ON THE INTERACTION OF DATA REPRESENTATION AND ROUTING IN SENSOR NETWORKS

Răzvan Cristescu*, Baltasar Beferull-Lozano†, Martin Vetterli†,‡, Deepak Ganesan§ and Jugoslava Acimovic†

* CMI, Caltech, Pasadena CA 91125, USA
† Dept. of IC, EPFL, Lausanne CH-1015, Switzerland
‡ Dept. of EECS, UC Berkeley, CA 94720, USA
§ University of Massachusetts, Amherst, MA 01003, USA
razvanc@caltech.edu, {baltasar.beferull, martin.vetterli}@epfl.ch

ABSTRACT

We consider data gathering by a network with a sink node and a tree communication structure, where the goal is to minimize the total transmission cost of transporting the information collected by the nodes, to the sink node. This problem requires a joint optimization of the data representation at the nodes and of the transmission structure. First, we study the case when the measured data are correlated random variables, both in the lossless scenario with Slepian-Wolf coding, and in the high-resolution lossy scenario with optimal rate-distortion allocation. We show that the optimal transmission structure is the shortest path tree, and we find in closed-form the rate and distortion allocation. Second, we study the case when the measured data are deterministic piecewise constant signals, and data is described with adaptive level wavelet-based multiresolution representation. We show experimentally that, when computation is decentralized, there is an optimal network division into node groups of adaptive size. Finally, we also analyze the node positioning problem where, given a correlation structure and an available number of sensors, the goal is to place the nodes optimally in terms of minimizing the transmission cost; our results show that important gains can be obtained compared to a uniformly distributed sensor positioning.

1. INTRODUCTION

1.1. Motivation

Consider a typical sensor network scenario [8], where sensors measure a data field (e.g. temperature) and the results of their measurements have to be transported across the network, to a certain designated node called the *sink* (see Fig. 1). This is referred to as *data gathering*, and it is a relevant problem in various sensor network settings, where data from the network is needed at a central base station node, for storage, monitoring or control purposes.

1.2. Measured Data and Network Characteristics

There are several important issues specific to sensor networks measuring and transporting data [8]. First, the measured data have certain redundancy characteristics. For instance, if the measured data are random variables (e.g. temperature), the values at nodes are correlated and the data structure is given by the spatial correlation. In a different setting, if the data is a deterministic quantity (e.g. threshold values for seismic data), then in most cases it can be
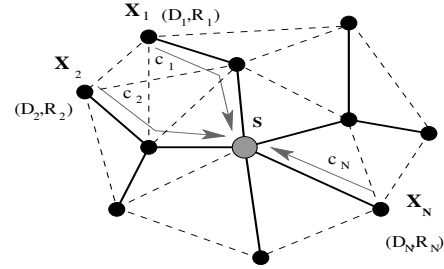

Fig. 1. In this example, data from nodes $X_1, X_2, \ldots, X_N$ need to arrive at sink S. A rate supply R_i is allocated to each node X_i, and, in the case of lossy coding, the distortion at that node is D_i. In thick solid lines, a chosen tree transmission structure is shown. In thin dashed lines, the other possible links are shown. The path from node i to the sink is shown in gray line, and its weight is c_i.

represented as a piecewise smooth function, with structure given by its spectral content, as a function of the node position.

Second, the limited coverage and transmission capabilities of the sensor nodes induce limited connectivity and communication patterns in the network graph. Nodes usually have knowledge only about other sensors situated in a limited neighborhood, so efficient joint representation of data by groups of nodes has to be done in a decentralized manner. Also, due to the battery power limitations, most nodes cannot send their data directly to the sink, and therefore data has to be relayed by other nodes. This implies that efficient routing is necessary, and moreover it has to be decentralized as well. Also, depending on the coding strategy that determines the amount of internode communication, the task of data representation at nodes may or may not separate from the task of routing that data across the network.

Third, the actual node positions influence both the accuracy of the measured data and the power efficiency of the network. For instance, placing most nodes close to the sink will improve their lifetime since they only have to transmit data on small distances; however, this will leave areas that are far from the sink uncovered, which means high inaccuracy in the overall data measurement. On the contrary, an even distribution of nodes over the measured field will provide a good data accuracy, but at the same time, the power consumption is large. Moreover, finding an optimized tradeoff has to take into account both data representation and routing.

