and independently deployed in each cluster, where we have $E[x_{LN}] = E[x_{CH}] = E[y_{LN}] = E[y_{CH}] = \frac{1}{2}\sqrt{\frac{L^2}{k}}$, and $E[(x_{LN})^2] = E[(x_{CH})^2] = E[(y_{LN})^2] = E[(y_{CH})^2] = \frac{1}{3k}L^2$. Therefore, the average squared distance from an LN to its CH within a cluster can be calculated as:

$$\begin{aligned} r^2 &= E[(x_{LN} - x_{CH})^2 + (y_{LN} - y_{CH})^2] \\ &= E[(x_{LN})^2] - 2E[x_{LN}]E[x_{CH}] + E[(x_{CH})^2] + E[(y_{LN})^2] - 2E[y_{LN}]E[y_{CH}] + E[(y_{CH})^2] \\ &= \frac{1}{3k}L^2 \end{aligned} \quad (2)$$

Without loss of generality, we suppose that the BS location (x_{BS}, y_{BS}) is prefixed at $(0,0)$ to simplify calculation. For a coarse grained analysis, we assume identical expected

distance from CHs to the BS. Therefore, the average squared distance from a CH to the BS is similarly given by:

$$R^2 = E[(x_{CH} - x_{BS})^2 + (y_{CH} - y_{BS})^2] = \frac{2}{3}L^2 \,. \tag{3}$$

The TEC per round, denoted by E_{Tot}, is the sum of the energy consumption E_{LN} of all LNs for data transmission and the energy consumption E_{CH} of all CHs for data receiving, processing, and transmission in one round, which can be defined as:

$$E_{Tot} = E_{LN} + E_{CH}. \tag{4}$$

Since E_{LN} only contains transmission energy cost E_t and the total number of LNs in the network is $n-k$, E_{LN} can be estimated as:

$$E_{LN} = (n-k)E_t = (n-k)(E_{elec} + \varepsilon_{fs} r^2) \,. \tag{5}$$

Similarly, E_{CH} is the total energy cost for the transfer of one unit of data from each CH to the BS in one round, which includes the energy cost E_r for receiving, E_p for processing, and E_t for transmission. Each of $n-k$ LNs transfers one unit of data to its corresponding CH, which performs processing (aggregation and compression) on the received data and its own sensing data, and sends the compressed aggregated result to the BS. Since there are total n units of input data (including $n-k$ units of data received from LNs and k units of data collected by k CHs themselves), the total energy consumed by k CHs is defined by:

$$\begin{aligned} E_{CH} &= (n-k)E_r + nE_p + \alpha k E_t \\ &= (n-k)E_{elec} + nE_p + \alpha k(E_{elec} + \varepsilon_{mp} R^4) \,, \end{aligned} \tag{6}$$

Using Eqs. 2, 3, 4, 5, and 6, we obtain the TEC per round E_{Tot} as follows:

$$E_{Tot} = (2n - 2k + \alpha k)E_{elec} + nE_p + (n-k)\varepsilon_{fs}\frac{L^2}{3k} + \alpha k \varepsilon_{mp}\frac{4L^4}{9} \,. \tag{7}$$

The TEC per round E_{Tot} can be minimized by selecting an optimal value of k, which is a solution to the first derivation of Eq. 7. Following that, the optimal number k of CHs can be calculated as (the negative solution is ignored):

$$k = \sqrt{\frac{3nL^2\varepsilon_{fs}}{9(\alpha-2)E_{elec} + 4\alpha L^4 \varepsilon_{mp}}} \,. \tag{8}$$

We further verify that the solution to the second derivation of Eq. 7 is positive. Therefore, we conclude that the value of k defined in Eq. 8 results in the minimum TEC in WSNs with uniform node distribution. Once the optimal number of CHs is obtained, their locations can be determined based on Voronoi tessellation among uniformly distributed sensors.

