Distributed Data Gathering in Multi-Sink Sensor Networks with Correlated Sources

Kevin Yuen, Baochun Li, Ben Liang

Department of Electrical and Computer Engineering
University of Toronto, Ontario, Canada
{yuenke, bli}@eecg.toronto.edu, liang@comm.utoronto.ca

Abstract. In this paper, we propose an effective distributed algorithm to solve the minimum energy data gathering (MEDG) problem in sensor networks with multiple sinks. The problem objective is to find a rate allocation on the sensor nodes and a transmission structure on the network graph, such that the data collected by the sink nodes can reproduce the field of observation, and the total energy consumed by the sensor nodes is minimized. We formulate the problem as a linear optimization problem. The formulation exploits data correlation among the sensor nodes and considers the effect of wireless channel interference. We apply Lagrangian dualization technique on this formulation to obtain a subgradient algorithm for computing the optimal solution. The subgradient algorithm is asynchronous and amenable to fully distributed implementations, which corresponds to the decentralized nature of sensor networks.

Key words: Sensor networks, data correlation, distributed algorithm, minimum energy, optimal rate allocation, transmission structure

1 Introduction

Many applications for sensor networks, such as target tracking [1] and habitat monitoring [2], involve monitoring a remote or hostile field. Sensor nodes are assumed to be inaccessible after deployment for such applications and thus their batteries are irreplaceable. Moreover, due to the small size of sensor nodes, they carry limited battery power. Thus, energy is a scarce resource that must be conserved to the extent possible in sensor networks.

In this context, we are interested in solving the MEDG problem in multi-sink sensor networks with correlated sources. The first part of the problem objective is to find an optimal rate allocation on the sensor nodes, such that the aggregated data received by the sink nodes can be decoded to reproduce the entire field of observation. If the data collected by the sensor nodes are independent, then the rate allocation can be trivially determined – each sensor node can transmit at its data collection rate. However, sensor nodes are often densely deployed in sensor networks, hence the data collected by nearby sensor nodes are either redundant or correlated. This data correlation can be exploited to reduce the amount of data transmitted in the network, resulting in energy savings.

The second part of the problem objective is to find an optimal transmission structure on the network graph, such that the total energy consumed in transporting the data from the sensor nodes to the sink nodes is minimized. If the wireless links have unlimited bandwidth capacities, then each sensor node can transmit its collected data via the minimum energy path. However, as in any practical network, there are capacity limitations on the links and interference among competing signals. As a variation of wireless ad hoc networks, sensor networks have the unique characteristic of location-dependent contention. Signals generated by nearby sensor nodes will compete with each other if they access the wireless shared-medium at the same time. It is shown in [3] that the two parts of the problem objective can be achieved independently if capacity constraints do not exist. But in the presence of capacity constraints, the MEDG problem becomes complicated because the decision on the rate allocation will affect the decision on the transmission structure, and vice versa.

In this paper, we propose an efficient algorithm to solve the MEDG problem. The problem is carefully formulated as a linear optimization problem that can be solved with a distributed solution. This is important since centralized solutions require the participating nodes to repeatedly transmit status information across the network to a central computation node, thus they are not feasible for real-time calculations when energy constraints are present. To design a practical algorithm, we have assumed a realistic data correlation model and considered the effect of location-dependent contention. The formulation is relaxed with Lagrangian dualization technique and solved using the subgradient algorithm. The resulting algorithm is asynchronous, distributed, and supports large-scale sensor networks with multiple sink nodes.

Data gathering with correlated sources in sensor networks and resource allocation with capacity constraints in wireless networks have been separately studied in previous literature. The main contribution of this paper is to propose a solution to the MEDG problem that considers both topics simultaneously, and copes with the dependent relationship between the rate allocation and the transmission structure. To the best of our knowledge, no previous works have addressed the MEDG problem with all of the factors above.

