Lifetime Optimized Hierarchical Architecture for Correlated Data Gathering in Wireless Sensor Networks

Tran Minh Tam [1], Hung Q. Ngo [2], P.T.H. Truc [3], Sungyoung Lee [4]

Department of Computer Engineering, Kyung Hee University

South Korea, 446-701

[1]tmtam@oslab.khu.ac.kr, [2]nqhung@oslab.khu.ac.kr, [3]pthtruc@oslab.khu.ac.kr, [4]sylee@oslab.khu.ac.kr

***Abstract*—In-network aggregation is essential for correlated data gathering in wireless sensor networks which are resource-constraint in terms of energy, computation and storage. In this paper, we consider the problem of building a minimum cost hierarchical architecture for correlated data gathering with in-network aggregation, which is formulated as a min-sum optimization problem. To solve the problem, we first develop a minimum-cost distributed algorithm which involves only simple message-passing rules. The algorithm is then tuned to be energy-aware so that high-energy sensor nodes are preferably selected to become cluster heads (CHs), which act as encoding and relaying nodes for the raw sensing data from their corresponding one-hop member nodes. After the cluster formation phase, joint-entropy coding technique with explicit communication (specifically, foreign coding) is applied at every CH to remove possible data redundancy (due to the spatial data correlation) for in-network aggregation. Simulations show that the network lifetime can be significantly extended using our minimum cost cluster-based approach.**

I. Introduction

In this paper [1], we focus on a generic type of applications called data gathering, in which all nodes periodically produce information by sensing a geographic area and transmit them to a sink for processing. The fact is that an inherent data redundancy corresponding to a degree of spatial correlation always exists, leading to a significant waste of energy during networking tasks. In-network data aggregation, which is able to remove such data redundancy, thus minimizing energy consumption, is a desirable work that many researchers have pursued.

Generally, the correlated data gathering problem with in-network aggregation is formulated as a joint-optimization problem of rate allocation and transmission structure. The former is to find the minimum data encoding rate at each node, while ensuring the data collected by the sinks can be decoded to reproduce the original data [1]. The latter is to build optimal routes from every node to the sink(s) such that the total energy consumption is minimized. There are two popular source coding paradigms used: distributed source coding (e.g. Slepian-Wolf coding) and joint-entropy coding with explicit communication [2], [3]. With Slepian-Wolf coding – a *multi-input* coding strategy, the rate allocation optimization is complicated but an optimal transmission structure appears to be a shortest path tree (SPT) of which algorithms have been well-developed [4]. However, performing Slepian-Wolf coding at all nodes is difficult because it involves a large number of bins, requires *synchronous* communication model, recoding at intermediate nodes and a global knowledge of the network (i.e. the distance between every pair of nodes) available at each node. Moreover, wireless network is error-prone so that even only one packet is lost will make decoding all measurements of involving nodes impossible [5]. Joint-entropy coding with explicit communication, a *single-input, asynchronous* source coding scheme that requires a complex algorithm to find an optimal transmission structure, makes the rate allocation problem much simpler to solve. In this scheme, a data source is encoded by only the side information received from at least one of neighboring nodes and no waiting for belated information at intermediate nodes is needed. Rickenbach et al. [6] classified it into two opposite techniques: *foreign coding* and *self coding*. With self coding, a node is able to encode its own raw data only in the presence of side information from at least another node. In contrast, foreign coding is the technique in which raw data originating at one node is encoded by another node. We can easily observe that foreign coding achieves better compression rate in a *many-to-one* (i.e. cluster-based) architecture than self coding [6].

