

JOINT SENSOR SELECTION AND ROUTING FOR DISTRIBUTED ESTIMATION IN WIRELESS SENSOR NETWORKS

Eduardo Portu-Repollés and Baltasar Beferull-Lozano

Group of Information and Communication Systems (GSIC)
Escuela Técnica Superior de Ingeniería (ETSI) - Instituto de Robótica
Universidad de Valencia (UV), 46071, Valencia, Spain

ABSTRACT

We consider a wireless sensor network (WSN) deployed over a large geographical area, where a querying node wishes to perform a estimation of a localized phenomenon. We formulate the problem as a joint optimization of sensor selection and routing structure where we minimize the estimation distortion subject to a total communication power constraint for the WSN. Two scenarios are analyzed: measurement forwarding and estimation-and-forward at the nodes. We show that the optimization problems corresponding to these scenarios are both NP-hard and we propose two approximation algorithms. First, we present a sensor selection algorithm for a predefined routing structure based on a primal relaxation and then, a greedy approximation algorithm that jointly optimizes the sensor selection and routing structure. Numerical results show good performance of these algorithms in both estimation scenarios.

1. INTRODUCTION

Wireless Sensor networks (WSN) have recently attracted a large amount of research [1][2]. In such networks, tiny and inexpensive devices with sensing and communication capabilities are deployed densely over large geographical areas to perform a wide variety of tasks, ranging from monitoring and detection to parameter estimation or tracking [1]. As the devices are usually battery-powered, only limited energy is available to perform these tasks, thus power-efficient algorithms need to be developed. When comparing sensing, computation and communication costs, the latter has been shown to be the most energy consuming phase[2].

Even though sensor networks usually cover a large geographical area, many interesting phenomena are in fact localized. Consider, for instance, applications such as forest fire detection or the estimation of a parameter from a spatially localized point source. Intuitively, in these settings, sensor nodes that are located far away from the source will have significantly less informative measurements than the nodes that are closer to it. Direct transmission of the measurements from all the nodes to a fusion center would certainly result in an inefficient use of the energy resources.

We will consider the scenario where the network is performing the distributed estimation of a deterministic parameter generated by a localized source. There is a clear need to tackle the problem of jointly determining which set of sensors needs to be activated, which operations have to be made at each node, and to whom the information should be routed

to. These decisions have to take into account both the cost of communication between the nodes, as well as the quality of the cooperative task to be performed, in this case, that the estimation quality should satisfy a certain degree of performance.

There has already been a certain amount of research in the area of sensor selection for wireless sensor networks. The work in [3] has investigated mechanisms for selecting sensors for cooperative target localization and tracking. Decisions are made based on the expected conditional entropy of the posterior target localization distribution. In [4, 5, 6], joint sensor selection and rate allocation schemes are developed. The authors in [7] propose a centralized algorithm that deals with correlated measurements by selecting at each time step the most informative sensor. Complexity and optimality results for families of utility functions are presented in [8].

Most of this previous work do not take into account the multihop nature of the WSN. Power savings are derived from the fact that we are activating a limited number of sensors (cardinality of set), but the cost of actually communicating a measurement through a path of the multihop network to the querying node is simplified or ignored. However, it is indeed very different to activate a sensor far away from the querying node than activating a node that is close to it. The incorporation of the communication costs in the sensor selection algorithms leads, in fact, to a more complex joint sensor selection and routing problem, as the communication costs depend on the routing structure, creating a strong interplay between the routing and the aggregated estimation at each node. To the best of our knowledge, this problem has not been considered yet in the literature.

The main contributions of this paper are the following. We formally define the joint sensor selection and routing problem, and give characterize its complexity. We also provide two algorithms: a sensor selection algorithm based on a simple relaxation for a fixed routing structure, and a distributed greedy approximation algorithm for the sensor selection and routing problem. The paper is structured as follows. In Section 2, we present the network, signal models and the estimation scenarios that we use in our work. In Section 3, the joint sensor selection and routing problem is formulated and results on complexity are shown. In Section 4, the algorithms are presented, whose numerical performance is shown in Section 5.