2 Problem Formulation

2.1 Network Model

The wireless sensor network is modeled as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of directed wireless links. Let S_N denote the set of sensor nodes and S_K denote the set of sink nodes. Then, $V = S_N \cup S_K$. The rate allocation assigns each sensor node $i \in S_N$ with R_i, which refers to a non-negative data collection rate. All sensor nodes have a fixed transmission range of r_{tx}. Let d_{ij} denote the distance between node i and node j. A directed link $(i, j) \in E$ exists if $d_{ij} \leq r_{tx}$. Each link is associated with a weight $e_{ij} = d_{ij}^2$, referring to the energy consumed per unit flow on link (i, j). All links are assumed to be symmetrical, where $e_{ij} = e_{ji}$. Moreover, f_{ij} represents the flow rate of link

(i, j). We have assumed that each sensor node has knowledge of its own location. Here, the rate vector $[R_i]_{\forall i \in S_N}$ and the flow vector $[f_{ij}]_{\forall (i,j) \in E}$ are the variables that can be adjusted in order to minimize the following optimization objective.

2.2 Optimization Objective

Given a rate allocation and a transmission structure, the flow rate on each link, denoted by f_{ij}, can be found and the total energy consumed on each link equals to $e_{ij} \cdot f_{ij}$. The objective of the MEDG problem is to minimize the total energy consumed in the network:

$$\textbf{Minimize} \quad \sum_{(i,j) \in E} e_{ij} \cdot f_{ij} \ . \tag{1}$$

2.3 Flow Conservation Constraints

For each sensor node $i \in S_N$, the total outgoing data flows must equal to the sum of the total incoming data flows and the non-negative data collection rate R_i. Since the sensor nodes relay all incoming data flows, only the sink nodes can absorb the data flows.

$$\sum_{j:(i,j) \in E} f_{ij} - \sum_{j:(j,i) \in E} f_{ji} = R_i, \quad \forall i \in S_N \ . \tag{2}$$

2.4 Channel Contention Constraints

The channel contention constraints model the location-dependent contention among the competing data flows. We build the constraints based on the protocol model [4] of packet transmission. According to the protocol model, all links originating from node k will interfere with link (i, j) if $d_{kj} < (1 + \triangle) d_{ij}$, where the quantity $\triangle > 0$ specifies a guard zone. We derive Ψ_{ij} for each link $(i, j) \in E$ as the cluster of links that cannot transmit as long as link (i, j) is active. The notation of cluster is used here as a basic resource unit, as compared to individual links in the traditional wireline networks. In sensor networks, the capacity of a wireless link is interrelated with other wireless links in its cluster. Therefore, data flows compete for the capacity of individual clusters, which is equivalent to the capacity of the wireless shared-medium. A flow vector $[f_{ij}]_{\forall (i,j) \in E}$ is supported by the wireless shared-medium if the channel contention constraints below hold:

$$f_{ij} + \sum_{(p,q) \in \Psi_{ij}} f_{pq} \leq C, \quad \forall (i, j) \in E \ , \tag{3}$$

where C is defined as the maximum rate supported by the wireless shared-medium. Note that the channel contention constraints are generic, since they can accommodate other models of packet transmission instead of the protocol model.

2.5 Rate Admissibility Constraints

Slepian-Wolf coding is introduced in [5]. It is an important work in exploiting data correlation among correlated sources. With Slepian-Wolf coding, sensor nodes are assumed to have correlation information of the entire network, and they can encode their data with only independent information. The Slepian-Wolf region specifies the minimum encoding rate that the sensor nodes must meet in order to transmit all independent information to the sink nodes. It is satisfied when any subset of sensor nodes encode their collected data at a total rate exceeding their joint entropy. In mathematical terms:

$$\sum_{i \in \mathbf{Y}} R_i \geq H(\mathbf{Y}|\mathbf{Y}^C), \quad \mathbf{Y} \subseteq S_N \ . \tag{4}$$

The rate admissibility constraints are non-linear since they grow at an exponential rate in relation to the number of nodes.

