# Networked Slepian-Wolf: Theory and Algorithms*

Răzvan Cristescu[1], Baltasar Beferull-Lozano[1], and Martin Vetterli[1,2]

[1] Laboratory for Audio-Visual Communications (LCAV),
Swiss Federal Institute of Technology (EPFL),
Lausanne CH-1015, Switzerland
[2] Department of EECS, University of California at Berkeley,
Berkeley CA 94720, USA
{Razvan.Cristescu,Baltasar.Beferull,Martin.Vetterli}@epfl.ch
http://ip7.mics.ch/

**Abstract.** In this paper, we consider the minimization of a relevant energy consumption related cost function in the context of sensor networks where correlated sources are generated at various sets of source nodes and have to be transmitted to some set of sink nodes. The cost function we consider is given by the product [rate] × [link weight]. The minimization is achieved by jointly optimizing the transmission structure, which we show consists of a superposition of trees from each of the source nodes to its corresponding sink nodes, and the rate allocation across the source nodes. We show that the overall minimization can be achieved in two concatenated steps. First, the optimal transmission structure has to be found, which in general amounts to finding a Steiner tree and second, the optimal rate allocation has to be obtained by solving a linear programming problem with linear cost weights determined by the given optimal transmission structure. We also prove that, if any arbitrary traffic matrix is allowed, then the problem of finding the optimal transmission structure is NP-complete. For some particular traffic matrix cases, we fully characterize the optimal transmission structures and we also provide a closed-form solution for the optimal rate-allocation. Finally, we analyze the design of decentralized algorithms in order to obtain exactly or approximately the optimal rate allocation, depending on the traffic matrix case. For the particular case of data gathering, we provide experimental results showing a good performance in terms of approximation ratios.

## 1 Introduction

### 1.1 Problem Motivation

Consider networks that transport supplies among nodes. This is for instance the case in sensor networks that measure some environmental data. Nodes are

* The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communications Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322.

H. Karl, A. Willig, A. Wolisz (Eds.): EWSN 2004, LNCS 2920, pp. 44–59, 2004.

supplied amounts of measured data which need to be transmitted to end sites, called sinks, for control or storage purposes. An example is shown in Fig.1, where there are $N$ nodes with sources $X_1, \ldots, X_N$, two of them being the sinks denoted by $S_1, S_2$, and a graph of connectivity with edges connecting certain nodes. We will use interchangeably the notions of network entity and its graph representation across the paper. Sources corresponding to nodes in the sets $V^1, V^2$ need to transmit their data, possibly using other nodes as relays, to sinks $S_1, S_2$ respectively. A very important task in this scenario is to find a rate allocation at nodes and a transmission structure on the network graph that minimizes a cost function of interest (e.g. flow cost [data size] $\times$ [link weight], total distance, etc.). This implies a joint treatment of source coding and optimization of the transmission structure.

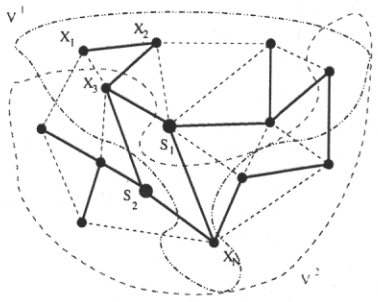



**Fig. 1.** An example of a network. Sources transmit their data to the sinks. Nodes from the $V^1, V^2$ set of nodes need to arrive at sink $S_1, S_2$, respectively. A rate supply $R_i$ is allocated to each node $X_i$. In solid lines, a chosen transmission structure is shown. In dashed lines, the other possible links are shown

The problem is trivial if the data measured at nodes are statistically independent: each node codes its data independently, and well developed algorithms can be used to solve the minimum cost flow problem [5].

However, in many situations, data at nodes are *not* independent, such as in typical sensor networks. It can be thus expected that approaches that take into account the correlation present in the data can improve over existing algorithms, with regard to optimizing many cost metrics of interest.

