On the Optimal Randomized Clustering in Distributed Sensor Networks

Ali Dabirmoghaddam[a], Majid Ghaderi[b], Carey Williamson[b]

[a] *Department of Computer Engineering*
University of California, Santa Cruz, CA 95064, USA
[b] *Department of Computer Science*
University of Calgary, AB T2N 1N4, Canada

Abstract

Cluster-based data gathering is widely used in wireless sensor networks, primarily to overcome scalability issues. While clustering is not the most efficient means of gathering data, many clustering algorithms have attempted to provide energy efficiency as well. In this paper, we first demonstrate that the general problem of optimal clustering with arbitrary cluster-head selection is NP-hard. Next, we focus on randomized clustering in which sensor nodes form clusters in a distributed manner using a probabilistic cluster-head selection process. In order to find tractable and efficient solutions, we develop a mathematical framework that carefully captures the interplay between clustering and data correlation in the network. We further generalize this model to allow heterogeneous-sized clusters in different regions of the network. According to this model, we observe that clusters tend to become larger further from the sink. We also present simulation results to quantify the energy savings of joint clustering and compression. The results demonstrate that: 1) optimal selection of cluster sizes with respect to the correlation among sensor data has a significant impact on energy consumption of the network, and 2) while non-uniform clustering slightly improves the energy efficiency of the network, simple uniform clustering is remarkably efficient and provides comparable results for energy savings.

Keywords: energy-efficiency, clustering, data correlation, data compression, wireless sensor networks.

Email addresses: alid@soe.ucsc.edu (Ali Dabirmoghaddam), mghaderi@ucalgary.ca (Majid Ghaderi), carey@cpsc.ucalgary.ca (Carey Williamson)

1. 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]. Some WSN applications require dense deployment of sensor nodes in harsh and remote environments where human access is impossible or inadvisable. Such networks typically have many nodes, because of their geographic size, as well as the need for robustness to node failures. Such deployments require efficient architectures that can easily scale with network size without significant loss in performance.

Clustering is a well-established technique that has primarily been adopted 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.

Apart from providing a scalable structure, another advantage that clustering can offer is local data compression. Since in most applications, sensor nodes are deployed densely within the environment, significant 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 temperature monitoring, measurements reported by proximally-located sensors are likely to be very similar. This dependence can be exploited to eliminate redundancies and reduce the volume of data 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 fusion center). In this sense, cluster-based data gathering schemes can construct a hierarchy of nodes in multiple levels to route the data from sources to the sink. The most trivial implementation includes a bi-level structure comprising cluster members and CHs. In a similar fashion, CHs can form tier-2 super-clusters whose members

are tier-1 CHs and one of them may act as a tier-2 CH as well. Following this strategy, data compression can be performed in multiple levels. However, as we shall see later, with spatial data correlation, the dependency between observations rapidly decays with their geographical distance. As a result, the amount of reduction in message size by applying more levels in the hierarchy would be negligible. Therefore, in this paper, we focus on a bi-level hierarchy in which data compression is only performed at the CH level. The model we develop, however, can be extended to multi-level networks as well.

We should emphasize that *cluster-based data gathering* and *correlated data gathering* have both been extensively studied in the past, though separately. Specifically, the joint problem of *optimal clustering* and *correlated data gathering* is not fully addressed in the existing literature. Once again, it is noteworthy to highlight that clustering is essentially adopted as a means to achieve scalability in large WSNs and in that sense, is not intended to serve as the most efficient method of data gathering in WSNs with correlated data. Besides, as we shall show later, the problem of optimal clustering for minimizing network energy consumption is computationally intractable. Still, viable frameworks can be constructed and optimized to generate clusters that provide maximum energy efficiency while enabling scalability, as well.

There have been a number of works that studied optimal (energy-efficient) clustering, but ignored the effect of data correlation and compression on optimal cluster sizing [4–10]. A pioneering example of energy-aware clustering protocols is LEACH [4] in which each node has a pre-determined chance of becoming CH based on some probability function. The basic idea of LEACH was quickly adopted and extended in many different directions by the research community. EEHC [6], MOCA [9] and GESC [10] for instance are randomized clustering protocols which are based on a similar foundation as LEACH. In all such works, although data fusion is performed to reduce the size of communicated data in the network, no notion of data compression is taken into account while forming the clusters.

On the other hand, some researchers considered optimal data compression in WSNs without explicitly focusing on the clustering aspect of the problem [11–14]. A seminal analysis of energy-efficient correlated data gathering is presented by Cristescu *et al.* [11]. In that work, the authors consider Slepian-Wolf Coding (SWC) [15], a well-known method of distributed source coding, for which establishing the routing tree is easy, yet the data coding is complex and requires global network knowledge for optimal implementation.

The authors prove that joint optimization of rate allocation and transmission structure in distributed networks is NP-complete. Aside from energy-conservation, efficient data gathering has also been investigated from other perspectives, such as minimizing latency (*e.g.*, GroCoca [16]) or improving throughput and scalability (*e.g.*, SelectCast [17] and DDA [18]).

There are only a few sporadic works that study optimal clustering in the presence of data correlation [19–22]. For instance, [19] and [20] 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. Due to the complexities of modeling the joint data compression in correlated data fields, the authors make some simplifying assumptions, *e.g.*, trivial network topologies (linear or grid) and *fixed rate* of data reduction per source after compression, that inevitably influence the reliability and accuracy of the outcomes under realistic situations.

Furthermore, a de facto approach sought after by researchers studying clustering with data compression (*e.g.*, LEACH [4], EEHC [6], NOLBC [20] and MOCA [9] to name a few) attempts to find a *globally optimal* cluster size that minimizes the total network energy consumption. In all such works, for simplicity of model and analysis, the problem has intentionally been restricted to find a *uniform* clustering pattern that results in clusters that contain, on average, the same number of nodes. However, the fundamental question being overlooked here is whether uniform clustering is optimal for total energy consumption. In fact, although all foregoing proposals result in some form of energy-efficient topology, their methodology for tackling the problem inherently lacks the flexibility to form independently-sized clusters in different areas of the network. This paper challenges the existing belief by introducing a comprehensive model that collectively considers the joint impact of all important network attributes in forming clusters.

In particular, in a precursor study [23], we demonstrated that for a simple single-cluster network model, the optimal size of the cluster is directly proportional to its geographical distance from the sink. Such a proposition intuitively promotes a *non-uniform* clustering strategy with larger clusters at further distances from the sink. In this work, for the first time, we examine the foregoing hypothesis under more realistic conditions and establish that although the optimal cluster size grows with the distance from the sink, in practice, uniform clustering - if carefully done - can perform reasonably close to any optimal non-uniform clustering scheme. We verify such unexpected behavior using both mathematical analysis and simulation validation

throughout this paper.

In short, our contributions in this paper can be summarized as follows:

- We provide a formal and general definition of the problem of optimal clustering in distributed sensor networks with arbitrary CH selection and prove that this problem is NP-hard.
- In an attempt to contrive efficient heuristic solutions for this problem, we focus on *randomized clustering* and develop a framework to quantify the energy consumption of randomized uniform clustering with data compression. We further generalize this scheme to allow non-uniform clusters in the network.
- Using numerical analysis and simulation experiments, we validate our models and demonstrate that a simple randomized uniform clustering, on average, provides comparable results for energy consumption to the more complicated non-uniform counterparts, even though the corresponding optimal cluster sizes found in two cases are remarkably different.

The remainder of the paper is organized as follows. Section 2 provides mathematical preliminaries. Section 3 presents a formal proof that optimal clustering with data compression is NP-hard. Section 4 discusses optimal uniform clustering, while Sections 5 and 6 explore non-uniform clustering. Section 7 provides some numerical simulations and Section 8 discusses the results. Finally, Section 9 concludes the paper.

2. System Model and Assumptions

We assume that individual sensor nodes within the WSN are statistically identical information sources, whose readings follow a zero-mean normal distribution with variance σ^2. The set of observations within a cluster can thus be represented by a *multi-variate Gaussian distribution*. This assumption makes our analysis easier since the analytical properties of Gaussian sources are well-known. Furthermore, Gaussian sources are the worst case in terms of the required number of bits for coding [13]. Thus, the results from Gaussian fields can be interpreted as a bound for other types of sources. Similar assumptions have been used in prior related work [11, 24].

2.1. Distributed Randomized Clustering

In many WSN applications, sensor nodes are randomly dispersed over the area of interest in an uncontrolled manner (*e.g.*, using a helicopter) and form an ad-hoc network. Such a spontaneous structure requires appropriate mechanisms to be able to self-organize itself into an efficient, scalable and fault-tolerant architecture in a distributed manner and without reliance upon any central administrative entity.

A distributed randomized clustering is able to address all the foregoing concerns in a WSN as long as certain elements are observed during its construction. In such a scheme, nodes become CHs based on a probability function. CHs publicly advertise themselves within their proximal neighborhood and the non-CH nodes join their geographically closest CH member. This observation is important to ensure cost-efficient cluster assignments and avoid overlapping clusters. In this sense, with distributed clustering, we essentially perform a Voronoi tessellation of the network with CHs representing the Voronoi nuclei. Cluster members send their readings to their CH and thenceforth, CH is the only node being in charge of collecting and reporting the cluster data to the sink. This procedure relieves individual cluster members from maintaining and consistently updating complex data structures for routing purposes, making the network structure more scalable and robust.