Since non-linear constraints are generally difficult to solve, it is desirable to remove them from the formulation. Moreover, the rate admissibility constraints require each sensor node to have global correlation information, which is not scalable in large networks. In this paper, we adapt a localized version of Slepian-Wolf coding from [6] to relax the rate admissibility constraints, such that only local correlation information is required at each sensor node. Here, we describe the localized Slepian-Wolf coding:

- Define a neighbourhood for each sensor node.
- Find the nearest sink node for each sensor node using a distributed shortest path algorithm, such as the Bellman-Ford algorithm [7]. Each sensor node refers to its nearest sink node as the destination sink node.
- For each sensor node i:
  - Find within its neighbourhood, the set N_i of sensor nodes that have the same destination sink node as node i, and are closer to that destination sink node than node i.
  - The Slepian-Wolf region is satisfied when node i transmits at rate $R_i = H(i|N_i)$.

Instead of global correlation information, the localized Slepian-Wolf coding only considers the correlation that a node has with its neighbourhood members. Based on a spatial data correlation model, it is natural to assume the nodes that are not in the neighbourhood contribute very little or nothing to the amount of compression. With a sufficient neighbourhood size, the localized coding should have a performance similar to global Slepian-Wolf coding. In this paper, we include the one-hop neighbours of the sensor nodes in their neighbourhoods.

2.6 Linear Programming Formulation

Combining the optimization objective with the introduced constraints, the MEDG problem can be modeled as a linear programming formulation.

$$\textbf{Minimize} \sum_{(i,j) \in E} e_{ij} \cdot f_{ij} \ , \tag{5}$$

$$\textbf{Subject to:} \quad \sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \ , \tag{6}$$

$$f_{ij} + \sum_{(p,q)\in\Psi_{ij}} f_{pq} \leq C, \quad \forall (i,j) \in E \ , \tag{7}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ . \tag{8}$$

3 Distributed Solution: The Subgradient Algorithm

3.1 Lagrangian Dualization

The MEDG formulation resembles a resource allocation problem, where the objective is to allocate the limited capacities of the clusters to the data flows originating from the sensor nodes. Previous research works in wireline networks [8, 9] have shown that price-based strategy is an efficient mean to arbitrate resource allocation. In this strategy, each link is treated as a basic resource unit. A shadow price is associated with each link to reflect the traffic load of the link and its capacity. Based on the notation of maximal cliques, Xue *et al.* [10] extend the price-based resource allocation framework to respect the unique characteristic of location-dependent contention in wireless networks. Due to the complexities in constructing maximal cliques, the notation of cluster as defined in Section 2 is used as the basic resource unit. Each cluster is associated with a shadow price, and the transmission structure is determined in response to the price signals, such that the aggregated price paid by the data flows is minimized. It is revealed from previous research that at equilibrium, such price-based strategy can achieve global optimum.

To solve the MEDG formulation with a price-based strategy, we relax the channel contention constraints (3) with Lagrangian dualization technique to obtain the Lagrangian dual problem:

$$\textbf{Maximize } \mathrm{LS}(\beta) \ , \quad \textbf{s.t.} \quad \beta \geq 0 \ . \tag{9}$$

By associating price signals or Lagrangian multipliers β_{ij} with the channel contention constraints, the Lagrangian dual problem is evaluated via the Lagrangian subproblem $\mathrm{LS}(\beta)$:

$$\textbf{Minimize} \quad \sum_{(i,j)\in E} e_{ij} \cdot f_{ij} + \beta_{ij} \cdot (f_{ij} + \sum_{(p,q)\in\Psi_{ij}} f_{pq} - C) \ , \tag{10}$$

$$\textbf{Subject to:} \quad \sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \ , \tag{11}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ . \tag{12}$$

We further define Φ_{ij} as the set of clusters that link (i,j) belongs to. Recall Ψ_{pq} is the cluster of links that cannot transmit when link (p,q) is active. For any link (i,j) that interferes with link (p,q), link (i,j) belongs to the cluster of link (p,q).

