On the Optimality of Non-Uniform Clustering in Wireless Sensor Networks

Ali Dabirmoghaddam Majid Ghaderi Carey Williamson

Department of Computer Science, University of Calgary

{adabirmo, mghaderi, carey}@cs.ucalgary.ca

Abstract

In wireless sensor networks, cluster-based data gathering has been pursued as a means to achieve network scalability as well as energy efficiency. By dividing a network into clusters, data aggregation and compression can be conveniently implemented in each cluster resulting in significant reduction in overall network energy consumption. Although many clustering algorithms have been proposed in the literature for minimizing energy consumption in sensor networks, a comprehensive and systematic analysis of optimal clustering subject to inherent network attributes such as data correlation, node density, and distance to the sink is still lacking. In particular, existing clustering schemes are designed to form uniform clusters in the network, where, on average, clusters have the same size. In this paper, we exploit spatial data correlation present among sensor readings to form optimal-sized clusters that minimize the total energy cost of the network. We develop a generalized multi-region network model that captures the interplay between clustering and data correlation, and postulate that a heterogeneous clustering scheme, with larger clusters at further distances from the sink, is more energy-efficient than uniform clustering. Based on this model, we develop a distributed randomized clustering algorithm that generates optimal-sized non-uniform clusters throughout the network. Simulation results confirm the superiority of our novel algorithm against uniform clustering schemes.

Index Terms

Energy-efficiency, clustering, data correlation, wireless sensor networks.

On the Optimality of Non-Uniform Clustering in Wireless Sensor Networks

I. Introduction

A *Wireless Sensor Network* (WSN) is formed by a large collection of cooperative micro-electronic sensing devices that are equipped with wireless communication capability. These autonomous self-configurable networks have given rise to many types of applications, from disaster management to home automation, and from health control to military missions [1]. The small dimensions of wireless sensors mean that they usually operate on limited-capacity batteries that are difficult or impossible to be replaced. Therefore, energy has always been a constrained resource for seamless operation of such networks. A highly active research topic in the area of WSNs is thus fine tuning network operations to consume less energy.

Clustering is a well-established technique that has been adopted primarily to address scalability issues in WSNs [2]. With clustering, sensor nodes are grouped into small disjoint sets that are coordinated by one of the cluster members known as *Cluster-Head* (CH). The CH is in charge of managing the internal activities of the cluster, such as scheduling nodes for intermittent subject monitoring and data transmission. Another side advantage that clustering can reward is local data compression. Since in most applications, sensor nodes are deployed densely within the environment, significant amount of redundancy is likely to be present among the readings from adjacent sensors. For instance, in a camera sensor network, the same event may be detected by multiple camera sensors in a local neighborhood [3]. Likewise, for scalar data gathering, such as remote temperature monitoring, measurements reported by proximally-located sensors are likely to be very close. This data dependence can be exploited to eliminate the redundancies and reduce the volume of information transmitted in a WSN. In a cluster-based sensor network, individual sensors transmit their observations to their corresponding CH. The CH compresses the whole cluster data and transmits a representative condensed message (subject to some tolerable distortion level) to the sink (the designated collection station).

Many clustering algorithms have been proposed for WSNs during the past few years [3]–[9]. Although many of them result in some form of energy-efficient clusters, they do not carefully

take into account the physical properties of the network, such as data dependency, node density and distance from the sink when forming clusters.

In this paper, we show that in correlated data fields, *non-uniform clustering* is more effective in minimizing network energy consumption as opposed to existing belief that uniform clustering results in optimal energy consumption [4], [5]. In particular, we take advantage of the inherent network properties such as data correlation and node density to propose an energy-efficient *distributed* clustering algorithm. Based on this algorithm, sensor nodes become CH according to a probability model that is a function of correlation degree, node density and distance to the sink. With this model, nodes in close vicinity of the sink are more likely to be elected as CH than those in far distances. Therefore, our novel clustering algorithm results in a heterogeneous tessellation of the network that outperforms the existing uniform clustering schemes in terms of energy consumption.

The main contributions of this paper can be summarized as follows:

- We develop an energy model for a WSN that accurately captures the *joint* effects of clustering and compression in the network. The model is appreciably general in that it considers non-homogeneous clusters in various regions of the network.
- Based on this model, we devise a clustering algorithm that generates optimal-sized clusters throughout the network in a distributed manner.
- Using numerical analysis and simulation experiments, we validate our model and demonstrate that non-uniform clustering outperforms the existing uniform clustering schemes in WSNs with correlated data.

The remainder of the paper is organized as follows. Section II surveys recent literature on WSN optimization. Mathematical preliminaries used to tackle our problem are reviewed in Section III. Sections IV and V are devoted to our mathematical energy model. Our distributed clustering algorithm is presented in Section VI. Section VII validates our model using numerical analysis and simulation experiments, while Section VIII concludes this paper.

II. Related Work

Many clustering methods for WSNs have been proposed over the past few years. A comprehensive and rather up-to-date survey appears in [2] classifying tens of clustering algorithms from various perspectives, such as convergence rate, cluster stability, and location and mobility

awareness. The current trend for clustering algorithms mostly concentrates on improving the network scalability, stabilizing the network topology and facilitating the routing arrangements. In spite of many energy-aware clustering strategies in the literature, a detailed analysis of optimal cluster formation subject to the intrinsic network properties, such as data correlation, node density and distance from the sink is still lacking. The existing approaches either ignore the joint effect of such attributes on clustering or make simplifying assumptions that are incorrect or far from reality.

A seminal analysis of energy-efficient correlated data gathering is presented in [10]. In that work, the authors argue that the joint optimization of distributed rate allocation and transmission structure in a sensor network is NP-complete. However, they propose and evaluate several near-optimal solutions for this problem. Although the authors do not explicitly focus on cluster-based data gathering in their work, their analysis provides a firm theoretical base for distributed rate allocation and transmission optimization in presence of data correlation.

A classic example of energy-aware clustering protocols for WSNs is LEACH [4]. In LEACH, each node has an equal chance of becoming a CH based on some probability function. Non-CH nodes join the CH that needs the least communication energy to be reached. CHs perform local data fusion to compress the cluster data and then form a hierarchical structure to route the information to the sink. While scalable and light-weight, LEACH does not provide a model for computing the optimal probability of CH selection. This problem is addressed in [5], where the authors quantify the optimal probability of CH selection with respect to the network dimensions and node density. Although representing a step forward in optimal clustering, no notion of data correlation and compression is considered in [5] when forming the clusters. Also, the effect of distance (between the clusters and the sink) is again neglected resulting in uniformly-sized clusters.