Several well-known distributed clustering algorithms have been built on this general framework, each of which seeking to optimize the network performance from a particular perspective and subject to different assumptions. For example, Heinzelman *et al.* [4] consider a pre-specified probability function for CH selection that is oblivious of the data correlation degree in the network; Bandyopadhyay and Coyle [6] and Younis *et al.* [7] neglect the impact of data compression while forming their clusters; whereas Ghiasi *et al.* [5] solve the optimal distributed clustering problem for a pre-specified number of clusters in the network.

In the present work, we relax all such assumptions and propose a general model that allows clusters to form freely in different regions of the network. To ensure maximum energy efficiency, we assume that CHs compress the collected data from the cluster members and only submit one condensed redundancy-free message to the sink during each data collection cycle. The clusters induced by our model are optimized to enable maximum data compression while minimizing the cost of data collection and reportage subject to the cluster size and distance from the sink.

2.2. Data Correlation Model

In a Gaussian field of N sources, the pair-wise data dependency between sensor readings can be expressed using a symmetric positive-definite *covariance matrix* $\Sigma = [\sigma_{ij}]_{N \times N}$. Depending on the physical properties of the random field under study, several types of covariance models can be defined [25]. The information collected from physical events often has an exponential autocorrelation function [24]. Therefore, in this paper, we use a special type of Power Exponential correlation model with the elements of the covariance matrix given by:

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

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

2.3. Data Compression Model

In order to discretize the continuous-valued sensor 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. 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 $\Delta = \sqrt{12D}$ [23]. The entropy of the quantized sources, denoted by $H(S_N^D)$, is then given by [14]:

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

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 (i.e., isolated CHs or individual cluster members), Equation (2) reduces to:

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

In this paper, for the sake of brevity, we use b_n and b_1 to respectively denote number of bits required for encoding the entire cluster data after compression and that of an individual source. These quantities are calculated from Equations (2) and (3), respectively.

Table 1: Table of notations

Symbol	Usage
b_1	Size of an individual sensor reading (bits)
b_n	Size of data from a cluster of n nodes after compression (bits)
$\mathcal{C}$	Cost of cluster-based data collection
$\bar{\mathcal{C}}(n)$	Amortized energy cost of a cluster of size n
D	Distortion level (bit/symbol)
$\mathcal{D}$	RV for the distance of a cluster to the sink
E^*	Optimal network energy consumption
ext	Shorthand for exterior region
int	Shorthand for interior region
$\phi(n)$	The compression ratio function for a cluster of size n
ℓ	RV for the distance of a cluster member to the CH
$\mathcal{L}$	RV for the cumulative distance of nodes in a cluster to their CH
m	Number of regions in the network
$\mathcal{N}$	RV for the number of nodes in a cluster
p_i	Probability of CH selection in region i
$\langle p_1^*, \cdots, p_m^* \rangle$	Vector of optimal CH probabilities in regions 1 through m
ρ	Node density (nodes/unit area)
r_i	Width of region i
R	The network radius
$\mathcal{R}$	Radio range of a node
s	RV for the number of clusters in the network
W	Normalized data correlation degree

2.4. Energy Model

Cluster members observe some spatial stochastic process, quantize their observations, and transmit them to the CH (or sink), either directly (single-hop) or via intermediate sensor nodes (multi-hop). 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 P required to transmit b bits over distance d is given by [26]:

$$P = \gamma \, b \, d^2 \ , \tag{4}$$

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 [1] For simplicity and without loss of generality, hereafter we assume that $\gamma = 1$ J/bit/m^2. While in real world, transmission energy and sensor communication range follow more complicated patterns, such simplified assumptions enable us to develop tractable models that provide useful insights and approximate results on the performance of WSNs.

For convenient reference, Table 1 summarizes the most frequently used notations introduced above as well as the ones to follow. "RV" is used as an abbreviation for random variable.

3. Optimal Clustering

There have been many prior works on *optimal* clustering in a WSN [5, 6, 19–21]. The problem of OPTIMAL CLUSTERING is to discover a clustering of the network such that the total energy required for collecting data from the whole network is minimized as compared with other possible clustering patterns. In this paper, we first demonstrate that the general problem of OPTIMAL CLUSTERING with arbitrary CH selection is NP-hard. Then we construct a framework to tame the complexity of the problem and provide some tractable heuristic solutions for it.

First, let us begin with a formal definition of our problem.

Definition 1. *Network Clustering*
Given a network of nodes as an undirected graph $H = (W, F)$, where W denotes the set of nodes and F is the set of possible connections between node pairs within radio range of each other, the goal is to select a subset of nodes $W' \subseteq W, W' \neq \emptyset$ as CHs that form a Voronoi tessellation of the network.

Optimization Problem: Given H, the set of rates, and the internode distances, determine a clustering of the network that results in the minimum energy consumption. We call such clustering of the network the OPTIMAL CLUSTERING.

[1] We ignore the energy spent on receiving a message as it is independent of the distance over which the message is delivered.

As a matter of convenience, in the course of our proof, we shall restrict our attention to the following decision problem.

Decision Problem: Given H, the set of rates, internode distances, and a positive real number B, is there a clustering of the network whose energy consumption is no more than B?

We observe that, so long as our energy function is relatively easy to evaluate, the foregoing decision problem is no harder than the corresponding optimization problem. In other words, if we could solve the optimization problem in polynomial time, we would readily have an answer for the decision version, simply by comparing the output of the optimization problem with the given bound B.

Theorem 1. OPTIMAL CLUSTERING *is* NP*-hard.*

PROOF. We show that the problem of finding a p-MEDIAN, which is known to be NP-complete [27], is polynomial-time reducible to the problem of OPTIMAL CLUSTERING.

Given an undirected graph $G = (V, E)$, we associate each node $v \in V$ with a positive number $s(v)$ called the weight of v, and each edge $e \in E$ with a positive number $l(e)$ denoting its length. Let $X_p \subseteq V$ be a subset of p vertices. We define the distance between any vertex $v \in V$ and the set X_p as follows:

$$D(v, X_p) = \min_{x_i \in X_p} \{D(v, x_i)\} \ ,$$

where $D(v, x_i)$ denotes the length of the shortest path between v and x_i. The *distance-sum* of the set X_p is given by:

$$\mathcal{C}(X_p) = \sum_{v \in V - X_p} s(v) \cdot D(v, X_p) \ .$$

The set X_p^* is called a p-MEDIAN of G if $\mathcal{C}(X_p^*) = \min_{X_p \subseteq V} \{\mathcal{C}(X_p)\}$. The decision version of the p-MEDIAN problem is to determine whether there exists any $X_p \subseteq V$ such that $\mathcal{C}(X_p) \leq C$, where C is a given target bound.

Now, let us concentrate on OPTIMAL CLUSTERING. Consider a network as an undirected graph $H = (W, F)$. Each node $w \in W$ encodes its observations at a rate $r(w)$. A pair of nodes are within radio range of each other if there exists an edge $f \in F$ that corresponds to them. Let $k(f)$ denote the length of this edge. To obtain a better understanding of the problem,

we break it into two subproblems, namely, INTRA-CLUSTERING and INTER-CLUSTERING. For any given set of CHs, the INTRA-CLUSTERING refers to the problem of collecting data from within the clusters and forwarding it to the corresponding CHs given an energy budget of $\beta_1 \in \mathbb{R}^+$. The INTER-CLUSTERING problem, likewise, describes the process of data forwarding from CHs to the sink with a budget of $\beta_2 \in \mathbb{R}^+$. The OPTIMAL CLUSTERING problem involves the joint optimization of these two subproblems such that $\beta_1 + \beta_2 \leq B$, where B denotes the target energy bound.

First, we focus on the problem of INTRA-CLUSTERING. Let X_{ch} be an arbitrarily chosen subset of p nodes to act as CHs. We define the squared distance between any node w and a set X_{ch} by

$$d^2(w, X_{ch}) = \min_{x_i \in X_{ch}} \{d^2(w, x_i)\} \ ,$$

where $d^2(w, x_i)$ is the square of the Euclidean distance between w and $x_i \in X_{ch}$. We define our cost function for intra-cluster data collection as

$$\mathcal{C}_1(X_{ch}) = \sum_{w \in W - X_{ch}} r(w) \cdot d^2(w, X_{ch}) \ .$$

Similarly, we define the cost function for INTER-CLUSTERING problem as follows:

$$\mathcal{C}_2(X_{ch}) = \sum_{x_i \in X_{ch}} r(x_i) \cdot d^2(x_i, \text{sink}) \ .$$

Our goal is to find an optimal subset X^*_{ch} such that $\mathcal{C}_1(X^*_{ch}) + \mathcal{C}_2(X^*_{ch}) \leq B$.

Now consider an instance of the P-MEDIAN problem described by an undirected graph $G = (V, E)$, the set of weights $s(v), \forall v \in V$, the set of lengths $l(e), \forall e \in E$ and the target bound C. We construct a polynomial transformation from such instance of P-MEDIAN to an instance of OPTIMAL CLUSTERING of $H = (W, F)$ by letting $W := V$ and $F := E$. Also, we let $r(w) = s(v), \forall w \notin X_{ch}$; $r(w) = 0, \forall w \in X_{ch}$; $k(f) = l^2(e), \forall f \in F$; and target bound $B = C$. This transformation can be done in $O(|V| + |E|)$. It simply cancels out the cost of data collection from CHs and simplifies the OPTIMAL CLUSTERING as an instance of INTRA-CLUSTERING. It is now clear that any solution of the OPTIMAL CLUSTERING provides a solution for P-MEDIAN. Thus, P-MEDIAN $\leq_P$ OPTIMAL CLUSTERING concluding that OPTIMAL CLUSTERING cannot be solved in polynomial time unless P = NP.