2. PROBLEM SETTING

Let us assume that we have a spatially localized event represented by a certain parameter to be estimated and that there is a certain querying node, whose goal is to get the best possible estimate. Motivated by the fact that power is a severe

This work was supported by the spanish MEC Grants TEC2006.10218 "SOFIWORKS", CONSOLIDER-INGENIO 2010 CSD2008-00010 "COMONSENS", the European STREP Project "SENDORA" Grant no. 216076 and by the Basque Government Grant BFI06.93.

constraint in WSNs, our problem consists of choosing jointly the optimal set of sensors and multihop tree routing structure in order to get the best possible estimate at the querying node for a given constraint in the total power to be spent by the WSN. The distortion measure will depend on the signal model and estimator used, and the cost of delivering a measurement to the querying node will depend on the network topology.

2.1 Network Model

We model the wireless sensor network as a directed graph $G = (V, E)$, with $N + 1 = |V|$ nodes including a special root node $r \in V$, labeled node $N + 1$, that plays the role of a querying node where the final estimate is to be obtained. An edge v between nodes i and j of the graph represents a communication link between this pair of nodes. This model is inherently assuming that the interference between nodes is negligible. Nevertheless, this approximate network model has been used in previous work, since the drawn conclusions are still very useful for the design of real WSNs [9]. Moreover, this assumption is realistic in a scenario where the communication across the different links is orthogonal or where the nodes are equipped with directional antennas.

In order to transmit the data to the querying node, node i will send it through a specified multihop path on the WSN. This path (routing structure) has to be optimized in our problem. The cost incurred when using a specific link v between nodes i and j is assumed to be a nondecreasing function of the distance between the corresponding nodes $f_c(d_{i,j})$, where $d_{i,j}$ represents the distance between nodes i and j . This approximate communication cost model has been experimentally supported [10].

2.2 Signal Model

For the sake of simplicity, we will consider a simple linear model¹ where a node i makes scalar observations $y_i \in \mathbb{R}$ of an unknown deterministic parameter $\theta \in \mathbb{R}$ distorted by a scalar $h_i \in \mathbb{R}$ and corrupted by additive Gaussian noise:

$$y_i = h_i \theta + n_i \quad i = 1, \dots, N \quad (1)$$

Moreover, $n_i \in \mathbb{R}$ are taken to be independent and identically distributed $n_i \sim \mathcal{N}(0, \sigma^2)$. We will assume that the scalar h_i captures the attenuation inherent to the physics of the propagation of the signal of interest from the source to the sensing node (i.e. an acoustic or electrical signal). This value of h_i will be a non-increasing function of the distance from the node i to the source of the parameter that we are interested to measure. It is intuitive to think that a sensor closer to the region of interest will receive a greater average power of the signal than a sensor that is far away. It is furthermore assumed to be known by the node, by using some sort of pre-calibration through pilot signals. The random variable n_i will capture the disturbance of the sensing device and the possible interferences that are present in the medium.

2.3 Estimation Strategies

Given measurements of the form of (1), the optimal estimator is given by the well-known BLUE (Best Linear Unbiased Estimator) [11]. The natural measure of estimation quality is

¹The generalization of the results of this paper to more complex estimation scenarios is part of our current research and will be presented elsewhere.

given by the mean squared error (MSE). In our setting, the estimator is given by:

$$\hat{\theta} = \frac{\sum_{k=1}^N h_k y_k}{\sum_{k=1}^N h_k^2} \quad (2)$$

and the associated MSE has the expression:

$$\text{MSE}_{\hat{\theta}} = \left(\frac{\sum_{k=1}^N h_k^2}{\sigma^2} \right)^{-1} \quad (3)$$

Moreover, the BLUE estimator has an important advantage: it can be easily implemented in a sequential fashion when the measurement noises are independent [11]. When a node is sending its measurement to the fusion center through a multihop path, it can either simply forward the data it receives from its children nodes in the routing tree, or fuse the measurements and forwarding only the aggregated estimate. This leads us to define two scenarios (Figure 1):

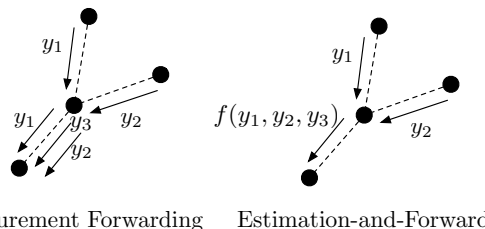


Figure 1: Illustration of the two possible scenarios, depending on the type of computation an intermediate node performs

- A. *Measurement forwarding at intermediate nodes:* In this scenario, the nodes simply forward the measurements they receive towards the querying node along the chosen multihop routing tree. The querying node is responsible to perform the final estimation, and no aggregated estimation is calculated in the intermediate nodes.
- B. *Estimation-and-forward at intermediate nodes:* In this scenario, a sequential estimation approach is considered. For a given routing structure, an intermediate node takes the estimates he receives from the active links and fuses its measurement with it. Then, it forwards this resulting fused estimation to its parent on the routing tree.