Error-tolerant WSNs can be approximated using lossy estimators. Several local estimations of the process under observation can be aggregated into a single outgoing flow that is routed towards the sink as studied in [8]. The authors, however, do not take the correlation between observations into account in their analysis. Therefore, the fused outgoing flow contains redundancies resulting from spatial proximity of original observations and is not optimal. A recent work manifests a thorough analysis of the trade-off between eliminating data redundancy and maintaining estimation accuracy [11] although the paper does not consider clustering.

The impact of spatial data correlation on optimal cluster sizing is also studied in some previous works. For instance, [6] and [12] model and analyze various configurations of a simple linear network topology and formulate the optimal cluster size with respect to the number of locally similar observations and distance to the sink. Due to the complexity of modeling the joint data compression in correlated data fields, the authors make some simplifying assumptions (*e.g.*, a *fixed rate* of data reduction per source due to compression) that inevitably influence the outcomes. In this work, in contrast, we take advantage of widely used techniques from information theory to model data correlation/compression. Our comprehensive analysis reveals new aspects of optimal clustering that were not known prior to this research.

All aforementioned approaches result in *uniform* clustering of the network with all clusters containing, on average, the same number of nodes. As we shall see later in this paper, the optimal cluster size is significantly affected by the correlation degree and cluster distance to the sink. Therefore, there is no unique optimal cluster size that performs equally well for various degrees of data correlation and over different regions of the network. In a previous work, we demonstrated that in correlated data fields, the optimal size of clusters is directly proportional to the cluster distance to the sink [9]. Our previous analysis, however, includes a simple single-cluster model that is analyzed solely to obtain some insight about the problem as it does not represent the behavior of a real network containing many clusters. In this paper, we focus on a general multi-cluster network model and study optimal non-uniform clustering for two-dimensional sensor networks. As we show, however, the analysis of the optimal clustering is significantly more challenging in this model.

III. System Model and Assumptions

We assume that individual sensor nodes within the WSN are statistically identical information sources. Assuming that sensor readings follow a zero mean normal distribution with variance σ^2, the set of observations within a cluster can be represented by a *multi-variate Gaussian distribution*. This assumption makes our analysis easier as the analytical properties of Gaussian sources are well-studied in the literature. Furthermore, Gaussian sources are the worst case in terms of the number of bits required to represent the field [13]. Thus, the results from Gaussian fields can be interpreted as a bound for sources of other types as well. We note that similar assumptions have been used in related work as well [10], [14].

In the following subsections, we formalize the data correlation and compression models used in our analysis.

A. Data Correlation Model

In a Gaussian field comprising N sources, a symmetric positive-definite *covariance matrix* $\Sigma = [\sigma_{ij}]_{N \times N}$ expresses the data dependency between each pair of sensor readings. For spatially correlated data fields, this dependency is often assumed to be a non-negative decreasing function of Euclidean distance, d, between sources. The limiting values are 1 at $d = 0$, and 0 at $d = \infty$. As the Euclidean distance between two sources increases, the correlation between them monotonically approaches zero.

Depending on the physical properties of the random field, several types of covariance models can be defined. A number of them are reviewed in [15]. The information collected from physical events often has an exponential autocorrelation function [14]. Therefore, in this paper, we use a special type of Power Exponential correlation model for which the elements of the covariance matrix are defined as:

$$\sigma_{ij} = \sigma^2 \exp(-\alpha d_{ij}^2) \ , \tag{1}$$

where α is the correlation exponent and d_{ij} denotes the Euclidean distance between sensor nodes i and j. For brevity, we define the parameter $W = \exp(-\alpha)$. W is a normalized parameter (***i.e.***, $0 < W < 1$) representing the degree of correlation. The limiting values, $W = 0$ and $W = 1$, respectively represent uncorrelated and highly correlated data fields.

B. Data Compression Model

As stated in the preceding subsection, sensor readings are drawn from a normal distribution. In order to discretize the continuous readings, the cluster members locally quantize their observations and transmit them to the CH. Since the originally transmitted data is quantized, the reconstructed version of data at the CH is subject to some distortion D.

For discrete information sources, the entropy function yields the minimum number of bits required to encode the source. However, for a continuous source, due to the infinite precision, the number of bits required to encode the source is also infinite. Hence, we assume that sensor readings, denoted by S, are discretized by a uniform quantizer of step size Δ. To achieve the

target distortion D, we set $\Delta = \sqrt{12D}$ [9]. The entropy of the quantized sources denoted by $H(S^D)$ is then given by [16]:

$$H(S^D) \approx \frac{1}{2}\log_2(\frac{\pi e}{6D})^{\varrho(\Sigma)}|\Sigma|^+ \ , \tag{2}$$

where $|\Sigma|^+$ and $\varrho(\Sigma)$ denote the product of non-zero eigenvalues and the rank of Σ, respectively. Equation (2) gives the lower-bound for the net size of the joint cluster data after quantization/compression. For individual sources (isolated CHs or individual cluster members), Equation (2) is simplified as follows:

$$H(S_1^D) \approx \frac{\sigma^2}{2}\log_2(\frac{\pi e}{6D}) \ . \tag{3}$$

IV. Non-Uniform Clustering

In this section, using the mathematical preliminaries discussed in the previous section, we develop a general 2-D network model with multiple clusters and examine the effect of data compression and distance on optimal cluster sizing and energy consumption.

We consider a planar disk-shaped network of radius R and assume that sensor nodes are scattered over the network area randomly according to a Poisson distribution of intensity ρ. For simplicity of analysis, let us assume that the sink is placed at the center of the disk. To study the effect of distance on the optimal size of the clusters, we split the network into two concentric ring-shaped areas: namely, the *interior* and the *exterior* regions (See Fig. 1). [1]

The radius of the interior region, r_{int}, is a fraction of the total network radius. That is to say,

$$r_{int} = \kappa R \ , \qquad 0 < \kappa < 1 \ . \tag{4}$$

We study a probabilistic clustering model in which nodes become CH with some probability p. This probability in the interior region (denoted by p_{int}) is independent of that for the exterior region (denoted by p_{ext}). Therefore, in any of the described regions, non-CH and CH nodes can be considered as two independent Poisson processes Π_0 and Π_1 with intensities $\rho_0 = (1-p)\rho$ and $\rho_1 = p\rho$, respectively (for the interior region, $p = p_{int}$, while $p = p_{ext}$ for the exterior region).

[1] By convention, in this section, we use subscripts int and ext to denote the analytical properties of the interior and exterior regions, respectively.