Algorithm 1. Distance-based Crowdedness Clustering
Input: a sensor network $G = (V,E)$ with n LNs randomly deployed in a $L \times L$ (m^2) square region and one BS deployed inside or outside the region.
Output: the optimal number k and location of CHs with minimum TEC.

Calculate all-pair distances $d_{i,j}$, for $v_i, v_j \in V$, in an array A_d;
Initialize minimum TEC $TEC_{min} = +\infty$;
for all distances $d_{i,j} \in A_d$ **do**
  Set cut-off distance $d_{cut} = d_{i,j}$;
  Set v_m as a neighbor of v_n if $d_{m,n} \leq d_{cut}$ for all $m,n \in V$;
  Sort all $v \in V$ according to the number of neighbors in a decreasing order and place them in an array A_v;
  Insert all $v \in V$ in an unclustered sensor queue Q_u;
  Initialize a clustered sensor queue $Q_c = \emptyset$;
  Initialize the number of clusters $n_{clusters} = 0$;
  while $Q_u \neq \emptyset$ **do**
    Retrieve $v_k \in Q_u$ from A_v and designate it as a CH;
    Form a cluster C_k of v_k and its neighbors $v_l \in Q_u$;
    Insert all $v \in C_k$ in Q_c;
    Remove all $v \in C_k$ from Q_u;
    $n_{clusters}$ ++;
  end while
  Calculate the TEC;
  if $TEC_{min} >$ TEC **then**
    TEC_{min} = TEC;
    $k = n_{clusters}$;
  end if
end for
return k and location.

4.2 Algorithm Design for General Distribution

Though uniform distribution is suited for achieving energy balance, it may not always be feasible for practical sensor network applications, especially for those deployed in large and harsh environments not accessible to humans. In such environments, sensors may be airborne or dropped by other appropriate means that could lead to a more general node distribution.

We develop a heuristic algorithm, Distance-based Crowdedness Clustering (DCC), based on a cut-off distance and the concept of crowdedness to solve the CH optimization problem in WSNs under general node distribution. Here, the "cut-off distance" is the threshold that decides which cluster an LN belongs to: if the distance from an LN to the CH is within the cut-off distance, the LN is considered to be located inside this cluster; otherwise, it is not. The concept of "crowdedness", which is closely related to the number of neighbor nodes of each sensor node, is used to describe the density of sensor nodes in a given region.

In the DCC algorithm, we first calculate the distances $d_{i,j}(i, j \in V)$ between all pairs of sensor nodes, each of which is selected as the cut-off distance. Using this distance, we designate the sensor with the largest number of neighbor nodes as a CH and form a cluster of this CH with all its unclustered neighbors. We repeatedly designate CHs with most neighbor nodes in the rest of the LNs using the same cut-off distance until there is no LN left unclustered. We calculate the TEC using Eq. 7 in Section 4.1 for all cut-off distances, from which, the minimum one is selected as the final result. The pseudocode of DCC is given in Algorithm 1. The complexity of this algorithm is $O(n^3)$.

5 Performance Evaluation

5.1 Implementation and Experimental Settings

The proposed DCC algorithm is implemented in C++ and runs on a Windows XP desktop equipped with a 3.0 GHz CPU and 2 Gbytes memory. We conduct an extensive set of experiments using a wide variety of simulated sensor networks, in which two deployment scenarios are considered: uniform and general distributions. We generate these simulation datasets by varying the size of network regions and the number of initially deployed sensor nodes. The parameters used in the testing sensor networks and the sensor radio characteristics of the energy cost models for wireless communication [12] are tabulated in Table 1. For testing purposes, we consider a fixed value 0.8 for the compression ratio α, which has an impact on the clustering process.