1.3. Metrics

There are certain specific metrics of interest for this type of applications, namely power efficiency and accuracy of the data reconstruction at the sink. In sensor networks, the power efficiency of the network depends on both the rate allocation at nodes and on the routing strategy (the paths chosen to transmit the data). Namely, the power consumed by a node is usually proportional to the product [rate] × [path weight], where the [rate] term represents the data amount (in bits) sent by a node, and the [path weight] is an increasing function of the euclidean distance between nodes.

The accuracy of data reconstruction depends on the distortion allowed at measuring nodes and on the node placement, and influences the data representation as well. Namely, the desired accuracy of data representation at nodes determines the rate necessary to accommodate the corresponding distortion, and thus how much power is needed to transmit that rate. Due to spatial representation reasons, the accuracy is influenced by the node placement as well.

To summarize, there is a strong interconnection between data representation and routing, and the metrics relevant for sensor network scenarios. The goal of this work is to study the interaction among these important issues, for designing practical efficient and accurate joint measurement and transmission strategies.

1.4. Related Work

Progress towards practical implementation of Slepian-Wolf coding [10] has been achieved in [1, 9]. A joint treatment of data aggregation and transmission structure is considered in [7], but their model does not take into account possible collaborations for joint coding among nodes. The rate-distortion region of coding with high-resolution for arbitrarily correlated sources has been found in [11]. In some scenarios, uncoded transmission is optimal [6].

1.5. Main Contributions and Organization of the Paper

The main contribution of this work is an unified treatment of data representation, routing and node placement in sensor networks, for the optimization of various metrics of interest. In Section 2 we present the network and signal models analyzed in this paper. Section 3 studies data gathering of random processes, namely the cases of lossless and high-resolution lossy coding, and addresses the node placement problem. In Section 4 we study the communication costs in data gathering of deterministic signals with wavelet-based multiresolution processing. We conclude with Section 5.

2. PROBLEM SETTING

2.1. Network Model

Consider a network of N nodes. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be the vector formed by the values representing the sources measured at nodes $1, \ldots, N$. The information measured at nodes has to be transmitted through the links of the network to the designated base station (see Fig. 1). We will assume that the interference among nodes is negligible, and there are no capacity constraints on the links[1]. Such assumptions are realistic in the case of wired networks or if the antennas are unidirectional. For such scenarios, the optimal gathering structure is a tree.

In some parts of this work, for the sake of simplicity, we use the one-dimensional network model in Fig. 2 rather than the two-dimensional model in Fig. 1.

[1]The case of omnidirectional interfering wireless channels is beyond the scope of this paper.

Fig. 2. One-dimensional grid network.

2.2. Signal Model

2.2.1. Random Signals: Gaussian Random Field

We will consider first the case where $\mathbf{X} = (X_1, \ldots, X_N)$ is a vector formed by random variables representing the sources measured at nodes $1, \ldots, N$. The samples taken at nodes are spatially correlated and independent in time. We assume that the random variables are continuous and that there is a high-resolution quantizer in each sensor. In the lossless source coding case, a rate allocation $\{R_i\}_{i=1}^{N}$ (bits) has to be assigned at the nodes. In addition, if the data can be transmitted in a lossy manner, a distortion allocation $\{D_i\}_{i=1}^{N}$ has to be assigned, so that the quantized measured information samples are described with certain total D and individual $\{D_i^{max}\}_{i=1}^{N}$ distortion constraints.

For the sake of clarity, we use as example a zero-mean *jointly Gaussian model* $\mathbf{X} \sim \mathcal{N}^N(0, \mathbf{K})$, with unit variances $\sigma_{ii} = 1$:

$$f(\mathbf{X}) = \frac{1}{\sqrt{2\pi} \det(\mathbf{K})^{1/2}} e^{-\left(\frac{1}{2}(\mathbf{X})^T \mathbf{K}^{-1}(\mathbf{X})\right)}$$

where $\mathbf{K}$ is the covariance matrix of $\mathbf{X}$, with elements depending on the distance between the corresponding nodes (e.g. $K_{ij} = \exp(-a d_{i,j}^\beta), \beta \in \{1, 2\}$, where $d_{i,j}$ is the distance between nodes i and j [2, 4]). Although we will show numerical evaluations performed using the Gaussian random field model, our results are valid for any spatially correlated random processes, whose correlation decreases with distance.

