A Distributed Framework for Correlated Data Gathering in Sensor Networks

Kevin Yuen, Ben Liang, Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario, M5S 3G4, Canada
yuenke@eecg.toronto.edu, liang@comm.utoronto.ca, bli@eecg.toronto.edu

Abstract

We consider the problem of correlated data gathering in sensor networks with multiple sink nodes. The problem has two objectives. First, we would like to find a rate allocation on the correlated sensor nodes, such that the data gathered by the sink nodes can reproduce the field of observation. Second, we would like to find a transmission structure on the network graph, such that the total transmission energy consumed by the network is minimized. The existing solutions to this problem are impractical for deployment because they have not considered all of the following factors: (i) distributed implementation, (ii) capacity and interference associated with the shared-medium, and (iii) realistic data correlation model. In this paper, we propose a new distributed framework to achieve minimum energy data gathering, while considering these three factors. Based on a localized version of Slepian-Wolf coding, the problem is modeled as an optimization formulation with a distributed solution. The formulation is first relaxed with Lagrangian dualization and then solved with the subgradient algorithm. The algorithm is amenable to fully distributed implementations, which corresponds to the decentralized nature of sensor networks. To evaluate its effectiveness, we have conducted extensive simulations under a variety of network environments. The results indicate that the algorithm supports asynchronous network settings, sink mobility, and duty schedules.

Index Terms

Sensor networks, data correlation, distributed algorithm, minimum energy, optimal data gathering, sink mobility.

I. Introduction

Recent technological advances have enabled the production of low-cost sensor nodes. These sensor nodes are small in size, and are equipped with limited sensing, processing, and transmission capabilities. They can be deployed in large numbers to construct a sensor network with the ability of distributed wireless sensing. The collaborative effort of these sensor nodes can achieve significant improvement over traditional sensors due to their improved accuracy and ease of deployment. In practice, sensor nodes are densely deployed in an ad hoc fashion over the area of interest. After their deployment, the sensor nodes collect data from their surroundings, encode the data, and transmit them to the sink nodes via wireless channels. In addition to collecting data, intermediate sensor nodes can be used as relays for other sensors distant from the sink nodes. Sink nodes are specialized nodes that are responsible for gathering collected data and serve as gateways between the sensor network and the wired or wireless backbone network.

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.

A. Problem Description and Design Goals

In this context, the first objective of the correlated data gathering problem is to find a rate allocation on the sensor nodes, such that the aggregated data collected by the sink nodes can be decoded to reproduce the field of observation. The rate allocation assigns each sensor node an encoding rate, which is equivalent to its data transmission rate. If the data collected by the sensor nodes are statistically independent, then the rate allocation can be trivially determined — each sensor node can transmit at its data collection rate. However, sensor nodes are densely deployed in sensor networks. Nearby sensor nodes have overlapping sensing ranges and their collected data 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. To achieve minimum energy data gathering, the optimal rate allocation should minimize the encoding rates, while ensuring the rates are sufficient to represent all independent data generated by the sensor nodes.

The second objective of the correlated data gathering problem is to find a transmission structure on the network graph, such that the total energy consumed in transporting the collected data from the sensor nodes to the sink nodes is minimized. If the network has unconstrained bandwidth capacity, then this objective can be simply achieved – each sensor node can transmit its collected data via the minimum energy path. However, in any practical networks, there are capacity limitation on the transmission medium and interference among the competing signals. In wireline networks, there is time-dependent contention, where two signals compete with each other if they both arrived at a receiver at the same time. The effect of interference in wireline networks is well studied, but they are not applicable in the context of sensor networks. As a variation of wireless ad hoc networks, sensor networks have the unique characteristic of location-dependent contention in addition to time-dependent contention. Signals will compete with each other if multiple sensor nodes in nearby vicinity access the wireless shared-medium simultaneously. To derive feasible solutions, the capacity and the interference associated with the shared-medium must be considered when constructing the optimal transmission structure.

It is shown in [3] that if the bandwidth capacity of the network is unconstrained, the two problem objectives can be achieved independently in two steps. First, according to the correlation model, the optimal rate allocation can be determined. Then the optimal transmission structure can be constructed by combining the minimum energy paths of the sensor nodes. However, when capacity constraints exist, the problem becomes complicated because the two objectives are dependent. Given a transmission structure, if the rate allocation is modified, then some of the links selected by the transmission structure may become congested due to the increased traffic flows. To alleviate this congestion, the transmission structure has to be adjusted. On the other hand, there are different coding schemes that exploit data correlation in the literature. They can be generally divided into two categories, which are distributed source coding and joint entropy coding with explicit communication. In practice, coding schemes from both categories determine the rate allocation based on a given transmission structure. Consequently, the decision on the rate allocation affects the decision on the transmission structure, and vice versa. One of the highlights of this paper is to take this dependence into account, and design an algorithm that jointly optimizes the rate allocation and the transmission structure simultaneously, while satisfying the capacity constraints.

In addition to the problem objectives, we have included several design goals when constructing the framework. The ultimate purpose of this paper is to create a solution to the correlated data gathering problem that is practical for deployment. More importantly, the framework should be compatible, allowing other energy-saving mechanisms to be built on top of the framework to further extend the lifetime of data gathering sensor networks.

- *Multi-sink support:* To facilitate efficient data gathering, it is envisioned that future sensor networks will consist of multiple sink nodes. By providing multi-sink support, the framework becomes feasible for deployment in large-scale sensor networks.
- *Distributed solutions:* With centralized solutions, participating nodes need to transmit detailed status information repeatedly across the network to a central computation node. Although centralized approaches can generate results closer to the global optimum, they are generally not feasible in energy-constrained sensor networks.
- *Asynchronous network settings:* Due to the ad hoc infrastructure of sensor networks, it is expensive in terms of communication overhead to synchronize the nodes. If the framework is applicable in asynchronous network settings, it can avoid the scaling limitation posed by synchronous solutions.
- *Sink mobility:* Because of its multi-hop nature, the appearance of energy holes in static sensor networks seems unavoidable. Sensor nodes positioned around the sink nodes deplete their energy faster because they are frequently acting as relays. A natural way to counter energy holes is to introduce sink mobility, where sink nodes move within the network as they gather data from the

sensor nodes.

- *Duty schedules:* To achieve maximum network lifetime, load balancing among the sensor nodes must be enforced. This can be accomplished with the introduction of duty schedules, where sensor nodes switch their operating status (on/off) to control and match their energy consumption rates. However, duty schedules give rise to network dynamics, since the sensor nodes may join and leave the network at run-time.

B. Main Contributions

Data gathering with correlated sources in sensor networks, and resource allocation with capacity constraints in wireless links, were studied separately in past literature. The main contribution of this paper is to propose a solution to the data gathering problem that considers both topics simultaneously. The proposed solution copes with the dependence that exists between the two problem objectives, as it jointly optimizes the rate allocation and the transmission structure. Furthermore, the optimization formulation is specifically designed to have a distributed solution.

Since the aim of the problem is to minimize energy consumption, it is a natural decision to employ optimization techniques. We model the problem as a non-linear optimization formulation. According to the protocol model [4] of packet transmission in wireless networks, the formulation considers the capacity limitation of the network and the effect of location-dependent contention. As a result, our solution is guaranteed to be supported by the wireless shared-medium. Since non-linear formulations are generally difficult to solve, we relax the formulation to become linear by adapting a localized version of Slepian-Wolf coding. Based on Lagrangian dualization, we utilize a price-based resource allocation strategy and solve the formulation with the subgradient algorithm. Price signals are communicated among the sensor nodes to reflect the congestion status of the network. The subgradient algorithm is amenable to distributed implementations, making it feasible for practical deployment. Moreover, we conduct extensive simulations to validate that our solution supports asynchronous network settings, sink mobility, and duty schedules. To the best of our knowledge, no previous works have addressed the correlated data gathering problem considering all of the factors above.

C. Paper Organization

The remainder of this paper is organized as follows. In Section II, we present the non-linear optimization formulation for the correlated data gathering problem. In Section III, we describe the localized Slepian-Wolf coding scheme and present the linear formulation. In Section IV, we construct an efficient distributed algorithm to solve the formulation with Lagrangian dualization and the subgradient algorithm. In Section

V, we discuss implementation issues related to the algorithm. Numerical results from simulations are presented in VI. Finally, we discuss related work in Section VII, and conclude the paper in Section VIII.

II. Problem Formulation

A. Network Model and Assumptions

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. S_N denotes the set of sensor nodes and S_K denotes 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 transmission rate. All sensor nodes have a fixed transmission range 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). 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 optimization objective.

