Cluster-Based Correlated Data Gathering in Wireless Sensor Networks

Ali Dabirmoghaddam Majid Ghaderi Carey Williamson
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada T2N 1N4
{adabirmo, mghaderi, carey}@cpsc.ucalgary.ca

***Abstract*—We consider the problem of optimal cluster-based data gathering in Wireless Sensor Networks (WSNs) when nearby readings are spatially correlated. Due to the dense nature of WSNs, data samples taken from nearby locations are statistically similar. We show how this data correlation can be exploited to reduce the amount of data to be transmitted in the network and thus conserve energy. While much attention in recent years has been paid to analyzing and optimizing cluster-based WSNs from various perspectives, the problem of energy-efficient clustering of WSNs in presence of data correlation is not yet fully explored. In this paper, we model a single-cluster network and analytically characterize the optimal cluster size subject to its distance from the sink as well as the degree of correlation. Contrary to existing approaches, our findings show that heterogeneous-sized clusters, where the clusters further from the sink are larger, are more energy-efficient. We also propose a heuristic greedy clustering algorithm to find a near-optimal solution to the problem of energy-efficient clustering. Simulation results confirm the effectiveness of having heterogeneous-sized clusters in WSNs.**

***Index Terms*—energy-efficiency, clustering, data correlation, data compression, wireless sensor networks.**

I. Introduction

A *Wireless Sensor Network* (WSN) consists of a large number of sensor nodes that cooperate to monitor environmental conditions (e.g., temperature, precipitation, radioactivity) in a given geographic area [1]. The sensor nodes have the capability to collect (sense) data from the environment, and cooperate with other sensor nodes to relay the data to a central processing center, known as the sink, using multi-hop wireless communication. WSNs are used in a wide range of agricultural, environmental, industrial, manufacturing, military, and security monitoring applications.

The minute size of sensor nodes (typically the size of a small coin) means that they operate using limited-capacity batteries. One example is *Berkeley's Smart Dust* with a volume of no more than a few cubic millimeters that can store on the order of 1 Joule of energy [2]. Nevertheless, critical WSN applications require long-term operation in remote, unattended, and even hostile environments in which it is often difficult or impossible to replace the sensor batteries. In the past few years, a considerable volume of research has studied various methods of energy conservation in WSNs.

Since data transmission can account for up to 70% of the power consumed in typical sensor nodes [3], substantial energy savings are possible if the volume of communicated data is reduced using compression. While data compression itself requires additional processing, the amount of energy required for CPU operations in sensor nodes is orders of magnitude lower than that for data transmission [4].

In dense deployments of sensor nodes in a WSN, it is expected that the readings from nearby nodes are strongly correlated [5]. For example, in a camera sensor network for monitoring the environment, it is likely that multiple proximally-located camera sensors detect the same phenomenon. In this case, it usually suffices to send one instance of the observation to the sink. This property can be exploited to aggregate and compress the collected data before sending it to the sink [6].

Clustering is a well-established technique for reducing data collection costs in WSNs [7]. In this technique, sensor nodes are grouped into disjoint sets, with each set managed by a designated *Cluster-Head* (CH), selected from among the sensor nodes. The cluster members send their collected observations (which are likely to be highly correlated) to their CH. The CH suppresses the local redundancies and communicates the compressed data to the sink possibly via multi-hop transmission. This approach avoids sending redundant data to the sink, and thus helps save energy. Organizing sensor nodes to form such cluster-based topologies is a widely accepted solution for energy conservation. In addition to the opportunity for local data compression, this approach can coordinate the activities of cluster members, and address scalability issues (e.g., routing and communication costs) in large WSNs. Hence, this class of WSNs is potentially viewed as the most energy-efficient and long-lived class of sensor networks [5].

Numerous clustering algorithms for WSNs have been proposed in the literature. These algorithms vary in their objectives, which may include load balancing, fault-tolerance, increased connectivity, reduced delay, and maximal network longevity. A good survey appears in [7].

While the existing energy-aware clustering algorithms ignore the effect of data correlation on the optimal cluster sizing and its impact on saving energy, in this paper, we propose and evaluate a novel WSN clustering strategy that exploits data correlation. That is, the nodes within each cluster have strong internal correlation, while the inter-cluster data dependence is negligible. To this end, we carefully analyze the mutual effect of cluster size and the distance from the sink on reducing total network energy consumption. Although it is computationally difficult to find optimal-sized clusters, we propose a model to obtain a near-optimal solution for forming energy-efficient clusters in the network. In a nutshell, the main contributions of this paper are as follows:

- We develop a model to incorporate the effect of spatial data correlation while forming energy-efficient clusters in the network.
- Unlike conventional clustering algorithms that result in uniform clusters of almost the same size, we advocate heterogeneous-sized clusters in the network, where the clusters further from the sink are larger than those located close to the sink.
- We justify the proposed clustering strategy using analysis and simulation.

The remainder of the paper is organized as follows. Section II reviews recent literature on WSN optimization. While Section III presents a simple motivating example for our work, an analytical model for spatial data dependency is reviewed in Section IV. Subsequently, we briefly discuss and compare two existing methods for modeling lossless/lossy data compression. In Section V, we propose a simple one-cluster network model and examine the joint effect of data correlation, distance from the sink, and network density on optimal cluster size. In Section VI, we propose a heuristic greedy clustering algorithm that confirms our findings about heterogeneous cluster sizes. Finally, Section VII concludes the paper and suggests some interesting areas for future work.