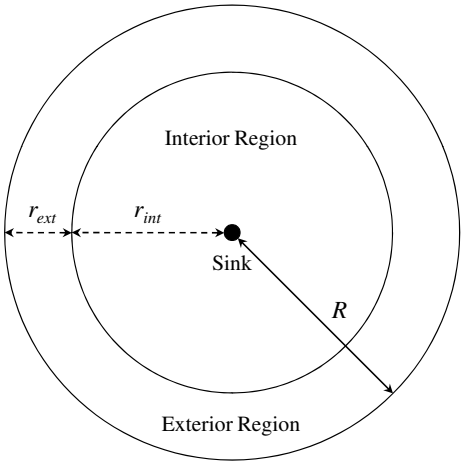


Fig. 1. A disk-shaped network split into 2 regions.

Once the CHs are specified, each region is partitioned into clusters resembling Voronoi cells with CHs representing the nuclei. Non-CH nodes are then assigned to the CH that is geographically closest to them, forming a Voronoi tessellation of the region. Clearly, since the probability p is fixed throughout each region, all the clusters (Voronoi cells) in a region are expected to contain the same number of nodes (on average).

The total expected number of sensor nodes in the network is $\rho\pi R^2$. Among them, a fraction of $p_{int} + p_{ext}$ are CHs. The expected number of clusters in the interior region is given by:

$$\mathbb{E}[\mathcal{N}_{int}] = p_{int} \cdot \rho\pi\kappa^2 R^2 \ , \tag{5}$$

and likewise, for the exterior region:

$$\mathbb{E}[\mathcal{N}_{ext}] = p_{ext} \cdot \rho\pi(1-\kappa^2) R^2 \ . \tag{6}$$

Throughout our analysis, we use a bit-hop metric to quantify the network energy consumption. In order to compute the expected transmission energy within any given cluster, we need to measure the total number of hops taken to communicate sensor readings to the CH. Let $\mathcal{L}$ be the cumulative Euclidean distance of all the cluster members to their CH (the Voronoi cell nucleus). Foss and Zuyev [17] have derived the following relation for a bivariate Poisson process as described above:

$$\mathbb{E}[\mathcal{L}] = \frac{\rho_0}{2\rho_1^{3/2}} = \frac{1-p}{2\rho^{1/2}p^{3/2}} \ . \tag{7}$$

Let $\mathcal{R}$ denote the radio range of a sensor node. We assume that all sensor nodes have the same radio range $\mathcal{R}$. Consequently, the total number of hops traversed within the cluster is readily calculated as $\lceil \mathbb{E}[\mathcal{L}]/\mathcal{R} \rceil$.

Using this information, in the following subsections, we carefully analyze the mean cost of data collection from within the clusters (intra-cluster cost) and also between the CHs and the sink (inter-cluster cost).

A. Intra-Cluster Data Collection Cost

Cluster members observe some spatial stochastic process in time, quantize their observations and transmit them to their CH via multi-hop transmission. This subsection analyzes the energy expended by the cluster-members to transmit their observations to the CHs. We call this energy the *intra-cluster data collection cost*, and derive it for the interior and exterior regions separately.

Since we assume that all sensor nodes have the same radio range, $\mathcal{R}$, the energy, ε, required to transmit one bit of information from a node to any other node in its radio coverage is fixed and proportional to the square of node's radio range [18]:

$$\varepsilon = \gamma \mathcal{R}^2, \tag{8}$$

where γ is a constant that represents the minimum power level required for successful transmission of one bit of data over one unit of distance. For simplicity and without loss of generality, hereafter we assume that $\gamma = 1$ J/bit/m^2. The energy spent on receiving a message is only a function of the message size and is independent of the distance over which the message is delivered. Since all sensor observations are quantized into data packets of the same size, we simply ignore the energy usage for data reception throughout our analysis.

In an arbitrary cluster in the interior region, $\lceil \mathbb{E}[\mathcal{L}_{int}]/\mathcal{R} \rceil$ gives the total number of hops traversed to deliver the readings to the CH. Let b_1 denote the size of a single quantized observation (in terms of bits) given by Equation (3). The mean total cost of data collection from an arbitrary cluster in the interior region is given by:

$$\mathbb{E}[\mathcal{C}^*_{int}] = b_1 \varepsilon \lceil \frac{\mathbb{E}[\mathcal{L}_{int}]}{\mathcal{R}} \rceil \approx b_1 \mathcal{R} \mathbb{E}[\mathcal{L}_{int}] \ . \tag{9}$$

Therefore, the mean total cost of intra-cluster data collection for the interior region is given by:

$$\mathbb{E}[\mathcal{C}^{\text{intra}}_{int}] \approx b_1 \mathcal{R} \mathbb{E}[\mathcal{N}_{int}] \mathbb{E}[\mathcal{L}_{int}] \ . \tag{10}$$

Applying the same argument, for the exterior region, it is obtained that:

$$\mathbb{E}[\mathcal{C}_{ext}^{\text{intra}}] \approx b_1 \mathcal{R}\mathbb{E}[\mathcal{N}_{ext}]\mathbb{E}[\mathcal{L}_{ext}] \ . \tag{11}$$

B. Inter-Cluster Data Collection Cost

Once the CH collects the data from all cluster members, it eliminates the redundancies present in the data by means of a lossless compression and transmits the compressed data to the sink over the *shortest path tree (SPT)*. The *inter-cluster data collection cost* refers to the energy spent by the CHs to communicate their compressed data to the sink.

The mean number of nodes in a cluster, n, is inversely proportional to the probability of being a CH in the region to which the cluster belongs. That is, $n = 1/p$. As discussed in Section III-B, the size of the compressed cluster data subject to some distortion level D can be quantified by the joint entropy of the cluster. For a cluster of size n, let b_n denote the size of the message (in bits) that CH transmits to the sink (b_n can be computed from Equation (2)). In order to compute the amount of energy required for this transmission, we only need to know the distance between the CH and the sink.