Most of the work so far focused on how to build an optimal tree for aggregating the data to a *single stationary sink*. However, if the sink is mobile, these tree-based approaches are intractable because tree re-building, which is a costly task, is required as the sink moves over time. Moreover, if there are multiple sinks, it is much harder to come up with an optimal solution involving a set of trees. On the other hand,

[1]This research was supported by the MKE (Ministry of Knowledge Economy, Korea) under the ITRC (Information Technology Research Center) support program supervised by the IITA-2008-C1090-0801-0002 (Institute of Information Technology Advancement); and by the MIC (Ministry of Information and Communication, Korea) under the ITFSIP (IT Foreign Specialist Inviting Program) supervised by the IITA-C1012-0801-0003. Also, this work is financially supported by the Ministry of Education and Human Resources Development (MOE), the Ministry of Commerce, Industry and Energy (MOCIE) and the Ministry of Labor (MOLAB) through the fostering project of the Lab of Excellency. Corresponding author: Sungyoung Lee

clustering is a mechanism to organize the sensor network into a connected hierarchy of cluster head (CH) and member nodes. Clustering can not only aid in reducing energy consumption and increasing network lifetime [7], [8], and surveys [9], [10], [11] but also providing scalability which implies the need for load balancing, efficient resource utilization, and in-network data fusion. Furthermore, clustering can benefit us in many other aspects: sleep scheduling, low-latency and energy-efficient communication, small possibility of reconstruction error at sink [12], cheaper route discovery and maintenance cost due to the overlay network of CHs. Because there is no load-balancing mechanism derived for the tree-based approaches, the nodes which are near to the sink is likely to receive and transmit much more data than nodes which are far from the sink. As a consequence, some of them will run out of energy more quickly than others and connectivity lost can become a major issue.

By applying cluster-based foreign coding, the rate allocation problem becomes simple; however, the minimum-cost clustering task turns out to be a hard optimization problem. In this paper, we propose a distributed clustering algorithm for correlated data gathering to minimize the transmission cost. We first formulate the problem as an NP-hard min-sum optimization problem. We then develop simple, heuristic message-passing rules to achieve a near-optimal solution. Our cluster-based approach not only overcomes the mentioned disadvantages of tree-based approaches but also be tuned to provide a good trade-off between minimum-cost transmission and node residual energy [13], thus effectively prolong the network lifetime by re-clustering as a load-balance mechanism. Our cluster-based data aggregation approach is also very scalable and simple to implement in a real environment. By experiments, we see that a small number of iterations are required until the algorithm converges. We also show that the network lifetime is further enhanced compared with an optimal tree-based Minimum Energy Gathering Algorithm (MEGA) [6].

The rest of this paper is organized as follows. After discussing about the related work and their limitations in Section II, we state the network model, correlation model; and then the correlated data gathering problem setup in Section III. In Section IV, we introduce a distributed algorithm to establish a minimum cost cluster-based architecture for correlated data gathering with the foreign coding strategy. In Section V, we evaluate the performance of our approach compared with that of MEGA. Finally, we give our conclusion and future work in Section VI.

II. Related Work

In-network data aggregation has been the focus of much research work. An overview of techniques and algorithms developed for optimal in-network data aggregation can be found in the recent survey paper [11]. However, the idea of exploiting data correlation for in-network aggregation in WSNs, especially with joint-entropy coding with explicit communication, has just been considered recently [13], [6], [14], [15], [16].

With tree-based approach, Rickenbach et al. [6] proposed two algorithms. With foreign coding, the authors build a directed minimum spanning tree for node encoding and a Shortest Path Tree (SPT) for routing in a algorithm named MEGA. With self coding, a Shallow Light Tree (SLT) that unifies the properties of Minimum Steiner Tree and SPT. In [16], the authors analyzed the tradeoff between SPT and Traveling Salesman Path (TSP) and proposed five heuristic approximation algorithms, which is proven close to optimal by numerical simulations, for building a correlated data gathering tree that exploit data correlation by using self coding strategy.

Using clustering approach, the authors of LEACH [7] and HEED [8] mainly focus on reducing the overhead and improving scalability but do not consider the existence of spatial data correlation. Wang et al. [12] propose a Distributed Optimal-Compression Clustering protocol for constructing clusters of nodes so that Slepian-Wolf coding can be performed locally within each cluster. The advantage of this approach is the reduction of computation complexity and relay failures that affect on the data reconstruction at the sink because it only requires a localized knowledge of the network structure. However, it is based on a sequential greedy method for solving the clustering problem as an NP-hard Minimum Weight Set Covering problem. This problem requires an unavoidable combinatorial explosion of computations to find the best value of a heuristic function based on the conditional entropies of a set of nodes. Moreover, this is an energy-unaware algorithm thus some nodes with a little amount of remained energy can be chosen as CHs.