II. Related Work

Cristescu, *et al.* analyze the effect of applying a well-known method of distributed source coding, Slepian-Wolf Coding (SWC) [8], for data compression in WSNs and proved that the shortest path tree yields the optimal gathering tree for any fixed rate allocation [9]. SWC requires side information about both the entire network topology and the exact data correlation model to allocate the optimal set of data rates to sensor nodes. This information, however, is hard to achieve in practice. Moreover, applying optimal SWC in WSN results in an imbalanced rate allocation, with the highest load imposed on nodes near sink. Energy depletion for the nodes near the sink can disrupt the operation of the entire network. Various approximation algorithms are proposed and evaluated in [9] for the optimal SWC. The authors, however, argue that the optimal rate allocation problem is an *NP*-complete problem.

LEACH [10] is a classic probabilistic clustering algorithm. It aims to distribute the traffic load evenly among sensor nodes and reduce the network energy consumption. The LEACH algorithm proceeds in rounds. In each round, each sensor node independently decides whether or not to become a CH according to a probability function. On average, this function makes each node become a CH for a similar period of time, assuring fair balancing of energy consumption among all nodes. Although LEACH performs local data fusion to compress the cluster information, it does not consider data correlation when forming optimal-sized clusters. Moreover, since the probability of becoming a CH is fixed, LEACH results in clusters that are on average of the same size throughout the entire network.

Pattem *et al.* consider the impact of spatial data correlation on three different class of routing schemes: Distributed Source Coding (DSC), Routing Driven Compression (RDC), and Compression Driven Routing (CDR) [11]. They show that RDC is best suited for WSNs with low correlation, while CDR is most appropriate for highly-correlated data fields. Although the authors consider data correlation as an important factor in reducing network energy consumption, their analysis is fundamentally based on a static cluster size throughout the entire network. A near-optimal value for this cluster size is then suggested that works equally well across a wide range of correlation degrees. In contrast, our findings show that the optimal cluster sizes vary with respect to the cluster distance to the sink, and the degree of correlation. Therefore, there is no identical and globally optimal size for the whole clusters that minimize the entire network energy consumption.

Cluster-based sensor networks generally outperform non-clustered WSNs [5]. The authors show that clusters should comprise nodes with highly correlated data-readings. They evaluated their model for a simple linear distributions of nodes, and formulated the optimal size of the clusters as a function of distance to the sink, and the number of nodes with similar data-readings. For simplicity of calculations, they assumed a globally-fixed aggregation factor for all clusters. As we will see later in this paper, data correlation is typically assumed to be a decreasing function of distance. Hence the level of data aggregation/compression depends on the distribution of nodes inside the network and thereby cannot be the same for all nodes.

A distributed, randomized WSN clustering algorithm with a hierarchy of clusters proposed in [12]. In that paper, the optimal probability of becoming a CH is computed. Also, the maximum number of hops allowed for each non-CH node to reach the designated CH is quantified. They develop a distributed hierarchical clustering algorithm based on these pre-computed optimal values and show that increasing the number of clustering levels in the hierarchy results in better energy savings. While the paper provides an energy-efficient method of network clustering, it does not consider the correlation between nearby reported observations while collecting information.

Li and AlRegib consider the problem of energy-efficient cluster-based distributed estimation in WSNs, where the major goal is to determine the optimal cluster size and the number of clusters to minimize the total energy cost of the network [13]. In their approach, sensor nodes send quantized versions of their observations to their respective CHs. CHs make a local estimation of the cluster data using a lossy estimator named BLUE, and directly send the estimate to the sink. The sink uses another estimator called quasi-BLUE to make the final estimation based on all the received signals. Although their proposed clustering algorithm shows significant energy savings, the effect of data correlation is again neglected.

III. Motivating Example: A Linear Network

Given a network of nodes, there are many ways to group the nodes together to form independent clusters. The general problem of minimum cost clustering is a more complicated instance of the *minimum cost network correlated data gathering* problem, which is known to be *NP*-complete [9].

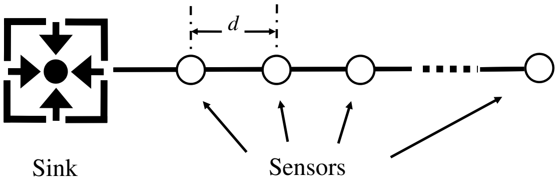


Fig. 1. A linear sensor network topology

We focus on a simple scenario of a small linear network for which we are able to exhaustively compute all possible permutations of the clusters and their energy consumption (see Fig. 1). To this end, we deterministically place n sensor nodes at equal distances apart along a horizontal line segment with the sink at the left end. Clearly, even for small networks, the number of cluster permutations rapidly grows as n increases; thus, here we just consider a network of size $n = 10$. Let $(a_1, a_2, \ldots a_m)$ represent a cluster configuration in which there are m clusters and a_i denotes the size of the i^{th} cluster, for $i \in 1, \ldots, m$. For example, tuple $(3, 1, 2, 4)$ represents a network with 4 clusters where the cluster closest to the sink has 3 nodes.

We assume that each sensor node within the cluster sends only a single instance of its readings to the CH. Cluster-heads are located at the left edge of each cluster. This assumption ensures that the collected information is transmitted towards the sink over the shortest possible path. The CH aggregates and compresses the collected readings and transmits a single representative message to the sink. The size of the compressed message depends on the joint entropy of the cluster, as will be discussed later in Section IV.

Data transmissions between the cluster members and the CHs, and also from the CHs to the sink are using multi-hop communication basis. In this approach, we assume that the radio range of sensors only covers their immediate neighbor(s). This ensures the minimum transmission energy while guaranteeing the network connectivity. Therefore, in order to communicate the data to the CH, a sensor needs to make use of the intermediate sensor nodes to reach the CH. Likewise, the CH relies upon intermediate sensors to forward the cluster information to the sink.

Based on the above assumptions, we calculate the total network energy consumption of each clustering pattern when three different degrees of data correlation (low, medium and high) are present. To configure the degree of data correlation, we assume the dependency between sample readings exponentially decreases with distance at a fixed rate. The higher is this rate, the lower is the degree of correlation.