### 1.2 Correlated Data

The source coding approach that takes maximum advantage of the data correlation, at the expense of coding complexity, is based on the important work of [19]. In that work, Slepian and Wolf showed that when nodes measure correlated data, these data can be coded with a total rate not exceeding the joint entropy, even *without* nodes explicitly communicating with each other (under some constraints on the minimal rates given by the Slepian-Wolf region). Their result

provides the whole achievable *rate region* for the rate allocation, that is *any* rate in that region is achievable. We describe in more detail the Slepian-Wolf coding in Sect. 2.2.

In addition to encoding the data, these data usually need to be transmitted over the network from the sources to the sinks. In such situations, it is important to study the influence that the transmission structure used to transport the data has on the rate allocation at the nodes. In this work, we consider a joint treatment of both the rate allocation and the chosen transmission structure. We show that when separable joint cost functions are considered (e.g. the [data size] $\times$ [link weight] metric), and Slepian-Wolf coding is used, then the problem of joint optimization separates, in the sense that first an optimal transmission structure needs to be determined, and second the optimal rate allocation is found on this transmission structure. However, the rate allocation is determined by the transmission structure, and thus the respective optimizations are not independent. The optimal rate allocation is in general unique, except in some degenerate cases. Since nodes have limited processing capability and/or battery power, it is necessary that the rate allocation and transmission structure optimization are done locally at each node, in a decentralized manner, by using information available only from nodes in the neighborhood.

In particular, let us consider the case of a network of sensors taking measurements from the environment [2], [12], [15]. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be the vector formed by the random variables representing the sources measured at the nodes $1, \ldots, N$. We assume that the random variables are continuous and that there is a quantizer in each sensor (with the same resolution for all sensors). There are also a number of sinks to where data from different subsets of nodes have to be sent. A rate allocation $(R_1, \ldots, R_N)$ (bits) has to be assigned at the nodes so that the quantized measured information samples are described losslessly. Notice that it is also possible to allocate different rates at each node, depending on which sink it sends its data to, but this involves important additional coding overhead, which might not be always feasible. We consider both cases in this paper. We assume that the spatial correlation between samples taken at the nodes depends in our setting only on the distance distribution across space. In this work, we assume that contention is solved by the upper layers. The transmission topology in our model is assumed to be an undirected graph with point-to-point links. Practical approaches use nearest neighbor connectivity, avoiding thus the complexity of the wireless setting. Since battery power is the scarce resource for autonomous sensors, a meaningful metric to minimize in the case of sensor networks is the total energy consumption. This is essentially given by the sum of products [data size] $\times$ [link weight] for all the links used in the transmission. The weight of the link between two nodes is a function of the distance $d$ of the two nodes (e.g. $kd^{\alpha}$ or $k\exp(\alpha d)$, with $k$, $\alpha$ constants of the medium).

The novelty of our approach stems from the fact that we consider jointly the optimization of both source coding and transmission structure in the context of sensor networks measuring correlated data. To the best of our knowledge, this is

the first research work that addresses jointly Slepian-Wolf lossless source coding and network flow cost optimization.

### 1.3 Related Work

Progress towards practical implementation of Slepian-Wolf coding has been achieved in [1], [13], [14]. Bounds on the performance of networks measuring correlated data have been derived in [11], [18]. However, none of these works takes into consideration the cost of transmitting the data over the links and the additional constraints that are imposed on the rate allocation by the joint treatment of source coding and transmission.

The problem of optimizing the transmission structure in the context of sensor networks has been considered in [9], [16], where the energy, and the [energy] $\times$ [delay] metric are studied, and practical algorithms are proposed. In these studies, the correlation present in the data is not exploited for the minimization of the metric.

A joint treatment of data aggregation and the transmission structure is considered in [8]. The model in [8] does not take into account possible collaborations among nodes. In our work, we consider the case of collaboration between nodes because we allow nodes to perform (jointly) Slepian-Wolf coding.