The second scenario has several interesting advantages over the first one. First of all, this second scheme is more power efficient as an active node in a route has only to forward the fused estimation (one information data to be transmitted), instead of forwarding its own measurement plus the measurements from the other nodes that are further away from the querying node on the routing tree. Moreover, we have the fact that the intermediate nodes in the route have an estimation of the parameter, which gets better as the node is closer (in number of hops) to the querying node.

3. SENSOR SELECTION AND COMMUNICATION FOR DISTRIBUTED ESTIMATION

Once we have defined the communication cost and the sensing model, the problem can now be well formulated. Let us

assign a variable x_k to each of the sensors. These variables are used to denote the status of the sensor. They take the value $x_k = 1$ if sensor i is active and $x_k = 0$ otherwise. Our problem can be formulated as an optimization problem where we want to minimize the distortion (MSE) of our estimation subject to a total power constraint:

$$\begin{aligned} \min_{\{x_k, \text{parent}_k\}} \quad & \text{MSE}_{\hat{\theta}} = \left(\frac{\sum_{k=1}^N x_k h_k^2}{\sigma^2} \right)^{-1} \\ \text{subject to} \quad & \sum_{k=1}^N x_k c_k \leq P \\ & x_k \in \{0, 1\} \end{aligned} \quad (4)$$

with variables x_k and parent_k , the latter one representing the parent node of node k . In this formulation, c_k defines the communication cost of activating node k . Moreover, we have to ensure that the routing structure determined by the set $\{\text{parent}_k\}$ is actually a tree rooted at the querying node $N+1$.

3.1 Measurement Forwarding

In this first scenario, the estimation and routing tasks clearly decouple as the intermediate nodes in the route only forward the measurements to the querying node, that is, at each intermediate node, since there is no fusion of estimates, there is no reduction of transmission data and the measurements generated at each node have to travel all the way up to the querying node. Therefore, for each given sensor, the optimal way to send its measurement to the querying node is to do it over the shortest path tree. In this simple case, the problem (4) simplifies and becomes:

$$\begin{aligned} \max_{\{x_k\}} \quad & \text{MSE}_{\hat{\theta}}^{-1} = \frac{\sum_{k=1}^N x_k h_k^2}{\sigma^2} \\ \text{subject to} \quad & \sum_{k=1}^N x_k c_k = \sum_{k=1}^N x_k \sum_{(i,j) \in \text{SPT}_k} f_c(d_{i,j}) \leq P \\ & x_k \in \{0, 1\} \end{aligned} \quad (5)$$

Notice that parent_k is no longer a variable as the optimal routing structure is already fixed (SPT).

3.2 Estimate-and-Forward

This scenario is very interesting because it creates a strong interplay between sequential estimation and routing. Moreover, it is more challenging to solve as now the routing structure is clearly tied to the sensor selection. Notice that the cost incurred by selecting a certain node k is simply the cost of communicating the information from this node to its parent, as the information is fused at each node, and only one information piece is forwarded to the next node in the tree. In this case, the problem in (4) becomes:

$$\begin{aligned} \max_{\{x_k, \text{parent}_k\}} \quad & \text{MSE}_{\hat{\theta}}^{-1} = \frac{\sum_{k=1}^N x_k h_k^2}{\sigma^2} \\ \text{subject to} \quad & \sum_{k=1}^N x_k c_k = \sum_{k=1}^N x_k f_c(d_{k, \text{parent}_k}) \leq P \\ & x_i \leq x_j, \text{ where } j = \text{parent}_i \\ & x_k \in \{0, 1\} \end{aligned} \quad (6)$$

where parent_k is chosen such that the resulting tree is rooted at the querying node. The second constraint ensures that no node is selected if its parent on the tree is not selected.