We exhaustively enumerate and evaluate all possible cluster configurations, starting with a single cluster of size 10, and then considering two-cluster configurations of sizes $(1, 9)$, $(2, 8)$, $(3, 7)$, $\ldots (9, 1)$, and then three-cluster configurations, and so on, finishing with 10 individual clusters of size 1. Table I summarizes the top 10 cluster patterns with the lowest energy requirements.

TABLE I
Top 10 cluster patterns with lowest energy consumption.

	Low Correlation		Medium Correlation		High Correlation	
rank	pattern	energy	pattern	energy	pattern	energy
1	(2,3,5)	193.94	(2,3,5)	179.83	(1,2,2,5)	102.48
2	(2,4,4)	193.95	(1,4,5)	179.96	(2,2,2,4)	102.48
3	(3,3,4)	193.96	(1,3,6)	180.15	(2,3,5)	103.25
4	(1,2,3,4)	193.97	(1,1,3,5)	180.24	(3,2,5)	103.93
5	(1,4,5)	194.01	(2,4,4)	180.26	(2,4,4)	104.55
6	(1,1,3,5)	194.02	(2,2,6)	180.45	(1,2,3,4)	104.70
7	(1,1,4,4)	194.03	(4,6)	180.50	(2,2,6)	105.07
8	(3,2,5)	194.04	(3,2,5)	180.54	(1,3,2,4)	105.38
9	(1,2,2,5)	194.05	(3,3,4)	180.60	(2,2,4,2)	105.88
10	(2,1,3,4)	194.05	(1,1,4,4)	180.67	(2,1,2,5)	106.06

The results in Table I indicate that data compression plays a significant role in reducing the total network energy consumption. For example, the results for the high correlation scenario show that energy consumption is reduced by almost 50% compared to low correlation case. Furthermore, the importance of appropriate clustering is also evident in the high correlation scenario. For instance, the tenth-best clustering, $(2, 1, 2, 5)$, has 3.5% higher energy consumption than the best cluster configuration, $(1, 2, 2, 5)$. and the very worst cluster configuration studied, $(1, 1, \ldots 1)$, is over 50% worse (not shown). For the low correlation scenario, the top 10 cluster configurations all have comparable energy consumption (within 0.06% of each other).

Studying the cluster sizes in Table I is also insightful. The most prevalent configurations have 3 or 4 clusters (only one 2-cluster configuration appears in Table I). More importantly, most of these cluster patterns have monotonically increasing cluster size as you move away from the sink. There are a few exceptions in each column (e.g., the $(3, 2, 5)$ cluster configuration), but even in these cases, the size of the rightmost cluster is always larger than the size of the leftmost cluster. This motivates exploring new methods of clustering that, unlike existing algorithms, produce heterogeneous-sized clusters.

In the following section, we review the mathematical background of the problem and formally investigate the importance of data compression in energy-constrained WSNs.

IV. Formal Problem Definition

The individual sensor nodes within the WSN are considered statistically identical information sources. We assume that sensor readings are normally distributed with mean zero and variance σ^2. Thus, the set of all observations (in a cluster of size N) can be formalized as a zero-mean *multi-variate Gaussian distribution* [1].

Since the properties of Gaussian sources are already well-explored in the literature, this assumption makes our calculations easier. Furthermore, in terms of the number of bits required to represent the field, Gaussian distribution is the worst case [14]. Thus, our results can be applied as a bound for other sources as well.

Gaussian fields can be represented with a symmetric positive-definite *covariance matrix* $\Sigma = [\sigma_{ij}]_{N \times N}$. Each element, σ_{ij}, expresses the data dependence between readings from sensor nodes i and j.

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

A. Data Correlation Model

In many distributed information systems, it is often assumed that sample observations are spatially correlated. Such correlation is generally formalized by a *covariance function*, which is a non-negative decreasing function of Euclidean distance. The limiting values are 1 at $d = 0$ and 0 at $d = \infty$, where d represents the Euclidean distance between two sample readings. In other words, as the Euclidean distance between two sample observations increases, the correlation between them monotonically approaches zero.

A general model for spatial correlation is proposed in [15]. Denoting the random field of interest by $\{S(u), u \in \mathcal{D} \subseteq \mathbb{R}^l\}$, the covariance between two sample observations at locations u and v is expressed by:

$$\mathrm{cov}\,\{S(u), S(v)\} = \sigma^2 K_\vartheta(\|u - v\|) \ , \tag{1}$$

where σ^2 is the variance of each sample observation, $\|\cdot\|$ denotes the Euclidean distance, and $K_\vartheta(\|u - v\|) = \mathrm{corr}\,\{S(u), S(v)\}$ denotes an isotropic correlation function with $\vartheta = (\theta_1, \cdots, \theta_c)' \in \Theta \subset \mathbb{R}^c$ as the set of parameters controlling the range of correlation and smoothness/roughness of the random field [15].

Depending on the inherent characteristics of the random field, several types of covariance models can be defined. The most common kinds are *Spherical*, *Power Exponential*, *Rational Quadratic*, and *Matérn* [15].

In this paper, we use the Power Exponential model for which the correlation function over a distance d is defined as:

$$K_\vartheta^{PE}(d) = \exp(-(d/\theta_1)^{\theta_2}) \ , \tag{2}$$

where $\theta_1 > 0$ and $\theta_2 \in (0, 2]$.

More specifically, in our analyses, we use a special type of the Power Exponential, known as Squared Exponential correlation model. Adopting the Squared Exponential correlation model, we express the elements of the covariance matrix as

$$\sigma_{ij} = \sigma^2 \exp(-a d_{ij}^2) \ , \tag{3}$$

where $a = \theta_1^{-2}$ is the correlation exponent and d_{ij} denotes the Euclidean distance between sensor nodes i and j.