### 1.4 Main Contributions

In this paper, we address the problem of Slepian-Wolf source coding for general networks and patterns of traffic, namely, in terms of our graph representation, for general (undirected) graphs and different sets of source nodes and sinks. We consider the flow cost metric given by [data size]$\times$[link weight], and we assess the complexity of the resulting joint optimization problem, for various network settings. We first prove that the problem can be always be separated into the tasks of transmission structure optimization and rate allocation optimization. For some particular cases, we provide closed-form solutions and efficient approximation algorithms. These particular cases include correlated data gathering where there is only one sink node. In the general case, we prove that the problem is NP-complete.

The rest of this paper is organized as follows. In Sect. 2, we state the optimization problem and describe the Slepian-Wolf source coding approach and the optimal region of rate allocations. In Sect. 3 we study the complexity for the case of a general traffic matrix problem and we prove that finding the optimal transmission structure is NP-complete. We show that, if centralized algorithms were allowed, finding the optimal rate allocation is simple; however, in our sensor network setting, the goal is to find distributed algorithms, and we show that in order to have a decentralized algorithm, we need a substantially large communication overhead in the network. In Sect. 4, we fully solve an important particular case, namely the correlated data gathering problem with Slepian-Wolf source coding. In Sect. 5 we consider other particular cases of interests. Finally, we present some

numerical simulations in Sect. 6. We conclude and present directions of further work in Sect. 7.

## 2 Problem Formulation

### 2.1 Optimization Problem

Consider a graph $G = (V, E)$, $|V| = N$. Each edge $e \in E$ is assigned a weight $w_e$. Nodes on the graph are sources of data. Some of the nodes are also sinks. Data has to be transported over the network from sources to sinks. Denote by $S_1, S_2, \ldots, S_M$ the set of sinks and by $V^1, V^2, \ldots, V^M$ the set of subsets $V^j \subseteq V$ of sources; data measured at nodes $V^j$ have to be sent to sink $S_j$. Denote by $S^i$ the set of sinks to which data from node $i$ have to be sent. Denote by $E^i \subseteq E$ the subset of edges used to transmit data from node $i$ to sinks $S^i$, which determines the transmission structure corresponding to node $i$.

**Definition 1 (Traffic matrix).** *We call the* traffic matrix *of a graph $G$ the $N \times N$ square matrix $T$ that has elements given by:*

$$T_{ij} = 1, \text{ if source } i \text{ is needed at sink } j,$$
$$T_{ij} = 0, \text{ else.}$$

With this notation, $V^j = \{i : T_{ij} = 1\}$ and $S^i = \{j : T_{ij} = 1\}$.

The overall task we consider is to assign an optimal rate allocation $R_i^*, i = 1, \ldots, N$ for the $N$ nodes and to find the optimal transmission structure on the graph $G$ that minimizes the total flow cost [data size] $\times$ [link weight]. Thus, the optimization problem is:

$$\{R_i^*, d_i^*\}_{i=1}^N = \arg_{\{R_i, d_i\}} \min \sum_{i=1}^N R_i d_i \tag{1}$$

where $d_i$ is the total weight of the transmission structure chosen to transmit data from source $i$ to the set of sinks $S^i$:

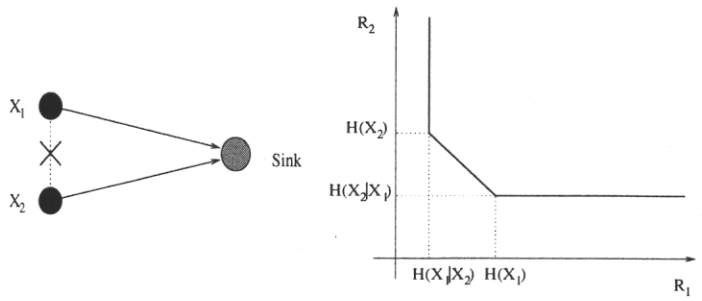
$$d_i = \sum_{e \in E^i} w_e.$$

Notice that finding the optimal $\{d_i\}_{i=1}^N$ is equivalent to finding the optimal transmission structure.