To exploit the benefits of clustering, we first develop a minimum-cost distributed clustering algorithm that terminates after a few iterations, we then apply foreign coding for data compression at every CH before relaying them through a SPT in an overlay network of CHs to the sink.

III. Problem Formulation

A. Network Model

Consider N stationary, location-unaware sensors (e.g. MICA motes [17]) uniformly distributed on an area of interest and left-unattended after deployment. They form a network which is represented as a connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ = {1,...,N} is the set of vertices and $\mathcal{E}$ is the set of edges. Each node i produces a reading $\mathcal{R}_i$ and all the readings form a $\mathcal{N}$-dimensional vector of data sources. A fixed sink is located at one end of the area to gather all the data monitored by the sensors. Every node can save energy by using a low transmission power level for short-range communication (intra-cluster communication) but a higher level for relaying the data through a long distance. The weight of an edge (i, j) in the graph represents the transmission cost c_{ij}, which is a function of energy spent for transmitting one unit of data from one node to another depending on the distance between them. Nodes i and j are neighbors if they are connected by an edge, i.e. $(i, j) \in \mathcal{E}$. We define an *open neighbor set* of node i is defined as $\mathcal{N}(i) = \{j|(i, j) \in \mathcal{E}\}$, a *closed neighbor set* of

node i as $\mathcal{N}[i] := \mathcal{N}(i) \cup \{i\}$, and $\mathcal{N}(i)\backslash j$ denotes the set obtained by excluding j from $\mathcal{N}(i)$. Finally, each node i is assigned a cost c_i representing the cost for relaying one unit of data packet to the sink.

B. Correlation Model

In reality, the data observed by one sensor is correlated with the data of its neighbor nodes according to a specific structure. Let the parameter η_{ij} be the compression rate, the reduced data rate by compressing raw data of node i at node j using data available at node j (i.e. foreign coding). This parameter depends on the distance d_{ij} between i and j and the underlaying correlation function. In this paper, we consider two correlation models to prove that our algorithm works well with different kinds of monitored data:

1) The sensing data of the monitored area is assumed to be Gaussian such that the correlation between every pair of nodes can be modeled by a covariance matrix [16]. We use the Power Exponential model for the correlation coefficient such that the reduced data rate can be assigned by the following expression:

$$\eta_{ij} = 1 - e^{-\gamma d_{ij}^2}$$

where γ is a constant indicating the degree of correlation. This correlation model can capture a wide variety of physical phenomenon (i.e. electromagnetic waves) in practice [18] and can simulate various degrees of correlation.

2) The *Inverse Distance* correlation model specified in [6] where the correlation coefficient between two nodes is:

$$\eta_{ij} = 1 - 1/(1 + d_{ij})$$

Each sensor is supposed to estimate the set of η_{ij} (each corresponds to a neighbor node) using distance estimation or through several message exchanges of sensing data [16].

C. Problem Setup

The total cost in a cluster-based data aggregation application is the sum of intra-cluster transmission cost for sending raw data from cluster members to their CHs, and the cost for relaying the compressed data from CHs to the sink(s). Our minimum cost hierarchical architecture problem can be regarded as a *min-sum labeling* problem: *Identifying a subset of nodes in the network to label as CHs, then assigning each of the remaining nodes to the CH with minimum transmission cost, so that the total transmission cost of the data aggregation application is minimized.*

The cluster (or label, used interchangeably) to which a node is assigned can be considered as a hidden variable. Let $X := \{x_1, x_2, ..., x_N\}$ be a set of N such hidden variables, in which, for each i, x_i takes on values (node IDs) in the closed neighbor set $\mathcal{N}[i]$. The transmission cost ζ_i for node i to deliver one unit of data packet to the sink can be the cost c_i if i acts as a CH; or the transmission cost c_{ij} for i to transmit one unit of data to some neighboring tentative CH j plus the relay cost ηc_j for sending the encoded data from j to the sink. Mathematically, the transmission cost of each node now can be defined as:

$$\zeta_i(x_i) \propto \begin{cases} c_i & if\, x_i = i \\ c_{ij} + \eta_{ij} c_j & if\, x_i = j,\, j \in \mathcal{N}(i) \end{cases} \quad (1)$$

where η_{ij} is the compression ratio achieved by foreign coding.