In the interior region, the mean distance of nodes to the sink (center of the network) is computed as (See Fig. 2):

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{int}] &= \int_0^{r_{int}} \mathcal{D}_{int}\mathbb{P}\{\mathcal{D}_{int}\}\, d\mathcal{D}_{int} \\
&= \int_0^{r_{int}} \mathcal{D}_{int} \cdot \frac{2\pi\mathcal{D}_{int}}{\pi r_{int}^2}\, d\mathcal{D}_{int} \\
&= \frac{2}{\kappa^2 R^2} \times \frac{\mathcal{D}_{int}^3}{3}\Big|_0^{\kappa R} \\
&= \frac{2}{3}\kappa R \ .
\end{aligned} \tag{12}$$

Considering that the mean number of hops to reach the sink from the interior region is given by $\lceil \mathbb{E}[\mathcal{D}_{int}]/\mathcal{R} \rceil$, the mean total cost of transmitting data from all the CHs in the interior region to the sink is readily calculated as:

$$\mathbb{E}[\mathcal{C}_{int}^{\text{inter}}] \approx b_{n_{int}} \mathcal{R}\mathbb{E}[\mathcal{N}_{int}]\mathbb{E}[\mathcal{D}_{int}] \ . \tag{13}$$

Likewise, the expected cost of inter-cluster data collection for the exterior region is:

$$\mathbb{E}[\mathcal{C}_{ext}^{\text{inter}}] \approx b_{n_{ext}} \mathcal{R}\mathbb{E}[\mathcal{N}_{ext}]\mathbb{E}[\mathcal{D}_{ext}] \ , \tag{14}$$

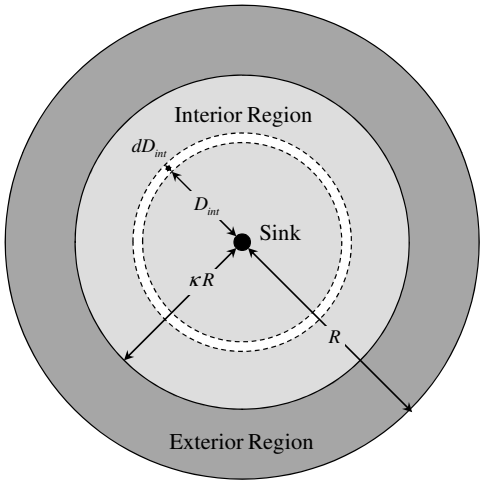


Fig. 2. Computing the mean distance of nodes in the interior region to the sink.

where,

$$\begin{aligned}\mathbb{E}[\mathcal{D}_{ext}] &= \int_{r_{int}}^{R} \mathcal{D}_{ext} \cdot \frac{2\pi\mathcal{D}_{ext}}{\pi(R^2 - r_{int}^2)}\, d\mathcal{D}_{ext} \\ &= \frac{2}{3} \cdot \frac{1+\kappa+\kappa^2}{1+\kappa} R \ .\end{aligned} \tag{15}$$

C. Total Data Collection Cost

The total cost of collecting data from the entire network is the sum of inter-cluster and intra-cluster costs over both regions:

$$\mathbb{E}[\mathcal{C}_{\text{total}}] = \mathbb{E}[\mathcal{C}_{int}^{\text{intra}}] + \mathbb{E}[\mathcal{C}_{int}^{\text{inter}}] + \mathbb{E}[\mathcal{C}_{ext}^{\text{intra}}] + \mathbb{E}[\mathcal{C}_{ext}^{\text{inter}}] \ . \tag{16}$$

While the boundary between the two regions is fixed, $\mathbb{E}[\mathcal{C}_{\text{total}}]$ is a function of p_{int} and p_{ext}. Let p_{int}^* and p_{ext}^* denote the optimal values of p_{int} and p_{ext} that minimize the total network energy consumption.

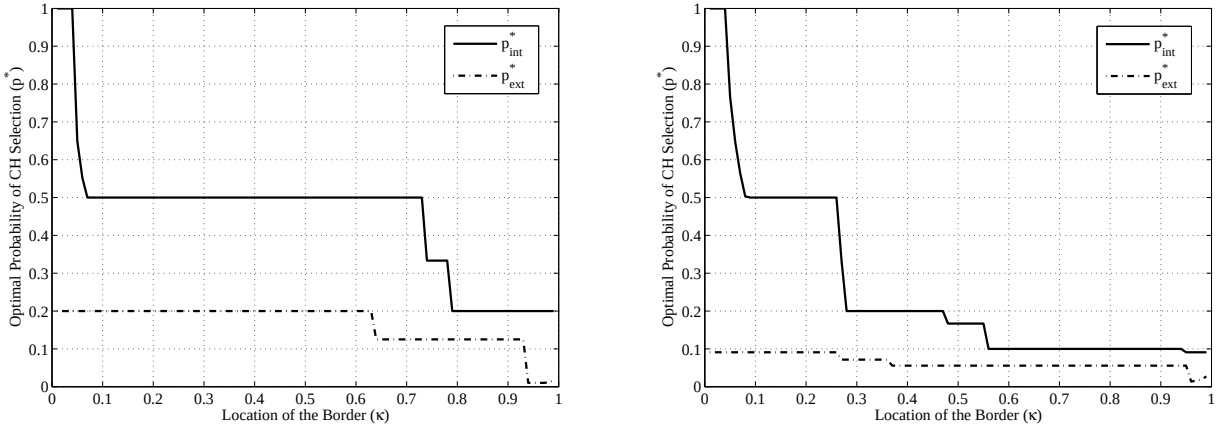
D. Experimental Analysis

In this subsection, we present some concrete numerical results to show the utility of our model and investigate the effect of various parameters on network energy consumption.

We scatter sensor nodes on a network of radius 15 with a density of 0.75 nodes per unit area (roughly, a total of 530 nodes on average). By letting κ change from 0 to 1, we gradually

move the boundary between the two regions through the entire network area. For any particular placement of the border, we then perform an exhaustive search over the interval $[0 \quad 1] \times [0 \quad 1]$ to find the pair (p^*_{int}, p^*_{ext}) that minimizes Equation (16).

Fig. 3 illustrates the optimal probabilities of CH selection in interior and exterior regions for any value of κ between 0 to 1 for medium and high degrees of data correlation.



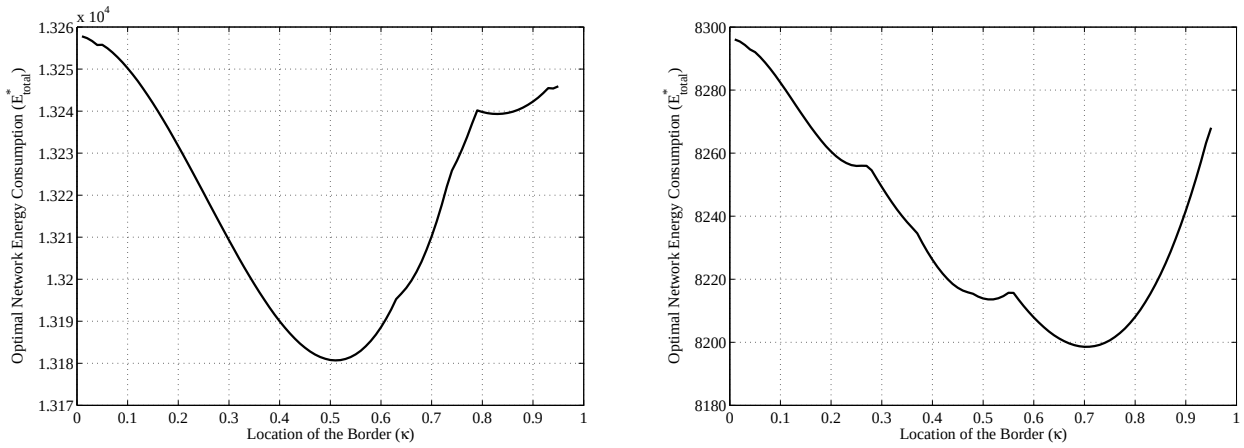
(a) Medium correlation ($W = 0.5$) (b) High correlation ($W = 0.9$)

Fig. 3. Optimal probability of CH selection vs. location of the border.