In the next Sect. 2.2 we show that, when Slepian-Wolf coding is used, the tasks of finding the optimal $\{d_i\}_{i=1}^N$ and respectively $\{R_i\}_{i=1}^N$ are separated, that is, one can first find the optimal transmission structure, which can be shown to be always a tree, and then find the optimal rate allocation. As a consequence, after finding the optimal transmission structure, (1) can be posed as a linear programming problem in order to find the optimal rate allocation. We study the complexity of solving the overall problem under various scenarios.

### 2.2 Slepian Wolf Coding

Consider the case of two random sources $X_1$ and $X_2$ that are correlated (see Fig. 2(a)). Intuitively, each of the sources can code their data at a rate equal to at least their corresponding entropies, $R_1 = H(X_1)$, $R_2 = H(X_2)$, respectively. If they are able to communicate, then they could coordinate their coding and use together a total rate equal to the joint entropy $R_1 + R_2 = H(X_1, X_2)$. However, Slepian and Wolf [19] showed that two correlated sources can be coded with a total rate $H(X_1, X_2)$ even if they are *not* able to communicate with each other. This can be also easily generalized to the $N$-dimensional case. Fig. 2(b) shows the Slepian-Wolf rate region for the case of two sources.




(a) Two correlated sources $X_1, X_2$ send their data to one sink

(b) The Slepian-Wolf region shows the achievable pairs of rates that can be allocated to sources $X_1$ and $X_2$ for lossless data coding

**Fig. 2.** Two correlated sources, and the Slepian-Wolf region for their rate allocation

Consider again the example shown in Fig. 1. Assume that the set of sources that send their data to sink $j$, that is the set of nodes denoted $\{X_{j1}, \ldots, X_{j|V^j|}\} \in V^j, j = 1, 2$, know in advance the correlation structure in that set $V^j$ (which depends only on the distance in our model). This is a reasonable assumption to make for localized data requests from the sources (that is, when nodes in $V^j$ are geographically close to each other). Then, nodes in $V^j$ can code their data jointly, without communicating with each other with a total rate of $H(X_{j1}, X_{j2}, \ldots, X_{j|V^j|})$ bits, as long as their individual rates obey the Slepian-Wolf constraints related to the different conditional entropies [6], [19].

**Proposition 1 (Separation of source coding optimization and transmission structure optimization).** *When Slepian-Wolf coding is used, the transmission structure optimization separates from the rate allocation optimization, in terms of the overall minimization of (1).*

*Proof.* Once the rate allocation is *fixed*, the best way to transport any amount of data from a given node $i$ to the set of sinks $S^i$ does not depend on the value of the rate. This is true because we consider separable flow cost functions, and the rate supplied at each node does not depend on the incoming flow at that node. Since this holds for any rate allocation, it is true for the minimizing rate allocation and the results follows. □

For each node $i$, the optimal transmission structure is in fact a tree that spans node $i$ and the sinks $S^i$ to which its data are sent [5]. Thus, the whole optimization problem can be separated into a spanning tree optimization for each source, and the rate allocation optimization. Then, after the optimal tree structure is formed, (1) becomes a problem of rate allocation that can be posed as a linear programming (LP) problem under the usual Slepian-Wolf linear constraints:

$$\begin{aligned} &\min_{\{R_i\}_{i=1}^N} \sum_{i=1}^{N} R_i d_i^* \\ &\text{under constraints:} \\ &\sum_{l \in \mathbf{Y}_j} R_l \geq H(\mathbf{Y}_j | V^j - \mathbf{Y}_j), (\forall) V^j, \mathbf{Y}_j \subseteq V^j, \end{aligned} \tag{2}$$

that is, first the optimal weights $\{d_i^*\}_{i=1}^N$ are found (which determine in general uniquely the optimal transmission structure), and then the optimal rate allocation is found using the fixed values $\{d_i^*\}_{i=1}^N$ in (2). Note that there is one set of constraints for each set $V^j$.