Corollary 1. OPTIMAL CLUSTERING *remains* NP*-hard even if no data compression is done in the network.*

Since INTRA-CLUSTERING is hard, regardless of whether or not any data compression is performed at the CH level, *i.e.*, CHs merely forward the aggregated data to the sink, finding the optimal clustering structure remains NP-hard.

Having shown that the OPTIMAL CLUSTERING is inherently intractable, we seek to develop a framework that enables forming arbitrary-sized clusters that provide "good" energy consumption. In particular, we focus on a special class of clustering algorithms that are simple and can be implemented in a distributed manner. Such algorithms are randomized in the sense that each node independently decides to become a CH according to some probability p. The main problem to be addressed is then how to determine the optimal probability of CH selection (p) for different nodes, which is the problem to be investigated in the remainder of this paper. Henceforth, the concept of *optimality* is only discussed in the context of solutions that are *heuristically optimal* and should not be interpreted in its strict mathematical sense.

4. Randomized Uniform Clustering

In this section, using the mathematical preliminaries discussed in the previous section, we develop a model for the cost of data collection in a cluster-based sensor network and investigate the effect of cluster size on energy usage.

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 process of intensity ρ. For simplicity of analysis, let us assume that the sink is placed at the center of the disk. However, the actual placement of the sink is immaterial to our results. We study a randomized clustering model in which nodes become CH with some probability p. Therefore, by thinning of Poisson processes, 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. 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.

For a Voronoi process related to a bivariate Poisson process, Foss and Zuyev [28] have derived the following closed-forms for $\mathcal{N}$, the number of Π_0 particles in each Voronoi cell and $\mathcal{L}$, the cumulative length of all segments

connecting Π_0 particles to the Voronoi nucleus in each cell.

$$\mathbb{E}[\mathcal{N}] = \frac{\rho_0}{\rho_1} \ , \qquad \mathrm{Var}(\mathcal{N}) = \frac{\rho_0}{\rho_1} + 0.280\,\frac{\rho_0^2}{\rho_1^2} \ ,$$
$$\mathbb{E}[\mathcal{L}] = \frac{\rho_0}{2\rho_1^{3/2}} \ , \qquad \mathrm{Var}(\mathcal{L}) = \frac{\rho_0}{\pi\rho_1^2} + 0.147\,\frac{\rho_0^2}{\rho_1^3} \ .$$

Adopting their results and considering Π_0 and Π_1 particles in each Voronoi cell as cluster members and CHs respectively, we can easily infer the following expression for the average distance between a cluster member and its corresponding CH.

$$\mathbb{E}[\ell] = \frac{\mathbb{E}[\mathcal{L}]}{\mathbb{E}[\mathcal{N}]} = \frac{1}{2\sqrt{\rho_1}} = \frac{1}{2\sqrt{p\rho}} \ .$$

4.1. Single-Hop Communication

Direct transmission to the sink is used in some WSN applications to avoid the complexities of routing and Medium Access Control (MAC) [29]. In this scheme, individual sensors quantize their observations into messages of length b_1 (computed from Equation (3)) and transmit them to their CH. According to Equation (4), energy consumption is a quadratic function of the distance over which data transmission occurs. We know that $\mathcal{L}$ is a random variable defined as the summation of the distances between all cluster members and their CH. Let random variable ℓ_i denote the distance between the i^{th} cluster member and the CH. We know that ℓ_i's are iid. The number of nodes in a cluster, $\mathcal{N}$, is also a random variable. The law of total variance requires that

$$\begin{aligned}
\mathrm{Var}(\mathcal{L}) &= \mathbb{E}[\mathrm{Var}(\mathcal{L}|\mathcal{N})] + \mathrm{Var}(\mathbb{E}[\mathcal{L}|\mathcal{N}]) \\
&= \mathbb{E}\Big[\mathrm{Var}\Big(\sum_{i=1}^{\mathcal{N}} \ell_i \Big| \mathcal{N}\Big)\Big] + \mathrm{Var}\Big(\mathbb{E}\Big[\sum_{i=1}^{\mathcal{N}} \ell_i \Big| \mathcal{N}\Big]\Big) \\
&= \mathbb{E}[\mathcal{N}\mathrm{Var}(\ell)] + \mathrm{Var}(\mathcal{N}\mathbb{E}[\ell]) \\
&= \mathrm{Var}(\ell)\mathbb{E}[\mathcal{N}] + \mathbb{E}[\ell]^2\mathrm{Var}(\mathcal{N}) \ .
\end{aligned} \tag{5}$$

Rearranging Equation (5) and considering that $\mathrm{Var}(\ell) = \mathbb{E}(\ell^2) - \mathbb{E}(\ell)^2$ gives

$$\mathbb{E}(\ell^2) = \frac{\mathrm{Var}(\mathcal{L})}{\mathbb{E}[\mathcal{N}]} + \Big(1 - \frac{\mathrm{Var}(\mathcal{N})}{\mathbb{E}[\mathcal{N}]}\Big)\mathbb{E}[\ell]^2 \ . \tag{6}$$

Equation (6) gives the average squared distance of nodes to their CH that comes in handy for estimating the total intra-cluster energy cost (between cluster members and the CHs).

Once the CH collects the data from all cluster members, it eliminates the redundancies present in the data using lossless compression, and transmits the compressed data to the sink over the shortest path. The *inter-cluster data collection cost* refers to the energy spent by the CHs to perform this task. In order to estimate the inter-cluster cost, we need to measure the average squared distance from the clusters to the sink, $\mathbb{E}[\mathcal{D}^2]$. This can easily be calculated as

$$\mathbb{E}[\mathcal{D}^2] = \int_0^R x^2 \cdot \frac{2\pi x}{\pi R^2}\,\mathrm{d}x = \frac{1}{2}R^2 \ .$$

The mean number of nodes in a cluster is inversely proportional to the probability of being a CH in the region to which the cluster belongs. As discussed in Section 2.3, 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 (in bits) of the message that the CH transmits to the sink (note that b_n can be computed from Equation (2) for $n = 1/p$).

The average total network energy consumption, $\mathbb{E}[\mathcal{C}_{sh}]$, can be broken into the energy spent for intra-cluster and inter-cluster (between CHs and the sink) data collection. In symbols,

$$\mathbb{E}[\mathcal{C}_{sh}] = \mathbb{E}[s]\Big(b_1\mathbb{E}[\mathcal{N}]\mathbb{E}[\ell^2] + b_n\mathbb{E}[\mathcal{D}^2]\Big) \ , \tag{7}$$

where, $\mathbb{E}[s] = \rho p \pi R^2$ is the expected number of clusters in the network.

4.2. Multi-Hop Communication

In this scenario, we use a bit-hop metric to quantify the network energy consumption. Let $\mathcal{R}$ denote the radio range of a sensor node. Since we assume that all sensor nodes have the same radio range, the energy required to transmit one bit of information from a node to any other node in its radio coverage (one hop distance) is fixed and proportional to the square of the node's radio range, $\mathcal{R}^2$. Although this communication policy ignores the energy differences due to transmission over variable-range hops, it is more practical for implementation.

In order to compute the expected transmission energy, we need to estimate the total number of hops taken to communicate sensor readings to

the CHs or the sink. Within any given cluster, the total number of hops traversed is at least $\lceil \mathbb{E}[\mathcal{L}]/\mathcal{R} \rceil$. Likewise, for inter-cluster data transmission, $\lceil \mathbb{E}[\mathcal{D}]/\mathcal{R} \rceil$ gives the minimum number of hops to deliver the cluster data to the sink, where

$$\mathbb{E}[\mathcal{D}] = \int_0^R x \cdot \frac{2\pi x}{\pi R^2} \, \mathrm{d}x = \frac{2}{3} R$$

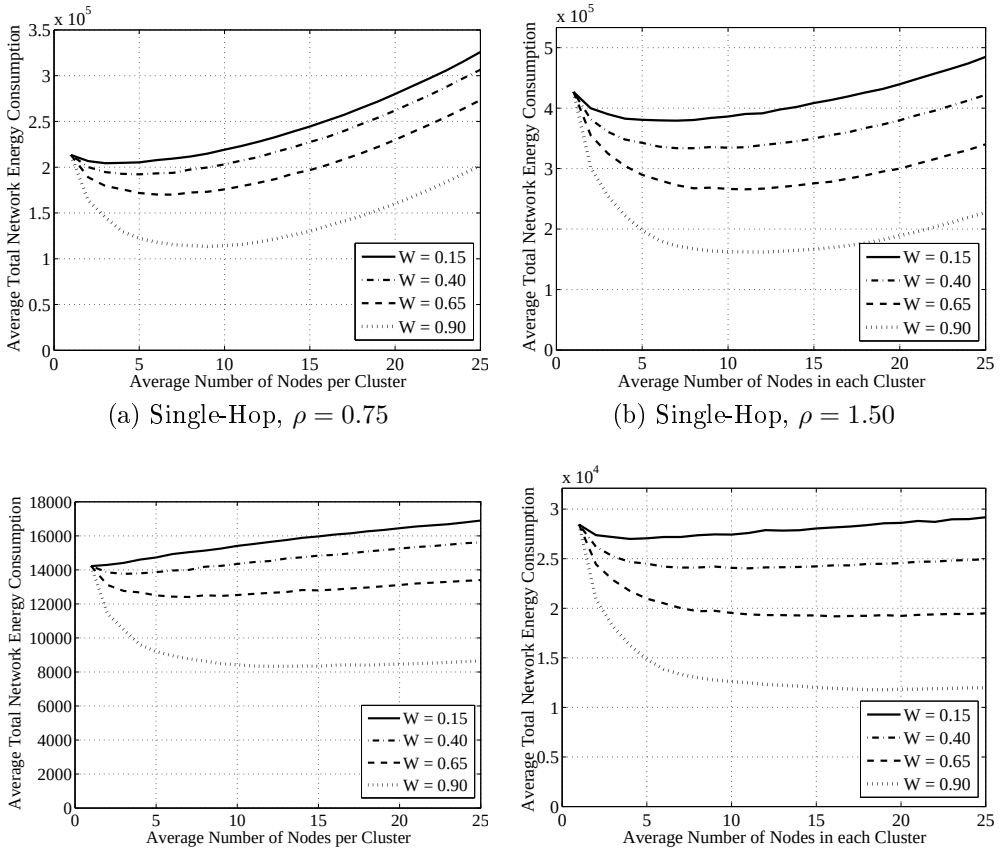
gives the average cluster distance from the sink. One may argue that the suggested approach for calculating the number of hops underestimates the actual steps required to deliver the data to the destination in a real network. We emphasize that, in this paper, we are interested in dense networks, since the data correlation in the network would be negligible otherwise. In such networks, the shortest path between a pair of nodes is closely approximated by a straight line segment between them. A similar assumption has been made in other prior work (*e.g.*, [6]). Furthermore, the good agreement between our mathematical model and the Monte Carlo simulations in Section 7 supports this claim.