There are various models for sensor networks. In this work, we mainly focus on a sensor network environment where:

- A spatial data correlation model [5] is assumed, where the sensor nodes can achieve various amount of data aggregation based on their distance of separation. In contrast, a perfect data correlation model is assumed in [6], [7], and [8], where intermediate sensor nodes can aggregate any number of incoming packets into a single packet. Although the perfect data correlation model can represent higher energy savings, it is generally not practical in most application scenarios.
- The transmission power is automatically managed by the sensor nodes. During a transmission, the sensor nodes have the ability to adjust their transmission power depending on the distance of transmission. Consequently, the energy consumed per unit flow on a link is a function of its distance. Moreover, the transmission power is assumed to be allocated in a specific way, such that the capacity of the wireless shared-medium is constant across the entire network. Power allocation is out of the scope of this paper and is left as a future research direction.
- Depending on the application of the sensor network, its data delivery model can be continuous, event-driven, or query-driven [9]. We have assumed a continuous data delivery model for illustration, where the sensor nodes periodically sense their surroundings and always have data to transmit in each round of communication. In the event-driven or query-driven delivery model, data are transmitted to the sink nodes when the sensor nodes detects an event or receives a query. We emphasize that, since our proposed solution supports duty schedules, it can be easily extended to accommodate these data

delivery models.

- The objective of the correlated data gathering problem is to minimize the total transmission energy consumed by the network. While this objective does not guarantee to maximize the lifetime of each individual sensor node, it can achieve better energy efficiency, thus extending the network lifetime. In this paper, sensor networks are assumed to have a high density of sensor nodes. This implies that the failure of an individual sensor node (possibility due to energy exhaustion) will not have a critical impact on the coverage or connectivity of the network. Moreover, our solution can be combined with load balancing mechanisms to achieve fairness in energy consumption.

B. Data Correlation Model

Since the sensor nodes are continuous and not discrete sources, the theoretical tool to analyze the problem is Rate Distortion Theory [10] [11]. 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 $d(S, \hat{S}) = ||S - \hat{S}||^2$ and with the constraint

$$E(||S - \hat{S}||^2) < D \ , \tag{1}$$

a Gaussian source is the worst case and needs most bits to be represented compared with other sources [10]. For the purpose of illustration, we let S be a spatially correlated random Gaussian vector $\sim N(\mu, \Sigma)$. In this case, the rate distortion function of S is

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

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

$$D_n = \min(K, \lambda_n) \ , \tag{3}$$

and K is chosen such that

$$\sum_{n=1}^{N} D_n = D \ . \tag{4}$$

This is known as “reverse water filling” [11]. In our analysis, we express the amount of data correlation as a function of the distance between the sensor nodes. Particularly, we let $\Sigma_{ij} = W^{d_{ij}^2}$, where W is a correlation parameter that represents the amount of 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, $H(X) \approx R(\Sigma_X, d \cdot |X|)$.

C. Optimization Objective

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

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

In addition to transmission energy, the objective can be modified to optimize other metrics of interest with the structure [link cost] $\times$ [data size]. Similar optimization objective can be found in [3].

D. Flow Conservation Constraints

For each sensor node $i \in S_N$, the total outgoing traffic flows must equal to the sum of the incoming traffic flows and the non-negative data transmission rate allocated to node, R_i. These constraints enforce lossless transmission, which implies no data packets will be discarded by any intermediate sensor nodes. In this paper, sensor nodes utilize Slepian-Wolf coding to exploit data correlation. As a result, all data packets generated by the sensor nodes contain independent data and they must be received by the sink nodes.

$$\sum_{j:(i,j) \in E} f_{ij} - \sum_{j:(j,i) \in E} f_{ji} = R_i, \quad \forall i \in S_N \ . \tag{6}$$

E. Channel Contention Constraints

To generate solutions that are supported by the wireless shared-medium, we introduce channel contention constraints in our formulation. The purpose of these constraints is to model the location-dependent contention that exists among the competing data flows. To accomplish this task, we need to identify when a transmission is successfully received by its intended recipient. In the literature, there exists two models for packet transmission in wireless networks [4]. They are generally referred to as the *protocol model* and the *physical model*, and they are presented as follows:

- *The Protocol model*: This model determines if a packet transmission is successful by considering the spatial location of the nodes. A packet transmission from node i to j is successful if for all node k with $d_{kj} < (1 + \triangle) d_{ij}$, node k is not transmitting. The quantity $\triangle > 0$ specifies a guard zone. In this paper, the interference range is assumed to be identical to the transmission range. Thus, $\triangle = 0$.
- *The Physical model*: This model is related to the physical layer and considers the signal power received at the receiver node. A packet transmission from node i to j is successful if the signal to interference ratio (SIR) is greater than a threshold, $SIR_{ij} \geq SIR_{thresh}$.

In this paper, we focus on the protocol model of packet transmission. Based on the protocol model, any links originating from node k will interfere with link (i,j) if $d_{kj} < (1+\triangle)d_{ij}$. Utilizing this model, 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. Here, the notation of cluster is treated as a basic resource unit, as compared to individual links in traditional wireline networks. In wireline networks, data flows compete for the capacity of individual links. However, in the case of sensor networks, the capacity of a wireless link is interrelated with other wireless links in its vicinity. Consequently, 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 [12].

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

where C is defined as the maximum rate supported by the wireless shared-medium.

In practice, there are various methods that can be employed to construct the clusters [13]. For instance, if each node is provided with its own location information, in coordinates or in relation to the other nodes, then the clusters can be formed by considering the protocol model. An alternative is for a node to form local topology knowledge based on overheard transmissions in its surroundings. The exact method used to construct the clusters is beyond the scope of this paper.

In addition to the protocol model of packet transmission, the channel contention constraints can be simply tailored to adapt a particular MAC protocol by adjusting the derivation of the clusters Ψ_{ij}. For instance, in an IEEE 802.11 MAC protocol based network, if link (i,j) is active, then Ψ_{ij} should include all links that are originating from node k that satisfies $d_{kj} < (1+\triangle)d_{ij}$ or $d_{ki} < (1+\triangle)d_{ij}$. The sending node i is also required to be free of interference since it needs to receive the link layer acknowledgments from the receiving node j.

F. Rate Admissibility Constraints

Due to data correlation, data collected by nearby sensor nodes are often redundant. Since transmitting redundant data across the network consumes unnecessary energy and decreases the useful throughput of the network, it is desirable to eliminate all redundancy. In the literature, there are many coding schemes that can be employed to exploit data correlation. They can be generally divided into two categories, which are distributed source coding and joint entropy coding with explicit communication [3]. For coding with explicit communication, sensor nodes aggregate their data with side information received from other nodes. In this scenario, it is shown that the optimal rate allocation can be simply determined since it only relies on the side information, but building the optimal transmission structure becomes NP-hard. In contrast, distributed source coding allows each sensor node to generate independent data packets, assuming the sensors have knowledge of the global correlation structure. Although distributed source coding requires increased coding complexity and knowledge of the correlation structure, it is theoretically the most efficient coding scheme. It can achieve maximum energy savings for lossless transmission, since no redundant data are ever transmitted. Moreover, it can be implemented in distributed, asynchronous network environments. Therefore, we employ distributed source coding to solve the correlated data gathering problem.

We employ Slepian-Wolf coding as introduced in [14], which is a fundamental research study in distributed source coding. The Slepian-Wolf region specifies the minimum encoding rates that the sensor nodes must meet in order to transmit all independent data 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:

$$\sum_{i \in \mathbf{Y}} R_i \geq H(\mathbf{Y}|\mathbf{Y}^C), \quad \mathbf{Y} \subseteq S_N \ . \tag{8}$$

G. Non-Linear Programming Formulation

Combining the optimization objective with the introduced constraints, the correlated data gathering problem can be modeled as an optimization problem.

Minimize

$$\sum_{(i,j) \in E} e_{ij} \cdot f_{ij} \ . \tag{9}$$

Subject to:

$$\sum_{j:(i,j) \in E} f_{ij} - \sum_{j:(j,i) \in E} f_{ji} = R_i, \quad \forall i \in S_N \ , \tag{10}$$

$$f_{ij} + \sum_{(p,q) \in \Psi_{ij}} f_{pq} \leq C, \quad \forall (i,j) \in E \ , \tag{11}$$

$$\sum_{i \in \mathbf{Y}} R_i \geq H(\mathbf{Y}|\mathbf{Y}^C), \quad \mathbf{Y} \subseteq S_N \ , \tag{12}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ , \tag{13}$$

$$R_i \geq 0, \quad \forall i \in S_N \ . \tag{14}$$

Since the rate admissibility constraints grow at an exponential rate in relation to the number of nodes, this optimization formulation is non-linear. In the following sections, we introduce a linear re-formulation of this problem, through localized Slepian-Wold coding, and further propose a price-based framework to provide a solution that is distributed among the individual sensor nodes.

TABLE I

LOCALIZED SLEPIAN-WOLF CODING

1) Define a neighbourhood for each sensor node.
2) Find the nearest sink node for each sensor node using a distributed shortest path algorithm, such as the Bellman-Ford algorithm [16], [17]. Each sensor node refers to its nearest sink node as its destination sink node.
3) For each sensor node i:
   a) 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.
   b) The Slepian-Wolf region is satisfied when node i transmits at rate $R_i = H(i|N_i)$.

