A k-Means-Based Formation Algorithm for the Delay-Aware Data Collection Network Structure

Pat-Yam Tsoi, Chi-Tsun Cheng†, and Nuwan Ganganath
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
Email: †chi-tsun.cheng@polyu.edu.hk

***Abstract*—A wireless sensor network (WSN) consists of a large number of wireless sensor nodes that collect information from their sensing terrain. Wireless sensor nodes are, in general, battery-powered devices with limited processing and transmission power. Therefore, the lifetime of WSNs heavily depends on their energy efficiency. Multiple-cluster 2-hop (MC2H) network structure is commonly used in WSNs to reduce energy consumption due to long-range communications. However, networks with the MC2H network structure are commonly associated with long data collection processes. The delay-aware data collection network structure (DADCNS) is proposed to shorten the duration of data collection processes without sacrificing network lifetime. In this paper, a k-means-based formation algorithm for the DADCNS, namely DADCNS-RK, is proposed. The proposed algorithm can organize a network into the DADCNS, while minimizing the total communication distance among connected sensor nodes by performing k-means clustering recursively. Simulation results show that, when comparing with other DADCNSs formed by different algorithms, the proposed algorithm can reduce the total communication distances of networks significantly.**

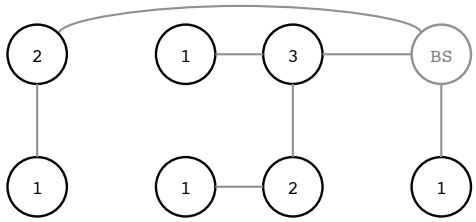
***Index Terms*—wireless sensor networks, delay-aware, data collection process, resources management, k-means algorithms**

I. Introduction

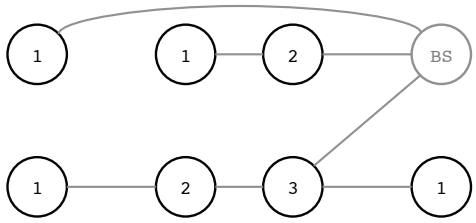
A typical wireless sensor network (WSN) consists of wireless sensor nodes and a remote base station (BS). The BS can be a fixed node or a mobile node, which connects the WSN to an existing communication infrastructure. For prolonging network lifetime, a network is usually divided into several clusters by means of clustering [1]. In each cluster, one of the sensor nodes is chosen as cluster head (CH) and the rest in the same cluster are regarded as cluster members (CM). The CH will receive all the data packets generated from its CMs directly or in a multi-hop manner. In WSNs, the amount of energy used in data transmission is directly related to the communication distance between a sender and a receiver. Longer the communication distance, more energy being dissipated by the sensor nodes. Hence, sensor nodes involved in long distance communications will die out quickly. This leads to a structure change of the WSN. Means to avoid having long communication links are illustrated in the following examples.

Consider a network N as shown in Fig. 1 , which has $|N| = 7$ nodes. Suppose the nodes are organized into the DADCNS using the bottom-up approach proposed in [2]. In such approach, a node or a sub-cluster will try to pair up with its nearest party of the same size. Such *greedy* behaviors work well for small-scale networks. However, this approach can easily be trapped in a local optimum and yield a non-ideal network arrangement. As shown in Fig. 1(a), the cluster on the left has to reach the base station (BS) via a long communication link. The top-down approach in [2] first considers the network as fully connected and tries to construct the DADCNS by removing as many long links as possible. This approach can avoid isolating those nodes at the two lower corners (see Fig. 1(b)). However, some long communication links may still exist.

By exploiting the location information of the sensor nodes, it is possible to yield the DADCNS with shorter communication links. Consider the same network as discussed in Fig. 1, one can easily divide the network into two parts by means of clustering. As geographical separations among nodes within a cluster are relatively shorter, forming DADCNSs inside those



(a) A network organized into the DADCNS using the bottom-up approach in [2].



(b) A network organized into the DADCNS using the top-down approach in [2] and [3].

Fig. 1. Networks with $|N| = 7$ nodes organized using the DADCNS. Circles with numbers represent wireless sensor nodes while circles with labels "BS" represent base stations. The numbers inside the circles indicate their transmission schedules.