Using this approximation, the total energy spent on data transmission in the multi-hop scenario is given by

$$\begin{aligned} \mathbb{E}[\mathcal{C}_{mh}] &= \mathbb{E}[s]\mathbb{E}[\mathcal{N}] b_1 \mathcal{R}^2 \left\lceil \frac{\mathbb{E}[\mathcal{L}]}{\mathcal{R}} \right\rceil + \mathbb{E}[s] b_n \mathcal{R}^2 \left\lceil \frac{\mathbb{E}[\mathcal{D}]}{\mathcal{R}} \right\rceil \\ &\approx \mathcal{R}\mathbb{E}[s] \Big(b_1 \mathbb{E}[\mathcal{N}]\mathbb{E}[\mathcal{L}] + b_n \mathbb{E}[\mathcal{D}]) \Big) \ . \end{aligned} \tag{8}$$

4.3. Numerical Analysis

Equations (7) and (8) describe the average total network energy usage as functions of various network properties, such as node density, data correlation degree, and cluster size. One important objective here is to find the optimal cluster size that minimizes the average network energy consumption. To this end, we numerically analyze the given energy functions. We consider a disk-shaped network of radius 15 on which nodes are scattered according to a Poisson process with an intensity of either 0.75 or 1.50. We change the data correlation degree from $W = 0.15$ (low) to $W = 0.90$ (high) and study the effect of changing the cluster size on the total network energy consumption. We examine both single-hop and multi-hop communication strategies. In single-hop communication, nodes adjust their power level appropriately to reach their destination. In the multi-hop scheme, nodes always transmit at full power, covering a radio range of 0.75 units in our simulations.



(a) Single-Hop, $\rho = 0.75$ (b) Single-Hop, $\rho = 1.50$

(c) Multi-Hop, $\rho = 0.75$ (d) Multi-Hop, $\rho = 1.50$

Figure 1: Average total network energy consumption in uniform clustering with different correlation degrees

Fig. 1 illustrates the average network energy consumption for different sizes of clusters and correlation degrees. In both single-hop and multi-hop scenarios, the stronger the data correlation is, the larger the optimal size of the cluster becomes. This observation is quite intuitive in the sense that by forming larger clusters, more redundancy can be removed (provided that a reasonable degree of data correlation exists among the original observations).

The impact of changing the cluster size on total energy consumption is more pronounced in single-hop communication than in multi-hop scheme. This is mainly due to the fact that the energy function is proportional to the square of the distance over which data transmission is done and this distance

for the single-hop communication is often longer than that for the multi-hop case. When a new node is added to a cluster, the cluster is able to save some energy via data compression. On the other hand, the data provided by the new node first has to be sent to the CH and then from the CH to the sink. If no data compression is performed, this transaction clearly is more energy-intensive than if the node individually transmits its data to the sink (possibly over a shorter path). Likewise, even with data compression, the amount of reduction per message achieved via making larger clusters should compensate for the extra energy spent on data communication on longer distances. With multi-hop communication, however, since all nodes transmit at the same power level, this issue becomes less crucial. In particular, when the data correlation degree is high, cluster sizes show a wider range of values. This is also the reason why the optimal cluster size in multi-hop communication gets larger than that of single-hop approach as data correlation degree increases. For example, when $W = 0.90$ and $\rho = 0.75$, with multi-hop communication, the energy consumption of clusters of size 9 to 25 are within 5% of the optimal, whereas in single-hop communication, such optimal range is only from 7 to 12.

5. Randomized Non-Uniform Clustering

Our previous uniform clustering model provides some useful insights as to how various degrees of data correlation and different transmission policies affect the optimal cluster sizing and energy consumption. However, the major downside of such a uniform clustering model is its inability to form variable size clusters in different regions of the network. In fact, by forcing the clusters to contain similar number of nodes, our model neglects any potential impact that *distance* can pose on optimal cluster sizing.

In previous work [23], we demonstrated that in correlated data fields, the optimal size of clusters is directly proportional to the cluster distance to the sink. Our previous analysis, however, was based on a very simple single-cluster model.

In this section, we concentrate on the effect of distance on forming optimal sized clusters in a realistic network made of possibly many clusters. We develop an elaborate model that allows clusters of arbitrary size to form freely in different regions of the network.

To be consistent with our previous model, we start with the same network topology as described in Section 4. In order to study the impact 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. 2a).

By convention, in this section, we use subscripts *int* and *ext* to denote the analytical properties of the interior and exterior regions, respectively. The radius of the interior region, r_{int}, is a fraction of the total network radius. That is to say,

$$r_{int} = \kappa R \ , \qquad 0 < \kappa < 1 \ . \tag{9}$$

We continue with our probabilistic clustering strategy. However, we let the probability of CH selection in the interior region (denoted by p_{int}) be 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).

The expected number of clusters in the interior region is:

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

and likewise, for the exterior region:

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

In this analysis, we only consider the multi-hop communication policy, since it is more general and practical than the single-hop scheme. In order to compute the intra-cluster data collection cost in the interior region, we act in the same way as our uniform clustering model. The intra-cluster data collection cost for such a cluster is given by

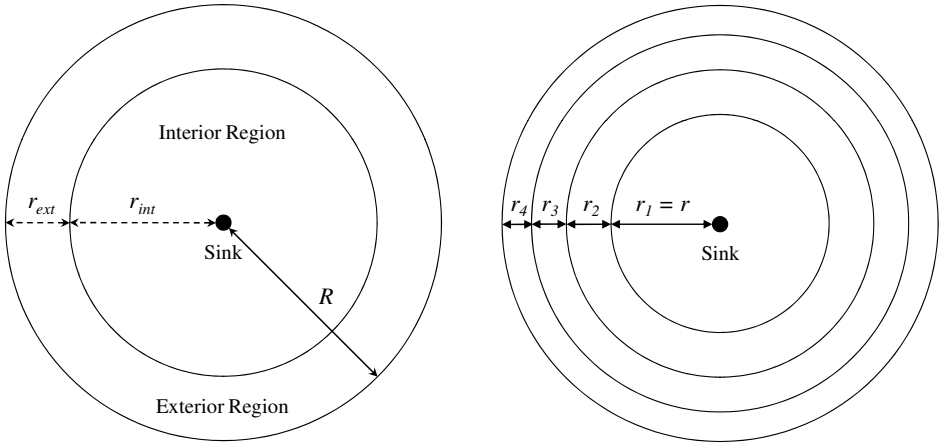
$$\mathbb{E}[\mathcal{C}^*_{int}] = b_1 \mathcal{R}^2 \lceil \frac{\mathbb{E}[\mathcal{L}_{int}]}{\mathcal{R}} \rceil \approx b_1 \mathcal{R} \mathbb{E}[\mathcal{L}_{int}] \ .$$

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

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

Similarly, for the exterior region, it is obtained that:

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



(a) A dual-region network (b) A multi-region network

Figure 2: Non-uniform clustering in a disk-shaped network.