III. LOCALIZED SLEPIAN-WOLF CODING

Non-linear optimization formulations are generally difficult to solve, hence it is desirable to remove the non-linear constraints from the formulation. Moreover, the rate admissibility constraints require each sensor node to have knowledge of the global correlation structure. This poses limitation on the scalability of our solution. In this paper, we adopt an approximated version of Slepian-Wolf coding from [15] to relax the rate admissibility constraints, such that only local correlation information is required at each sensor node. The approximation gives a definition for a neighbourhood. For each sensor node, its neighbourhood contains other sensors that are located in its surroundings. When a sensor node is determining its data transmission rate, it considers its data correlation with other sensors in its neighbourhood, instead of the entire network. Based on the spatial data correlation model, it is natural to assume the sensors that are not in the neighbourhood contribute very little or nothing in reducing the transmission rates. With a sufficient neighbourhood size, this approximation should have a performance comparable to global Slepian-Wolf coding. In this paper, we include the one-hop neighbours of the sensor nodes in their neighbourhoods.

Extending from the approximation, we present a localized Slepian-Wolf coding scheme in Table I. This coding scheme supports sensor networks with multiple sinks, and it is amenable to distributed implementation. The localized Slepian-Wolf coding scheme specifies that each sensor node i should encode its data at a rate equal to the conditioned entropy. The conditioning is performed only on N_i, a subset of sensors within the neighbourhood of sensor i that are closer to sensor i's destination sink node than sensor i itself. Instead of the global correlation structure, sensor nodes using this coding scheme are only required to have knowledge of the correlation structure within their neighbourhoods. As a result, the localized Slepian-Wolf coding scheme overcomes the scalability limitation imposed by global Slepian-Wolf coding.

The performance of the localized Slepian-Wolf coding scheme depends on the transmission structure. The sensor nodes must realize their destination sink nodes before they can determine their achievable data transmission rates. If the capacity of the network is unconstrained, then the sensor nodes can simply determine their destination sink nodes based on relative spatial information. The sink node that is located closest to the sensor will be chosen as its destination sink node. However, when capacity constraints exist, a sensor node may not be able to transmit its collected data to its closest sink node due to data congestion. To avoid data congestion, our solution allows the sensor nodes to switch their destination sink nodes during run-time. Hence, the transmission structure is dynamic as it is adjusted according to the rate allocation and the data congestion experienced by the wireless links. On the other hand, to accommodate the dynamic transmission structure, the localized Slepian-Wolf coding scheme dynamically determines the appropriate rate allocation during run-time. Consequently, our solution jointly optimizes the rate allocation and the transmission structure, which are dependent upon each other. We believe this approach will provide substantial improvements over the traditional approaches in solving the correlated data gathering problem.

It is now possible to model the correlated data gathering problem as a linear programming formulation. The non-linear rate admissibility constraints are relaxed, but the Slepian-Wolf region is still satisfied. The sensor nodes are required to transmit at the conditioned entropy specified by the localized Slepian-Wolf coding scheme. The linear programming formulation, also denoted as the *primal problem*, is expressed as follows:

Minimize

$$\sum_{(i,j)\in E} e_{ij} \cdot f_{ij} \quad . \tag{15}$$

Subject to:

$$\sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \quad , \tag{16}$$

$$f_{ij} + \sum_{(p,q)\in\Psi_{ij}} f_{pq} \leq C, \quad \forall (i,j) \in E \quad , \tag{17}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \quad . \tag{18}$$

IV. Distributed Solution: A Price-Based Framework

Many algorithms have been proposed in past literature to solve linear optimization formulations, such as simplex, ellipsoid, and interior point methods. These algorithms are efficient in the sense that they can solve large instance of optimization formulations in a few seconds. However, they have the disadvantage of being inherently centralized, which implies that they are not applicable for distributed implementations. In this section, we present our distributed solution to the proposed linear optimization formulation. The formulation is relaxed with Lagrangian dualization, and then solved with the subgradient algorithm. In addition, we discuss the asynchronous network model that is utilized in this paper.

A. Lagrangian Dualization

With the localized Slepian-Wolf coding scheme, we are able to determine the optimal rate allocation. Our next step towards solving the linear programming formulation is to obtain the optimal transmission structure, given the rate allocation. This part of the problem resembles a resource allocation problem, where the goal is to allocate the limited capacity of the wireless shared-medium to the data flows originating from the sensor nodes.

In the literature, [18] and [19] have shown that price-based resource allocation strategy is an efficient means to arbitrate resource allocation in wireline networks. With price-based strategy, prices are computed as signals to indicate the relation between the supplies and demands of a resource. In these works, each wireless link is treated as a basic resource unit. A shadow price is associated with each wireless link to reflect the relation between the traffic load of the link and its bandwidth capacity. Based on the notation of maximal cliques, Xue *et al.* [20] extend the price-based resource allocation framework to respect the unique characteristic of location-dependent contention in wireless ad hoc networks.

In this paper, the notation of clusters as defined in Section II is utilized as the basic resource unit. Each cluster is associated with a shadow price, and the signals compete for the capacity of the clusters. 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 shown from the previous works that at equilibrium, such price-based resource allocation strategy can achieve global optimum, leading to the optimal utilization of the resource. To solve the linear programming formulation with a price-based strategy, we first relax the channel contention constraints (7) with Lagrangian dualization and obtain the Lagrangian dual problem:

$$\textbf{Maximize } \text{LS}(\beta) \ , \quad \textbf{Subject to:} \quad \beta \geq 0 \ . \tag{19}$$

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

Minimize

$$\sum_{(i,j)\in E} e_{ij} \cdot f_{ij} + \beta_{ij} \cdot (f_{ij} + \sum_{(p,q)\in \Psi_{ij}} f_{pq} - C) \ . \tag{20}$$

Subject to:

$$\sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \ , \tag{21}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ . \tag{22}$$

Here, we introduce a new notation Φ_{ij} as the set of clusters that link (i,j) belongs to. Recall Ψ_{pq} refers to 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 $(i,j) \in \Psi_{pq}$. Then, the Lagrangian subproblem can be remodelled using this notation:

Minimize

$$\sum_{(i,j)\in E} f_{ij}(e_{ij} + \beta_{ij} + \sum_{(p,q)\in\Phi_{ij}} \beta_{pq}) - \beta_{ij} C \ . \tag{23}$$

Subject to:

$$\sum_{j:(i,j)\in E} f_{ij} - \sum_{j:(j,i)\in E} f_{ji} = H(i|N_i), \quad \forall i \in S_N \ , \tag{24}$$

$$f_{ij} \geq 0, \quad \forall (i,j) \in E \ . \tag{25}$$

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 its 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. Thus, the actual cost for accessing link (i,j) should equal to total cost for accessing link (i,j) and all the 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 any distributed shortest path algorithm, such as the well-known Bellman-Ford approach. 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.

B. Subgradient Algorithm

We now describe the subgradient algorithm, which is an efficient iterative algorithm to solve the Lagrangian dual problem. The algorithm starts with a set of initial non-negative Lagrangian multipliers $\beta_{ij}[0]$. Since the Lagrangian multipliers are price signals that reflect the congestion status of the clusters, a possible choice for the initial Lagrangian multipliers can be zeroes, assuming there is no data congestion in the network. In this case, the initial shortest paths chosen by the sensor nodes will be the minimum energy paths without any adjustments on the link weights.

During each iteration k, given current Lagrangian multiplier values $\beta_{ij}[k]$, we solve the Lagrangian subproblem by finding the shortest path from each sensor node to its nearest sink node, where the weight of each link equals to the sum of its energy cost, its Lagrangian multiplier, and the Lagrangian multipliers of the clusters that this link belongs to. Using the new primal values $f_{ij}[k]$ obtained from the Lagrangian subproblem, we update the Lagrangian multipliers by:

$$\beta_{ij}[k+1] = \max(0, \beta_{ij}[k] + \theta[k](f_{ij}[k] + \sum_{(p,q)\in\Psi_{ij}} f_{pq}[k] - C)) \ , \tag{26}$$

where θ is a prescribed sequence of step sizes. The equation above states that the Lagrangian multipliers vary depending on the value of $(f_{ij} + \sum_{(p,q)\in\Psi_{ij}} f_{pq} - C)$, which represents the amount of the capacity violation within a cluster. When the violation of a cluster is positive, there are data flows traveling in the cluster that are not supported by the wireless shared-medium. The Lagrangian multiplier for the cluster then increases according to the amount of violation to reflect the congestion. Conversely, when the violation for a cluster is negative, there is free bandwidth in the cluster that is not utilized by the data flows. Therefore, the Lagrangian multiplier for the cluster decreases, in order to attract more data flows to occupy the free bandwidth.