2.2.2. Deterministic Signals: Piecewise Constant Signals

Second, for the deterministic case, we will consider a particular class of signals, namely time-invariant continuous piecewise constant signals with a finite number of (uniformly distributed) discontinuities placed at a-priori unknown positions. This type of signals appears in many practical problems, for instance in thresholding/alert scenarios. We assume that between each two discontinuities, the signal takes a value in the interval $[a, b]$, with a, b real numbers.

3. DATA GATHERING OF RANDOM PROCESSES

Consider data gathering of random processes. For a given network with connectivity graph $G = (V, E)$, we formulate our problem as follows:

$$\{\{R_i^*, c_i^*\}_{i=1}^{N}, ST^*\} = \arg \min_{R_i, c_i, ST} \sum_{i \in V} R_i c_i \qquad (1)$$

under constraints

$$\sum_{i \in \mathcal{X}} R_i \geq H(\mathcal{X} | \mathcal{X}^C), \forall \mathcal{X} \in V \qquad (2)$$

where $\{R_i\}_{i=1}^{N}$ is the rate allocation at nodes, ST is a spanning tree of G, and c_i is the total weight of the path connecting node i to S on the spanning tree ST; (2) are the Slepian-Wolf constraints on rates, for joint data representation at nodes.

3.1. Optimal Transmission Structure is *SPT*

Constraints (2) imply that nodes can code with any rate that obeys the constraint region *without explicitly exchanging data*. As a consequence, we can state the following theorem [2]:

Theorem 1 – *Separation of the joint optimization of source coding and transmission structure:*
The overall joint optimization (1) can be achieved by first optimizing the transmission structure with respect to only *the link weights c_i, and then optimizing the rate allocation for the given transmission structure under the constraints (2).*

Proof: By definition, for any given node, the cost function in (1) is separable as the product of a function that depends only on the rate allocated at that node, and another function that depends only on the link weights. Once the rate allocation is *fixed*, the best way (least cost) to transport any amount of data from a given node i to the sink S does not depend on the value of the rate R_i. Since this holds for any rate allocation, it is also true for the minimizing rate allocation and the result follows.∎

Theorem 1 implies that the optimal transmission structure that optimizes (1) is the shortest path tree, given by the superposition of the shortest paths from all nodes to the sink (ST^*=SPT). Denote by $c_i^* = d_{SPT}(i, S)$ the total weight of the path from node i to the sink S. For the rest of this paper, suppose without loss of generality that nodes are ordered in a list with increasing values of the weights corresponding to the shortest paths from each node to the sink, that is, $c_1^* \leq c_2^* \leq \cdots \leq c_N^*$.

3.2. Lossless Data Gathering

3.2.1. Rate Allocation

We can show that the solution of the optimal rate allocation under Slepian-Wolf constraints is [2]:

$$R_i^* = H(X_i|X_{i-1}, X_{i-2}, \dots), i = 1 \dots N, \qquad (3)$$

where $H(\cdot)$ is the entropy. That is, for optimal rate allocation, nodes code by conditioning on all the other nodes that are closer to the sink on the SPT.

3.2.2. Node Placement

Further, we consider the related problem where a given number of nodes is placed in a field such that the sensed data can be reconstructed at the sink within specified distortion bounds while minimizing the energy consumed for communication. Such a placement provides important power performance improvements as compared to uniform placement (see Fig. 3); exploiting data correlation by using Slepian-Wolf coding further improves the results [5].

3.3. Lossy Data Gathering

3.3.1. Rate-Distortion Allocation

Let us further consider the case when data at nodes is lossy coded with high-resolution [11], and the information measured by the nodes should be available at the sink within certain total and individual distortion bounds. A rate/distortion allocation $\{(R_i, D_i)\}_{i=1}^N$ (bits) has to be assigned at the nodes so that the quantized measured information samples are described with certain total D and
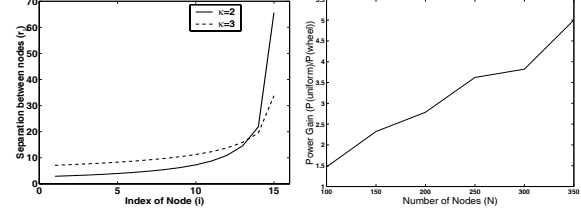
(a) Optimal placement for different pathloss exponents. (b) Power gain as a function of number of network nodes.