In this paper, we consider multi-hop inter-cluster communication for relaying data to the sink. The estimation of transmission cost from one node to the sink is known to be non-trivial and application specific. However, if we assume geographic routing for the data routing between a pair of sensors, the cost model is simplified by the assumption that the transmission cost is proportional to the hop distance between them; and if the nodes are deployed densely enough, this hop-distance is proportional to the Euclidean distance [19] (see Fig. 1). Therefore, the cost for amplifying the radio signal is assumed to be proportional to a path loss exponent of 2 if the distance is less than a threshold d_0 as in [7]; otherwise, the cost is proportional to the sum of square distances accumulated through multi-hop path. The above equation can be reformulated as follows:

$$\zeta_i(x_i) \propto \begin{cases} \delta_{is}^2 & if\, x_i = i \\ d_{ij}^2 + \eta_{ij} \delta_{js}^2 & if\, x_i = j,\, j \in \mathcal{N}(i) \end{cases} \quad (2)$$

where d_{ij} is the Euclidean distance between node i and node j, $\delta_{js}^2 = \lfloor d_{js}/d_0 \rfloor \times d_0^2 + (d_{js} - \lfloor d_{js}/d_0 \rfloor \times d_0)^2$ in which d_{js} is the Euclidean distance between node j and the sink, and $\lfloor x \rfloor$ gives the greatest integer less than or equal to x In a

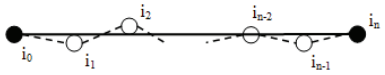


Fig. 1. In a dense network with geographic routing: $d_{i_0 i_n} \propto \sum_{k=1}^{n} d_{i_{k-1} i_k}$

cluster-based WSN, the cluster heads tend to lose their energy more quickly than their member nodes due to heavy network traffic they have to bear. In order to make the network load-balanced, we need to re-elect other high-energy nodes to act as CHs periodically. However, re-election (or re-clustering, used interchangeably) is a waste of energy, thus we need to reduce the re-clustering rate. By scaling the transmission cost function with the node relative residual energy, we can provide a good trade-off between node residual energy and transmission cost for our cluster-based correlated data gathering scheme as below:

$$\zeta_i(x_i) \propto \begin{cases} \frac{e_0}{e_i} \delta_{is}^2 & if\, x_i = i \\ \frac{e_i}{e_j} (d_{ij}^2 + \eta_{ij} \delta_{js}^2) & if\, x_i = j,\, j \in \mathcal{N}(i) \end{cases} \quad (3)$$

where e_0 is the initial energy of a node; e_i, e_j are the residual energy of node i and node j, respectively.

Let the N-tuple $\mathbf{x} := (x_1, x_2, ..., x_N)$ denote the *configuration* (or *labeling*) of the whole network. Since the system is specified via its configuration, this approach is also known as *behavioral modeling*. The N-tuple $\mathbf{x}$ can be a *valid* or *invalid*

configuration [20]. For example, if node i selects j as its CH (i.e., $x_i = j$), but node j is not correctly labeled as a CH (e.g., $x_j = k \neq j$), then this is an invalid configuration. We use the constraint function $\theta_i(x_i, x_1^i, ..., x_I^i)$ to enforce valid configurations between the label x_i of sensor i and the labels of its 1-hop neighboring nodes, denoted as $x_{1:I}^i$, with $I = |\mathcal{N}(i)|$. For the min-sum configuration problem, the constraint function gives a penalty of 0 or ∞ for a valid or invalid configuration respectively, defined as follows:

$$\theta_i(x_i, x_1^i, ..., x_I^i) := \begin{cases} \infty, & \text{if } x_i \neq i \text{ but } \exists i' \in \mathcal{N}(i) : x_{i'}^i = i \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