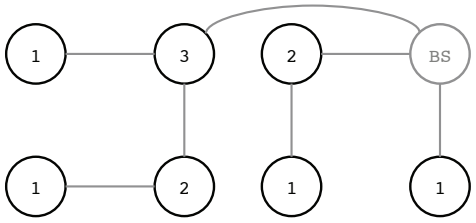


Fig. 2. A network with $|N| = 7$ nodes organized using the DADCNS with the help of a clustering technique. Circles with numbers represent wireless sensor nodes while the circle with a "BS" label represents a base station. The numbers inside the circles indicate their transmission schedules.

clusters separately are less likely to yield long communication links. Fig. 2 is showing a network organized using the DADCNS with the help of a clustering technique, where the network is first vertically divided at the middle into clusters with sizes 4 and 3. The cluster on the right is further divided into clusters with sizes 2 and 1.

In this paper, a k-means-based formation algorithm for the DADCNS is proposed. It utilizes location information of the sensor nodes, such that the DADCNS can be constructed while its associated data links can be kept as short as possible. Simulation results show that the proposed algorithm can greatly shorten communication links without degrading the data collection performance of the DADCNS. The rest of the paper is arranged as follows. Related work is reviewed in Section II. A network formation algorithm based on a k-means algorithm is proposed in Section III. The proposed algorithm is analyzed in Section IV. In Section V, performances of the proposed algorithm are evaluated using computer simulations. The results are further studied and discussed in Section VI. Finally, concluding remarks are given in Section VII.

II. Related Work

Intensive research [1], [4]–[6] has been conducted on reducing energy consumption by forming clusters with appropriate network structures. Heinzelman et al. proposed a clustering algorithm called LEACH [1]. Since then, network formation algorithms based on clustering techniques are developed intensively. Nguyen et al. proposed M-LEACH [7] by improving LEACH. In [8], Jung et al. proposed a network formation algorithm with considerations of both the residual energy of sensor nodes and the number of neighbors around each node when selecting CHs. In addition, Maraiya et al. developed ECHSSDA in which efficient CH selection was proposed [9]. Ducrocq et al. developed BLAC [10], the very first distributed clustering algorithm providing non-overlapping multi-hop clusters with energy concerns. In [3], Cheng and Ganganath are the first who attempted to exploit the geographical locations of sensor nodes in order to facilitate the formation of DADCNS. In their first attempt, a network is divided into sub-clusters by a k-means algorithm. Whenever a sub-cluster is having a cluster size of 2^k, $k \in \mathbb{Z}^+$, such sub-cluster will not be divided any further. Instead a top-down approach proposed in [2] will be executed inside the sub-cluster and it will organize the sub-cluster into the DADCNS. Nevertheless, the formation of DADCNS needs to be further investigated to minimize the communication distance among connected sensor nodes.

III. The Proposed Algorithm

The essence of the proposed DADCNS-RK algorithm is clustering nodes in a network, by using a k-means algorithm recursively, such that their within-cluster geographical separations are minimized. The DADCNS can be maintained by imposing constraints on cluster sizes. Procedures are listed as follows.

Step-1 Initialize the algorithm with a network N together with a centroid C_{P} of its *parent* network. If N is the uppermost network, C_{P} will be replaced by the coordinates of the BS. Calculate the centroid of N and denote it as C. Divide the network into two sub-networks using k-means algorithm (i.e. setting $k = 2$), such that $N = N_1 \cup N_2$ and $N_1 \cap N_2 = \emptyset$. Without lost of generality, assume $|N_1| \geq |N_2|$.

Step-2 Since an ordinary k-means algorithm has no constrain on cluster sizes, N_1 and N_2 will go through a network resizing sub-routine to ensure that

$$|N_1| = 2^{\lceil \log_2(\frac{|N|}{2}) \rceil} \text{ and } |N_2| = |N| - |N_1|.$$