Thus, for any links (i,j) and (p,q), $(p,q) \in \Phi_{ij}$ iff $(i,j) \in \Psi_{pq}$. The Lagrangian subproblem can be remodelled using this notation:

$$\textbf{Minimize} \sum_{(i,j)\in E} f_{ij}(e_{ij} + \beta_{ij} + \sum_{(p,q)\in\Phi_{ij}} \beta_{pq}) - \beta_{ij}C \ , \tag{13}$$

$$\textbf{Subject to:} \sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \ , \tag{14}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ . \tag{15}$$

The objective function of the remodelled Lagrangian subproblem specifies that the weight of each link is equal to the sum of its energy and capacity cost. And the capacity cost is equal to the Lagrangian multiplier of the link plus the sum of the Lagrangian multipliers in Φ_{ij}. This is intuitive since when link (i,j) is active, any links in the set Φ_{ij} cannot transmit due to interference. So the actual price to pay for accessing link (i,j) should equal to the total price for accessing link (i,j) and all links in Φ_{ij}.

Since the capacity constraints are relaxed, we observe that the solution of the remodelled Lagrangian subproblem requires each sensor node to transmit its data along the shortest path that leads to its nearest sink node. As a result, the Lagrangian subproblem can be solved with a distributed shortest path algorithm, such as the Bellman-Ford algorithm [7]. Recall from the localized Slepian-Wolf coding scheme, a sensor node will co-encode with another sensor node only if they have the identical nearest sink node. Consequently, for any solution generated by the Lagrangian subproblem, data flows due to sensor nodes that have co-encoded with each other will be absorbed by an identical sink node.

3.2 Subgradient Algorithm

Many algorithms have been proposed to solve optimization problems, such as simplex, ellipsoid and interior point methods. These algorithms are efficient in the sense that they can solve large instance of optimization problems in a few seconds. However, they have the disadvantage of being inherently centralized, which implies that they are not applicable for distributed deployment. In this subsection, we describe the subgradient algorithm, a distributed solution to the Lagrangian dual problem.

The algorithm starts with a set of initial non-negative Lagrangian multipliers $\beta_{ij}[0]$. In our simulations, we set $\beta_{ij}[0]$ to zeros, assuming no congestion in the network. During each iteration k, given current Lagrangian multiplier values $\beta_{ij}[k]$, the Lagrangian subproblem is solved. Using the new primal values $[f_{ij}[k]]_{\forall(i,j)\in E}$ obtained from the Lagrangian subproblem, we update the Lagrangian multipliers by:

$$\beta_{ij}[k+1] = \max(0, \beta_{ij}[k] + \theta[k](f_{ij}[k] + \sum_{(p,q)\in\Psi_{ij}} f_{pq}[k] - C)) \ , \tag{16}$$

where θ is a prescribed sequence of step sizes. If the step sizes are too small, then the algorithm has a slow convergence speed. If the step sizes are too large, then

β_{ij} may oscillate around the optimal solution and the algorithm fails to converge. However, the convergence is guaranteed [11], when θ satisfies the conditions $\theta[k] \geq 0$, $\lim_{k\to\infty} \theta[k] = 0$, and $\sum_{k=1}^{\infty} \theta[k] = \infty$. In this paper, we use the sequence of step sizes, $\theta[k] = \frac{a}{b+ck}$, where a, b, and c are positive constants.