Table 1. WSN communication parameters

Parameter	Value
$L \times L$	200×200 (m^2)
E_{elec}	50 (nJ/bit)
ε_{fs}	10 $(pJ/bit/m^2)$
ε_{mp}	0.0013 $(pJ/bit/m^4)$
E_p	5 $(nJ/bit/signal)$
α	0.8

5.2 Case Study for Uniform Distribution

We first investigate the optimization problem on the number and location of CHs in a sensor network within a square region of 200×200 (m^2) under uniform distribution. Fig. 1 plots the TEC optimization curve that is calculated by Eq. 7 in the network of $n = 200$ sensor nodes in response to the number k of CHs varying from 1 to 50.

From Eq. 8 in Section 4.1, we obtain the optimal number of CHs to be $k = 6$, which is consistent with the optimal one observed in Fig. 1. The TEC increases as the number of CHs moves away from the optimal point. We further study a case with a larger WSN of $n = 900$ sensor nodes. Again, the optimal number 13 observed in Fig. 2 is consistent with the one produced by Eq. 8. The unimodal property of the TEC optimization curves

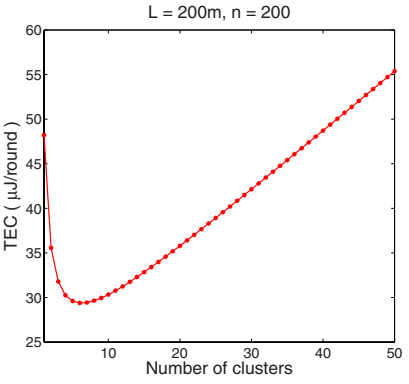

Fig. 1. Analytical calculation of TEC per round with $n = 200$

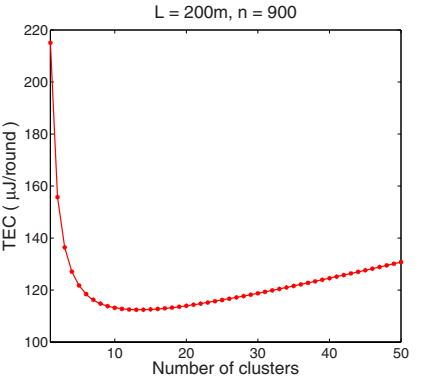

Fig. 2. Analytical calculation of TEC per round with $n = 900$

justifies the correctness of our derivation for the optimal number of CHs in WSNs under uniform node distribution.

To evaluate the robustness of our solution, we perform more experiments on networks under uniform distribution with 10 different problem sizes from small to large ones: in the first case, the total number of initial sensor nodes is set to be 10; in the rest 9 cases, it increases from 100 to 900 nodes at an interval of 100 nodes. The optimization results, i.e. the optimal number of CHs, the expected distance from an LN to its associated CH, and the corresponding minimum TEC in each case, are plotted in Fig. 3. We observe that the expected distance from an LN to its CH varies from 32 m to 116 m and it decreases when the optimal number of CHs increases.

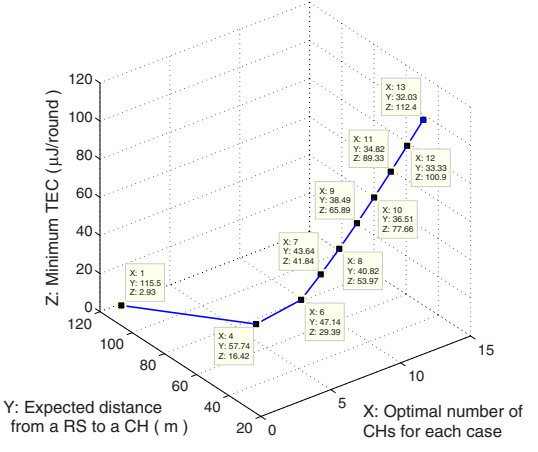

Fig. 3. Analytical calculation of minimum TEC per round in 10 cases ranging from small to large scales

5.3 Case Study for General Distribution

To better illustrate the optimization process of DCC in WSNs under general node distribution, we calculate and plot the TEC per round under different cut-off distances at each optimization step for a network of $n = 200$ sensor nodes deployed in the same region, as shown in Fig. 4. This three-dimensional optimization curve also features a unimodal property: the TEC is minimized with an optimal number of CHs at a certain cut-off distance.