It is worth noting that the constraint functions just serve as a mathematical modeling of the problem at hand; they do not put any burdens on the computation of the algorithm. The problem of choosing a minimum-cost hierarchical architecture for data aggregation now becomes the problem of finding the min-sum labeling among the *valid* configurations, defined as:

$$\mathbf{x}_{opt} := \arg\min_{\mathbf{x}} \left[\sum_{i \in \mathcal{V}} \zeta_i(x_i) + \sum_{i \in \mathcal{V}} \theta_i(x_i, x_1^i, ..., x_I^i) \right] \quad (5)$$

It is known that exactly minimizing the overall cost is computationally intractable, since a special case of this problem is the NP-hard k-mean problem in data clustering; for large problem we can only find approximate solutions which are heuristic in nature. We propose a new approach for finding a near-optimal solution by recursively applying the min-sum message-passing algorithm [21].

IV. LOCDA: Lifetime Optimized Correlated Data gAthering with Foreign Coding in Clustered WSNs

In this section, we describe the simplified message-passing rules of our Lifetime Optimized Correlated Data Gathering (LOCDA) protocol. Factor graphs [21] can be used to represent a complicated global function, which can be factored into simpler "local" functions, each of which depends on a subset of the variables. In a factor graph, message-passing algorithms can compute, either exactly or approximately, various function marginalization and maximization using simple message passing rules. More details on factor graph derivation for the min-sum labeling problem given in Eq. 5 and message simplification can be found in our prior work [22]. The two types of messages exchanged between the sensor nodes in the network graph $\mathcal{G}$ are:

- Request message $\mu_{x_i \to \theta_j}$ sent from sensor i to its neighbor $j \in \mathcal{N}[i]$, reflects the accumulated level of *suitability* for sensor i to select neighbor j as its CH, taking into account other neighboring CH candidates j' of i.

$$\mu_{\theta_j \to x_i} = \max \left[0 \ , \ \mu_{x_j \to \theta_j} + \sum_{j' \in \mathcal{N}(j) \setminus i} \min \left(0, \mu_{x_{j'} \to \theta_j} \right) \right] \quad (6)$$

- Reply message $\mu_{\theta_i \to x_j}$ sent from sensor i to its neighbor $j \in \mathcal{N}[i]$, reflects the accumulated level of *willingness* of sensor i to act as a CH for sensor j, taking into account the requests from other neighbors j' of i.

$$\mu_{x_i \to \theta_j} = \zeta_i(j) - \min_{j' \in \mathcal{N}[i] \setminus j} \left[\zeta_i(j') + \mu_{\theta_{j'} \to x_i} \right] \ , \ \forall j \in \mathcal{N}[i] \quad (7)$$

Each request/reply message contains a single number such that it allows a node to marshal all the request and reply messages into a vector message $COMPACT$ and send to all of its neighbors by a single broadcast. The messages can be initialized arbitrarily. In our implementation we initialize them to zeros.

The proposed one-hop clustering algorithm is fully distributed and can be efficiently implemented in real sensors because it involves only simple computations using information available via message broadcast without any routing mechanism. The near optimal set of CHs emerges from this recursive message-passing procedure. At any iteration, each node can evaluate its intermediate CH candidate by identifying the sensor ID in its *closed* neighbor set given the expression:

$$CH_i = \arg\min_{j \in \mathcal{N}[i]} [\zeta_i(j) + \mu_{\theta_i \to x_j}] \quad (8)$$

Each node begins to run the algorithm with a Cluster Head Election process (see Pseudocode 1) and this process is terminated when the maximum number of iterations $maxIter$ is reached. This is a key parameter that needs to be carefully selected in real implementation, since the more number of recursions, the better approximation of the optimal clustering, at the cost of more messages to be exchanged. We found through simulations that $maxIter = 5$ is a reasonable upper bound (see Fig. 2).

PSEUDOCODE 1: CLUSTERHEAD ELECTION

1. **repeat**
2. update out-going messages;
3. broadcast($COMPACT$);
4. collectAllCompactMessages();
5. **until** $TERMINATE$

After the Cluster Head Election period, each CH broadcasts an advertisement to its neighboring nodes. A node, which does not become a CH, then chooses and joins the neighboring CH with least cost as shown in Pseudocode 2.