The subgradient algorithm is an efficient tool for solving the Lagrangian dual problem. However, it has the disadvantage that an optimal solution, or even a feasible solution to the primal problem (the linear MEDG formulation) may not be available. We adapt the primal recovery algorithm introduced by Sherali *et al.* [11] to recover the primal optimal solution f_{ij}^*. At iteration k of the subgradient algorithm, the primal recovery algorithm composes a primal feasible solution $f_{ij}^*[k]$ via the solutions generated by the Lagrangian subproblem:

$$f_{ij}^*[k] = \sum_{m=1}^{k} \lambda_m^k f_{ij}[m] \ , \tag{17}$$

where $\lambda_m^k = \frac{1}{k}$ are convex weights. In this paper, for each iteration, the Lagrangian subproblem generates a rate allocation and a transmission structure. The primal recovery algorithm specifies that the solution to the MEDG problem (the optimal rate allocation and transmission structure) should equal to a convex combination of the solutions that are generated by the Lagrangian subproblem. Note that since each solution generated by the Lagrangian subproblem satisfies the Slepian-Wolf region, the convex combination of the solutions also satisfies the Slepian-Wolf region. In the kth iteration, we can calculate $f_{ij}^*[k]$ by:

$$f_{ij}^*[k] = \frac{k-1}{k} f_{ij}^*[k-1] + \frac{1}{k} f_{ij}[k] \ . \tag{18}$$

3.3 Distributed MEDG Algorithm

We now present our distributed algorithm for the MEDG problem. Each directed link (i, j) is delegated to its sender node i, and all computations related to link (i, j) will be executed on node i.

1. Choose initial Lagrangian multiplier values $\beta_{ij}[0]$, $\forall (i, j) \in E$.
2. For the kth iteration, determine the weight of each link as $(e_{ij} + \beta_{ij}[k] + \sum_{(p,q)\in\Phi_{ij}} \beta_{pq}[k])$.
3. Compute the shortest path from each sensor node to its nearest sink node using the distributed Bellman-Ford algorithm. Sensor nodes refer to their nearest sink node as their destination sink node.
4. For each sensor node i, determine its rate allocation according to the localized Slepian-Wolf coding scheme introduced in Section 2.
5. Based on the rate allocation and the transmission structure obtained, compute $f_{ij}[k+1]$ and $f_{ij}^*[k+1]$, for all links $(i, j) \in E$.
6. Update Lagrangian multipliers $\beta_{ij}[k+1] = \max(0, \beta_{ij}[k] + \theta[k](f_{ij}[k] + \sum_{(p,q)\in\Psi_{ij}} f_{pq}[k] - C))$, where $\theta[k] = \frac{a}{(b+ck)}$, for all links $(i, j) \in E$.
7. For each link (i, j), send $\beta_{ij}[k+1]$ to all links in Ψ_{ij} and send $f_{ij}[k+1]$ to all links in Φ_{ij}.
8. Repeat steps 2 to 7 until convergence.

4 Performance Evaluation

4.1 Data Correlation Model

Since the sensor nodes are continuous and not discrete sources, the theoretical tool to analyze the problem is Rate Distortion Theory [12]. Let S be a vector of n samples of the measured random field returned by n sensor nodes. Let $\hat{S}$ be a representation of S, and $d(S, \hat{S})$ be a distortion measure. With the mean square error (MSE) as the distortion measure, i.e., $d(S, \hat{S}) = ||S - \hat{S}||^2$, and the constraint $E(||S - \hat{S}||^2) < D$, a Gaussian source is the worst case, since it requires the most bits to be represented when compared with other sources [13]. For the purpose of illustration, we let S be a spatially correlated Gaussian random vector $\sim N(\mu, \Sigma)$. In this case, the rate distortion function of S is

$$R(\Sigma, D) = \sum_{n=1}^{N} \frac{1}{2} \log \frac{\lambda_n}{D_n} \ , \tag{19}$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N$ are the ordered eigenvalues of the correlation matrix Σ and

$$\sum_{n=1}^{N} D_n = D \ , \ D_n = \begin{cases} K & \text{if } K < \lambda_n \ , \\ \lambda_n & \text{otherwise} \ , \end{cases} \tag{20}$$

and K is chosen such that $\sum_{n=1}^{N} \min(K, \lambda_n) = D$. In our analysis, we let $\Sigma_{ij} = W^{d_{ij}^2}$, where W is a correlation parameter that represents the amount of data correlation between spatial samples. W should be less than one such that Σ is a semi-positive definite matrix. Given any subset of nodes X and the distortion per node d, we can construct its correlation matrix Σ_X and approximate its entropy with its rate distortion function, $H(X) \approx R(\Sigma_X, d \cdot |X|)$.