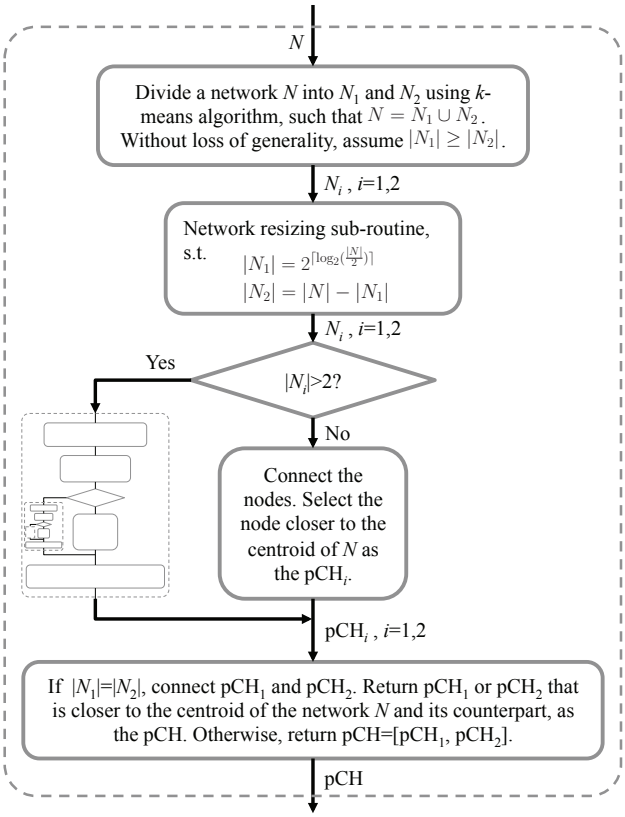


Fig. 3. The flow chart of the proposed DADCNS-RK for constructing networks with a multiple-tree structure.

Step-3 For $i = 1, 2$, if $|N_i| > 2$, set $N_i \rightarrow N$ and $C \rightarrow C_{\text{P}}$. Return to Step-1 and further divide the network recursively. Otherwise, connect the nodes (if any) in N_i. Return the one closer to the centroid C as a potential cluster head pCH_i of the sub-cluster.

Step-4 If $|N_1| = |N_2|$, both sub-networks N_1 and N_2 can form fully-filled DADCNS (i.e. $|N_i| = 2^k,\ k \in \mathbb{Z}^+$). Each sub-network should have a single potential cluster head pCH. Connect pCH_1 with pCH_2 and return the one closer to the centroid C as the pCH of the merged network. Otherwise, return all pCHs in N_1 and N_2 as $\text{pCH} = [\text{pCH}_1,\ \text{pCH}_2]$.

The procedures of the DADCNS-RK algorithm are summarized in Fig. 3. In Step-2, the aim of the network resizing sub-routine is to move nodes between N_1 and N_2 such that $|N_1| = 2^{\lceil \log_2(\frac{|N|}{2}) \rceil}$ and $|N_2| = |N| - |N_1|$. Pseudo-codes of the sub-routine are given in Algorithm 1.

Data: N_1 and N_2, where $|N_1| \neq 2^k, k =\in \mathbb{Z}^+$
Result: $|N_1| = 2^{\lceil \log_2(\frac{|N|}{2}) \rceil}$ and $|N_2| = |N| - |N_1|$.
while $|N_1| \neq 2^{\lceil \log_2(\frac{|N|}{2}) \rceil}$ **do**
  if $|N_1| < 2^{\lceil \log_2(\frac{|N|}{2}) \rceil}$ **then**
    Calculate C_1;
    Move a node from N_2 to N_1 that is closest to C_1;
  else
    Calculate C_2;
    Move a node from N_1 to N_2 that is closest to C_2;
  end
end

Algorithm 1: The network resizing sub-routine

In Step-3, when $|N_i| = 2,\ i = 1, 2$, the DADCNS-RK algorithm will always join the two nodes in N_i together. The one that is closer to the centroid C will be selected as pCH_i. The main reason is to try selecting a pair of pCH_i from N_1 and N_2 that are having a relatively shorter separation. Such technique can help reducing the total communication distance of the constructed network. However, if $|N_i| = 1,\ i = 1, 2$, the only node in the network will be denoted as the pCH_i.

In Step-4, the sizes of the two sub-clusters will only be equal if their parent network has a network size of $N = 2^k,\ k \in \mathbb{Z}^+$. Assuming N_1 and N_2 are both organized using DADCNS, joining pCH_1 with pCH_2 will still maintain the DADCNS in the merged outcome.

The DADCNS-RK algorithm will end with a single pCH if $|N| = 2^k,\ k \in \mathbb{Z}^+$. Otherwise, it will deliver a number of pCHs of sub-clusters with different sizes. All these pCHs will be connected to the BS directly. If multiple-clusters are not allowed, Step-4 should be modified as the shaded box shown in Fig. 4. In the modified version, pCH_1 should always be connected with pCH_2. Afterward, the pCH that is closer to the centroid C_{P} will be selected as the pCH of the merged outcome.