We can also cast this problem in an equivalent network flow formulation as follows. Let \mathbf{A} be the $R^{N+1 \times |E|}$ (node $N+1$ is the sink) incidence matrix of the network, where $|E|$ denotes the number of edges in the network. Moreover, let $\mathbf{f} \in R^{|E|}$ be the vector of information flows in the network, where f_{ij} denotes the flow going from node i to node j . $\mathcal{I}(i)$ is the set of incoming flows to node i . The auxiliary vector variables $\mathbf{s}, \mathbf{d} \in R^{N+1}$ represent the inflow/outflow operations at each of the nodes. With this notation, the problem can be equivalently formulated as:

$$\begin{aligned} \max_{\{x_k\}, \{f_{ij}\}} \quad & \sum_{k=1}^N x_k h_k^2 \\ \text{subject to} \quad & \mathbf{A}\mathbf{f} = \mathbf{s} - \mathbf{d} \\ & s_i = x_i \quad i = 1, \dots, N, \quad s_{N+1} = 0 \\ & d_i = \sum_{k \in \mathcal{I}(i)} f_{ki}, \quad d_{N+1} \geq 1 \\ & \sum_{(i,j) \in E} c_{ij} f_{ij} \leq P \\ & f_{ij} + f_{ji} \leq 1 \quad \forall i \neq j \\ & f_{ij}, x_i \in \{0, 1\} \end{aligned} \quad (7)$$

The first constraint ensures flow conservation in the network. The second one indicates that an active (selected) sensor generates flow into the network and that the sink node does not take any measurement. The third one captures the idea of data fusion at each node when using a sequential estimator (only one fused estimation exits the node) and ensures that the sink node has at least one incoming flow of information. The fourth constraint reflects the power constrained nature of the network (c_{ij} denotes the cost of sending information from node i to node j). Finally, the fifth constraint prevents length-two cycles in a link.

3.3 Complexity

Lemma 1. *The joint optimization problem of sensor selection and routing structure (5) and (7) are NP-Hard.*

Proof. Both problems are 0-1 integer linear programs, thus NP-Hard. This follows using a reduction from 3SAT problem [12]. \square

4. APPROXIMATION ALGORITHMS

The fact that our problem is NP-Hard, motivates us to find good approximation algorithms that can yield solutions close to the optimal value with a reasonable (polynomial) computational complexity. Next, we present two different algorithms.

4.1 Fixed-tree relaxation-based algorithm

The first algorithm assumes that a routing structure is already fixed, and that only sensor selection has to be performed. This algorithm is well suited to the measuring forwarding scenario presented in Section 3.1 where the optimal routing tree is the SPT or the case where we could be restricted, for the sake of simplicity, to a fixed simple routing structure in the WSN.

In this case, we perform a relaxation [13] of the problem (5) and (6), depending on the scenario that is considered, with the desired routing tree, by simply relaxing the constraints $x_k \in \{0, 1\}$ by $0 \leq x_k^{rel} \leq 1$. For instance, the relaxation of problem (5) becomes:

$$\begin{aligned} \max_{\{x_k^{rel}\}} \quad & \text{MSE}_{\hat{\theta}}^{-1} = \frac{\sum_{k=1}^N x_k^{rel} h_k^2}{\sigma^2} \quad (8) \\ \text{subject to} \quad & \sum_{k=1}^N x_k^{rel} c_k = \sum_{k=1}^N x_k^{rel} \sum_{(i,j) \in \text{Path-to-sink}_k} f_c(d_{i,j}) \leq P \\ & 0 \leq x_k^{rel} \leq 1 \end{aligned}$$

This is a simple linear program that can be easily solved in polynomial time [13]. Solving this problem provides us with an upper bound on the optimal value of our problem as we are optimizing over a larger constraint set that includes the original constraint set. At the same time, we can use the solution to the problem (8) in order to construct an approximation to the optimal of the original problem (5). This can be done sorting the optimal values $\{x_k^{*rel}\}$ in descending order and select the subset corresponding to the largest x_k^{*rel} that still satisfy the power constraint inequality. Notice that this approximation algorithm will give us a lower bound on the optimal value of the original problem. By calculating the gap between the upper and lower bound we can assess the performance of the algorithm.

Algorithm 1. Fixed-tree Relaxation-based Algorithm

- Solve linear program in (8)
- Sort the set optimal values $\{x_k^{*rel}\}$ in descending order
- Select the subset of sensors corresponding to the largest x_k^{*rel} that still satisfy the power constraint inequality
- Compute optimality gap

4.2 Greedy Approximation Algorithm

The algorithm presented in the previous section is tailored to the situations where the routing structure is fixed and the estimation is performed only at a fusion center (centralized algorithm). We now present a distributed greedy approximation algorithm that performs a joint sensor selection and routing. As the routing tree does not need to be fixed, this algorithm is well suited to the estimation-and-forward scenario.