The selection of step sizes plays an important role in the subgradient algorithm. 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 fail to converge. The convergence is guaranteed when θ satisfied the following conditions [21], regardless of the values of the initial Lagrangian multipliers:

$$\theta[k] \geq 0, \quad \lim_{k\to\infty} \theta[k] = 0, \quad \text{and} \ \sum_{k=1}^{\infty} \theta[k] = \infty \ . \tag{27}$$

C. *Distributed Algorithm*

Based on the localized Slepian-Wolf coding scheme and the subgradient algorithm, we construct our distributed algorithm to solve the correlated data gathering problem. Each cluster and wireless link is treated as an entity capable of processing, storing, and communicating information. In practice, each cluster and wireless link (i, j) is delegated to its sender node i, and all computations related to (i, j) will be executed on node i. The algorithm is summarized in Table II. In this algorithm, the price signals or the Lagrangian multipliers reflect the congestion status of the network. In addition, they act as a link of communication between the two problem objectives. When the algorithm converges, the solution generated will jointly optimize the rate allocation and the transmission structure.

We now give an illustrative example to demonstrate the convergence of the distributed algorithm. Fig. 1 illustrates a random sensor network with 90 sensors and 10 sinks, represented by asterisks and circles respectively. All nodes have a transmission range of 30 meters and the wireless links are represented by the dotted lines. The distributed algorithm is executed for 500 iterations. The solid lines represent the links chosen by the obtained transmission structure. The thickness of each solid line indicates the amount of aggregated data traveling on the link, while the sensor nodes transmit according to the

TABLE II

THE OPTIMIZATION PHASE

1) Choose initial Lagrangian multiplier values $\beta_{ij}[0]$, $\forall(i,j) \in E$.
2) For each cluster and link (i,j), repeat the following iteration until convergence (at times $t = 1, 2, \ldots$):

Cluster Price Update by cluster (i,j):

1) Receive flow rates $f_{pq}[t]$ from all links (p,q) in Ψ_{ij}.
2) Update Lagrangian multiplier $\beta_{ij}[t+1] = \max(0, \beta_{ij}[t] + \theta[t](f_{ij}[t] + \sum_{(p,q)\in\Psi_{ij}} f_{pq}[t] - C))$, where $\theta[t] = \frac{a}{(b+ct)}$.
3) Send $\beta_{ij}[t+1]$ to all links in Ψ_{ij}.

Link Rate Update by link (i,j):

1) Receive cluster prices $\beta_{pq}[t]$ from all clusters (p,q) in Φ_{ij}.
2) Determine the weight of link (i,j) as $(e_{ij} + \beta_{ij}[t] + \sum_{(p,q)\in\Phi_{ij}} \beta_{pq}[t])$.
3) If node i is a sensor node:
   a) Compute its nearest sink node using a distributed shortest path algorithm, such as the Bellman-Ford algorithm. Sensor i refers to its nearest sink node as its destination sink node.
   b) If modified, sensor i notifies other sensors in its neighbourhood about the identity and its distance to its new destination sink node.
   c) Based on the information received, sensor i finds within its neighbourhood the set N_i. N_i consists of sensor nodes that have the same destination sink node as sensor i, and are closer to that destination sink node than sensor i.
   d) Sensor i encodes its collected data at rate $R_i = H(i|N_i)$, and transmits the encoded data to its destination sink node via the shortest path.
4) Measure and record the current rate on the link, denoted as $f_{ij}[t+1]$.
5) Send $f_{ij}[t+1]$ to all clusters in Φ_{ij}.

obtained rate allocation. Evidently, the distributed algorithm minimizes energy consumption by exploiting data correlation. Sensor nodes that are distant from their corresponding sinks are assigned with lower transmission rates. For the duration of the experiment, the distributed algorithm generates a sequence of solutions. The total energy consumed by these solutions are recorded in Fig. 2. We observe from the figure that after an initial spike, the algorithm rapidly converges toward the optimal value within the first 50 iterations.

D. Asynchronous Network Model

Until now, we have assumed a synchronous implementation for the iterative subgradient algorithm. In this case, the local clocks on the nodes are synchronized, such that all of the nodes will simultaneously execute an iteration of the algorithm at every time instance ($t = 1, 2, 3, \ldots$). Bounded communication delay is assumed where price and rate updates will arrive at their destinations before the next time instance. As a result, each node is able to execute the algorithm based on the most recent price and rate information. However, in realistic ad hoc network environments, it is expensive in terms of communication overhead to synchronize local clocks across the entire network.

In asynchronous network environments, nodes with different computation speed will execute the iterative

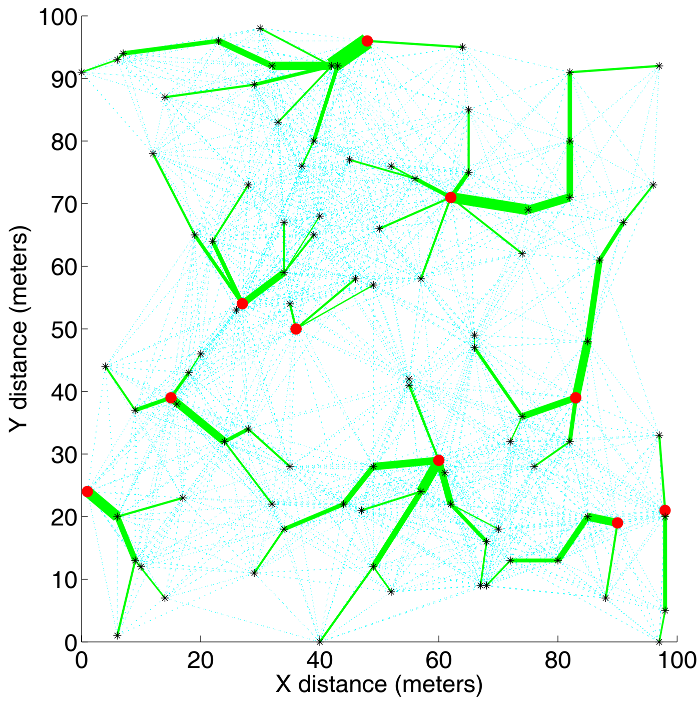

Fig. 1. A random topology with 100 nodes.

algorithm at varying paces. Consequently, the nodes may not always have the most recent price or rate information due to delayed or out-of-order updates. To accommodate these asynchronous updates, we introduce the *partial asynchronism model* that will be assumed in the practical implementation of our algorithm. The partial asynchronism model makes the following assumption.

There exists a positive integer B such that:

- For every cluster and wireless link (i, j), the time between consecutive updates is bounded by B for both price and rate updates.
- One-way communication delays between any two nodes is at most B time instances.

The partial asynchronism model is first discussed in [17]. Later, it is adopted by Low *et al.* [19] in wireline networks and Xue *et al.* [20] in wireless networks. In [20], a technique is proposed to improve the price-based resource allocation strategy to accommodate the partial asynchronous model. At time instance t, instead of the most recent information, a node may only received a sequence of recent updates. The concept of this technique is for the nodes to estimate the most recent price and rate information by computing the average of the sequence received from time $t - B$ to t. To improve the accuracy of the estimation, a moving average can be utilized with a heavier weight assigned to the more recent updates. From their work, it is shown that such strategy will converge the fastest when all the weight is assigned to the most recently received update. Moreover, with sufficiently small step size θ, the strategy will converge to the global optimum in asynchronous network environments. Since our optimization formulation is

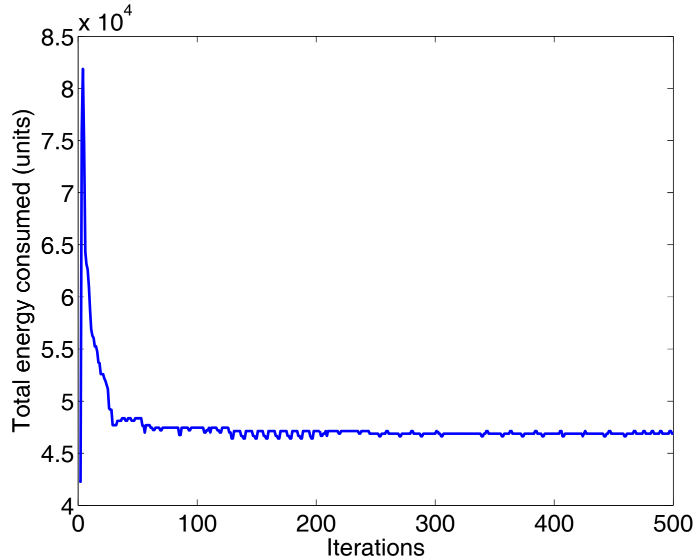

Fig. 2. Convergence behaviour of the distributed algorithm.

solved with a price-based strategy, it is natural for us to adapt this technique. In our implementation, each node estimates the price and rate information based on the most recently received update. In Section VI, we validate via simulations that our algorithm converges in asynchronous network environments.