PSEUDOCODE 2: CLUSTER FORMATION

1. find CH_{temp} using incoming messages;
2. **if** $CH_{temp} = myID$ //myID is a CH node
3. $CH_{final} \leftarrow myID$;
4. announceCH($myID$);
5. collectJoinCluster();
6. **else** //myID is a cluster-member node
7. collectAnnounceCH();
8. $\boldsymbol{CH}_{candidates} \leftarrow \{j | \text{incoming\_announceCH}\}$;
9. $CH_{final} \leftarrow j \in \boldsymbol{CH}_{candidates} | least\_cost$;
10. joinCluster($myID, CH_{final}$);
11. **end**

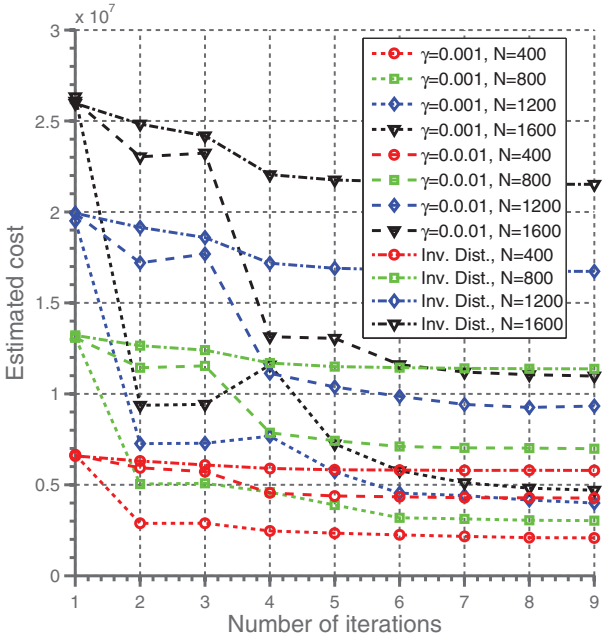


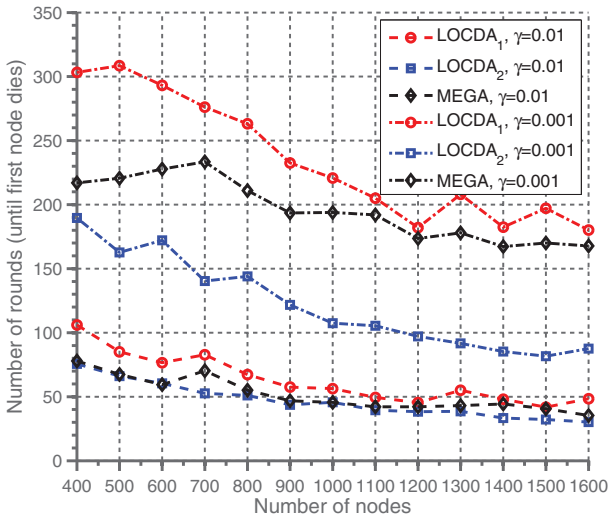
Fig. 2. Estimated Cost vs. Number of Iterations