Next, we focus on finding the inter-cluster data collection cost. In order to compute the energy required for this transmission, we only need to know the distance between the CH and the sink. Similar to our previous model, the mean distance of nodes in the interior region to the sink (center of the network) is computed as:

$$\mathbb{E}[\mathcal{D}_{int}] = \int_0^{r_{int}} x \cdot \frac{2\pi x}{\pi r_{int}^2} \, \mathrm{d}x = \frac{2}{3}\kappa R \; .$$

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:

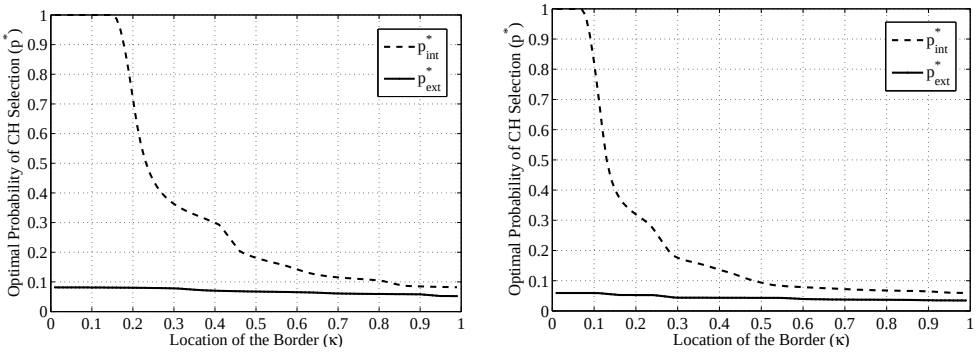
$$\mathbb{E}[\mathcal{C}_{int}^{\text{inter}}] \approx b_{n_{int}} \mathcal{R} \mathbb{E}[\mathcal{N}_{int}] \mathbb{E}[\mathcal{D}_{int}] \; .$$

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

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

where,

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



(a) $\rho = 0.75$ (b) $\rho = 1.50$

Figure 3: Optimal probability of CH selection vs. location of the border

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

$$\mathbb{E}[\mathcal{C}_{\text{total}}] = \mathbb{E}[\mathcal{C}_{int}^{\text{intra}}] + \mathbb{E}[\mathcal{C}_{int}^{\text{inter}}] + \mathbb{E}[\mathcal{C}_{ext}^{\text{intra}}] + \mathbb{E}[\mathcal{C}_{ext}^{\text{inter}}] \ . \tag{10}$$

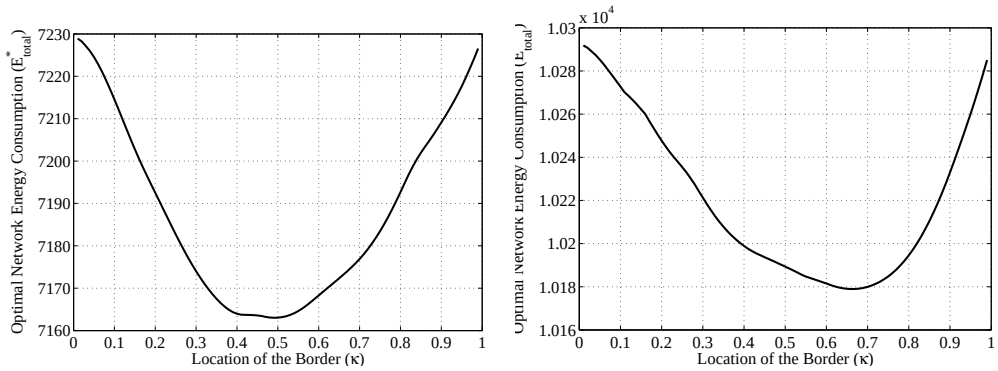
While the boundary between the two regions is fixed, $\mathbb{E}[\mathcal{C}_{\text{total}}]$ is a function of p_{int} and p_{ext}. We use p_{int}^* and p_{ext}^* to denote the optimal values of p_{int} and p_{ext} that minimize the total network energy consumption for all possible placements of the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

5.1. Experimental Analysis

We scatter sensor nodes on a network of radius 15, once with a density of 0.75 and once with 1.50 nodes per unit area. By varying κ from 0 to 1, we gradually move the boundary between the two regions across its full range. For any particular placement of the border, we then find the pair $\langle p_{int}^*, p_{ext}^* \rangle$ over the unit square that minimizes Equation (10).

Fig. 3 illustrates the optimal probabilities of CH selection in interior and exterior regions for any value of κ between 0 to 1. 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 that of the exterior region. That is, *clusters in the interior region are smaller than in the exterior region.*

Next, we analyze the effect of changing the border location on the network energy consumption. As Fig. 4 shows, the optimal position for the border



(a) $\rho = 0.75$ (b) $\rho = 1.50$

Figure 4: Optimal network energy usage vs. location of the border

is about $0.5R$ for $\rho = 0.75$, and $0.7R$ for $\rho = 1.5$. We note that in the former situation, the network is split into two equal-width regions, while in the latter, we have equal-area regions.

6. Generalized Non-Uniform Clustering

In this section, we extend our previous analysis to a general multi-region network model. The dual-region network analysis showed that splitting the network into two equal-area regions ($\kappa = 0.7R$) provides reasonably good energy efficiency. In this situation, nodes are equally divided between both regions. Therefore, we have a fair balancing of resource allocation over both regions. With our multi-region model, we also split the network into m concentric ring-shaped equal-area regions making each region contain the same number of nodes (on average). Hence, changing the cluster size throughout any region fairly affects the total energy consumption since all regions have almost the same number of nodes. We emphasize that our analysis is general and can easily be modified to fit other scenarios as well (*e.g.*, equal-width regions).

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. 2b). In region i, nodes become CH with a probability p_i. This probability is identical and independent of that of other regions.

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

is obtained as:

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

where $\mathbb{E}[\mathcal{N}_i] = p_i \rho \pi r^2$ 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.

Since the network is evenly divided into m regions all of the same area, we can easily obtain the following expression for the width of region i:

$$r_i = \left(\sqrt{i} - \sqrt{i-1}\right) r \ , \quad 1 \leq i \leq m \ . \tag{11}$$

Since all the clusters in region i are at a similar distance from the sink, the approximate cluster distances are:

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

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:

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

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:

$$\begin{aligned} \mathbb{E}[\mathcal{C}_{\text{total}}] &= \sum_{i=1}^{m} \mathbb{E}[\mathcal{C}_i^{\text{intra}}] + \mathbb{E}[\mathcal{C}_i^{\text{inter}}] \\ &= \rho \pi r^2 \mathcal{R} \sum_{i=1}^{m} p_i \left(b_1 \mathbb{E}[\mathcal{L}_i] + b_{n_i} \mathbb{E}[\mathcal{D}_i]\right) \ . \end{aligned} \tag{12}$$

Equation (12) 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,

$$\begin{aligned} \langle p_1^*, \cdots, p_m^* \rangle = \ & \underset{\{p_i\}}{\text{argmin}} \ \mathbb{E}[\mathcal{C}_{\text{total}}] \\ & \text{s.t. } 0 \leq p_i \leq 1, \ \forall i \in \{1, 2, \ldots, m\}, \end{aligned} \tag{13}$$

where $\langle p_1^*, \ldots, p_m^* \rangle$ are the optimal CH probabilities in regions 1 through m. Since p_i^*'s are independent, Equation (13) can be seen as the minimization of each summation term in Equation (12), separately. This can conveniently be done using existing numerical methods [30]. Some numerical examples are provided in the next section.

7. Simulation Experiments

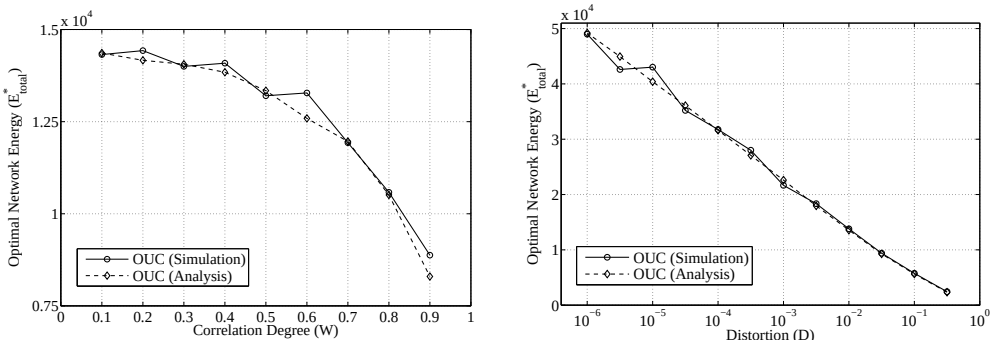
In this section, we look at the optimization problem introduced in Section 6, trying to find the best configuration for CH allocation over the network regions.

7.1. Simulation Environment

We use MATLAB for both our numerical and experimental analyses. The results reported for the model are the solutions of Equation (13) that are calculated in MATLAB. The simulation environment used in our experiments includes a disk network of radius 15 on which nodes are Poisson distributed with a density of 0.75 nodes per unit area (roughly, a total of 530 nodes, on average). The distortion level is set to 0.01 bits per sample. We assume multi-hop communication along the shortest path between pairs of nodes. The radio range of each node covers a radius of 0.75 units, and since all nodes transmit at the same power, the per-hop transmission cost is fixed per every bit of information sent.

7.2. Impact of Data Compression and Distortion on Energy Usage

In this subsection, we demonstrate how careful consideration of data correlation/compression in forming optimal-sized clusters helps reduce the total network energy consumption. For this experiment, we tentatively ignore the effect of distance on optimal cluster sizing and simply focus on a single-region network.



(a) The impact of data compression on network energy usage

(b) The impact of distortion level on network energy usage

Figure 5: Analysis of energy usage in a single-region network (uniform clustering)

Fig. 5a and 5b depict the network energy consumption of Optimal Uniform Clustering (OUC) in simulation versus the results obtained by the model. As evident from both figures, simulation results are fully consistent with the proposed model. In Fig. 5a, 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. Similarly, as Fig. 5b shows, increasing the tolerable distortion (D) also results in enhanced energy usage in the network. In order to ensure the fairness of CH selection through all areas of the network throughout our simulation experiments, 1000 random network configurations are generated per each value per independent variable (W or D) and the average energy-consumptions are reported.