V. Implementation Issues

The subgradient algorithm provides an efficient tool in obtaining a lower bound on linear programming formulations, such as our primal problem, via the solutions to the Lagrangian dual problem. However, it has the disadvantage that an optimal solution, or even a feasible solution to the primal problem may not be found. With such a subgradient optimization approach, methods such as primal penalty functions and tangential approximation schemes have been proposed for directly obtaining the primal solutions. These methods are not suitable for our purpose because they either require the optimization to be conducted in the joint primal-dual space, or introduce a significant additional computation overhead. In this section, we propose two implementations of our distributed algorithm that are aiming to overcome this problem. Moreover, we discuss how the distributed algorithm can be extended to handle network dynamics.

A. Implementation I: Primal Recovery

In [22], Sherali *et al.* introduce a primal recovery algorithm. The algorithm recovers the primal solutions directly from the solutions to the Lagrangian dual problem generated by the subgradient algorithm. We adapt this algorithm in our first implementation. The primal recovery algorithm restricts the step size strategy, and specifies that the primal solutions should equal to the convex combination of the solutions generated by the Lagrangian subproblem. At iteration k of the subgradient algorithm, we compose a primal solution $f^*_{ij}[k]$ via:

$$f^*_{ij}[k] = \sum_{m=1}^{k} \lambda^k_m f_{ij}[m] \ , \tag{28}$$

where $\theta[k]$ are the step sizes and λ^k_m are the convex combination weights given by:

$$\theta[k] = \frac{a}{b + ck}, \ \forall k, \quad \lambda^k_m = \frac{1}{k}, \ \forall m = 1 \ldots k, \ \forall k, \tag{29}$$

where a, b, and c are positive constants. This step size strategy also satisfies condition (27), hence the convergence of the subgradient algorithm is guaranteed. In the kth iteration, we can calculate the adjusted flow vector $f^*_{ij}[k]$ by:

$$f^*_{ij}[k] = \frac{k-1}{k} f^*_{ij}[k-1] + \frac{1}{k} f_{ij}[k] \ . \tag{30}$$

It is proven in [22] that when conditions (29) are met, any accumulation point of the sequence $f^*_{ij}[k]$ generated via (28) is feasible to the primal problem.

Although the primal recovery algorithm guarantees to generate feasible primal solutions, it has a major disadvantage. Since the generated primal solution depends on the previous solutions, the network must remain static. Any dynamics introduced to the network, such as sink mobility and duty schedules, may introduce obsolete links during the execution of the distributed algorithm. If any one of the previous solutions contains the obsolete links, then the generated primal solution becomes invalid. To accommodate these dynamics, we propose a heuristic approach in our second implementation.

B. Implementation II: Capacity Reservation

Recall that the subgradient algorithm provides quick lower bounds to linear programming formulations. In the context of this paper, as the subgradient algorithm converges, it generates a sequence of rate allocations and transmission structures as solutions to the Lagrangian dual problem. However, some of these solutions often violate the channel contention constraints, which is imposed by the primal problem, but relaxed in the dual problem. To analyze this behaviour, we conduct a simulation study on sensor networks with 90 sensors and 10 sinks. The capacity of the wireless shared-medium is set to 150 bits. The distributed algorithm is executed for 1000 iterations on 50 random topologies. For each random topology, we record the amount of capacity violation induced by the last solution generated by the algorithm, and the statistics is presented in Table III.

Evidently, the subgradient algorithm generates tight lower bounds, since the mean capacity violation is only a small fraction of the capacity offered by the shared-medium. Based on this behaviour, we observe that with high probability, the distributed algorithm can generate primal feasible solutions by reserving a suitable amount of capacity in advance. To reserve capacity, the distributed algorithm can be executed

TABLE III

STATISTICS ON CAPACITY VIOLATION IN BITS.

Maximum	Minimum	Mean	Standard deviation
21.78944	0	8.23610	4.87849

with the knowledge that the shared-medium can only support a fraction (e.g. 90%) of its actual capacity. Although this implementation does not guarantee primal feasible solutions, it does not introduce any additional computational complexity into the algorithm.

C. Handling Network Dynamics

With energy saving mechanisms such as sink mobility and duty schedules, the topology of a sensor network is inherently dynamic. Wireless links maybe added or removed from the topology. Moreover, since the energy cost of a wireless link is a function of its distance, it can vary with node movement. One of the main design goals for the distributed algorithm is to be compatible with these mechanisms and variations in order to achieve higher energy savings. To this end, we propose an extension to the algorithm for handling network dynamics.

In the distributed algorithm, each wireless link is treated as two separate entities, a link and a cluster. For each wireless link (i, j), its sender node maintains two lists, Ψ_{ij} and Φ_{ij}. The first list Ψ_{ij} contains the identities and the rates of the links that belong to cluster (i, j). And the second list Φ_{ij} contains the identities and the prices of the clusters that link (i, j) belongs to. To handle network dynamics, these lists must be updated as the topology changes.

We assume that the nodes are able to retrieve up-to-date topology information within their transmission range. At the beginning of each iteration, each participating node initiates the distributed algorithm by determining if it has new, obsolete, or modified links originating from itself. Afterwards, the nodes execute the maintenance phase given in Table IV. Finally, the nodes complete the iteration by executing the price and rate updates given in Table II. The purpose of the maintenance phase is to update the lists for the wireless links. It introduces several light-weighted control packets, and they are exchanged between the nodes and their local neighbourhoods.

VI. PERFORMANCE EVALUATION

A. Simulation Environments

With the C++ programming language, we have implemented the proposed distributed algorithm for solving the correlated data gathering problem. Our implementation includes both the optimization phase and the maintenance phase presented in Table II and Table IV. Data packets, control packets, and update

TABLE IV

THE MAINTENANCE PHASE

1) For each wireless link (i, j) delegated to node i:
   - If wireless link (i, j) is new or modified, node i sends a NOTIFY packet to node j. The NOTIFY packet contains the current price for cluster (i, j).
   - If wireless link (i, j) is obsolete, node i sends a DELLINK packet to all nodes p for the clusters $(p, q) \in \Phi_{ij}$. In addition, node i sends a DELCLUSTER packet to all nodes p for the links $(p, q) \in \Psi_{ij}$.
2) Upon receiving the following packets:
   - **NOTIFY**: Node j sends a CHECK packet to all potential nodes k that can be interfering with wireless link (i, j). Since the interference range is assumed to be equivalent to the transmission range in this paper, the potential nodes are limited to the nodes that are within node j's transmission range. The CHECK packet contains the price for cluster (i, j) extracted from the NOTIFY packet.
   - **CHECK**: Based on local information exchanged between the nodes, node k determines if its outgoing wireless links (k, l) are interfering with wireless link (i, j) according to the protocol model of packet transmission.
     - If wireless link (k, l) is interfering with wireless link (i, j). Check if cluster $(i, j) \in \Phi_{kl}$. If NO, add the identity and the price of cluster (i, j) to Φ_{kl}. Then, node k sends a ADDLINK packet to node i with the identity and rate of link (k, l).
     - If wireless link (k, l) is NOT interfering with wireless link (i, j). Check if cluster $(i, j) \in \Phi_{kl}$. If YES, remove cluster (i, j) from Φ_{kl}. Then, node k sends a DELLINK packet to node i with the identity of link (k, l).
   - **ADDLINK**: Cluster (i, j) adds the identity and the rate of link (k, l) to Ψ_{ij}.
   - **DELLINK**: If DELLINK packet is generated by node i, cluster (p, q) removes link (i, j) from Ψ_{pq}. If DELLINK packet is generated by node k, cluster (i, j) removes link (k, l) from Ψ_{ij}.
   - **DELCLUSTER**: Link (p, q) removes cluster (i, j) from Φ_{pq}.

packets are communicated between the participating nodes with a round robin scheduling algorithm. In practice, we expect the data packets can be scheduled with a weighted fair queueing algorithm [23]. As a result, the sensors can achieve guaranteed data transmission rates specified by the rate allocation.

In this section, we evaluate both implementations proposed in Section V with extensive experimental results. The experiments are conducted on a high-performance cluster with 50 dual-CPU servers. Unless stated, the experiments are performed on the random topology with 100 nodes presented in Fig. 1. The transmission range and the interference range are set to 30m. The capacity of the wireless shared-medium is set to 150 bits. The correlation parameter W and the per node distortion d is set to 0.99 and 0.01, respectively.

We study the distributed algorithm in three different simulation environments. In the *independent* environment, we neglect the effect of data correlation by substituting the localized Slepian-Wolf coding scheme with an independent coding scheme. In the *synchronous* environment, the participating nodes simultaneously execute an iteration of the algorithm at every time instance. The *asynchronous* environment is based on the partial asynchronism model, 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 instance. When the timer reaches 0, the sensor node executes an iteration of the algorithm before resetting the timer. For experiments involving network dynamics, we make a conservative estimation that the sensor network is capable of executing two iterations of the algorithm per second. This implies that the duration of a time instance equals to half of a second.