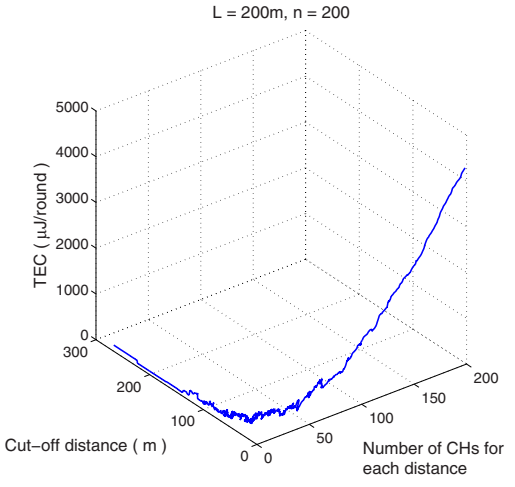

Fig. 4. TEC per round at each optimization step using DCC algorithm

For performance comparison, we adapt the classical k-means clustering algorithm to our problem and compare its performance with that of the proposed DCC algorithm using the same set of 10 problem sizes previously considered for uniform distribution. We first determine the optimal number k of CHs and the corresponding minimum TEC using the proposed DCC algorithm, and then use k-means algorithm to perform clustering based on the k value obtained by the DCC algorithm, referred to as deterministic k-means algorithm. For visual comparison, we plot the performance measurements of TEC produced by these two algorithms in Fig. 5. The results produced by the DCC algorithm outperform those produced by the deterministic k-means algorithm in all 10 cases we studied, which shows the performance superiority of the proposed DCC algorithm.

For illustration purposes, we lay out the node distribution and clustering of the sensor network with 100 sensor nodes in Fig. 6, in which the unclustered solid node denotes the BS. The clustering results obtained by the DCC algorithm are marked by the solid lines while the results obtained by the deterministic k-means algorithm are marked by the dashed lines. The DCC algorithm produces 20 clusters in this instance.

We further compare DCC with a clustering method completely based on the classical k-means algorithm, in which we iteratively use k-means algorithm to cluster n sensor nodes for n times by varying the parameter k from 1 to n, and select the clustering result with the minimum energy cost. We refer to this clustering method as adaptive k-means

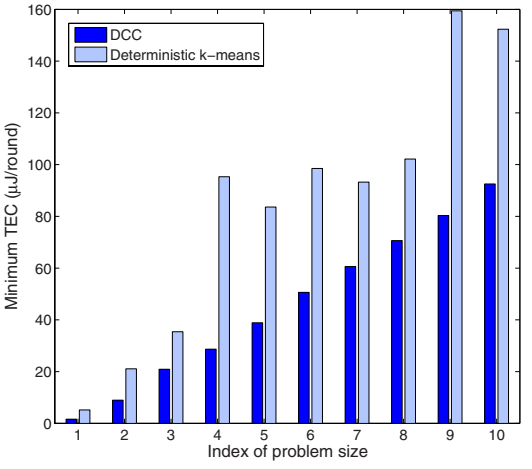

Fig. 5. Performance comparison between DCC and deterministic *k*-means algorithms in 10 cases ranging from small to large scales

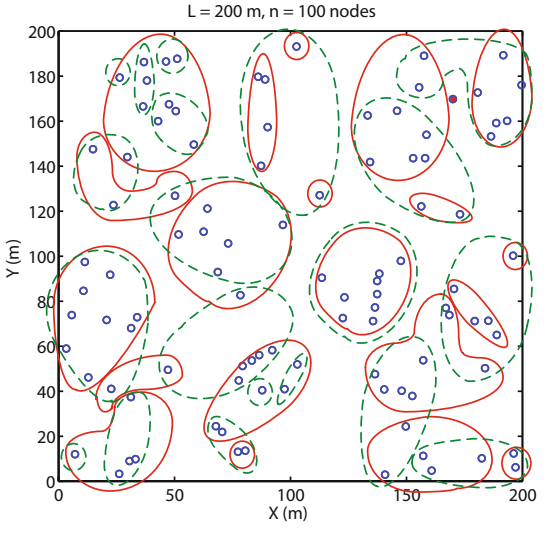

Fig. 6. Visualization of the clustering in a network of 100 nodes using DCC and deterministic *k*-means algorithms: the results of DCC are marked by solid lines and the results of deterministic *k*-means are marked by dashed lines