7.3. Impact of Data Correlation and Distortion on Optimal Cluster Sizing

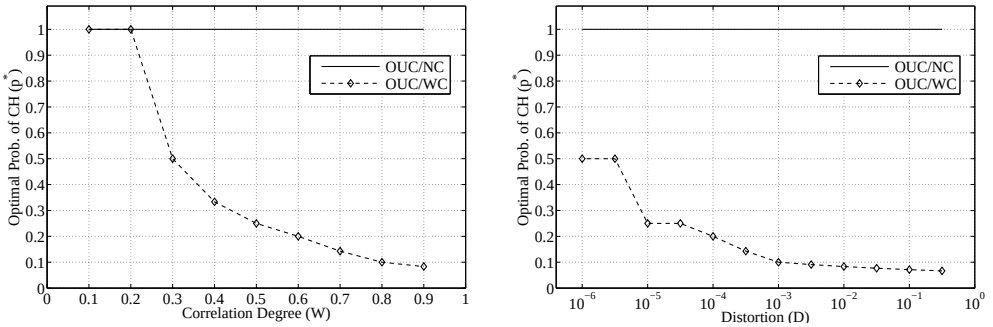
For the next experiment, we consider two scenarios:

1. Optimal Uniform Clustering with no Data Compression (OUC/NC): quantization on local observations; data aggregation at the CHs without compression.
2. Optimal Uniform Clustering with Data Compression (OUC/WC): 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, whereas 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 and also on potential energy savings when data correlation is present.

For both cases described above, namely, OUC/NC and OUC/WC, Fig. 6a and 6b respectively illustrate numerical analyses of the impacts of data dependence and distortion level on the optimal size of clusters.

As seen from both figures, when no data compression is performed at CH level (aggregation only), the optimal cluster size is always 1. This is reasonable in the sense that without data compression, no reduction in size of the cluster's aggregate data is attained. However, in clustering with data compression, as seen in Fig. 6a, increasing the correlation degree reduces the optimal probability of becoming CH in the network. In other words, the



(a) The impact of data correlation on optimal cluster sizing

(b) The impact of distortion level on optimal cluster sizing

Figure 6: Analysis of optimal cluster sizing in a single-region network (uniform clustering)

stronger the correlations are between the sensor observations, the larger the clusters become.

As shown in [11], 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 requires some nodes to send their readings through their pre-specified CH, which 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 CHs compensates for the extra energy spent due to the transmission of data over suboptimal paths. 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, which provides greater reduction in the size of the cluster data after compression. The optimal cluster sizes found in this experiment are 10 for $W = 0.8$ and 13 for $W = 0.9$.

Fig. 6b likewise demonstrates the impact of increasing the tolerable distortion level on optimal cluster sizing. As seen, when a higher level of distortion is allowed, readings from a broader local neighborhood can practically be compressed into a single message at the CH level; thus larger clusters become more affordable.

7.4. A Comparison of Uniform Clustering Schemes

In this subsection, we present a comparison of three uniform clustering schemes, namely Near-Optimal Location-Based Clustering [20] (NOLBC), Energy-Efficient Hierarchical Clustering [6] (EEHC) and Optimal Uniform Clustering (OUC) which we presented in this paper, with a particular focus on their corresponding energy usages. NOLBC is proposed as a heuristic scheme for approximating optimal cluster sizes as a function of number of sensors in the network. A somewhat different functional relationship is established between optimal probability of CH selection, network size and node density in EEHC where a multi-tier hierarchy of clusters is formed.

We believe that these two frameworks are similar to OUC in various aspects. First, they all are based on a randomized foundation and thus, can readily be implemented in real networks in a distributed manner. Secondly, energy-efficiency is the primary focus of all three schemes when forming the clusters. Thirdly, they all consider data correlation in order for removing data redundancies and saving energy. Based on all this, we believe that a side-by-side benchmark of these three schemes can be a fair and meaningful comparison.

Fig. 7 depicts the results of our simulations. For EEHC and OUC, results of both clustering with data compression and without data compression (identified by /WC and /NC suffixes respectively in the legend of Fig. 7) are provided. The purpose for including the latter is to provide a comparison baseline that highlights how much benefit is solely contributed by data compression itself. As seen, OUC generally yields better energy-efficiency compared to the other two. However, as data correlation degree increases, the results of all three schemes become more comparable.

The fundamental difference between the foregoing proposals (and their corresponding energy usages) lies in the extent to which they exploit data correlation. While all three schemes somehow implement data aggregation and compression, NOLBC and EEHC are oblivious of the impact of data correlation in forming optimal-sized clusters. In fact, in both schemes a fixed near-optimal cluster size is obtained to minimize the network energy consumption across the entire range of data correlation degrees. However, according to our findings in this paper, there exists a strong dependence between these two concepts. This observation motivates the idea of correlation-aware cluster sizing. It is interesting to note that in our simulations, NOLBC and EEHC construct clusters with fixed sizes of 32 and 11, respectively; while

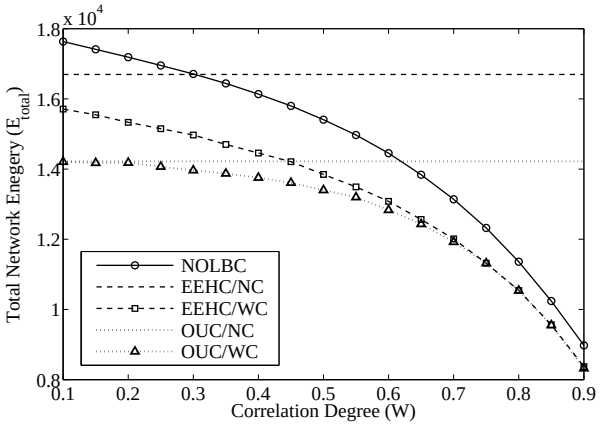


Figure 7: A comparison of energy consumption between different uniform randomized clustering schemes.

clusters formed by OUC, comprise a variable average ranging from 1.1 to 13.8 nodes per cluster as data correlation degree increases.

The energy differences between NOLBC, EEHC/WC and OUC/WC curves, as seen in Fig. 7, highlight the importance of careful adjustment of cluster sizes based on data correlation. That the differences between energy usages become less evident in presence of high correlation stems from the fact that clusters formed by NOLBC and EEHC are inherently large enough to provide maximum intra-cluster savings. In fact, it is in the absence of sufficient data correlation where having such excessively large clusters breaks the optimal routing structure (SPT) and induces additional transmissions over longer paths to the sink.

7.5. Non-Uniform Clustering in a Multi-Region Network

In this subsection, we first quantify the 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. The upper curve corresponds to a single-region network (uniform clustering), and the lower lines correspond to more regions, from 2 to 5, respectively (non-uniform clustering).

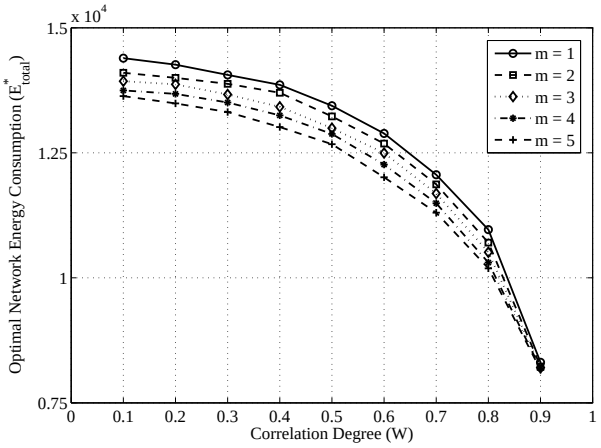
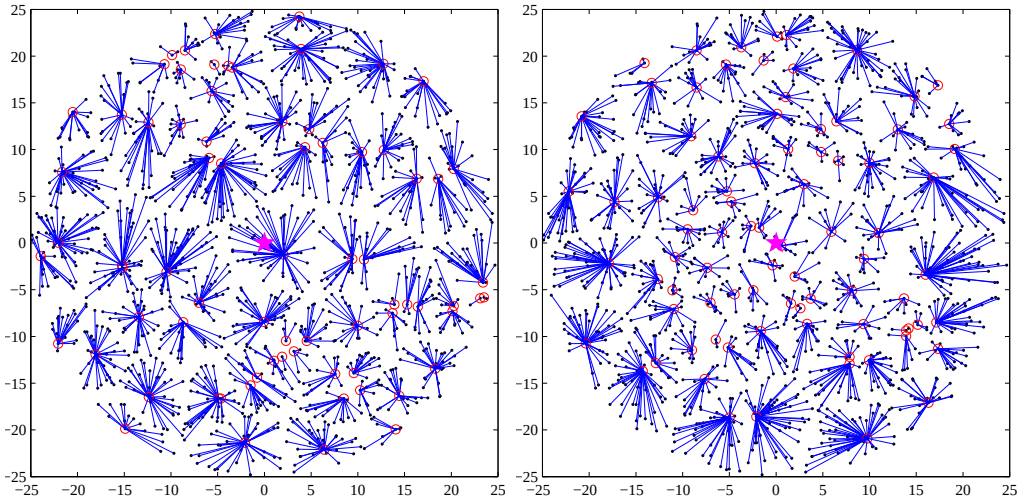


Figure 8: The energy consumption comparison between uniform clustering ($m = 1$) vs. non-uniform clustering ($m > 1$).