B. Convergence Speed

In our first study, we observe the convergence speed of the algorithm with different numbers of participating nodes. 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. We maintain a constant node density by scaling the area of the other sensor fields. This eliminates the effect caused by varying node density, and allows us to focus on the scalability of our algorithm. To attain convergence, we let the algorithm runs for 500 iterations, and the optimum is taken as the minimum total energy consumption achieved. For each experiment, the algorithm is executed in the *synchronous* simulation environment on 10 random topologies. For both implementations, we plot the mean number of iterations required to achieve 90% and 99% optimality in Fig. 3. The horizontal bars indicate one standard deviation below and above each mean. Supposing the numbers approximately follow a Gaussian distribution given by the Central Limit Theorem, each interval includes about 70% of the observations.

The figure reveals that the convergence behaviours of the two implementations are different. On average, the primal recovery algorithm increases the convergence time by 50%, but the standard deviations on the convergence time are smaller when compared with the capacity reservation scheme. This is an expected result since the primal recovery algorithm generates a solution by averaging all the previous solutions obtained in the subgradient algorithm. In contrast, the capacity reservation scheme always utilizes the most current solution obtained in the subgradient algorithm. As a result, it has a shorter convergence time, but it is also heavily influenced by the fluctuations introduced by the subgradient algorithm, leading to larger deviations. These convergence behaviours can be verified in Fig. 4, where we plot the sequences of solutions generated by the two implementations.

In general, we observe that the time needed to achieve 99% optimality remains almost the same for networks with 200 to 500 nodes. Moreover, we notice that the algorithm can achieve 90% optimality in less than half of the time needed to achieve 99% optimality. Recall that for both implementations, the solution generated in each iteration is primal feasible. Therefore, when it is not necessary to achieve the true optimum, we can obtain a near-optimal solution in a much shorter time. These results exhibit the excellent scalability of our algorithm as the network size increases.

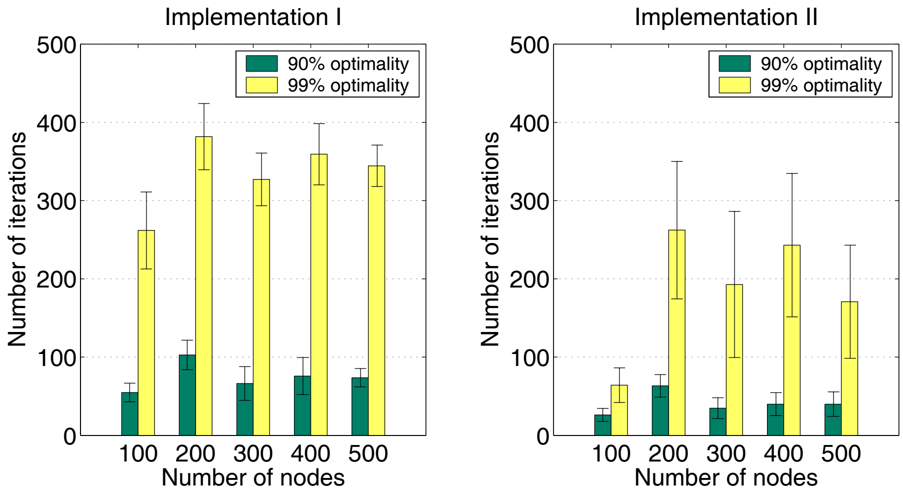

Fig. 3. Convergence speed in static networks. For each experiment, the horizontal bars indicate the one standard deviation below and above the mean.

C. Impact of Asynchronous Network Settings

With the *asynchronous* simulation environment, we evaluate the convergence behaviour of the distributed algorithm in asynchronous network settings. The algorithm is executed for 500 iterations with different time bounds $B = 1, 5, 10, 25$. For both implementations, the total energy consumption attained at each iteration is recorded in Fig. 4. In all experiments, the algorithm converges towards an identical optimal solution. This result indicates that our algorithm is able to achieve convergence in asynchronous network settings. 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.

D. Effect of Data Correlation

We investigate the effect of data correlation by comparing the *synchronous* simulation environment against the *independent* simulation environment. For each simulation environment, we execute the algorithm under three per node distortion values $d = 0.001, 0.01, 0.1$. As the correlation parameter W varies from 0.9 to 0.9999, the minimum total energy consumption achieved by the different simulation environments are recorded in Fig. 5. Intuitively, the energy consumed at high correlation ($W = 0.9999$) is much lower compared with the energy consumed at low correlation ($W = 0.9$). Overall, the two implementations achieve similar results, and the localized Slepian-Wolf coding scheme outperforms the independent coding scheme in all experiments. These results imply that even though the proposed algorithm utilizes only local information, it can achieve significant energy savings for a wide range of data correlation and distortion levels.

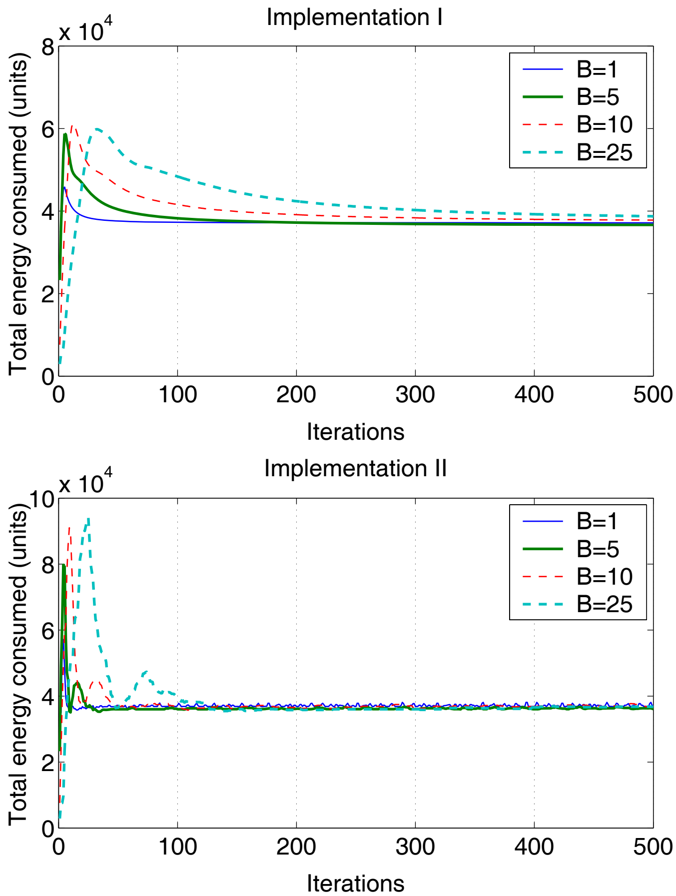

Fig. 4. Convergence behaviour in asynchronous network settings.

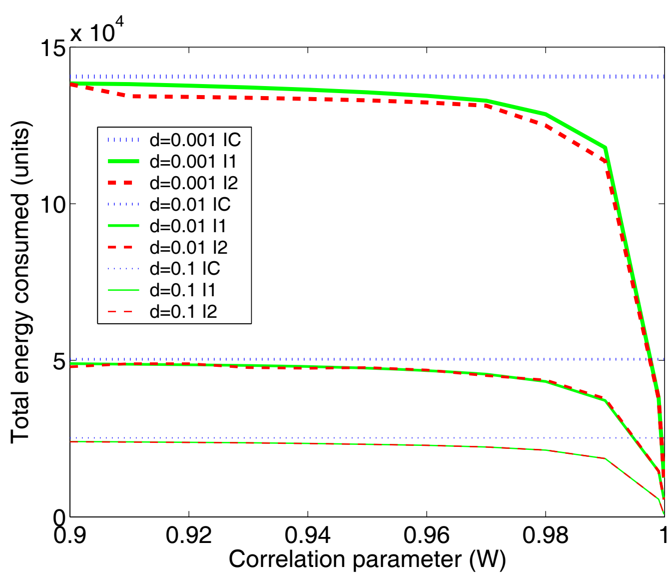

Fig. 5. Localized Slepian-Wolf coding vs. independent coding. IC: independent coding, I1: implementation I, I2: implementation II.

E. Adaptation to Sink Mobility

In this section, we study the impact of sink mobility over random networks. Our aim is to seek the mobility threshold such that the algorithm is not fast enough to remain at convergence. Based on the random topology with 100 nodes, we introduce sink mobility by moving the 10 sink nodes simultaneously. With periods of 50 and 100 seconds, the sink nodes move and remain static between alternating periods. For each mobile period, the sink nodes move in random directions with a specified average speed. The algorithm is executed in the *synchronous* simulation environment for 500 seconds. Fig. 6 plots the convergence behaviour of the algorithm for the different scenarios. From this figure, we observe that the algorithm can achieve new convergence after the network topology is modified. The results indicate that the algorithm converges sufficiently well when the sink nodes move at 0.1 m/s without pause. When the node speed increases to 0.5 m/s and 1 m/s, there are larger fluctuations in the attained total energy consumptions. Further increase in node speed may result in insufficient convergence time. In addition, we observe that the algorithm rapidly achieves and stays in convergence once the topology remains static. Obviously, the algorithm can support higher node speeds when the pause time increases.