IV. Analyses of the DADCNS-RK

The pCH of a network N with the DADCNS requires $\log_2 N$ time-slots to receive data from its CMs and take an additional time-slot to return the fused data to its parent node or the BS. Therefore, the duration of its data collection process (DCP) is expressed as $T_{\text{DCP}} = \log_2 N + 1$ [2]. In the proposed algorithm, a network of N is divided into $|N_1| = 2^{\lceil \log_2(\frac{|N|}{2}) \rceil}$ and $|N_2| = |N| - |N_1|$. Since both N_1 and N_2 are organized as DADCNS, the pCH of network N_1 will take $T_{\text{DCP\_1}} = \log_2 N_1 + 1 = T_{\text{DCP}} - 1$ to collect data from all its CMs. As $|N_2| \leq |N_1|$, pCH of N_2 will have $T_{\text{DCP\_2}} \leq T_{\text{DCP}} - 1$.

If multiple-cluster is not allowed or if $|N_1| = |N_2|$, the two pCHs will be connected and one of them will become the chief pCH of the merged cluster. Suppose $T_{\text{DCP\_1}} = T_{\text{DCP\_2}}$, both pCHs will finish their DCP using the same number of time-slots. One of the pCHs will take one time-slot to collect the fused data from the other pCH and therefore, T_{DCP} of the merged network $= \log_2 N + 1$. If $T_{\text{DCP\_1}} > T_{\text{DCP\_2}}$, pCH of N_2 can only return data to pCH of N_1 after $T_{\text{DCP\_1}}$ time-slots. Therefore, T_{DCP} of the merged network remains unchanged.

If multiple-cluster is allowed and $|N_1| \neq |N_2|$, the merged network is not a fully-filled DADCNS. Therefore, the two pCHs should not be connected even though N_1 and N_2

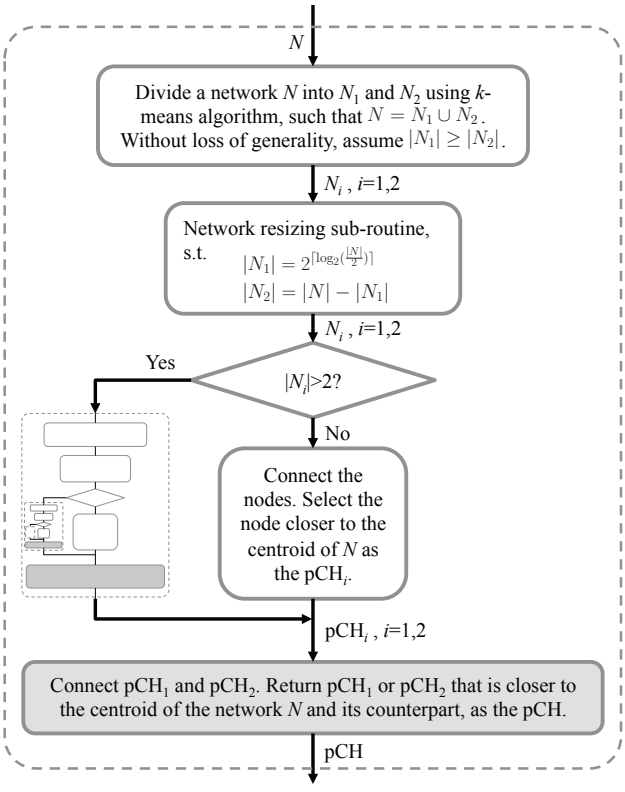


Fig. 4. The flow chart of the proposed DADCNS-RK for constructing networks with a single-tree structure.

are with the DADCNS. Having under-filled clusters leads to unnecessary idling in a data collection process, which should be avoided. If both N_1 and N_2 are filly-filled and $|N_1| > |N_2|$, the two clusters will have different T_{DCP} values. Their pCHs can, therefore, be connected to the BS without introducing any conflict in the transmission schedule. If N_2 is not fully-filled, according to the DADCNS-RK algorithm, the cluster will be further broken down into sub-clusters recursively until all its sub-clusters are fully-filled clusters of different sizes. All the pCHs of these clusters will return data to the BS using different time-slots. The T_{DCP} of the whole network is therefore governed by that of its largest cluster, i.e. $T_{\text{DCP}} = \lceil \log_2 \frac{|N|}{2} \rceil + 1 = \log_2 N$, which concurs with findings in [2], [3].