For brevity, we define the parameter $W = \exp(-a)$. W is a normalized parameter (*i.e.*, $0 < W < 1$) representing the degree of correlation, where $W = 0$ represents no correlation, and $W = 1$ represents high correlation (*i.e.*, globally identical observations).

B. Data Compression Model

Cluster members observe some spatial stochastic process at specific points in time. Let $S = \{s_i, i = 1, 2, \ldots, N\}$ be the set of N sample observations captured by N cluster members. Recall that we assume a continuous data model, where 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. Denoting the reconstructed version of S by $\hat{S}$, we consider the *mean-squared error* (MSE) between two observations as the measure of distortion. Since the maximum tolerable distortion at the receiver is D, it requires that:

$$E[\|S - \hat{S}\|^2] \leq D \ . \tag{4}$$

There are two approaches for computing the number of bits required to represent the quantized observation $\hat{S}$ subject to the given distortion D. Below, we briefly summarize the two approaches and describe their relation to each other.

1) Rate Distortion Theory: In Rate Distortion Theory, the minimum required number of bits to represent a multi-variate Gaussian source $G(0, \Sigma_{N \times N})$ subject to a distortion bound D per source is given by:

$$R(N, D) = \frac{1}{2} \sum_{n=1}^{N} \log_2(\frac{\lambda_n}{D_n}) \ , \tag{5}$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N$ are the eigenvalues of the covariance matrix Σ, and D_n's are expressed as follows:

$$\begin{aligned} D_n &= \begin{cases} \theta, & \text{if } \theta < \lambda_n \\ \lambda_n, & \text{otherwise,} \end{cases} \\ \sum_{n=1}^{N} D_n &= N \cdot D \end{aligned} \tag{6}$$

For sufficiently small values of D where $\sum_{n=1}^{N} \lambda_n \geq D$, there exists a $n_0 \leq N$ such that $\lambda_{n_0} > \theta \geq \lambda_{n_0+1}$. Consequently, θ can be computed from the following relation:

$$\theta = \frac{1}{n_0} \left(N \cdot D - \sum_{n=n_0+1}^{N} \lambda_n \right) \ . \tag{7}$$

2) Source Coding and Entropy: 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.

[1] Hereafter, we use the terms “multi-variate Gaussian distribution”, “Gaussian random field” and “Gaussian field” interchangeably.

If a continuous source is discretized by a uniform quantizer of step size Δ, the entropy of the quantized source denoted by $H(S^{\Delta})$ is given by [16]:

$$H(S^{\Delta}) \approx h(S) - \log_2 \Delta^N \ , \tag{8}$$

where $h(S)$ is the differential entropy of S given by

$$h(S) = -\int_S p(S) \log_2 p(S) dS \ . \tag{9}$$

For a continuous multi-variate Gaussian source, the entropy function is given by [16]:

$$h(S) = \frac{1}{2} \log_2 (2\pi e)^N |\Sigma| \ , \tag{10}$$

where $|\Sigma|$ denotes the determinant of the full-rank covariance matrix Σ (*i.e.*, the product of its eigenvalues, λ_n).

With our special correlation model, the covariance matrix Σ becomes singular, as $N \to \infty$. Under such conditions, it has been shown that the differential entropy, $h(S)$, is given by [17]:

$$h(S) = \frac{1}{2} \log_2 (2\pi e)^{\varrho(\Sigma)} |\Sigma|^+ \ , \tag{11}$$

where $|\Sigma|^+$ and $\varrho(\Sigma)$ denote the product of non-zero eigenvalues (*i.e., principal components*) and the rank of Σ, respectively. Therefore:

$$H(S^{\Delta}) \approx h(S) - \log_2 \Delta^{\varrho(\Sigma)} \ . \tag{12}$$

For a uniform quantizer with step size Δ, the per source distortion is given by $\Delta^2/12$ [18]. Hence, to achieve the distortion D, we require that:

$$\Delta^2 = 12D \ , \tag{13}$$

and, consequently

$$H(S^D) \approx \frac{1}{2} \log_2 (\frac{\pi e}{6D})^{\varrho(\Sigma)} |\Sigma|^+ \ . \tag{14}$$

In the following subsection, we study the effect of data compression on suppressing the local data redundancies and also the relation between the two aforementioned methods of modeling data compression.

C. On the Importance of Data Compression

To examine the importance of data compression in WSNs with correlated data, we model the savings that can be achieved by compressing the cluster data.

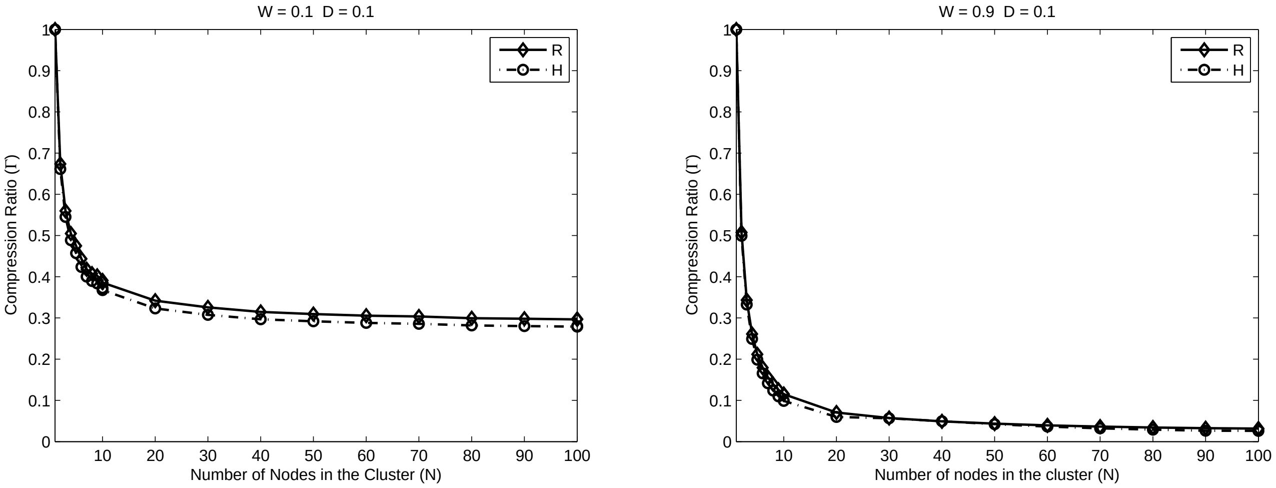
Consider a cluster with N nodes that are uniformly and randomly distributed spatially. For different degrees of correlation (W), we consider the metric *Compression Ratio*, $\mathit{\Gamma}$, that measures the amount of reduction that can be achieved by compressing the cluster information compared to the aggregate size of the cluster data without compression as follows:

$$\begin{aligned} \mathit{\Gamma}_H &= \frac{H(S_1^D, S_2^D, \ldots S_N^D)}{N \cdot H(S_*^D)} \ , \\ \mathit{\Gamma}_R &= \frac{R(N,D)}{N \cdot R(1,D)} \ . \end{aligned} \tag{15}$$

where $H(S_*^D)$ and $H(S_1^D, S_2^D, \ldots S_N^D)$ are the entropy of a single source and the joint entropy of the whole cluster subject to distortion D, respectively. Also, $R(1,D)$ and $R(N,D)$ respectively denote the rates at which a single source and the cluster entirely are coded with distortion D.

Fig. 2 illustrates the effect of data correlation on the Compression Ratio.

The first interesting observation is the similarity between the rate distortion and entropy curves. As seen from Fig. 2, both theories indicate almost the same reduction for a fixed level of distortion and data correlation. We note that rate distortion provides a theoretical lower bound on the number of bits required to represent continuous samples [17]. The entropy technique, on the other hand, is sub-optimal, yet more practical and easier to be implemented. Hence, in this paper, we use



(a) $W = 0.1$ (low correlation)

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

Fig. 2. Compression Ratio ($\mathit{\Gamma}$) (Rate Distortion (R) vs. Entropy Coding (H))

the entropy technique to model the size of the cluster after compression.

As seen in previous subsection, our analysis of the entropy-based compression is asymptotic in the sense that the effect of minor components of the covariance matrix are neglected. This is the reason why in Fig. 2, the entropy curve is slightly below the rate distortion curve. However, it is important to note that the difference between the two presented curves is the maximum possible error, as Gaussian fields represent the worst case scenario [17].

As evident from Fig. 2, the higher the degree of data correlation is, the greater are the savings from compressing the cluster data. Surprisingly, even with a very low degree of data correlation ($W = 0.1$), the cluster data can be compressed to one-third of its original size. In the highly correlated case ($W = 0.9$), the reduction is even greater, resulting in a huge saving in power consumption of the network.

As the number of nodes in the cluster increases, the Compression Ratio levels off. This observation can be explained as follows. When the number of nodes in the cluster is relatively small, the data dependency between sample observations is high, since cluster members are all geographically close to each other. In this case, adding more nodes to the cluster (with the same density throughout the cluster) dramatically improves the savings, since the observations from the new nodes are likely similar to those from other nodes in the cluster. As the cluster radius expands, and more nodes join the cluster, the similarity between observations from disparate nodes throughout the cluster reduces. In other words, readings from new members at the periphery of the cluster are only weakly correlated with the readings from most of the nodes in the cluster, and have negligible effect on the Compression Ratio.

V. A Simple Single Cluster Model

In this section, we develop a model to examine the effect of cluster size and distance from the sink on energy consumption. Consider a circular cluster of radius R at distance L from the sink in which nodes are uniformly distributed with average density ρ (see Fig. 3). Thus, the expected number of sensors in the cluster is $N = \rho\pi R^2$. For simplicity of calculations, we assume that the CH is located at the center of the cluster. This assumption is consistent with many clustering schemes (*e.g.*, LEACH [10] and EEHC [12]).

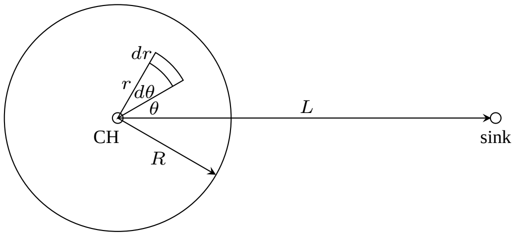


Fig. 3. A circular cluster of radius R at distance L from the sink.

Cluster members observe some spatial stochastic process, quantize their observations, and transmit them to the sink, either directly (single-hop) or via other sensor nodes (multi-hop). For successful data transmission, a minimum received power level γ is required. We assume a large-scale fading channel between each transmitter and receiver, in which the received power is inversely proportional to the square of the distance between the transmitter and the receiver. Therefore, the energy (E) required to transmit b bits over distance d is given by [19]:

$$E = \gamma b d^2 \ . \tag{16}$$

For simplicity and without loss of generality, hereafter we assume that $\gamma = 1$.

In our analysis, we consider two communication schemes (direct and cluster-based) and compute the total transmission power required to report all observations to the sink.

A. Direct Communication with Local Compression

In this scheme, each sensor in the disk individually quantizes and compresses its own observation and transmits the message directly to the sink. Clearly, this naïve form of communication is suboptimal; nonetheless, we use this scheme as a baseline to characterize the performance of the cluster-based method.