Fig. 3. Optimal placement for one-dimensional network (left); power improvement over uniform placement for the two-dimensional network (right).
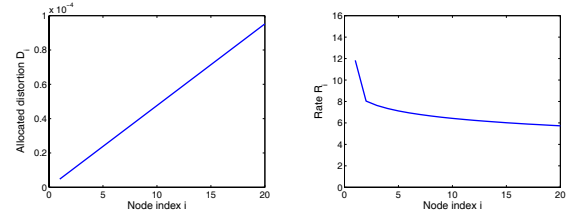
(a) Distortion allocation as function of node index. (b) Rate allocation as a function of node index.

Fig. 4. One-dimensional network, average distortion constraint: distortion and rate allocations as a function of the node index.

individual $D_i^{max}, i = 1, \dots, N$ distortions. Then, the most general form of our optimization problem is given as follows:

$$\{\{R_i^*, D_i^*, c_i^*\}_{i=1}^N, ST^*\} = \arg \min_{R_i, D_i, c_i, ST} \sum_{i=1}^N c_i R_i$$

under constraints
$$\sum_{i \in \mathcal{X}} R_i \geq h(\mathcal{X}|V \backslash \mathcal{X}) - \log \prod_{i \in \mathcal{X}} 2\pi e D_i, \forall \mathcal{X} \subset V$$

$$\sum_{i=1}^N D_i \leq D, \qquad D_i \leq D_i^{\max}, i = 1 \dots N..$$

In the high rate regime, uniform quantization and Slepian-Wolf coding is optimal [11]. Thus, a similar result as the one stated in Theorem 1 holds in the lossy case about the separation between transmission optimization and rate/distortion allocation, since nodes do not need to communicate explicitly to code data with a given rate/distortion allocation. Thus, the SPT is the optimal transmission structure in this case as well.

Next, assume that only the average distortion constraint is active. Analogously to (3) we obtain the closed-form solution of the rate-distortion allocation at nodes (see Fig. 4):

$$R_i^* = h(X_i|X_{i-1}, \dots, X_1) - \log \frac{2\pi e D c_i^*}{C}, i = 1 \dots N. \quad (4)$$

where $h(\cdot)$ is the differential entropy and $C = \sum_{i=1}^N c_i^*$. The case of individual distortion constraints can further be easily solved.

3.3.2. Node Placement

We now study the problem of optimal placement for two power efficiency targets of interests, namely total power and network life-

time[2], and compare the tradeoffs involved. For the one-dimensional example in Fig. 2, the optimal placement is:

$$w_i^* = \frac{L}{(\sum_{j=i}^N R_j^*)^\kappa \left(\sum_{l=1}^N \frac{1}{(\sum_{j=l}^N R_j^*)^\kappa} \right)}, i = 1 \ldots N,$$

where $\{w_i^*\}_{i=1}^N$ is the distance between nodes $i-1$ and i; $\kappa = 1$ for total power minimization, and $\kappa = 1/2$ for lifetime maximization. The optimal joint solution for the placement and rate allocation is obtained by using an iterative algorithm:

Algorithm 1: Placement and rate allocation.

- Initialize uniformly the node placement $\{w_i\}_{i=1}^N = L/N$.

- **until** convergence **do**:

    1. Given $\{w_i\}_{i=1}^N$, solve the power minimization problem for $\{R_j\}_{j=1}^N$.

    2. Re-write $\{w_i\}_{i=1}^N$ as a function of $\{R_j\}_{j=1}^N$.

Our current research is focused on optimization problems where one of the two targets of interest is minimized under upper bound constraints on the other target.

4. DATA GATHERING OF DETERMINISTIC SIGNALS

In this section we will perform a communication cost analysis for data gathering of piecewise constant signals [3], by studying an adaptive multilevel wavelet-based algorithm. Similarly to Section 3, the cost function is energy related, and given by the product [rate] × [path weight]; in this section, the [rate] term represents the number of bits needed to code the transform coefficient.

4.1. Haar Transform and Signal Approximation

Without loss of generality, we consider the M-level Haar wavelet transform, which provides an efficient data representation for piecewise constant processes[3]. For each level $k = 1 \ldots M$ of the transform, low-pass (LP) and high-pass (HP) coefficients are computed as sum and differences of lower level coefficients. These coefficients are sufficient for signal reconstruction in the network sink.