algorithm. Two WSNs of problem sizes of 200 and 900 sensor nodes, respectively, are tested using adaptive *k*-means algorithm, and their partial optimization curves are plotted in Figs. 7 and 8. The rest of the TEC values continuously increase as *k* increases.

We observe some small random variations on the performance curves produced by the adaptive *k*-means algorithm. These variations are mainly caused by the following

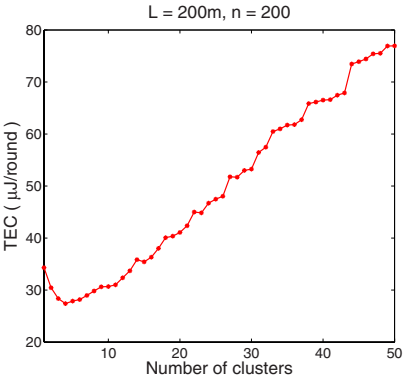

Fig. 7. TEC per round for each k value using adaptive k-means algorithm

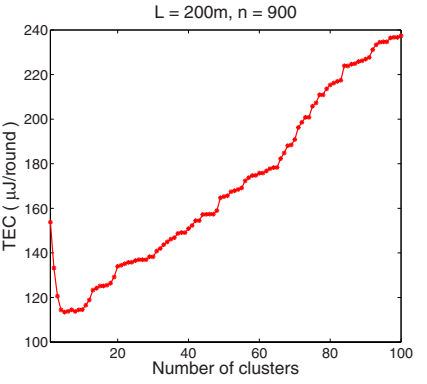

Fig. 8. TEC per round for each k value using adaptive k-means algorithm

two factors: (i) the k-means algorithm may be trapped in local optimal solutions only considering the distance from LNs to CHs, while the TEC has to consider the distances from LNs to CHs and from CHs to the BS; (ii) as widely evidenced, the quality of the final solution produced by the k-means algorithm also depends on the initially (randomly) selected set of CHs. We apply the adaptive k-means algorithm to all 10 cases of different problem sizes, and plot the performance measurements in terms of minimum TEC and optimal number of CHs produced by DCC, adaptive k-means, and deterministic k-means algorithms in Fig. 9. We observe that DCC achieves the best performance

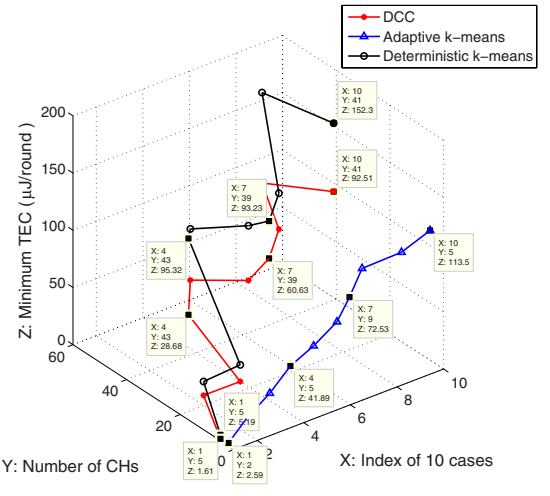

Fig. 9. Performance comparison among DCC, adaptive k-means and deterministic k-means algorithms in 10 cases ranging from small to large scales