The idea of the algorithm is as follows. We start from the node that first detects the phenomenon (on average the one with the largest SNR, i.e., the largest h_i). The best way to send this measurement to the querying node is using the corresponding path on the SPT. This motivates us to select this path as an initialization of the algorithm. All the intermediate nodes in the path are selected. Then, each of the selected nodes calculates an objective function for all its 1-hop neighbors. This objective function is a combination of the communication cost incurred when selecting the corresponding idle sensor and the information gain of that sensor. For each sensor k , the importance of the communication cost vs. the information gain is controlled through a weight, denoted as γ_k in the algorithm. While there is still available power, the algorithm selects the node (over all the idle neighbors of the selected sensors) that minimizes this objective function. After a new node is selected, a backtracking is performed. We check if an alternative route through the new selected sensor

is more power efficient (see Figure 2). Finally, we update the term γ_k that determines the relative importance of the communication cost and the information gain in the objective function of the algorithm. We use the following update scheme. If the node information gain of the last selected sensor is greater than the gain in the preceding step, the value of γ_{k+1} is lowered as we are heading in the correct direction to the location of the event and more importance is given to the power efficiency. Otherwise, the value of γ_{k+1} is increased to try to head towards the correct direction. The formal description of the algorithm is given as follows:

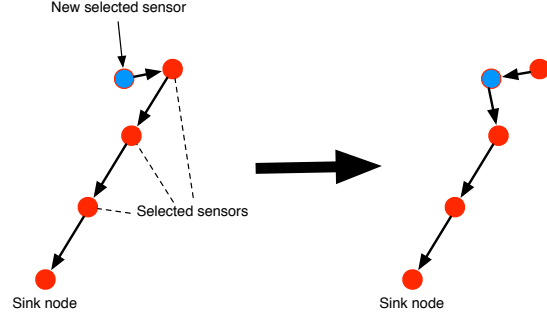


Figure 2: Successful backtracking operation performed by the algorithm

Algorithm 2. Greedy Approximation Algorithm

- Build Path_{d-SPT} from detection node d to the querying node
- $\mathbf{S} \leftarrow$ nodes in Path_{d-SPT} , $\mathbf{T} \leftarrow \text{Path}_{d-SPT}$
- $\mathbf{P} \leftarrow$ cost Path_{d-SPT}
- **While** $\mathbf{P} \leq P_{max}$
 - \mathcal{N}_i : node i neighbors
 - $\{i_k, l_k\} = \arg \min_{\{i \in \mathbf{S}, l \in \mathcal{N}_i, l \notin \mathbf{S}\}} (f_c(d_{i,l}) - \gamma_k h_l^2)$
 - $\mathbf{T} = \mathbf{T} \cup \{i_k, l_k\}$, $\mathbf{S} = \mathbf{S} \cup \{l_k\}$, $\mathbf{P} = \mathbf{P} + f_c(d_{i_k, l_k})$
 - **If** $\text{parent}_{i_k} \in \mathcal{N}_{l_k}$ & $f_c(d_{l_k, \text{parent}_{i_k}}) < f_c(d_{i_k, \text{parent}_{i_k}})$, **Then** update $\text{parent}_{l_k} = \text{parent}_{i_k}$, $\text{parent}_{i_k} = l_k$, $\mathbf{P} = \mathbf{P} - f_c(d_{i_k, \text{parent}_{i_k}}) + f_c(d_{l_k, \text{parent}_{i_k}})$
 - Update γ_k . **If** $h_{i_k} \geq h_{l_{k-1}}$ decrease γ_{k+1} . **Else**, increase γ_{k+1}
 - $k = k + 1$
- **EndWhile**

5. NUMERICAL RESULTS

In order to test the performance of our algorithms, we show numerical simulations. We test our algorithms using 100 different network topologies. Each topology consists of $N = 100$ nodes randomly placed in a square region centered at $[0,0]$. The network connectivity is based on a Euclidean distance model, and we assume that all the nodes have the same communication range. The sink node $N + 1$ is placed at $[0,0]$ and the target t is set to be at $[4,4]$. The cost of communication between nodes i and j is given by $f_c(d_{i,j}) = d_{i,j}^\alpha$ with $\alpha = 4$. Moreover, we take the measurement coefficient at node i to be inversely proportional to the distance from i to the target, i.e $h_i = 1/d_{i,t}$. For each topology, we have run our algorithms for a range of maximum power constraints.