As evident from this figure, p^*_{int} is always greater than p^*_{ext} for all values of κ. This suggests that, regardless of the position where the interior and exterior regions are separated, the probability of being CH in the interior region is always greater than in the exterior region. That is, *clusters in the interior region are smaller than in the exterior region*. It is important to note that throughout this experiment p_{int} and p_{ext} are chosen independently. Therefore, if according to the conventional clustering algorithms, a uniform clustering strategy is the optimal solution for our problem, it is required that $p^*_{int} = p^*_{ext}$. However, the simulation results suggest that $p^*_{int} > p^*_{ext}$ regardless of the size of the regions. The state transitions in Fig. 3 are due to the fact the cluster size, n, is a whole number; therefore, the optimal probability $p = 1/n$ can only take certain values over the interval $[0 \quad 1]$.

Another interesting observation is that for any particular placement of the border, the optimal probability of becoming CH when the correlation degree is high is lower than or equal to the case when the correlation degree is medium. This suggests that with more data correlation, the optimal

size of the clusters become larger. This is intuitive in the sense that the higher the data correlation degree is, the more redundancy is present among the readings from a local neighborhood. In other words, with high data correlation, the effective radius in which the amount of redundancy between readings from neighboring sensor nodes is substantial increases.



(a) Medium correlation ($W = 0.5$)

(b) High correlation ($W = 0.9$)

Fig. 4. Optimal network energy consumption vs. location of the border.

Next, we analyze the effect of changing the border location on the network energy consumption. As Fig. 4 indicates, placing the border at a radius between $0.50R$ and $0.70R$ gives satisfactory results for both medium and high degrees of data correlation. At $\kappa = R/\sqrt{2} \approx 0.70R$, the network is split into two equal-area regions which both contain, on average, the same number of nodes. Hence, changing the clusters size through any of the two regions affects the energy cost of only one half of the total network nodes and thus imposes a fair impact on the total network energy consumption. We use this heuristic as a basis to build our generalized multi-region network model upon in the following section. We, however, emphasize that our analysis is general and can easily be modified to fit other scenarios as well (*e.g.*, networks with equal-width regions, etc).

V. Generalized Non-Uniform Clustering

So far, using a dual-region network model, we have found that in a correlated data field, the distance from the sink directly affects the size of the clusters. We also analyzed the optimal

distance from the sink to separate the interior and exterior regions. In this section, we extend our analysis to a general multi-region network model. From our dual-region network analysis, we speculate that the network energy consumption is within a reasonable threshold from the optimal when both the interior and exterior regions have the same area. Thus, with our multi-region model, we split the network into m concentric ring-shaped equal-area regions making each region containing the same number of nodes (on average).

We assign each region with a number i from 1 to m from the innermost region all the way to the outermost one. The width of region i is denoted by r_i (See Fig. 5).

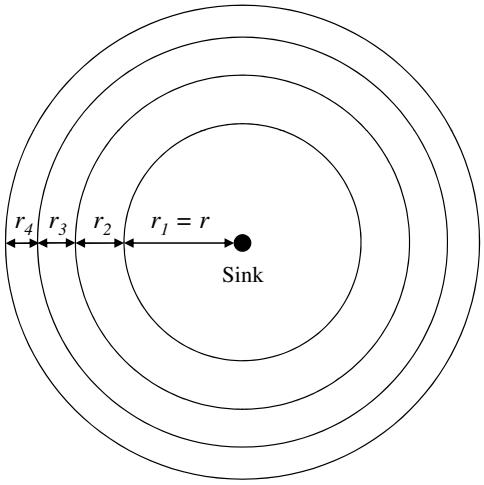


Fig. 5. A disk-shaped network split into 4 equal-area regions.

In region i, nodes become CH with a probability p_i. This probability is identical and independent of that of other regions. Non-CH nodes join their closest CH node to form a non-homogeneous Voronoi tessellation of the network. Clearly, all the Voronoi cells (clusters) in region i (that are roughly at the same distance from the sink) are expected to be of the same size.

Going through the same steps as the dual-region model, the mean intra-cluster energy cost for data gathering from all the clusters of region i is obtained as:

$$\mathbb{E}[\mathcal{C}_i^{\text{intra}}] \approx b_1 \mathcal{R} \mathbb{E}[\mathcal{N}_i] \mathbb{E}[\mathcal{L}_i] \ , \tag{17}$$

where,

$$\mathbb{E}[\mathcal{N}_i] = p_i \rho \pi r^2 \ , \tag{18}$$

is the mean number of clusters in region i, and $\mathcal{L}_i$ is the cumulative distance of nodes to the CH in any cluster in region i that was defined earlier by Equation (7).

Since the network is evenly divided into m regions of all the same area, we have:

$$\begin{aligned} \pi r^2 &= \pi \left(\sum_{k=1}^{i} r_k \right)^2 - \pi \left(\sum_{k=1}^{i-1} r_k \right)^2 , \\ r^2 &= r_i^2 + 2 \left(\sum_{k=1}^{i-1} r_k \right) r_i \ . \end{aligned} \tag{19}$$

Also, it is clear that:

$$\begin{aligned} \pi \left(\sum_{k=1}^{i-1} r_k \right)^2 &= (i-1)\pi r^2, \\ \sum_{k=1}^{i-1} r_k &= \sqrt{i-1}\ r \ . \end{aligned} \tag{20}$$

From Equations (19) and (20), we obtain the following expression for the width of region i:

$$r_i = \left(\sqrt{i} - \sqrt{i-1} \right) r \ . \tag{21}$$

As stated earlier, all the clusters in region i are almost at the same distance from the sink. Hence, the approximate cluster distance can be computed as follows:

$$\begin{aligned} \mathbb{E}[\mathcal{D}_i] &= \int_{\sqrt{i-1}\ r}^{\sqrt{i}\ r} \mathcal{D}_i \cdot \frac{2\pi \mathcal{D}_i}{\pi r^2}\, d\mathcal{D}_i \\ &= \frac{2}{3} r \left(i^{3/2} - (i-1)^{3/2} \right) \ . \end{aligned} \tag{22}$$