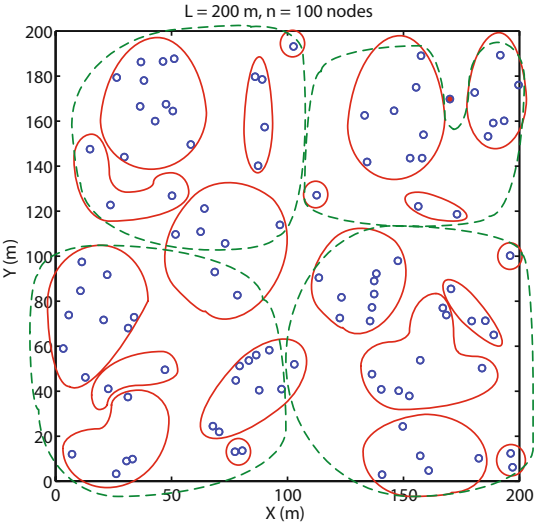

Fig. 10. Visualization of the clustering in a network of 100 nodes using DCC and adaptive *k*-means algorithms: the results of DCC are marked by solid lines and the results of adaptive *k*-means are marked by dashed lines

among the three algorithms in comparison and the adaptive *k*-means algorithm outperforms the deterministic *k*-means algorithm.

Using the same network instance of 100 sensor nodes, we visually compare the clustering results obtained by the DCC algorithm and the adaptive *k*-means algorithm, as shown in Fig. 10. The 20 clusters marked by the solid lines are obtained by the DCC algorithm (the same DCC clustering in Fig. 6), and the 4 clusters marked by dashed lines are obtained by the adaptive *k*-means algorithm. In Fig. 10, we observe that the DCC algorithm achieves a more reasonable clustering result than the *k*-means based algorithms in terms of local sensor density.

6 Conclusion and Future Work

We investigated the problem of selecting a subset of sensor nodes as CHs in WSNs to achieve minimum TEC. We considered two energy dissipation models, free space model for intra-cluster communication and multipath model for inter-cluster communication. We derived an analytical formula to determine the optimal number and location of CHs in WSNs under uniform distribution and proposed a heuristic clustering algorithm based on distance and crowdedness to optimize the number and location of CHs in WSNs under general node distribution. The simulation results illustrated the performance superiority of the proposed solution in comparison with the clustering schemes based on classical *k*-means algorithm.

In the present work, we only considered one-hop communications for both intra-cluster and inter-cluster data communication under uniform and general node

distributions. We will consider multi-hop routing methods and the balance of sensor nodes in each cluster in our future work. It would also be of our future interest to derive an analytical performance bound of energy cost for the clustering of WSNs under complex and general distributions.

Acknowledgment

This research is sponsored by National Science Foundation under Grant No. CNS-0721980 with University of Memphis.