V. Performance Evaluation and Analysis

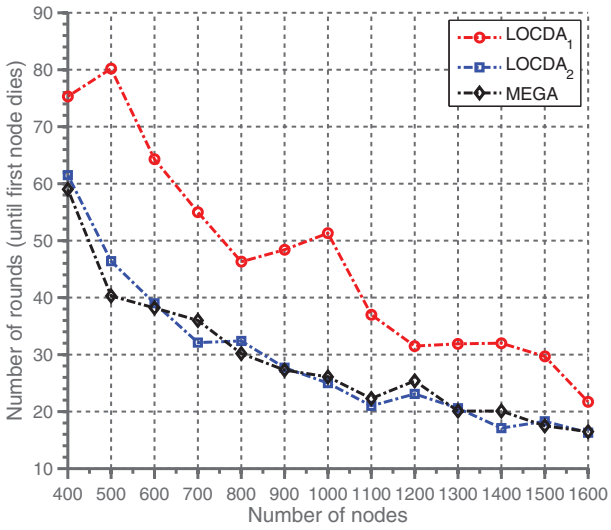
TABLE I
Simulation Parameters

Type	Parameter	Value
Network	Network Size	$(0,0)$ to $(300,300)$
	Sink Location	$(300,300)$
	Initial Energy	$2\ J$
Data Aggregation	Intra-cluster radius	$30\ m$
	Inter-cluster radius	$100\ m$
	Data packet size	$125\ bytes$
Radio Model	ϵ_{fs}	$10\ pJ/bit/m^2$
	ϵ_{mp}	$0.0013\ pJ/bit/m^4$
	E_{elec}	$50\ nJ/bit$
	E_{fusion}	$5\ nJ/bit/signal$
	Distance Threshold (d_0)	$75\ m$
	Max Iteration $maxIter$	5
Power Expo. Corr. Model	$\gamma = 0.001$	$50\ rounds$
	$\gamma = 0.01$	$25\ rounds$
Inv. Dist. Corr. Model		$15\ rounds$

We begin with an analysis of the network lifetime for a setup involving a large number of sensor nodes (400 to 1600 nodes) distributed densely and uniformly in a square area of 300×300 to evaluate the tradeoff between node residual energy and data correlation for building near-optimal but maximum lifetime hierarchical aggregation architecture. We compare our clustering scheme with the Minimum-Energy Gathering Algorithm (MEGA) described [6], which computes a *Minimum Spanning Arborescence* structure for aggregating raw data at intermediate nodes and a *Shortest Path Tree* for relaying the aggregated data from intermediate nodes to the sink. The energy consumption is calculated based on the energy dissipation parameters described in Table I. We choose the short radio range of 30 m and long radio range of 100 m to be compatible with the specification of MICA$_2$ motes. We



(a) Power Exponential Model



(b) Inverse Distance Model

Fig. 3. Network Lifetime comparison between LOCDA$_1$, LOCDA$_2$, and MEGA

consider two versions of our Lifetime Optimized Correlated Data Aggregation protocol (LOCDA) depending on which cost function is used:

1) Function of transmission cost and correlation scaled by the relative node residual energy as shown by expression (3) (LOCDA$_1$)
2) Function of transmission cost and correlation as shown by expression (2) (LOCDA$_2$)

We also vary the simulation parameters such as the node density, the degree of correlation and perform extensive, repeated simulations to fully examine scalability and robustness of our clustering protocol under two different correlation models specified in Section III-B. To provide the tradeoff between residual energy and data correlation, we use a simple re-clustering method for load-balancing, which means that the clustering process is repeated after a fixed number of rounds.

The average network lifetime is computed as the number of rounds until the first node dies. A round is defined as an atomic period during which every node sends its data towards the sink once. In Fig. 3(a), the average network lifetime of MEGA is a little bit higher than that of LOCDA_2 and the average network lifetime of LOCDA_1 is the highest in the case of Power Exponential correlation model with a low degree of correlation $\gamma = 0.01$. In the case of high degree of correlation $\gamma = 0.001$ (Fig. 3(a)) and Inverse Distance correlation model (Fig. 3(b)), the network lifetime of LOCDA_1 is relatively higher than that of MEGA when the number of nodes is not too large ($N \leq 1000$). The reason is that LOCDA_1 can be load-balanced by giving a tradeoff between the transmission cost and node residual energy, resulted in lower re-clustering rate, which is an energy wasted process. However, when the node density is larger, the average network lifetime of LOCDA_1 drops down rapidly and only comes towards the average network lifetime of MEGA. This phenomenon can be explained by this observation: the more member nodes each CH has to serve for data aggregation and relay, the more quickly it dies in spite of re-clustering for load-balancing. Through this analysis, we show that our cluster-based data aggregation approach can achieve good performance in terms of network lifetime and robustness.