Similar to the dual-region network model, the mean total cost of transmitting data from all the CHs in region i to the sink is calculated as:

$$\mathbb{E}[\mathcal{C}_i^{\text{inter}}] \approx b_{n_i} \mathcal{R} \mathbb{E}[\mathcal{N}_i] \mathbb{E}[\mathcal{D}_i] \ . \tag{23}$$

The mean total cost of data gathering from the whole network is the sum of the energy required for intra-cluster and inter-cluster data collection over all the regions. Thus, we obtain that:

$$\begin{aligned} \mathbb{E}[\mathcal{C}_{\text{total}}] &= \sum_{i=1}^{m} \mathbb{E}[\mathcal{C}_i^{\text{intra}}] + \mathbb{E}[\mathcal{C}_i^{\text{inter}}] \\ &= \rho \pi r^2 \mathcal{R} \sum_{i=1}^{m} p_i \left(b_1 \mathbb{E}[\mathcal{L}_i] + b_{n_i} \mathbb{E}[\mathcal{D}_i] \right) \ . \end{aligned} \tag{24}$$

Equation (24) suggests a closed-form relation for the mean total cost of data collection in the network with respect to the probabilities $p_i, i \in \{1, 2, \ldots, m\}$. The goal is to determine the set of optimal p_i's for which the total energy consumption is minimized. Formally stated,

$$\begin{aligned}(p_1^*, \cdots, p_m^*) = \ & \underset{\{p_i\}}{\operatorname{argmin}}\ \mathbb{E}[\mathcal{C}_{\text{total}}] \\ & \text{s.t. } 0 \leq p_i \leq 1,\ \forall i \in \{1, 2, \ldots, m\},\end{aligned} \tag{25}$$

where $(p_1^*, \ldots, p_m^*)$ is the optimal CH probability in regions 1 through m.

Equation (25) is a multi-variable optimization problem that can be solved via several well-known programming methods, such as hill-climbing, simulated annealing, and genetic algorithm. In Section VII, we use genetic algorithm to provide approximate solutions for this problem. Before discussing our experimental results, in the following section, we briefly describe a distributed clustering algorithm that uses the solutions of (25) to generate optimal clusters.

VI. Clustering Algorithm

Our proposed clustering algorithm has four steps as described below:

Step 1 The sink pre-computes the optimal probability allocation over the network regions and broadcasts a control packet containing the optimal probability of CH selection in each region throughout the network.

Step 2 Every node knows its location and hence can compute its distance to the sink. Having known this distance, nodes independently determine the region to which they belong using Equation (21). Therefore, on receipt of the control packet, the recipient node elects itself to become a CH as per the probability pre-specified in the packet.

Step 3 CH nodes advertise themselves in their local neighborhood to let other nodes know of their position.

Step 4 Non-CH nodes choose the geographically closest CH and join its cluster.

Depending on the node density, correlation degree, and network radius, the sink needs to first solve the optimization problem defined by Equation (25). As we shall see later in Section VII, the optimal number of regions depends on the density and radius of the network. However, splitting the network into 5 or 6 regions usually yields satisfactory results.

Once the optimal probabilities of CH selection over the regions are calculated, the sink broadcasts them all over the network and they will remain valid for the rest of the network

lifetime (given no significant change happening to the network settings). Thus, the clustering algorithm does not impose much control traffic on the network.

The following section presents some experimental results. First, the effect of data compression on optimal cluster sizing and also reducing the network energy consumption is studied for a uni-region network. Next, a multi-region network is considered and the optimal cluster sizes at different distances from the sink are quantified. Also, using simulation experiments, the optimal energy consumption for different values of m (number of regions) are reported.

VII. Simulation Experiments

In this section, we look at the optimization problem introduced in Section V, trying to find the best configuration for CH allocation over the network regions.

A. Simulation Environment

We use MATLAB for both our numerical and experimental analyses. The results reported for the model are the solutions of Equation (25) that are directly calculated using MATLAB's Genetic Algorithm and Direct Search Toolbox™. The simulation environment used in our experiments includes a network of radius 15 and density 0.75 nodes per unit area (roughly, a total of 530 nodes, on average). Each node elects itself as a CH according to the optimal probability corresponding to its region. Non-CH nodes then join the nearest CH. The size of the cluster data after compression is approximated by the joint entropy of the cluster, as defined by Equation (2). The distortion level is set to 0.01 bit per sample. We assume multi-hop communication along the shortest path between pairs of nodes. Node's radio range covers a radius of 0.75 unit and the per-hop transmission cost is fixed per every bit of information. We conduct our experiments for various degrees of data correlation (W) changing from 0.1 (low correlation) to 0.9 (high correlation) at a step of 0.1.

B. Impact of Data Compression on Network Energy Consumption

In this subsection, we demonstrate how careful consideration of data correlation/compression in forming optimal-sized clusters help reduce the total network energy consumption. At the moment, we tentatively ignore the effect of distance from the sink. Thus, we simply consider a network with a single region.

For any data correlation degree, nodes elect themselves as CH with the optimal probability, p^*, derived by the model. Throughout our simulation experiments, we generate 1000 random network configurations for each value of W and report the average energy-consumption to ensure the fairness of CH selection through all areas of the network. Fig. 6 depicts the optimal network energy consumption in simulation versus the results obtained by the model.

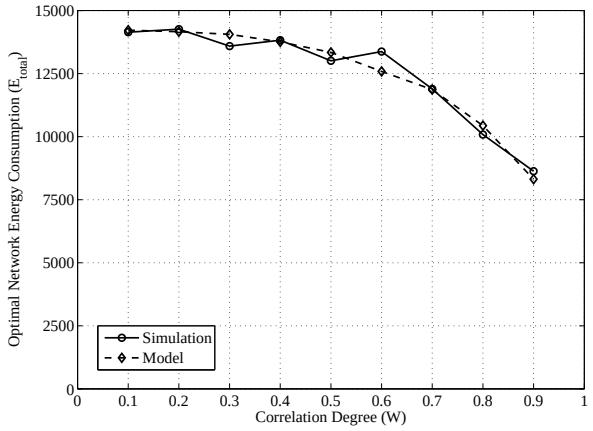


Fig. 6. The impact of data compression on network energy consumption in a uni-region network (Model Vs. Simulation).

As evident from Fig. 6, simulation results are fully consistent with the model. Increasing the correlation degree (W) throughout the field improves the network energy consumption such that a highly-correlated network is almost 42% more energy-efficient than a network with the same topology but low data correlation. This is expected as with more data correlation, more compression can be done at the CHs eliminating redundant data transmissions to the sink.