References

1. Abbasi, A.A., Younis, M.F.: A survey on clustering algorithms for wireless sensor networks. J. of Computer Communications 30(14-15), 2826–2841 (2007)
2. Bandyopadhyay, S., Coyle, E.J.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: Proc. of IEEE INFOCOM, vol. 3, pp. 1713–1723 (2003)
3. Das, G.K., Das, S., Nandy, S.C., Sinha, B.P.: Efficient algorithm for placing a given number of base stations to cover a convex region. J. of Parallel and Distributed Computing 66(11), 1353–1358 (2006)
4. Dasgupta, K., Kukreja, M., Kalpakis, K.: Topology-aware placement and role assignment for energy-efficient information gathering in sensor networks. In: Proc. of the 8th IEEE Int. Symposium on Computers and Communications, pp. 341–348 (2003)
5. Du, X., Xiao, Y.: Energy efficient chessboard clustering and routing in heterogeneous sensor network. Int. J. of Wireless and Mobile Computing, IJWMC (2007)
6. Ergen, S.C., Varaiya, P.: Optimal placement of relay nodes for energy efficiency in sensor networks. In: Proc. of IEEE Int. Conf. on Communications, vol. 8, pp. 3473–3479 (2006)
7. Guha, S., Khuller, S.: Approximation algorithms for connected dominating sets. Journal of Algorithmica 20(4), 374–387 (1998)
8. Heinzelman, W.R.: Application-specific protocol architectures for wireless networks. PhD thesis, Massachusetts Institute of Technology (June 2000)
9. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proc. of the 33rd Annual Hawaii Int. Conf. on System Sciences, vol. 2, pp. 3005–3014 (2000)
10. Hyder, A.K., Shahbazian, E., Waltz, E.: Multisensor Fusion. Kluwer Academic, Dordrecht (2002)
11. Jayasimha, D.N., Iyengar, S.S.: Information integration and synchronization in distributed sensor networks. IEEE Trans. Systems, Man, and Cybernetics 21(5), 1032–1043 (1991)
12. Kim, H., Kim, S.W., Lee, S., Son, B.: Estimation of the optimal number of cluster-heads in sensor network. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 87–94. Springer, Heidelberg (2005)
13. Kim, H., Seok, Y., Choi, N., Choi, Y., Kwon, T.: Optimal multi-sink positioning and energy-efficient routing in wireless sensor networks. In: Kim, C. (ed.) ICOIN 2005. LNCS, vol. 3391, pp. 264–274. Springer, Heidelberg (2005)
14. Lloyd, E.L., Xue, G.: Relay node placement in wireless sensor networks. IEEE Trans. on Computers 56(1), 134–138 (2007)
15. Mao, Y., Wu, M.: Coordinated sensor deployment for improving secure communications and sensing coverage. In: Proc. of the 3rd ACM Workshop on Security of Ad Hoc and Sensor Networks, Alexandria, USA, pp. 117–128 (2005)

16. Oyman, E.I., Ersoy, C.: Multiple sink network design problem in large scale wireless sensor networks. In: Proc. of Int. Conf. on Communications, June 2004, vol. 6, pp. 3663–3667 (2004)
17. Pan, J., Cai, L., Hou, Y.T., Shi, Y., Shen, S.X.: Optimal base-station locations in two-tiered wireless sensor networks. IEEE Trans. on Mobile Computing 4(5), 458–473 (2005)
18. Savvides, A., Han, C., Strivastava, M.: Dynamic fine-grained localization in ad-hoc networks of sensors. In: Proc. of ACM MOBICOM, pp. 166–179 (2001)
19. Selvakennedy, S., Sinnappan, S.: An energy-efficient clustering algorithm for multihop data gathering in wireless sensor networks 1(1) (April 2006)
20. Srinivas, A., Zussman, G., Modiano, E.: Mobile backbone networks - construction and maintenance. In: Proc. of the 7th ACM Int. Symposium on Mobile Ad Hoc Networking and Computing, Florence, Italy, pp. 166–177 (2006)
21. Tao, Y., Zheng, Y.: The combination of the optimal number of cluster-heads and energy adaptive cluster-head selection algorithm in wireless sensor networks. In: Proc. of Int. Conf. on Wireless Communications, Networking and Mobile Computing, September 2006, pp. 1–4 (2006)
22. Vincze, Z., Vida, R., Vidacs, A.: Deploying multiple sinks in multi-hop wireless sensor networks. In: IEEE Int. Conf. on Pervasive Services, July 2007, pp. 55–63 (2007)
23. Wang, L.C., Lin, C.M., Wang, C.W.: Optimizing the number of clusters in a wireless sensor network using cross-layer analysis. In: Proc. of IEEE Int. Conf. on Mobile Ad-hoc and Sensor Systems, pp. 585–587 (2004)
24. Xu, N., Cassandras, C.G.: Optimal cluster-head deployment in wireless sensor networks with redundant link requirements. In: Proc. of the 2nd Int. Conf. on Performance Evaluation Methodologies and Tools, pp. 1–9 (2007)
25. Yang, H., Sikdar, B.: Optimal cluster head selection in the LEACH architecture. In: 26th IEEE Int. Performance Computing and Communications Conf., New Orleans, LA, April 2007, pp. 93–100 (2007)