4.2 Simulation Environments

We study the distributed MEDG algorithm in three different simulation environments. In the *independent* environment, we neglect the effect of data correlation by substituting Slepian-Wolf coding with an independent coding scheme. In the *synchronous* environment, the participating nodes simultaneously execute an iteration of the algorithm at every time step. Bounded communication delay is assumed where price and rate updates will arrive at their destinations before the next time step. The *asynchronous* environment is based on the partial asynchronism model [10], which assumes the existence of an integer B that bounds the time between consecutive updates. To implement this environment, each sensor node maintains a timer with a random integer value between 0 and B. The timer decreases itself by 1 at every time step. When the timer reaches 0, the sensor node executes an iteration of the algorithm before resetting the timer. In this environment, update messages may be delayed or out-of-date.

The distributed MEDG algorithm is implemented with the C++ programming language. For all experiments, the transmission and interference range are

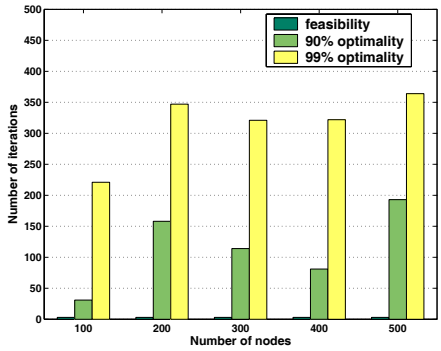

Fig. 1. Convergence speed of the distributed MEDG algorithm.

set to 30m and the capacity of the wireless shared-medium is set to 600 bits. Unless stated, the experiments are executed on a random topology with 100 nodes, the correlation parameter W and the per node distortion d are set to 0.99 and 0.0001, respectively.

4.3 Convergence Behaviour

We study the convergence behaviour of our algorithm under the *synchronous* environment. To this end, we generate five random sensor fields, ranging from 100 to 500 nodes in increments of 100 nodes, with 10% of the nodes randomly chosen as sink nodes. The sensor field with 100 nodes has an area of 100m $\times$ 100m. Other sensor fields are generated by scaling the area to maintain a constant node density. The convergence speed of the algorithm is shown in Fig. 1. The optimal value is taken as the convergence value of the algorithm. We observe that it takes about 220 iterations to converge to 99% optimality in a network with 100 nodes, and this number increases to about 360 for a network with 500 nodes. Due to the slow increase in the number of iterations, the scalability of our algorithm is not affected by the network size. In addition, we notice that the algorithm can achieve 90% optimality in about half the iterations required to achieve 99% optimality. Therefore, in practice, when it is not necessary to achieve the optimal solution, we can obtain a near-optimal solution in a much shorter time. This result illustrates that our distributed algorithm is efficient for real-time calculations.

4.4 Asynchronous Network Environments

To show that our algorithm is applicable in asynchronous network environments, we execute the algorithm under the *asynchronous* environment with different time bounds $B = 1, 2, 5, 10$. Each experiment is performed for 1000 time steps, and the total energy consumption attained at each time step is plotted in Fig. 2.

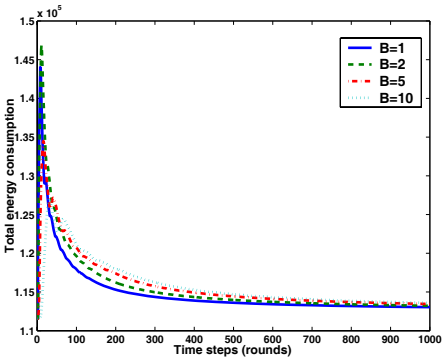

Fig. 2. Convergence in asynchronous network environments.