C. Impact of Data Correlation on Optimal Cluster Sizing

We consider two scenarios:

1) Clustering without Data Compression: quantization on local observations; data aggregation at the CHs without compression.
2) Clustering with Data Compression: quantization on local observations; joint cluster data compression at the CHs.

In the former scenario, CHs aggregate the cluster data and transmit it to the sink without compression, while in the latter, the CHs remove the redundancy present between data samples

and transmit a condensed version of the cluster data to the sink. Our goal is to investigate the effect of data correlation/compression on optimal cluster sizing as well as studying the energy savings when data correlation is present.

For simplicity, we still consider a uni-region network with homogeneous clusters. For both cases described above, namely, clustering with and without data compression, Fig. 7 illustrates a numerical analysis of the effect of data dependence on the optimal size of clusters.

As seen from Fig. 7, when no data compression is performed at CH level (aggregation only), irrespective of the intensity of data correlation present between data samples, the optimal cluster size is constant. This is reasonable in the sense that without data compression, no reduction in size of the cluster's aggregate data is attained. Therefore, nodes tend to individually quantize their observations and transmit their data to the sink over the shortest possible path. Thus, each node acts as a an isolated CH forming a cluster of size 1. ($p^* = 1.0$).

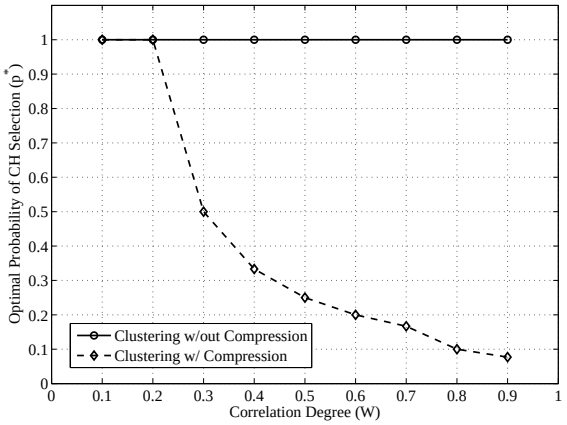


Fig. 7. The impact of data correlation on optimal cluster sizing in a uni-region network.

With clustering with data compression, however, increasing the correlation degree reduces the optimal probability of becoming CH in the network. In other words, the more the sensor observations are correlated, the larger the clusters become; since with more nodes joining the cluster, more reduction in the size of cluster's aggregate data is achieved at the CH.

As shown in [10], for any rate allocation, the shortest path tree (SPT) is the optimal routing structure for correlated data gathering. It is, however, interesting to note that forming clusters

require some nodes to send their readings through their pre-specified CH that is not necessarily part of the SPT rooted at the sink. Therefore, cluster formation is worthwhile only if the amount of compression ultimately achieved at the CH compensates for the extra energy spent due to the transmission of data over a suboptimal path. When the correlation degree is very low (*e.g.*, $W = 0.1$), no significant reduction in cluster data can be attained by forming clusters of multiple nodes. Rather, similar to clustering without compression, nodes tend to form isolated clusters of size 1 and individually transmit their data over the SPT. With a high correlation degree (*e.g.*, $W = 0.9$), however, more nodes tend to join each cluster to attain even a higher reduction in size of the cluster data after compression. The optimal size of clusters found in this numerical experiment are 10 for $W = 0.8$ and 13 for $W = 0.9$.

D. Better Energy Efficiency with Heterogeneous Clustering

In this subsection, we first quantify the amount of energy savings attained by using non-uniform clusters throughout the network. We also study the effect of distance on optimal cluster sizing by analyzing the solutions of a multi-region network.

For the network configuration described previously, Fig. 8 compares the optimal network energy consumption for various degrees of data correlation when different number of regions are used. As clearly seen, increasing the number of regions improves the network energy consumption. The upper curve corresponds to a uni-region network (uniform clustering), and the lower

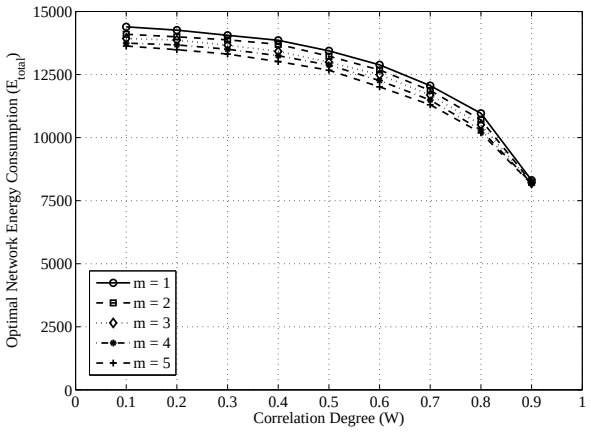


Fig. 8. The energy consumption comparison between uniform clustering ($m = 1$) vs. non-uniform clustering ($m > 1$).

lines correspond to more number of regions from 2 to 5, respectively (non-uniform clustering). The optimal solutions for such networks are derived using MATLAB's Genetic Algorithm and Direct Search Toolbox™. Although the difference between the energy consumptions is slight, the graph suggests that the uniform clustering cannot optimize the network energy consumption in presence of data correlation.

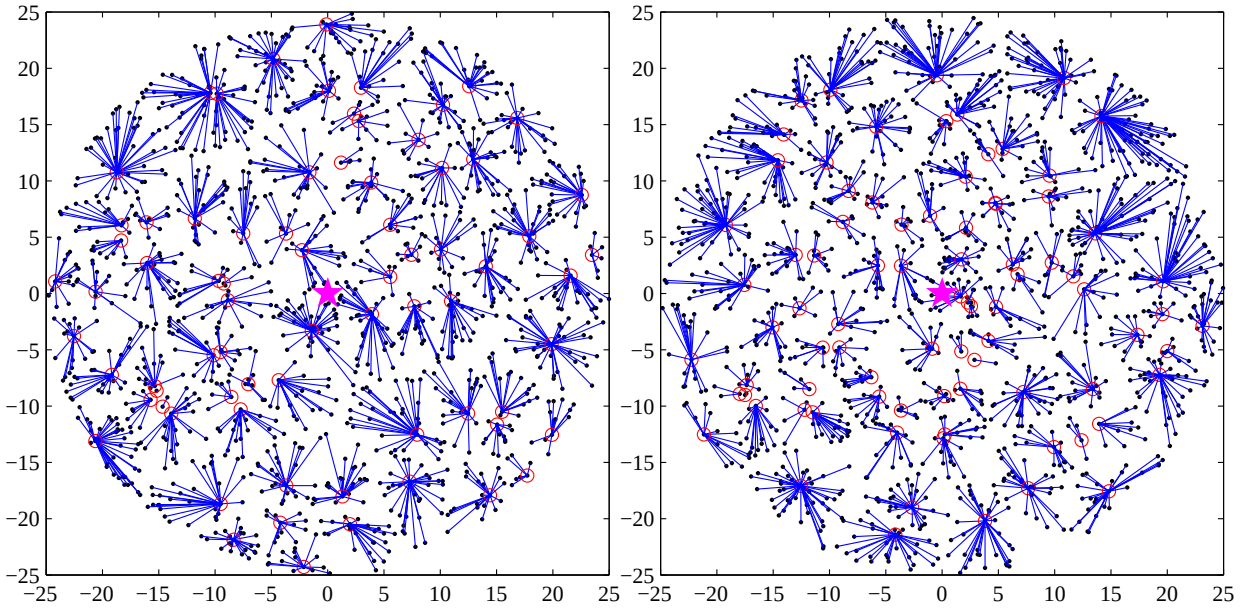
Table I presents the optimal probability allocation over various number of regions from 1. As evident from this table, the optimal probabilities of CH selection decrease from the innermost region all the way to the outermost one for all configurations. For a network with 5 regions, the optimal theoretical network energy consumption shows almost 1.5% improvement compared to a uni-region network. The simulation results demonstrate up to 6.8% enhancement under the same conditions. Although this is not a substantial improvement, it suggests uniform clustering is not optimal in terms of energy consumption for correlated data fields (contrary to the conventional clustering methods). For such networks, heterogeneous clusters with larger clusters at further distances from the sink yield better performance.