Table 2: Optimal probabilities of CH selection and their corresponding energy usage (model vs. simulation)

m	E^*_{mod}	E^*_{sim}	$\langle p^*_1, \ldots, p^*_n \rangle$
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	$(0.2002, 0.1001, 0.0715, 0.0556, 0.0556)$

Surprisingly, increasing the number of regions only slightly improves the network energy consumption. In order to interpret this unexpected behavior, let us have a look at Table 2 to see the optimal probability allocation over the regions of a certain realization. As evident from this table, the optimal CH probabilities decrease with the distance to the sink for all configurations. For a 5-region network, for example, the clusters of the outermost region are almost 4 times larger than the ones in the innermost region. However, the optimal theoretical network energy consumption for such a setting is only 1.5% better than that of uniform clustering in a single-region network. Also, the simulation results demonstrate less than 7% enhancement under the same conditions. In fact, both data correlation degree and distance



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

Figure 9: Realizations of two optimal clusterings

make the optimal size of the clusters larger. On the other hand, the larger a cluster becomes, the more energy has to be spent on collecting data from the cluster periphery. These two factors turn out to offset each other, yielding only marginal improvements. More precisely, adding more nodes to a cluster initially helps achieve higher data compression rates and better energy efficiency. Gradually, less and less energy savings are made as more nodes are attached to the cluster. At some point, the cluster gets "saturated". That is to say, the cluster reaches its limit in terms of maximum energy saving. At this point, additional nodes not only provide no extra savings, but also prove detrimental to the total energy consumption. Such phenomenon is often referred to as "diminishing returns". With optimal uniform clustering, not all clusters are saturated, but most of them are close to their limits. With optimal non-uniform clustering, all clusters can reach their capacity limit. However, the difference between the two stages is so small that in practice, optimal uniform clustering performs quite close to any optimal non-uniform clustering strategy.

Fig. 9 compares two optimal realizations of uniform clustering (single-region network) against non-uniform clustering (multi-region network) on an arbitrary network. As presented by Fig. 9, with non-uniform clustering,

the optimal cluster size grows with distance from the sink. Also, for this particular example, the non-uniform clustering saves 8.5% more energy than the uniform clustering.

8. Further Results and Discussion

In this section, our objective is to shed some light on why a basic uniform clustering provides comparable energy savings to non-uniform schemes, even though the average cluster sizes are remarkably different.

Consider a cluster of nodes with radius r at an arbitrary distance d from the sink (see Fig. 10). We want to see how the per-node data collection cost changes as we expand the cluster radius by Δr. For simplicity, in the following, we consider direct data transmission; however, as we showed earlier, since the relative energy savings for various cluster sizes in the multi-hop scheme is no better than that of the direct communication, we can consider the resulting savings as an upper bound for multi-hop communication, as well.

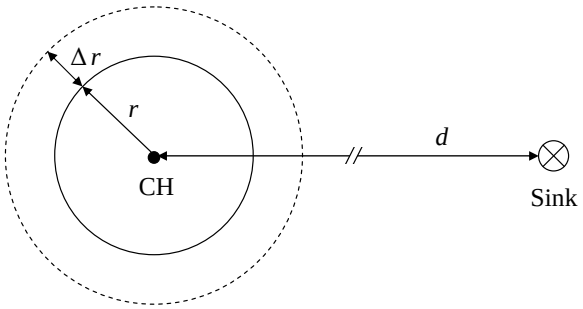


Figure 10: A cluster of radius r at distance d from the sink.

As explained in Section 4.1, we can derive the following expressions for the energy cost of data collection from an arbitrary cluster when it comprises n and $n + \Delta n$ nodes, respectively.

$$\begin{aligned} \mathcal{C}_{(n)} &\approx n b_1 (\frac{1}{2} r^2) + b_n d^2 \\ \mathcal{C}_{(n+\Delta n)} &\approx (n + \Delta n) b_1 \Big(\frac{1}{2} (r + \Delta r)^2 \Big) + b_{(n+\Delta n)} d^2 \ . \end{aligned} \tag{14}$$

Using Poisson approximation, we know that $n \approx \rho\pi r^2$ and $n + \Delta n = \rho\pi(r + \Delta r)^2$. Therefore, we can rewrite Equation (14) as follows.

$$\begin{aligned} \mathcal{C}_{(n)} &\approx n^2 \varepsilon b_1 + b_n d^2 \\ \mathcal{C}_{(n+\Delta n)} &\approx (n + \Delta n)^2 \varepsilon b_1 + b_{(n+\Delta n)} d^2 \ , \end{aligned} \tag{15}$$

where $\varepsilon = 1/(2\rho\pi) \approx 0.16\rho$ is a constant independent of n.

Now, let $\bar{\mathcal{C}}_{(n)}$ denote the *amortized energy cost of a cluster of size* n. We have that:

$$\bar{\mathcal{C}}_{(n)} = \frac{\mathcal{C}_{(n)}}{n} \ . \tag{16}$$

In fact, $\bar{\mathcal{C}}_{(n)}$ can be seen as the average energy usage of an arbitrary node when it is assigned to a cluster of size n. Clearly, by expanding the cluster size we want

$$\begin{aligned} \bar{\mathcal{C}}_{(n+\Delta n)} &\leq \bar{\mathcal{C}}_{(n)} \ \Rightarrow \\ (n + \Delta n)\varepsilon b_1 + \frac{b_{(n+\Delta n)}}{(n + \Delta n)} d^2 &\leq n\varepsilon b_1 + \frac{b_n}{n} d^2 \ \Rightarrow \\ \frac{\varepsilon \Delta n}{d^2} &\leq \frac{b_n}{n b_1} - \frac{b_{(n+\Delta n)}}{(n + \Delta n) b_1} \ . \end{aligned} \tag{17}$$

In previous work [23], we introduced the metric *compression ratio* that is defined as $\phi_{(n)} = b_n/(nb_1)$. As mentioned earlier, CHs only forward a condensed message representing the entire cluster information to the sink after removing the redundancies. The compression ratio is a normalized measure that indicates what fraction of the collected data from the cluster members is transmitted to the sink after compression, and in this sense, the less the compression ratio, the better. The limiting values are 1 when exactly the same copy is sent (*i.e.*, a cluster of size 1 or when no data correlation exists) and 0 for a highly correlated field as $n \to \infty$.

Using the notation of compression ratio and from Equation (17), we can readily infer that an additional node can be added to a cluster of size n as long as

$$|\Delta\phi_{(n)}| = |\phi_{(n+1)} - \phi_{(n)}| \geq \frac{\varepsilon}{d^2} \ . \tag{18}$$

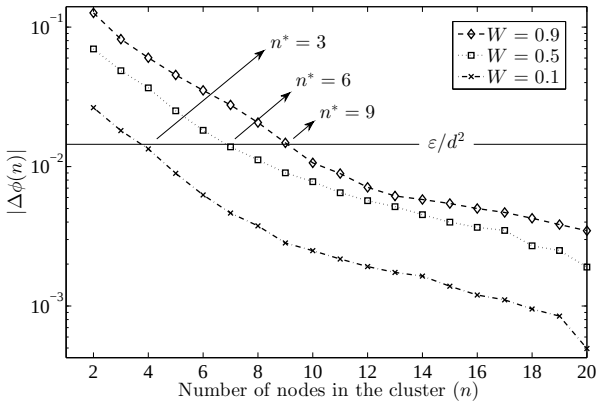


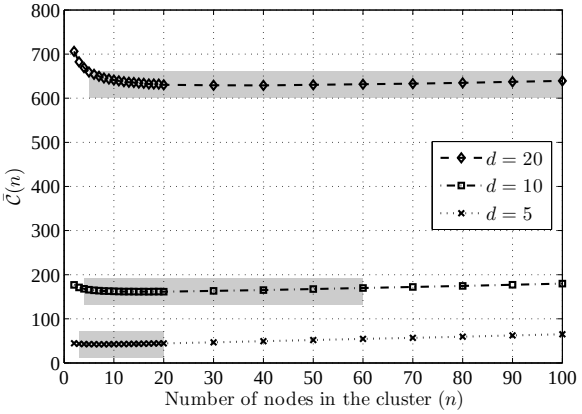
Figure 11: The required threshold for specifying the optimal cluster size.

As shown in [23], $\phi_{(n)} : \mathbb{N} \rightarrow [0 \quad 1]$ is a non-increasing convex function of n; therefore, $\forall n \in \mathbb{N} : 0 \leq |\Delta\phi_{(n)}| \leq 1$. However, as Equation (18) shows, assigning a new member to an existing cluster is cost-saving only if the resulting cluster's compression ratio is at least ε/d^2 less than that of the cluster excluding the new member. Knowing their approximate distance to the sink, CHs can use this criterion to decide whether or not comprising a new member is beneficial.

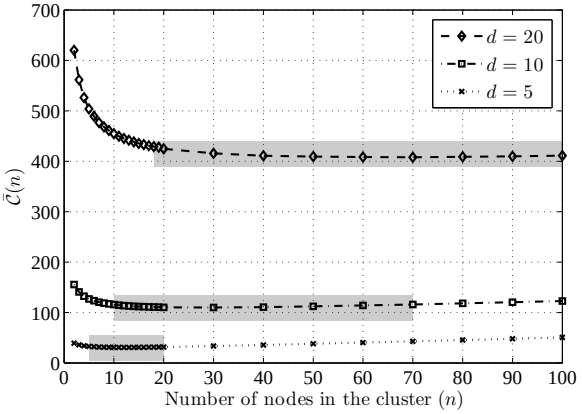
For different degrees of data correlation, Fig. 11 illustrates that $|\Delta\phi_{(n)}|$ monotonically decreases with the cluster size. The horizontal line shows the required difference of compression ratios for a cluster at $d = 3$ to expand. The intersections of the horizontal line with curves specify the thresholds for specifying the optimal cluster sizes (denoted by n^*). That is to say, by expanding the cluster size beyond this limit, the per-node cost of data collection increases. As the cluster gets further from the sink (*i.e.*, d increases), the constraint on the right-hand-side of the inequality (18) becomes looser, setting the horizontal line lower, implying that the optimal cluster size increases with distance (confirming our former results).