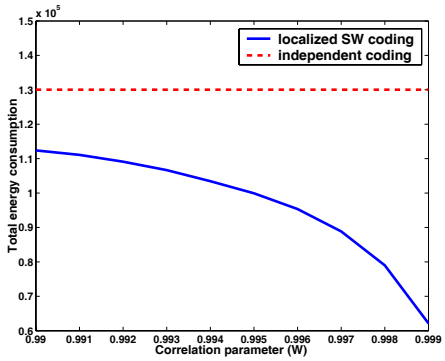

Fig. 3. Localized Slepian-Wolf coding vs. independent coding.

In all four experiments, the algorithm converges to an identical optimal solution, which indicates that it can achieve convergence in asynchronous network environments. Moreover, we conclude that the convergence speed of the algorithm is associated with the time bound B, since longer convergence time is required when B is large.

4.5 The Effect of Data Correlation

We investigate the effect of data correlation by comparing the *asynchronous* environment against the *independent* environment. As the correlation parameter W varies from 0.99 to 0.999, the total energy consumed by the different environments at convergence is recorded in Fig. 3. Clearly, the energy consumed at high correlation ($W = 0.999$) is much lower compared with the energy consumed at low correlation ($W = 0.99$). Overall, the localized Slepian-Wolf coding scheme outperforms the independent coding scheme by 15% to 50%. This result suggests that even though the algorithm utilizes only local information, it can achieve significant energy savings for a wide range of data correlation level.

5 Related Work

In [14], Kalpakis *et al.* have formulated the maximum lifetime data gathering and aggregation problem as an integer program. Although this formulation yields satisfactory performance, it makes the assumption of perfect data correlation, where intermediate sensor nodes can aggregate any number of incoming packets into a single packet. Perfect data correlation can also be found in [15], which analyzes the performance of data-centric routing schemes with in-network aggregation. We do not assume perfect data correlation in this paper since it may not be realistic in practical networks.

While our paper utilizes Slepian-Wolf coding, there are works that exploit data correlation with alternative techniques. Single-input coding is considered in

[3, 16], where intermediate nodes can aggregate their collected data with the side information provided by another node. Cristescu *et al.* [3] prove that solving the MEDG problem with single-input coding is NP-hard, even in a simplified network setting. Since single-input coding can only exploit data correlation between pairs of nodes, it will not perform as well as Slepian-Wolf coding. In contrast, data aggregation with multi-input coding is performed when all input information from multiple nodes is available. Goel *et al.* [17] consider the joint treatment of data aggregation and transmission structure with multi-input coding. Although multi-input coding exploits data correlation among multiple nodes, it requires the nodes to explicitly communicate with each other. Since Slepian-Wolf coding does not require such communication, it can be implemented in asynchronous network environments without timing assumptions.

Other closely related works are the ones involving Slepian-Wolf coding. In [18], Servetto *et al.* introduced the sensor reachback problem, which requires a single node in the sensor network to receive sufficient data to reproduce the entire field of observation. Slepian-Wolf coding is employed to meet this requirement. This paper inspires us to apply Slepian-Wolf coding in the MEDG problem, allowing the sink nodes to receive independent information from all sensor nodes. In [6], Cristescu *et al.* address the MEDG problem with Slepian-Wolf coding. However, their optimization problem does not consider the effect of wireless channel interference, hence the solution generated may not be supported by the wireless shared-medium.