VI. Conclusion and Future Work

In this work, we introduce a novel Lifetime Optimized Correlated Data Gathering approach for clustered WSNs. We first build an minimum cost hierarchical architecture in which high-energy nodes are chosen as CHs and then apply foreign coding to reduce data redundancy due to the spatial correlation. By giving a trade-off between node residual energy and minimum cost transmission and utilizing periodical re-clustering as a load-balancing mechanism, we show that our cluster-based approach is better than previous tree-based approach in terms of network lifetime. Our protocol is also very scalable, fully distributed and simple to implement in the real environment. In the future, we intend to develop an adaptive re-clustering algorithm which can be run in a localized, distributed manner to elect new CHs, better achieving load-balancing in the network.

References

[1] K. Yuen, B. Liang, and L. Baochun, "A distributed framework for correlated data gathering in sensor networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 578–593, Jul. 2008.

[2] D. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471– 480, Jul. 1973.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: John Wiley and Sons, Inc., 2006.

[4] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Networked Slepian-Wolf: theory, algorithms, and scaling laws," *IEEE Trans. Inf. Theory*, vol. 51(12), pp. 4057–4073, Dec. 2005.

[5] D. Marco and D. L. Neuhoff, "Reliability vs. efficiency in distributed source coding for field-gathering sensor networks," in *Proc. ACM IPSN'04*, Berkeley, California, USA, Apr. 2004, pp. 161–168.

[6] P. von Rickenbach and R. Wattenhofer, "Gathering correlated data in sensor networks," in *Proc. ACM DIALM-POCM'04*, Philadelphia, Pennsylvania, USA, Oct. 2004.

[7] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1(4), pp. 660–670, Oct. 2002.

[8] O. Younis and S. Fahmy, "HEED: A Hybrid, Energy-Efficient, Distributed clustering approach for ad hoc sensor networks," *IEEE Trans. Mobile Comput.*, vol. 3, no. 4, Oct.-Dec. 2004.

[9] J. Y. Yu and P. H. J. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Communications Surveys and Tutorials*, vol. 7, no. 1, pp. 32–48, 1st Quarter 2005.

[10] O. Younis, M. Krunz, and S. Ramasubramanian, "Node clustering in wireless sensor networks: Recent developments and deployment challenges," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 20–25, May/Jun. 2006.

[11] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: A survey," vol. 14(2), Apr. 2007, pp. 70–87.

[12] P. Wang, C. Li, and J. Zheng, "Distributed data aggregation using clustered slepian-wolf coding in wireless sensor networks," in *Proc. IEEE ICC'07*, Glasgow, Scotland, Jun. 2007, pp. 3616–3622.

[13] S. Pattem, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *Proc. ACM IPSN'04*, 2004, pp. 28–35.

[14] M. Lotfinezhad and B. Liang, "Effect of partially correlated data on clustering in wireless sensor networks," in *Proc. IEEE SECON'04*, Oct. 2004, pp. 172–181.

[15] J. Liu, M. Adler, D. Towsley, and C. Zhang, "On optimal communication cost for gathering correlated data through wireless sensor networks," in *Proc. ACM MobiCom'06*, 2006, pp. 310–321.

[16] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Network correlated data gathering with explicit communication: Np-completeness and algorithms," *IEEE/ACM Trans. Netw.*, vol. 14(1), pp. 41–54, Feb. 2006.

[17] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. E. Culler, and K. S. J. Pister, "System architecture directions for networked sensors," in *Proc. ASPLOS-IX*, Mar. 2000, pp. 93–104.

[18] M. C. Vuran and I. F. Akyildiz, "Spatial correlation-based collaborative medium access control in wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 316 – 329, Apr. 2006.

[19] S. De, "On hop count and Euclidean distance in greedy forwarding in wireless ad hoc networks," *IEEE Commun. Lett.*, vol. 9, no. 11, pp. 1000– 1002, Nov. 2005.

[20] J. C. Willems, "Models for dynamics," in *Dynamics Reported, Volume 2*, U. Kirchgraber and H. O. Walther, Eds. New York: Wiley, 1989, pp. 171–269.

[21] F. R. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Nov. 2001.

[22] H. Q. Ngo, T. M. Tam, Y. K. Lee, and S. Y. Lee, "A message-passing approach to min-cost distributed clustering in wireless sensor networks," in *Proc. IEEE ATC '08.*, Hanoi, Vietnam, Oct. 2008.