Let us further consider in this section the one-dimensional network. For the intervals where the signal is constant, all HP coefficients are zero. We assume that the power required to transmit zero-valued coefficients is negligible, and thus only the LP coefficients are transmitted. If, for instance, the process has a single discontinuity, then there is at most one non-zero HP coefficient in each multiresolution level k, which results in at most M non-zero HP coefficients for the whole representation.

4.2. Adaptive Network Segmentation

The algorithm starts with odd index sensors sending their data to the even index sensors, which compute the corresponding HP and LP wavelet coefficients. Each receiving sensor makes a decision about whether to further create a 2-sensor group, by comparing the costs of data gathering the wavelet coefficients of the new group with the sum of costs corresponding to the existing 1-sensor groups. After this operation, sensors $2, 4, \ldots, N$ have information about their neighbors with odd index. In the second step, they

---

[2]We address network lifetime optimization by considering the constraint that all nodes use equal power.

[3]This can be easily generalized to the case of piecewise polynomial signals by using the appropriate higher-order wavelet processing.
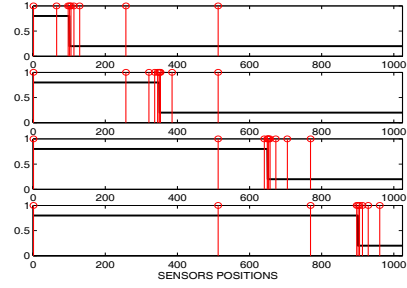

Fig. 5. Cost-efficient network segmentation for various positions of a discontinuity (the sink is at the right). Stem lines indicate the positions of nodes that send aggregated data directly to sink.

transmit data further to leader-nodes of 4-sensor groups. A new set of coefficients is computed at the leader-nodes (precisely, two new coefficients). The decision about collapsing the groups into one larger group is essentially determined by the values of the computed HP coefficients. The algorithm proceeds similarly for the other levels: for $k = 3, \ldots, M+1$ there are $\frac{N}{2^{k-1}}$ groups of size 2^{k-1}. For each group the total cost of data gathering is compared with the data gathering cost of nodes in that group corresponding to the previous algorithm steps. Further grouping is performed only if it decreases the cost. The algorithm stops when all data reach the network sink (see Fig. 5). Our experiments show that this procedure provides important gains with respect to raw data gathering, where no transform is used to transmit data at nodes.

5. CONCLUSIONS

We studied the interaction between data representation at nodes, routing and node placement, for gathering of redundant data in sensor networks. We analyzed both random spatially correlated processes, and deterministic signals. Our results show that a joint consideration of these issues provides important improvements in the overall data gathering power efficiency.

6. REFERENCES

[1] Aaron, A., Girod, B.: *Compression with Side Information Using Turbo Codes*, in Proc. DCC 2002.

[2] R. Cristescu, B. Beferull-Lozano, M. Vetterli: *On Network Correlated Data Gathering*, in Proc. INFOCOM 2004.

[3] A. Ciancio, A. Ortega: *A Distributed Wavelet Compression Algorithm for Wireless Sensor Networks using Lifting*, in Proc. ICASSP 2004.

[4] N. Cressie: *Statistics for Spatial Data*, Wiley 1991.

[5] D. Ganesan, R. Cristescu, B. Beferull-Lozano: *Power-Efficient Sensor Placement and Transmission Structure for Data Gathering under Distortion Constraints*, submitted to ACM Trans. on Sensor Networks, Oct. 2004.

[6] M. Gastpar, B. Rimoldi, M. Vetterli: *To Code, or Not to Code: Lossy Source-Channel Communication Revisited*, IEEE Tr. on Inf. Th., 2003.

[7] A. Goel, D. Estrin: *Simultaneous Optimization for Concave Costs: Single Sink Aggregation or Single Source Buy-at-Bulk*, ACM-SIAM Symp. on Discrete Alg., 2003.

[8] G. J. Pottie, W. J. Kaiser: *Wireless Integrated Sensor Networks*, Communications of ACM 2000.

[9] S. Pradhan, K. Ramchandran: *Distributed Source Coding Using Syndromes: Design and construction,* in Proc. DCC 1999.

[10] D. Slepian, J.K. Wolf: *Noiseless Coding of Correlated Information Sources*, IEEE Trans. Inf. Th., 1973.

[11] R. Zamir, T. Berger, *Multiterminal Source Coding with High Resolution*, IEEE Trans. on Inf. Th., 1999.