V. Simulations

The performances of the proposed algorithm are evaluated using computer simulations. In the simulations, the duration of a DCP (T_{DCP}) and the total squared communication distance (Ψ) are used as performance indicators [3]. T_{DCP} is expressed as the total number of time-slots required by a BS to collect data from all the nodes in the network.

The total squared Euclidean distance [2], [3], [11] is expressed as

$$\Psi = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} c_{ij} d_{ij}^2 + \sum_{k=1}^{N} c'_k d'^2_k. \qquad (1)$$

Here, c_{ij} is an indicator showing the existence of a connection between the i^{th} and the j^{th} nodes. If a connection exists, $c_{ij} = 1$, else $c_{ij} = 0$. Variable d_{ij} is representing the Euclidean distance between the i^{th} and the j^{th} nodes. Similarly, c'_k indicates the existence of a connection between the BS and the k^{th} node, while d'_k represents the Euclidean distance between the BS and the k^{th} node. The total squared Euclidean distance is a good estimation for the total energy consumption of a WSN [2].

A. Simulation Settings

Simulations were conducted in Matlab. In each simulation, a network with $|N|$ wireless sensor nodes are distributed randomly on a square sensing terrain with 50×50 m^2, which has its center and one of its corners located at (25, 25) m and (0, 0) m, respectively. The BS is located at the center of the terrain, which tries to collect data from all the nodes in the networks. In the simulations, performance of the original DADCNS will be used as references. The DADCNS will be constructed as a single cluster and multiple clusters using the top-down network formation approaches proposed in [2] and [3], respectively. In order to evaluate the effect of $|N|$ to the performances of networks with different network structures, $|N|$ is varied from 3 to 99 with a step-size of 3. In the simulations, all the network formation algorithms are implemented in a centralized manner. Results presented in this paper are the averaged values of 50 simulations.

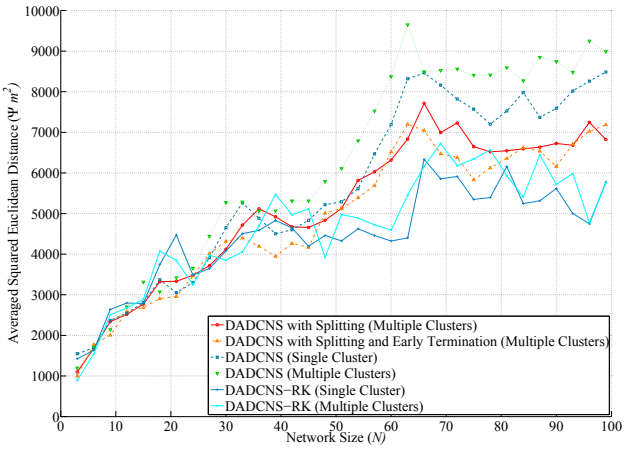


Fig. 5. The averaged squared Euclidean distance of networks with the DADCNS formed by different algorithms.

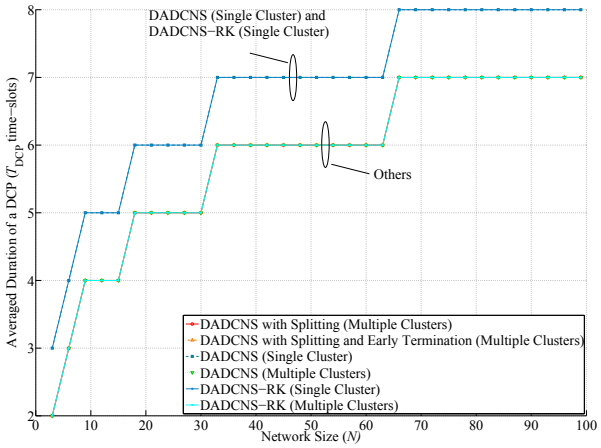


Fig. 6. The averaged duration of a data collection process in networks with the DADCNS formed by different algorithms. Note that except networks with the DADCNS (Single Cluster), results of networks with other structures are overlapping (the lower curve).

B. Simulation Results

Simulation results are shown in Fig. 5 and Fig. 6. In general, Ψ values of networks with different network formation algorithms increase with $|N|$. Networks with the proposed DADCNS-RK algorithm can achieve lower values of Ψ especially for scenarios with large $|N|$. T_{DCP} of networks with different network formation algorithms increase monotonically with $|N|$.