6 Conclusion

In this paper, we have presented an efficient solution for the MEDG problem in multi-sink sensor networks with correlated sources. The problem is carefully formulated as a linear optimization problem with a distributed solution. In the presence of capacity constraints, we show that finding the optimal rate allocation and transmission structure are two dependent problems, hence they must be addressed simultaneously. With a realistic model, the formulation exploits data correlation among the sensor nodes, accounts for location-dependent contention in the wireless shared-medium, and minimizes the total energy consumed by the network. Sensor nodes are required to transmit at a rate that satisfies the Slepian-Wolf region, which implies that the sink nodes will be able to reproduce the entire field monitored by the sensor network. The algorithm is amenable to fully distributed implementations, as the participating nodes are only needed to communicate with other nodes in their neighbourhood. The algorithm is asynchronous and provides multi-sink support, making it feasible for practical deployment in large-scale sensor networks. To the best of our knowledge, this is the first work that addresses the MEDG problem with data correlation and wireless channel interference simultaneously, especially when a price-based approach is employed to obtain a distributed solution.

References

1. Clouqueur, T., Phipatanasuphorn, V., Ramanathan, P., Saluja, K.K.: Sensor Deployment Strategy for Detection of Targets Traversing a Region. In: ACM Mobile Networks and Applications. Volume 8. (2003) 453–461
2. Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D.: Wireless Sensor Networks for Habitat Monitoring. In: Proc. of First ACM International Workshop on Wireless Sensor Network and Applications. (2002)
3. Cristescu, R., Beferull-Lozano, B., Vetterli, M.: On Network Correlated Data Gathering. In: Proc. of IEEE INFOCOM. (2004)
4. Gupta, P., Kumar, P.R.: The Capacity of Wireless Networks. IEEE Trans. Information Theory **46**(2) (2000) 388–404
5. Slepian, D., Wolf, J.K.: Noiseless Coding of Correlated Information Sources. IEEE Trans. on Information Theory **4**(IT-19) (1973) 471–480
6. Cristescu, R., Beferull-Lozano, B., Vetterli, M.: Networked Slepian-Wolf: Theory and Algorithms. In: Proc. of European Workshop on Wireless Sensor Networks. (2004)
7. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice Hall (1989)
8. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate Control in Communication Networks: Shadow prices, Proportional Fairness and Stability. In: Journal of the Operational Research Society. Volume 49. (1998) 237–252
9. Low, S.H., Lapsley, D.E.: Optimization Flow Control: Basic Algorithm and Covergence. In: IEEE/ACM Trans. on Networking. Volume 7. (1999) 861–874
10. Xue, Y., Li, B., Nahrstedt, K.: Optimal Resource Allocation in Wireless Ad Hoc Networks: A Price-Based Approach. In: to appear in IEEE Transactions on Mobile Computing. (2005)
11. Sherali, H.D., Choi, G.: Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs. In: Operations Research Letter. Volume 19. (1996) 105–113
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. New York: Wiley (1991)
13. Lotfinezhad, M., Liang, B.: Effect of Partially Correlated Data on Clustering in Wireless Sensor Networks. In: Proc. of the IEEE International Conference on Sensor and Ad hoc Communications and Networks (SECON). (2004)
14. Kalpakis, K., Dasgupta, K., Namjoshi, P.: Efficient Algorithms for Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks. Computer Networks Journal (2002)
15. Krishnamachri, B., Estrin, D., Wicker, S.: Modelling Data-centric Routing in Wireless Sensor Networks. In: Proc. of IEEE INFOCOM. (2002)
16. Rickenbach, P.V., Wattenhofer, R.: Gathering Correlated Data in Sensor Networks. In: Proc. of DIALM-POMC '04: Proceedings of the 2004 joint workshop on Foundations of mobile computing. (2004) 60–66
17. Goel, A., Estrin, D.: Simultaneous Optimization for Concave Costs: Single Sink Aggregation or Single Source Buy-at-Bulk. In: Proc. of the 14^{th} Symposium on Discrete Algorithms (SODA). (2003)
18. Barros, J., Servetto, S.D.: Network Information Flow with Correlated Sources. Submitted to the IEEE Transactions on Information Theory, November 2003 (Original title: The Sensor Reachback Problem) Revised (2005)