We next focus on how the amortized cost of the cluster changes with its size. In particular, we want to quantify the savings achieved by adding more nodes to the cluster while the same node density is maintained in the cluster. Expanding Equation (16), we can write

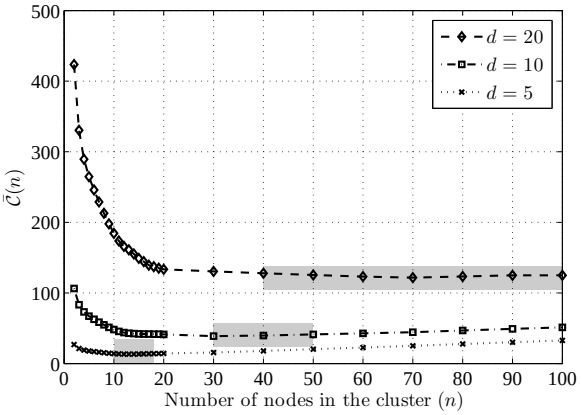
$$\bar{\mathcal{C}}_{(n)} = b_1(n\varepsilon + \phi_{(n)}d^2) \ . \tag{19}$$



(a) Low data correlation ($W = 0.1$)



(b) Medium data correlation ($W = 0.5$)



(c) High data correlation ($W = 0.9$)

Figure 12: The amortized cost of clustering vs. cluster size

The first term inside the brackets in Equation (19) (*i.e.*, $n\varepsilon$) corresponds to the intra-cluster data collection cost and is an increasing function of the cluster size (n). On the other hand, the second term (*i.e.*, $\phi_{(n)}d^2$) contributes to the amount of savings obtained via compressing the cluster data and thus, is a non-increasing function of the cluster size. Analogous to anti-parallel forces, these two terms pull the cluster boundaries in opposite directions. The latter is stronger when the cluster size is small, but it gradually becomes weaker as the cluster grows. Fig. 12 better explains this interesting behavior.

For a given degree of data correlation (W), Fig. 12 depicts the amortized cost ($\bar{\mathcal{C}}_{(n)}$) of a cluster at a certain distance (d) from the sink as a function of the cluster size (n). As clearly evident, expanding the cluster size first helps achieve a lower energy consumption per node. Such savings are more significant for clusters at further distances from the sink or when the data correlation degree is relatively high. By adding more nodes, the cluster eventually comes to its saturation limit. The amortized cost of the cluster begins to slightly increase by expanding the cluster size beyond this point. In fact, after the cluster gets saturated, the extra cost from having additional nodes in the cluster turns out to offset the savings due to achieving better data compression rates, such that the difference in the amortized cost of the cluster is barely noticeable after this point.

The shaded areas in Fig. 12 show the cluster sizes whose energy consumptions are within 5% of the optimal. As seen, for clusters further away from the sink, such optimal range is wider than for the closer ones. Moreover, for clusters at various distances, these optimal ranges are overlapping. In other words, even though the optimal cluster size significantly varies with distance, it is always possible to find a globally optimal cluster size that performs very well across the entire network. This result justifies why even a simple uniform clustering can perform reasonably close to the more complicated non-uniform schemes.

9. Conclusions and Future Work

In this paper, we showed that the general problem of Optimal Clustering is NP-hard. We proposed a novel framework for modeling cluster-based data gathering in WSNs and optimized it to produce the best possible clustering of the network in terms of energy consumption.

We presented the first analysis of non-uniform clustering in WSNs and demonstrated that heterogeneous-sized clusters are more energy-efficient in

WSNs with spatial data correlation. We further showed that due to the trade-offs induced by physical characteristics of clusters, optimal uniform clustering can also perform very well compared to the more complicated non-uniform counterparts.

In the specific network configurations considered in our simulations, the improvements achieved by non-uniform clustering are not significant. An avenue for further research is to study the specific topologies (including contrived and arbitrary configurations) which might better benefit from non-uniform clustering.

Analyzing the network lifetime and investigating potential mechanisms (*e.g.*, CH rotation) that can help fairly distribute the data collection load throughout the network is another interesting area of future study.

Last but not least, it is noteworthy to mention that our proposed framework is originally tailored for static configurations. Nonetheless, mobility is an ever-growing necessity in most recent trends of applications. Extension of the proposed scheme to cope with mobility and its related challenges is yet another important problem which remains for future work.

References

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A Survey on Sensor Networks, IEEE Communications Magazine 40 (2002) 102–114.

[2] A. Abbasi, M. Younis, A Survey on Clustering Algorithms for Wireless Sensor Networks, Computer Communications 30 (2007) 2826–2841.

[3] P. Wang, R. Dui, I. Akyildiz, Collaborative Data Compression Using Clustered Source Coding for Wireless Multimedia Sensor Networks, in: Proc. IEEE Conference on Computer Communications (INFOCOM), San Diego, CA, USA, 2010, pp. 1713–1723.

[4] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-Efficient Communication Protocol for Wireless Microsensor Networks, in: Proc. Hawaii International Conference on System Sciences (HICSS), volume 8, Maui, HI, USA, 2000, p. 8020.

[5] S. Ghiasi, A. Srivastava, X. Yang, M. Sarrafzadeh, Optimal Energy Aware Clustering in Sensor Networks, Sensors 2 (2002) 258–269.

[6] S. Bandyopadhyay, E. Coyle, An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks, in: Proc. IEEE Conference on Computer Communications (INFOCOM), San Francisco, CA, USA, 2003, pp. 1713–1723.

[7] O. Younis, S. Fahmy, Distributed Clustering in Ad-Hoc Sensor Networks: A Hybrid, Energy-Efficient Approach, in: Proc. IEEE Conference on Computer Communications (INFOCOM), volume 1, Hong Kong, 2004, pp. 629–640.

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

[9] A. Youssef, M. Younis, M. Youssef, A. Agrawala, Distributed Formation of Overlapping Multi-hop Clusters in Wireless Sensor Networks, in: IEEE Global Telecommunications Conference (GLOBECOM), San Francisco, CA, USA, 2006, pp. 1–6.

[10] N. Dimokas, D. Katsaros, Y. Manolopoulos, Energy-Efficient Distributed Clustering in Wireless Sensor Networks, Journal of Parallel and Distributed Computing 70 (2010) 371–383.

[11] R. Cristescu, B. Beferull-Lozano, M. Vetterli, On Network Correlated Data Gathering, in: Proc. IEEE Conference on Computer Communications (INFOCOM), Hong Kong, 2004, pp. 2571–2582.

[12] I. Koutsopoulos, M. Halkidi, Measurement Aggregation and Routing Techniques for Energy-Efficient Estimation in Wireless Sensor Networks, in: Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), Avignon, France, 2010, pp. 1–10.

[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), Yokohama, Japan, 2003, p. 174.

[14] A. Scaglione, S. Servetto, On the Interdependence of Routing and Data Compression in Multi-hop Sensor Networks, Wireless Networks 11 (2005) 149–160.

[15] D. Slepian, J. Wolf, Noiseless Coding of Correlated Information Sources, IEEE Transactions on Information Theory 19 (1973) 471–480.

[16] C. Chow, H. Leong, A. Chan, GroCoca: Group-based Peer-to-Peer Cooperative Caching in Mobile Environment, IEEE Journal on Selected Areas in Communications 25 (2007) 179–191.

[17] C. Wang, C. Jiang, S. Tang, X. Li, SelectCast: Scalable Data Aggregation Scheme in Wireless Sensor Networks, IEEE Transactions on Parallel and Distributed Systems 23 (2012) 1958–1969.

[18] S. Ji, Z. Cai, Distributed Data Collection in Large-Scale Asynchronous Wireless Sensor Networks Under the Generalized Physical Interference Model, IEEE/ACM Transactions on Networking 21 (2013) 1270–1283.

[19] N. Vlajic, D. Xia, Wireless Sensor Networks: To Cluster or Not To Cluster?, in: Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Buffalo-Niagara Falls, NY, USA, 2006, pp. 268–276.

[20] S. Pattem, B. Krishnamachari, R. Govindan, The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks, in: Proc. ACM International Conference on Information Processing in Sensor Networks (IPSN), Berkeley, CA, USA, 2004, pp. 28–35.

[21] H. Chen, 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), volume 4, Las Vegas, NV, USA, 2006, pp. 2318–2323.

[22] M. Lotfinezhad, B. Liang, Effect of Partially Correlated Data on Clustering in Wireless Sensor Networks, in: Proc. IEEE Conference on Sensor and Ad Hoc Communications and Networks (SECON), Santa Clara, CA, USA, 2004, pp. 172–181.

[23] A. Dabirmoghaddam, M. Ghaderi, 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), Miami Beach, FL, USA, 2010, pp. 163–171.

[24] M. Vuran, I. Akyildiz, Spatial Correlation-Based Collaborative Medium Access Control in Wireless Sensor Networks, IEEE/ACM Transactions on Networking 14 (2006) 316–329.

[25] J. Berger, V. de Oliveira, B. Sanso, Objective Bayesian Analysis of Spatially Correlated Data, Journal of the American Statistical Association 96 (2001) 1361–1374.

[26] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, An Application-Specific Protocol Architecture for Wireless Microsensor Networks, IEEE Transactions on Wireless Communications 1 (2002) 660–670.

[27] O. Kariv, S. Hakimi, An Algorithmic Approach to Network Location Problems. II: The p-MEDIANS, SIAM Journal on Applied Mathematics 37 (1979) 539–560.

[28] S. Foss, S. Zuyev, On a Voronoi Aggregative Process Related to a Bivariate Poisson Process, Advances in Applied Probability 28 (1996) 965–981.

[29] M. Lotfinezhad, B. Liang, E. Sousa, Adaptive Cluster-Based Data Collection in Sensor Networks with Direct Sink Access, IEEE Transactions on Mobile Computing 7 (2008) 884–897.

[30] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes, Cambridge University Press, New York, NY, USA, 2007.