VI. Discussions

As expected in section IV, the averaged duration of DCP in networks with the DADCNS formed by the proposed DADCNS-RK algorithm is the same as that of DADCNS formed by other algorithms. The reason is DADCNS-RK could attain the same network structure as DADCNS for both single-

cluster and multiple-cluster cases. The time slots needed for the BS to collect all data packets in the network could hence be unaffected. In terms of minimizing Ψ, the DADCNS formed by the proposed algorithm DADCNS-RK outperforms the DADCNS formed by other algorithms significantly when the network size N is larger than 30, which show that the proposed network formation algorithm is highly suitable for large-scale networks. For networks with $N < 30$, performances of all algorithms under test were close to each other. DADCNS-RK (Single Cluster) could achieve a better performance of reducing Ψ than DADCNS-RK (Multiple Clusters) in general. The main reason is that in a single-cluster structure, there is only one CH; while in a multiple-cluster structure, there are more than one CHs. The total communication distance between the BS and the CHs is expected to be higher in the later case.

VII. Conclusions

In this paper, a k-means-based formation algorithm for the DADCNS, namely DADCNS-RK, is proposed. To cater for different applications, two variations of DADCNS-RK are proposed such that a network can be constructed in either single-cluster or multiple-cluster styles. Performances of the proposed algorithm are evaluated based on the averaged squared Euclidean distance of the network and the averaged duration of a data collection process in the network. Networks with the proposed algorithm are compared with networks formed by the conventional DADCNS formation algorithms with and without splitting. Simulation results show that networks formed by the proposed DADCNS-RK algorithm can greatly reduce the averaged squared Euclidean distance of the network while keeping the averaged duration of a data collection process in the network the same as that of other DADCNSs.

Acknowledgment

This work is supported by the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University (Project G-UB45) and the Hong Kong PhD Fellowship Scheme.

References

[1] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. on Wireless Communications*, vol. 1, no. 4, pp. 660–670, October 2002.

[2] C.-T. Cheng, C. K. Tse, and F. C. Lau, "A delay-aware data collection network structure for wireless sensor networks," *Sensors Journal, IEEE*, vol. 11, no. 3, pp. 699–710, March 2011.

[3] C.-T. Cheng and N. Ganganath, "To split or not to split? from the perspective of a delay-aware data collection network structure," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*, Oct 2013, pp. 349–355.

[4] S. Lindsey and C. S. Raghavendra, "PEGASIS: Power-efficient gathering in sensor information systems," in *Proc. IEEE Conf. Aerospace*, vol. 3, Big Sky, Montana, USA, March 2002, pp. 1125–1130.

[5] H. O. Tan and I. Körpeoğlu, "Power efficient data gathering and aggregation in wireless sensor networks," *SIGMOD Rec.*, vol. 32, no. 4, pp. 66–71, Dec. 2003.

[6] O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, and P. Levis, "Collection tree protocol," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '09. New York, NY, USA: ACM, 2009, pp. 1–14.

[7] L. T. Nguyen, X. Defago, R. Beuran, and Y. Shinoda, "An energy efficient routing scheme for mobile wireless sensor networks," in *Wireless Communication Systems. 2008. ISWCS '08. IEEE International Symposium on*, Oct 2008, pp. 568–572.

[8] I. Jung, B. Lee, N. Ha, K. Cho, Y. Choi, M. Choi, B. Lee, and K. Han, "An energy efficient clustering method for wireless sensor networks," in *Proceedings of the 6th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications*, ser. EHAC'07. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2007, pp. 139–144.

[9] K. Maraiya, K. Kant, and N. Gupta, "Article: Efficient cluster head selection scheme for data aggregation in wireless sensor network," *International Journal of Computer Applications*, vol. 23, no. 9, pp. 10–18, June 2011, published by Foundation of Computer Science.

[10] T. Ducrocq, M. Hauspie, and N. Mitton, "Balancing energy consumption in clustered wireless sensor networks," *ISRN Sensor Networks*, vol. 2013, pp. 1–14, 2013.

[11] C.-T. Cheng, H. Leung, and P. Maupin, "A delay-aware network structure for wireless sensor networks with in-network data fusion," *Sensors Journal, IEEE*, vol. 13, no. 5, pp. 1622–1631, May 2013.