TABLE I

OPTIMAL PROBABILITIES OF CH SELECTION AND THEIR CORRESPONDING ENERGY CONSUMPTIONS (MODEL VS. SIMULATION)

m	E^*_{mod}	E^*_{sim}	$(p^*_1, \ldots, p^*_n)$
1	8292.96	8415.60	(0.0909)
2	8202.94	8318.63	$(0.1000, 0.0556)$
3	8196.12	8071.83	$(0.1001, 0.0714, 0.0556)$
4	8180.63	7890.37	$(0.1668, 0.0909, 0.0556, 0.0556)$
5	8169.44	7845.75	$(0.2002, 0.1001, 0.0715, 0.0556, 0.0556)$

Fig. 9 compares two optimal realizations of uniform clustering (uni-region network) against non-uniform clustering (multi-region network) on an arbitrary network. As presented by Fig. 9, with non-uniform clustering, optimal size of clusters grow as you move further from the sink. This is completely consistent with our previous analysis of a single cluster network [9].



(a) Uniform clustering ($m = 1$) (b) Non-uniform clustering ($m = 3$)

Fig. 9. Optimal probability of CH selection vs. location of the border.

VIII. Conclusion

In this paper, we proposed a general non-uniform cluster-based data gathering strategy that optimizes the total network energy consumption. Our model carefully takes into account the joint effect of data correlation and distance on forming optimal-sized clusters. Unlike the existing clustering approaches that produce uniform clusters throughout the entire network, we showed that heterogeneous-sized clusters are more energy-efficient in WSNs with spatial data correlation. We demonstrated this through both numerical analysis and simulation experiments.

Our network model can be generalized further by relocating the sink position and generating asymmetric network topologies. Also, since the distributed clustering algorithm proposed is randomized, cluster sizes may not be evenly balanced throughout the regions. Directions for the future research include delegating the CH role to the most appropriate cluster member as well as exchanging nodes between the clusters to form equal-sized clusters over the regions. Also, our current analysis is based on a fixed number of regions in the network. Choosing the optimal number of regions dynamically based on the inherent attributes of the network is another interesting area for future study.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on Sensor Networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, November 2002.

[2] A. Abbasi and M. Younis, "A Survey on Clustering Algorithms for Wireless Sensor Networks," *Computer Communications*, vol. 30, no. 14-15, pp. 2826–2841, 2007.

[3] P. Wang, R. Dui, and I. F. Akyildiz, "Collaborative Data Compression Using Clustered Source Coding for Wireless Multimedia Sensor Networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, March 2010, pp. 1713–1723.

[4] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," in *Proc. Hawaii International Conference on System Sciences (HICSS)*, vol. 8, 2000, p. 8020.

[5] S. Bandyopadhyay and E. J. Coyle, "An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, April 2003, pp. 1713–1723.

[6] N. Vlajic and D. Xia, "Wireless Sensor Networks: To Cluster or Not To Cluster?" in *Proc. International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, June 2006.

[7] H. Chen and S. Megerian, "Cluster Sizing and Head Selection for Efficient Data Aggregation and Routing in Sensor Networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 4, April 2006, pp. 2318 –2323.

[8] J. Li and G. AlRegib, "Energy-Efficient Cluster-Based Distributed Estimation in Wireless Sensor Networks," in *Proc. IEEE Military Communications Conference (MILCOM)*, October 2006, pp. 1 –7.

[9] A. Dabirmoghaddam, M. Ghaderi, and C. Williamson, "Cluster-Based Correlated Data Gathering in Wireless Sensor Networks," in *Proc. IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, August 2010.

[10] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On Network Correlated Data Gathering," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, March 2004, pp. 2571–2582.

[11] I. Koutsopoulos and M. Halkidi, "Measurement Aggregation and Routing Techniques for Energy-Efcient Estimation in Wireless Sensor Networks," in *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, June 2010.

[12] S. Pattem, B. Krishnamachari, and R. Govindan, "The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks," in *Proc. International Symposium on Information Processing in Sensor Networks (IPSN)*. New York, NY, USA: ACM, 2004, pp. 28–35.

[13] A. Scaglione, "Routing and Data Compression in Sensor Networks: Stochastic Models for Sensor Data that Guarantee Scalability," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 29 June - 4 July 2003, p. 174.

[14] M. C. Vuran and I. F. Akyildiz, "Spatial Correlation-Based Collaborative Medium Access Control in Wireless Sensor Networks," *IEEE/ACM Transactions on Networking*, vol. 14, no. 2, pp. 316–329, 2006.

[15] J. O. Berger, V. de Oliveira, and B. Sanso, "Objective Bayesian Analysis of Spatially Correlated Data," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1361–1374, 2001.

[16] A. Scaglione and S. Servetto, "On the Interdependence of Routing and Data Compression in Multi-hop Sensor Networks," *Wireless Networks*, vol. 11, no. 1-2, pp. 149–160, 2005.

[17] S. G. Foss and Z. S. A., "On a Voronoi Aggregative Process Related to a Bivariate Poisson Process," *Advances in Applied Probability*, vol. 28, no. 4, pp. 965–981, December 1996.

[18] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, “An Application-Specific Protocol Architecture for Wireless Microsensor Networks,” *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, 2002.