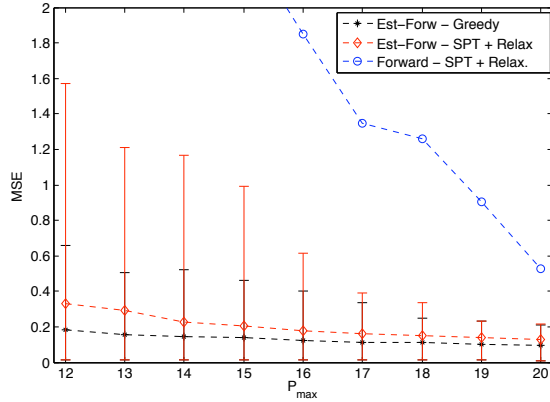


Figure 3: MSE estimation performance for the proposed algorithms

In Figure 3, we show the obtained values of MSE vs. total available power. First of all, as we already expected, we can observe that the estimation-and-forward strategy is much more energy efficient than the measurement forwarding one. As a performance assessment of the relaxation algorithm when the routing tree is fixed, we have calculated the mean sub-optimality gap, which is 4.5%.

Moreover, when considering the estimation-and-forward scenario, we see that the SPT routing structure is not optimal. Our greedy approximation algorithm outperforms the quasi-optimal solution with the SPT by a 25-45%, depending on the available power. An example of the resulting sensor selection and transmission structure after having run the greedy approximation algorithm in a test topology is shown in Figure 4. Moreover, we also show with error bars the set where the 90% of our simulation results have fallen into. The less the available power, the better our algorithm behaves with respect to the SPT routing. This is an important result in the context of WSNs, since these networks have usually severe power constraints.

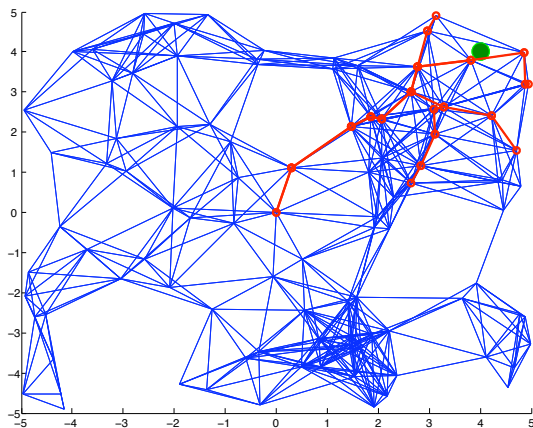


Figure 4: Example of the greedy approximation algorithm run in a specific topology for $P_{max}=16$

REFERENCES

- [1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.
- [2] G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Commun. ACM*, vol. 43, no. 5, pp. 51–58, 2000.
- [3] F. Zhao, J. Shin, and J. Reich, "Information-driven dynamic sensor collaboration," *Signal Processing Magazine, IEEE*, vol. 19, no. 2, pp. 61–72, Mar 2002.
- [4] J. Li and G. AlRegib, "Rate-constrained distributed estimation in wireless sensor networks," *Signal Processing, IEEE Transactions on*, vol. 55, no. 5, pp. 1634–1643, May 2007.
- [5] Junlin Li and G. AlRegib, "Maximizing network lifetime for estimation in multi-hop wireless sensor networks," *Computer Communications and Networks, 2008. ICCCN '08. Proceedings of 17th International Conference on*, pp. 1–6, Aug. 2008.
- [6] I. Rubin and Xiaolong Huang, "Capacity aware optimal activation of sensor nodes under reproduction distortion measures," *Military Communications Conference, 2006. MILCOM 2006*, pp. 1–8, Oct. 2006.
- [7] Z. Quan and A.H. Sayed, "Innovations-based sampling over spatially-correlated sensors," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 3, pp. III–509–III–512, 15–20 April 2007.
- [8] F. Bian, D. Kempe, and R. Govindan, "Utility-based sensor selection," *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on*, pp. 11–18, 19–21 April 2006.
- [9] R. Cristescu, B. Beferull-Lozano, M. Vetterli, and R. Wattenhofer, "Network correlated data gathering with explicit communication: Np-completeness and algorithms," *Networking, IEEE/ACM Transactions on*, vol. 14, no. 1, pp. 41–54, Feb. 2006.
- [10] Bhaskar Krishnamachari, *Networking Wireless Sensors*, Cambridge University Press, New York, NY, USA, 2005.
- [11] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
- [12] M. R. Garey and D. S. Johnson, *Computers and Intractability : A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*, W. H. Freeman, January 1979.
- [13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, March 2004.