Let b_1 denote the minimum number of bits required to encode each observation. For a sensor node at polar coordinate (r, θ), as shown in Fig. 3, the transmission cost for sending a message to the sink is $P(r, \theta) = b_1 d^2(r, \theta)$, where $d(r, \theta)$ is the Euclidean distance from (r, θ) to the sink. The total transmission power P_d consumed by all sensors is obtained by integrating over the disk as follows:

$$\begin{aligned} P_d &= \int_0^R \int_0^{2\pi} b_1 \left(r^2 + L^2 - 2rLcos\theta\right) \rho r d\theta dr \\ &= 2\pi\rho b_1 \left(\frac{L^2R^2}{2} + \frac{R^4}{4}\right), \end{aligned} \tag{17}$$

where the factor $\rho r d\theta dr$ represents the expected number of sensor nodes in the differential “rectangle” shown in Fig. 3.

B. Cluster-Based Communication with Joint Compression

In cluster-based communication, each sensor sends its quantized observation to the CH. The CH losslessly compresses all N messages and transmits the compressed version to the sink. Denoting the compressed message size by b_N, the total transmission cost is given by

$$\begin{aligned} P_c &= \int_0^R \int_0^{2\pi} \left(b_1 r^2\right) \rho r d\theta dr + b_N L^2 \\ &= 2\pi\rho b_1 \left(\frac{R^4}{4}\right) + b_N L^2. \end{aligned} \tag{18}$$

Equation (14) can be used to compute the minimum number of bits required to represent the size of an individual sensor reading (b_1), and the entire cluster data after compression (b_N) for a pre-specified target distortion D.

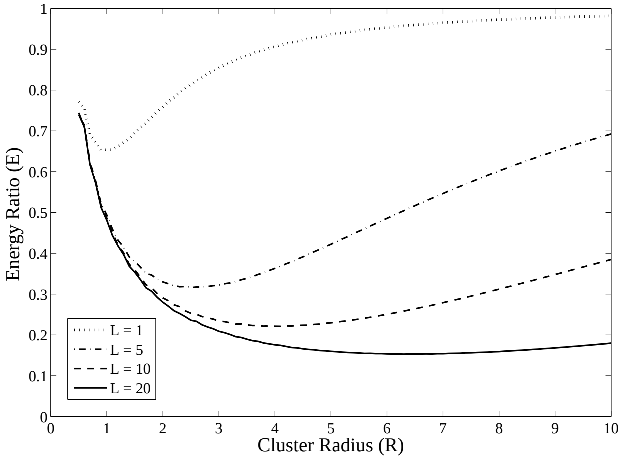


Fig. 4. The joint effect of the cluster size (N) and the distance from the sink (L) on the Energy Ratio (E) for $\rho = 6.25$ and $W = 0.75$.

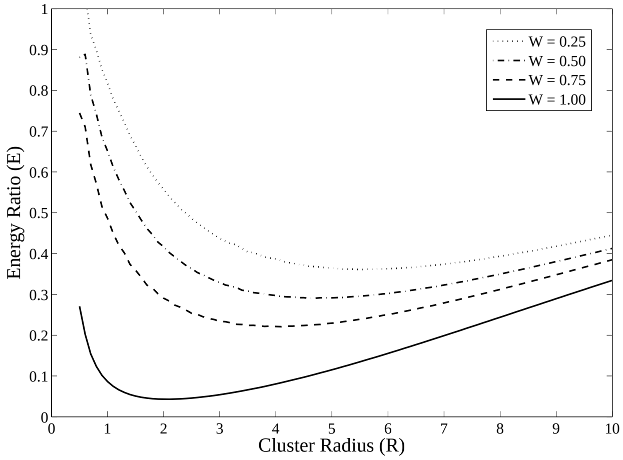


Fig. 5. The joint effect of the cluster size (N) and the correlation degree (W) on the Energy Ratio (E) for $\rho = 6.25$ and $L = 10$.

C. Optimal Cluster Size

In this subsection, we evaluate the impact of the cluster size on the total energy consumption. To this end, we define the performance metric *"Energy Ratio"* as the ratio of energy required to collect the network data in cluster-based method to the direct communication scheme.

$$E = \frac{P_c}{P_d} \ . \tag{19}$$

To compute our numerical examples of Energy Ratio, we place an arbitrary CH at distance L from the sink and dynamically expand the cluster radius R for a given node density ρ. For different distances L, we run our simulations for 1000 random cluster configurations and report the average Energy Ratio.

1) Impact of Distance on Optimal Cluster Size: Fig. 4 shows the relationship between energy consumption (vertical axis, lower is better) and cluster radius (horizontal axis) for different distances L (the four lines on the graph). The results in Fig. 4 suggest that, for any given distance L, there is a different optimal cluster radius, at which the Energy Ratio is minimized. For example, the upper line for $L = 1$ has its minimum energy consumption near a radius of $R = 1$, while the lower line for $L = 20$ has its minimum near $R = 7$.

In general, the farther the cluster is from the sink, the larger the optimal radius is. In other words, the optimal cluster size is not uniform throughout the network. This insight is important, since many existing energy-aware clustering schemes, such as LEACH [10] and EEHC [12], assume homogeneous-sized clusters that are uniformly distributed in the WSN.

2) Impact of Correlation Degree on Optimal Cluster Size: We next focus on the effect of data correlation degree on the optimal size of the clusters. Fig. 5 shows that as more degree of correlation increases, the optimal cluster radius shrinks. For higher degrees of correlation, nearby observations are tightly coupled with each other. In the extreme case where $W = 1.00$, all the observations are globally identical. Therefore, it would suffice if each sensor node reports its data to its immediate neighbor. Since both observations are identical, the neighboring node needs only forward one instance of the observation. Thus, the optimal cluster radius will decrease such that each node forms its own local neighborhood.