F. Adaptation to Duty Schedules

To extend network lifetime, it is essential to establish load balancing between the sensor nodes with mechanisms such as duty schedules. In our final study, we are interested in examining the dynamic behaviour of the distributed algorithm triggered by sensor joins and departures. We model the duty schedules as a two-state Markov chain shown in Fig. 7. The state transition probabilities α and β are adjusted in order to emulate different duty schedules. The experiments are performed in the *synchronous* simulation environment for 300 seconds. In the first 100 seconds, all sensor nodes remain active. Afterwards, the sensors switch their operating status based on the introduced duty schedules.

The results of the experiments are summarized in Fig. 8 and Fig. 9. In Fig. 8, we adjust the summation of α and β with a fixed transition ratio α/β of 5. The summation represents the frequency of state transitions experienced by the network. Note that the summation cannot be greater than 2 since each of the transition probabilities cannot exceed 1. We observe that as the frequency of state transitions increases, the topology of the network changes more rapidly, leading to larger fluctuations in the attained total energy consumptions. Fig. 9 illustrates the performance of the algorithm under different transition ratios with a fixed summation of 0.01. We have avoided combinations of transition ratio and summation that may lead to network partition. For example, if the transition ratio is less than 1, then active sensor nodes are more likely to shut themselves off than inactive sensor nodes turning themselves on. As the number of inactive sensor nodes increases, a partition in the sensor network would eventually occur. Moreover, we notice

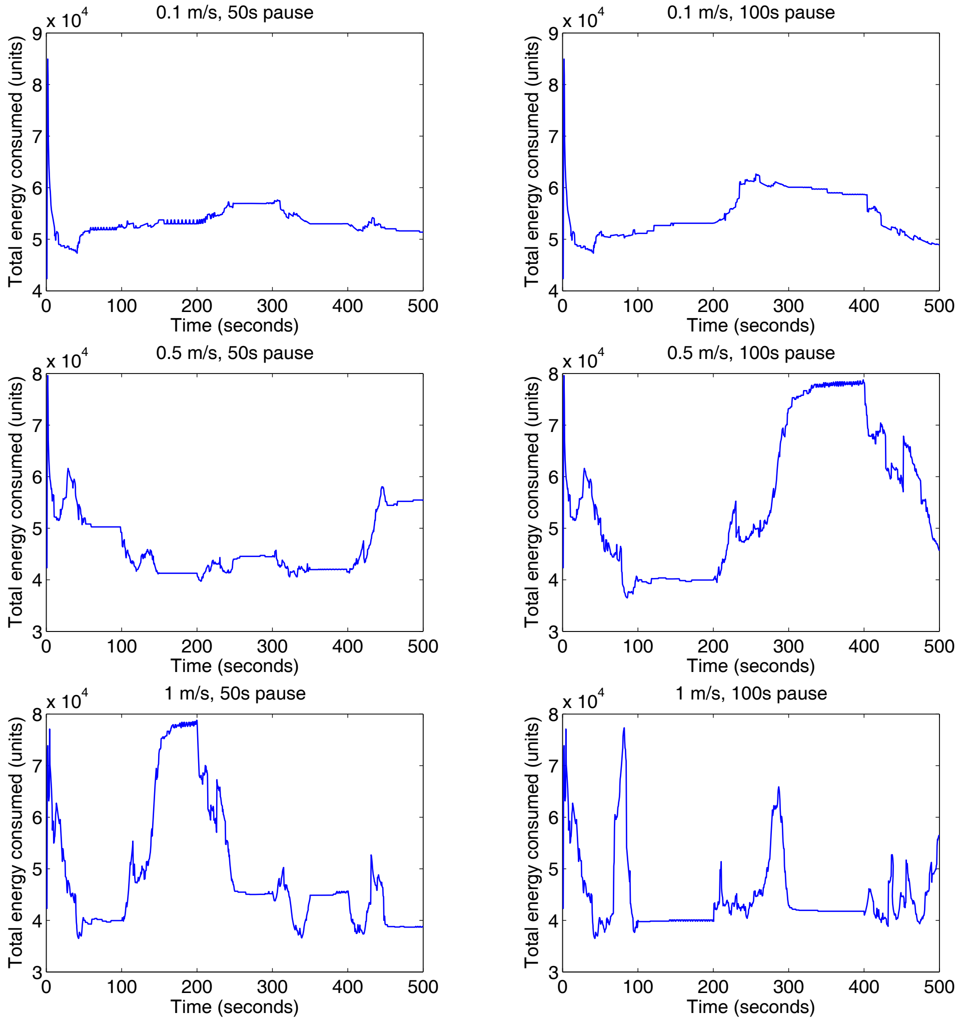

Fig. 6. Experiments with varying sink speeds and pause times.

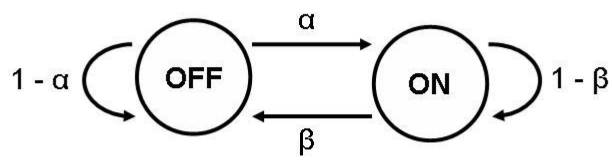

Fig. 7. A two-state Markov chain.

that with higher transition ratio, the network consumes more energy since more sensor nodes are active.

VII. Related Work

The problem of energy efficient routing in sensor networks has been investigated with mathematical optimization techniques in research studies including [24], [25], [26], and [27]. Chang *et al.* [24] have formulated a flow-based linear programming formulation to maximize the network lifetime. In [25], the optimization model minimizes energy consumption and takes into account the channel contention constraints associated with the wireless shared-medium. Krishnamachari *et al.* [26] propose another

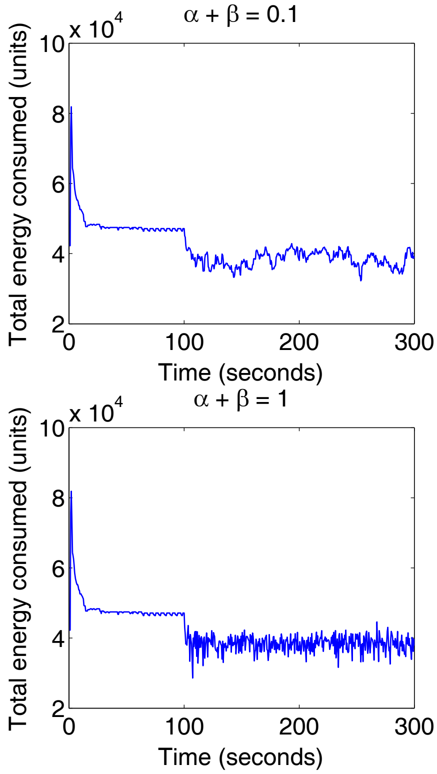

Fig. 8. Experiments with varying amount of state transitions.

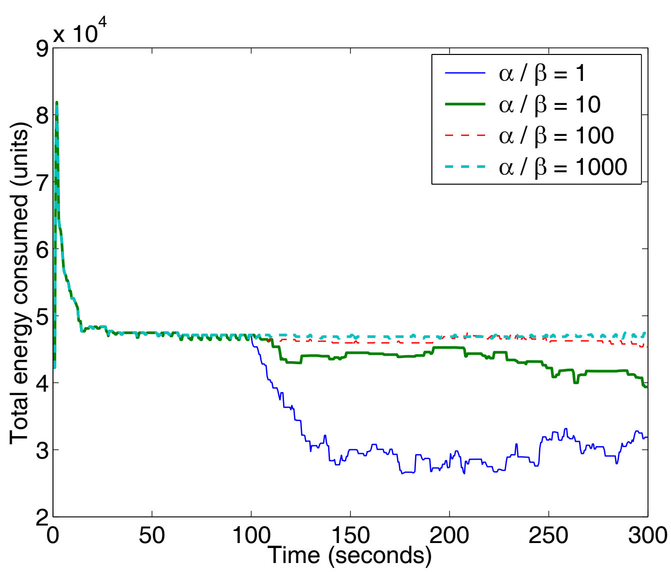

Fig. 9. Experiments with different transition ratios.

optimization formulation to maximize the raw data arrived at the sink nodes, subject to flow, fairness, energy, and capacity constraints. Boyd *et al.* [27] study the simultaneous routing and power allocation problem in wireless data networks using optimization techniques. In [26] and [27], the optimization problems utilize the physical model [4] of packet transmission in wireless networks to model the channel contention constraints. However, the resulting channel contention constraints are non-convex, which can lead to extremely difficult optimization problems. In our paper, we represent channel contention as linear constraints based on the protocol model [4]. More importantly, even though all of the above existing works generally save energy, they do not consider the additional energy savings that can be achieved by exploiting data correlation among the sensor nodes.