3) Impact of Network Density on Optimal Cluster Size: The next interesting analysis is the effect of network density on the optimal size of the clusters, for a given level of correlation. In this analysis, we fix the position of the cluster at distance L from the sink and study how the Energy Ratio changes as the network becomes denser. According to Fig. 6, the optimal cluster radius shrinks as the network density increases. This result is intuitive in the sense that adding more nodes to the network for a fixed degree of data correlation is expected to have the same effect as increasing the data correlation degree for a fixed density.

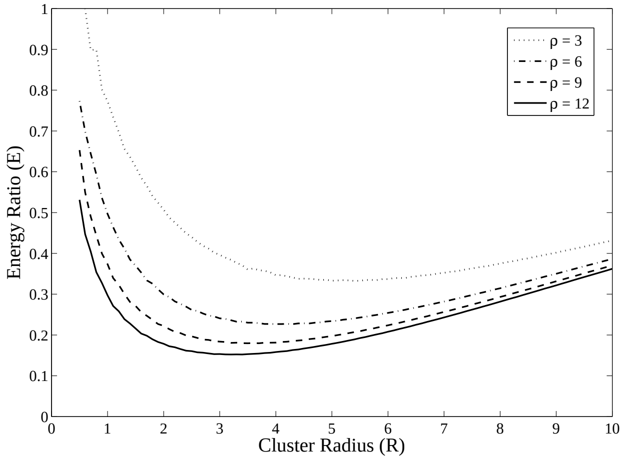


Fig. 6. The joint effect of the cluster size (N) and the node density (ρ) on the Energy Ratio (E) for $L = 10$ and $W = 0.75$.

In the following section, we propose a heuristic clustering algorithm that closely verifies our findings in this section.

VI. Simulation Experiments

To verify our results from the previous section, we simulate a centralized greedy clustering algorithm to find near-optimal solutions for the clustering problem in a 2-D network.

A. Greedy Clustering Algorithm

The greedy algorithm selects the WSN node with the highest cost to reach the sink, and then assigns it to the candidate CH for which the overall energy reduction is largest. The node joins the best candidate cluster and is marked. The algorithm iteratively performs this task until all the nodes in the WSN are marked. Algorithm 1 shows the details.

Algorithm 1 greedy_clustering($V, sink$)

```
$\tilde{V} = V$                                  // Initially, all nodes are unmarked.
**repeat**
    $\varepsilon_n, n \leftarrow \max\{$energy($v, sink$), $\forall v \in \tilde{V} \subseteq V\}$
    $CH \leftarrow n$                  // Set n as an isolated cluster of size 1.
    $\delta_{max} \leftarrow 0$        // Initialize maximum observed gain to zero.
    **for** every node $v \in V \ (v \neq n)$ **do**
        $\varepsilon_n^v \leftarrow$ energy($n$, $v$)
        $\varepsilon_v \leftarrow$ energy($v$, $sink$)
        assign($n$, $v$)               // Tentatively assign n to v's cluster.
        $\varepsilon'_v \leftarrow$ energy($v$, $sink$)
        $\delta \leftarrow (\varepsilon_n^v + \varepsilon'_v) - (\varepsilon_n + \varepsilon_v)$      // Compute the gain.
        **if** $\delta > \delta_{max}$ **then**
            $\delta_{max} \leftarrow \delta$          // Update max gain value observed.
            $CH \leftarrow v$        // Set $v$ as the best candidate CH for $n$.
        **end if**
        remove($n$, $v$)               // Undo this tentative assignment.
    **end for**
    assign($n$, $CH$)  // Permanently assign n to the best cluster.
    $\tilde{V} \leftarrow \tilde{V} - \{n\}$      // Remove n from further consideration.
**until** $\tilde{V} = \varnothing$                 // Repeat until all nodes are marked.
```

In Algorithm 1, V and $\tilde{V}$ denote the set of all WSN nodes and all unmarked WSN nodes, respectively. δ is the energy gain achieved by adding node n to a cluster and δ_{max} denotes the maximum energy gain attained by adding node n to any of the existing clusters. Subroutine *energy(u, v)* computes the required energy to transmit the data from node u to v. The amount of data to be transmitted is proportional to the number of members in the cluster of u (assuming u as a CH). Subroutine *assign(i, v)* adds node i to the set of members of the cluster whose CH is v. Subroutine *remove(i, v)* removes node i from the set of members of v. Statement $\tilde{V} \leftarrow \tilde{V} - \{n\}$ removes n from the set of unmarked nodes, assuring it will not be revisited in future iterations of the algorithm.

In Algorithm 1, all the nodes are initially unmarked, and thus considered as isolated clusters of size 1, which directly send their data to the sink. In each iteration of the algorithm, the node $n \in \tilde{V}$ with the highest energy requirement to reach the sink is chosen. Then, for every $v \in V$, the algorithm computes the gain achieved by adding node n to v. The node that achieves the largest gain is selected as the candidate CH for n to join. The procedure repeats until all the nodes in V are visited. Therefore, the time complexity of the algorithm is of $O(|V|^2)$.

B. Simulation Results

In our simulation, we uniformly scatter 400 sensor nodes in a symmetric WSN with the sink at the center (0,0). The physical placement of WSN nodes is the same in all experiments; only the data correlation degree changes. The output of our greedy algorithm on this network is depicted in Fig. 7.

Non-uniform cluster sizes are clearly evident in Fig. 7a. Clusters further from the sink are larger in radius, consistent with the results in Fig. 4 and Fig. 5. For higher degrees of correlation, the number of visually distinct clusters reduces, and there is greater affinity to proximal nodes. In the extreme case of Fig. 7d with $W = 1.0$, the resulting topology resembles a Minimum Spanning Tree (MST) for the network, though it still differs from (and is more energy-efficient than) the MST.