Data aggregation was introduced by Estrin *et al.* [28] as an essential paradigm for wireless routing in sensor networks. The concept is to exploit the data correlation among the sensor nodes by eliminating redundancy. Consequently, there are less transmissions in the network and thus saving energy. In [7], Kalpakis *et al.* have formulated the maximum lifetime data gathering problem as a linear programming formulation, taking data aggregation into consideration, and presented a polynomial-time algorithm to solve the problem. Although this optimization framework yields satisfactory performance, it makes the simplistic 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 [6], which analyzes the performance of data-centric routing schemes with in-network aggregation. In [8], Goel *et al.* consider the joint treatment of data aggregation and transmission structure. The problem of data gathering is addressed by using concave, non-decreasing cost functions to model the aggregation function utilized by the intermediate nodes. However, it also makes the assumption of perfect data correlation. The aggregation performance of a node only depends on the number of nodes providing incoming data, regardless of the correlation structure. The assumption of perfect data correlation is not made in this paper since it is not applicable in most application scenarios.

While this paper exploits data correlation with Slepian-Wolf coding, there are alternative approaches to take advantages of the correlation structure. In [3], [29], the correlated data gathering problem is considered with single-input coding schemes. With single-input coding, the data compression ratio at an intermediate node only depends on the side information provided by one other node. Cristescu *et al.* [3] prove that this optimization problem is NP-hard even in a simplified network setting, where the data compression ratio at the nodes does not depend on the quantity of side information, but only on its availability. Since single-input coding schemes only consider data correlation among pair of nodes, they will not perform as well as source coding schemes, which consider the joint data correlation of multiple nodes.

Multi-input coding schemes are often employed by routing schemes embedded with data aggregation, such as directed diffusion [30], LEACH [31] and PEGASIS [32]. Directed diffusion is a routing-drive algorithm that emphasizes source compression at each individual node and data aggregation occurs opportunistically when routes intersect. In the model of LEACH, nodes are chosen as cluster heads, which are then responsible for aggregating all data generated in their corresponding cluster into a single packet. Instead of clusters, the PEGASIS algorithm finds chains of nodes, and the head node of each chain aggregates data from other nodes in the chain. Although the multi-input coding schemes can exploit data correlation among multiple nodes, they require the participating nodes to explicitly communicate with each other. In contrast, Slepian-Wolf coding schemes do not require any explicit communication, hence they can be applied in asynchronous network settings where no timing assumptions are made. In addition, these routing schemes do not incorporate the effect of wireless interference in their design.

Other closely related works are the ones involving Slepian-Wolf source coding. In [33], Servetto *et al.* introduce the sensor reachback problem, which requires one of the nodes in the network to receive enough information to reproduce the entire field of observation. Slepian-Wolf coding is employed to meet the above requirement. This paper inspires us to apply Slepian-Wolf coding in the correlated data gathering problem, hence the sink nodes will be able to receive all independent data from the sensor nodes. In [15], Cristescu *et al.* address the correlated data gathering problem with Slepian-Wolf coding. However, since their formulation does not consider the capacity and interference associated with the wireless channels, their solution may not be supported by the shared-medium.

VIII. Conclusion

With the ability of distributed wireless sensing, sensor networks can be applied to a vast number of applications. However, before we can recognize the full potential of sensor networks, the problem of correlated data gathering must be solved under realistic assumptions. We conclude this paper with the belief that our proposed framework is an efficient means to accomplish this task. In this paper, we show that in the presence of capacity constraints, finding the optimal rate allocation and the optimal transmission structure are two dependent problems. By jointly optimizing both problems, our approach minimizes the total transmission energy consumed by the network. Furthermore, it exploits data correlation among the sensor nodes, and accounts for the effect of location-dependent contention in the wireless channels. To ensure scalability, our algorithm is amenable to distributed implementations, applicable in asynchronous network settings, and provides support for multi-sink sensor networks. To the best of our knowledge, there does not exist any previous work that have considered the correlated data gathering problem with data aggregation and wireless channel interference simultaneously, especially when a price-based strategy is employed to obtain a distributed algorithm to solve the problem.

REFERENCES

[1] T. Clouqueur, V. Phipatanasuphorn, P. Ramanathan, and K. K. Saluja, "Sensor Deployment Strategy for Detection of Targets Traversing a Region," in *ACM Mobile Networks and Applications*, 2003, vol. 8, pp. 453–461.

[2] A. Mainwaring, J. Polastre, R. Szewczyk, and D. Culler, "Wireless Sensor Networks for Habitat Monitoring," in *Proc. of First ACM International Workshop on Wireless Sensor Network and Applications*, September 2002.

[3] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On Network Correlated Data Gathering," in *Proc. of IEEE INFOCOM*, 2004.

[4] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Trans. Information Theory*, vol. 46, no. 2, pp. 388–404, March 2000.

[5] M. Lotfinezhad and B. Liang, "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)*, October 2004.

[6] B. Krishnamachri, D. Estrin, and S. Wicker, "Modelling Data-Centric Routing in Wireless Sensor Networks," in *Proc. of IEEE INFOCOM*, 2002.

[7] K. Kalpakis, K. Dasgupta, and P. Namjoshi, "Efficient Algorithms for Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks," *Computer Networks Journal*, 2002.

[8] A. Goel and D. Estrin, "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.

[9] S. Tilak, N. Abu-Ghazaleh, and W. Heinzelman, "A Taxonomy of Wireless Microsensor Network Models," in *ACM Mobile Computing and Communications Review (MC2R)*, June 2002.

[10] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*, Prentice-Hall series in information and system science. Englewood Cliffs, N.J. Prentice-Hall, 1971.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[12] B. Hamdaoui and P. Ramanathan, "Rate Feasibility under Medium Access Contention Constraints," in *Proc. of IEEE GLOBECOM*, December 2003.

[13] B. Liang and Z. J. Haas, "Virtual backbone generation and maintenance for ad hoc network mobility management," in *Proc. of IEEE INFOCOM*, March 2000.

[14] D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. on Information Theory*, vol. 4, no. IT-19, pp. 471–480, 1973.

[15] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Networked Slepian-Wolf: Theory and Algorithms," in *Proc. of European Workshop on Wireless Sensor Networks*, 2004.

[16] D. P. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1992.

[17] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, 1989.

[18] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate Control in Communication Networks: Shadow prices, Proportional Fairness and Stability," in *Journal of the Operational Research Society*, 1998, vol. 49, pp. 237–252.

[19] S. H. Low and D. E. Lapsley, "Optimization Flow Control: Basic Algorithm and Covergence," in *IEEE/ACM Trans. on Networking*, 1999, vol. 7, pp. 861–874.

[20] Y. Xue, B. Li, and K. Nahrstedt, "Optimal Resource Allocation in Wireless Ad Hoc Networks: A Price-Based Approach," in *IEEE Transactions on Mobile Computing*, April 2006, vol. 5.

[21] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, 1985.

[22] H. D. Sherali and G. Choi, "Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs," in *Operations Research Letter*, 1996, vol. 19, pp. 105–113.

[23] H. Luo, S. Lu, and V. Bharghavan, "A New Model for Packet Scheduling in Multihop Wireless Networks," in *Procs. of ACM MOBICOM*, 2000.

[24] J. H. Chang and L. Tassiulas, "Maximum Lifetime Routing in Wireless Sensor Networks," in *IEEE/ACM Transactions on Networking*, August 2004, vol. 12, pp. 22–31.

[25] B. Hamdaoui and P. Ramanathan, "Energy-efficient and MAC-aware Routing for Data Aggregation in Sensor Networks," in *Proc. of Sensor Network Operations*, 2004.

[26] F. Ordonez and B. Krishnamachari, "Optimal Information Extraction in Energy-Limited Wireless Sensor Networks," in *IEEE Journal on Selected Areas in Communications, special issue on Fundamental Performance Limits of Wireless Sensor Networks*, 2004.

[27] M. Johansson, L. Xiao, and S. Boyd, "Simultaneous Routing and Resource Allocation in CDMA Wireless Data Networks," in *Proc. of IEEE International Conference on Communications*, May 2003, vol. 1, pp. 51–55.

[28] B. Krishnamachari, D. Estrin, and S. Wicker, "The Impact of Data Correlation in Wireless Sensor Networks," in *Proc. of the 22nd International Conference on Distributed Computing Systems*, 2002, pp. 575–578.

[29] P. V. Rickenbach and R. Wattenhofer, "Gathering Correlated Data in Sensor Networks," in *Proc. of DIALM-POMC '04: Proceedings of the 2004 joint workshop on Foundations of mobile computing*, 2004, pp. 60–66.

[30] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed Diffusion for Wireless Sensor Networking," in *IEEE/ACM Transactions on Networking*, February 2003, number 1, pp. 2–16.

[31] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient Communication Protocol for Wireless Sensor Networks," in *Proc. of the Hawaii International Conference System Sciences*, January 2000.

[32] S. Lindsey and C. S. Raghavendra, "PEGASIS: Power Efficient GAthering in Sensor Information Systems," in *Proc. of the IEEE Aerospace Conference*, March 2002.

[33] J. Barros and S. D. Servetto, "Network Information Flow with Correlated Sources," in *IEEE Transactions on Information Theory*, January 2006, vol. 52, pp. 155–170.