The diverse behaviours in Fig. 7 show the important effect of spatial data correlation on optimal cluster formation. These simulation results are also consistent with our analytical results for this problem.

VII. Conclusion and Future Work

In this paper, we studied the joint effects of data correlation, distance, and network density on forming optimal-sized clusters that require less power than conventional approaches. We showed that unlike most of the existing clustering approaches that produce uniform clusters throughout the whole network, heterogeneous-sized clusters are more energy-efficient in WSNs with spatial data correlation.

Our current analyses are based on a simple single cluster model. Although we have verified the correctness of our hypotheses with the outputs from a heuristic greedy clustering algorithm as well as the optimal solution of a linear network, further systematic studies of more generalized multi-cluster networks are needed. We are also working on a distributed version of the proposed greedy clustering algorithm that can be implemented and tested on real WSNs.

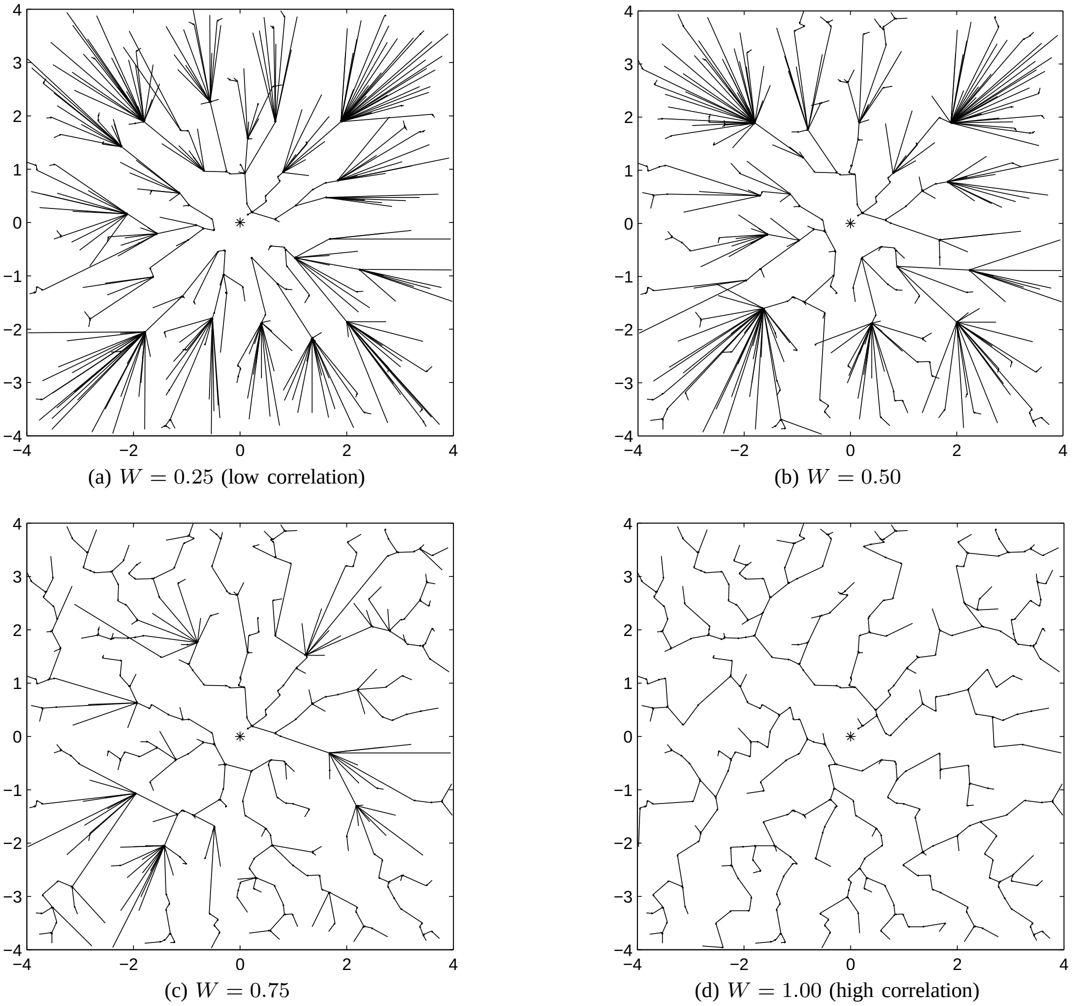
Acknowledgments

The authors gratefully thank the anonymous reviewers for their comments on an earlier version of this manuscript. Financial support for this research was provided by iCORE (Informatics Circle of Research Excellence) and NSERC (Natural Sciences and Engineering Research Council).

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] J. M. Kahn, R. H. Katz, and K. S. J. Pister, “Next Century Challenges: Mobile Networking for “Smart Dust”,” in *Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*. New York, NY, USA: ACM, 1999, pp. 271–278.



(a) $W = 0.25$ (low correlation)

(b) $W = 0.50$

(c) $W = 0.75$

(d) $W = 1.00$ (high correlation)

Fig. 7. Greedy clustering algorithm results for different degrees of data correlation.

[3] H. Çam, S. Özdemir, P. Nair, D. Muthuavinashiappan, and H. Ozgur Sanli, "Energy-Efficient Secure Pattern Based Data Aggregation for Wireless Sensor Networks," *Computer Communications*, vol. 29, no. 4, pp. 446–455, 2006.

[4] G. J. Pottie and W. J. Kaiser, "Wireless Integrated Network Sensors," *ACM Communications*, vol. 43, no. 5, pp. 51–58, May 2000.

[5] 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.

[6] 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.

[7] 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.

[8] D. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, January 1973.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[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] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

[17] 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.

[18] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[19] 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.