Moreover, note that (2) is an LP optimization problem under linear constraints, so if the weights $\{d_i^*\}_{i=1}^N$ can be determined, the optimal allocation $\{R_i\}_{i=1}^N$ can be found easily with a *centralized* simplex algorithm [10]. However, in Sect. 3 we will show that for a general traffic matrix $T$, finding the optimal coefficients $\{d_i^*\}_{i=1}^N$ is NP-complete. Moreover, in general, even if the optimal structure is found, it is hard to *decentralize* the algorithm that finds the optimal solution $\{R_i^*\}_{i=1}^N$ of (2), as this requires a substantial amount of global knowledge of the network.

In the following sections, for various problem settings, we first show how the transmission structure can be found (i.e. the values of $\{d_i^*\}_{i=1}^N$), and then we discuss the complexity of solving (2) in a decentralized manner.

## 3 Arbitrary Traffic Matrix

We begin the analysis with the most general case, that is when the traffic matrix $T$ is arbitrary, by showing the following proposition:

**Proposition 2 (The optimal transmission structure is a superposition of Steiner trees).** *Given an arbitrary traffic matrix $T$, then, for any $i$, the optimal value $d_i^*$ in (2) is given by the minimum weight tree rooted in node $i$ and which spans the nodes in $S^i$; this is exactly the minimum Steiner tree that has node $i$ as root and which spans $S^i$, which is an NP-complete problem.*

*Proof.* The proof is straightforward: data from node $i$ has to be send over the minimum weight structure to the nodes in $S^i$, possibly via nodes in $V - \{i, S^i\}$. This is a minimum Steiner tree problem for the graph $G$ with weights $w_e$, thus it is NP-complete. □

The approximation ratio of an algorithm that finds a solution for an optimization problem is defined as the guaranteed ratio between the cost of the found solution and the optimal one. If the weights of the graph are the Euclidean distances ($w_e = l_e$ for all $e \in E$), then the problem becomes the Euclidean Steiner tree problem, and it admits a PTAS [3] (that is, for any $\epsilon > 0$, there is a polynomial time approximation algorithm with an approximation ratio of $1 + \epsilon$). However, in general, the link weights are not the Euclidean distances (e.g. if $w_e = l_e^2$ etc.). Then finding the optimal Steiner tree is APX-complete (that is, there is a hard lower bound on the approximation ratio), and is only approximable (with polynomial time in the input instance size) within a constant factor $(1 + \ln 3)/2$ [4], [17].

The approximation ratios of the algorithms for solving the Steiner tree translate into bounds of approximation for our problem. By using the respective approximation algorithms for determining the weights $d_i$, the cost of the approximated solution for the joint optimization problem will be within the Steiner approximation ratio away from the optimal one.

Once the optimal weights $d_i^*$'s are found (i.e. approximated by some approximation algorithm for solving the Steiner tree), then, as we mentioned above, (2) becomes a Linear Programming (LP) problem. Consequently, it can be readily solved with a centralized program. The solution of this problem is given by the innermost corner of the Slepian-Wolf region that is tangent to the cost function (see Fig. 3 for an example with two nodes or sources). If global knowledge of the network is allowed, then this problem can be solved computationally in a simple way. However, it is not possible in general to find in closed-form the optimal solution determined by the corner that minimizes the cost function, and consequently the derivation of a decentralized algorithm for the rate allocation, as this involves exchange of network knowledge among the clusters.

Figure 4 shows a simple example (but sufficiently complete) which illustrates the difficulty of this problem. Suppose that the optimal total weights $\{d_i^*\}_{i=1}^3$ in (2) have been approximated by some algorithm. Then the cost function to be minimized is:

$$R_1 w_{11} + R_2(w_{21} + w_{22}) + R_3 w_{32}$$

with $d_1^* = w_{11}, d_2^* = w_{21} + w_{22}, d_3^* = w_{32}$, and the Slepian-Wolf constraints are given by:

$$\begin{aligned} R_1 + R_2 &\geq H(X_1, X_2) \\ R_1 &\geq H(X_1|X_2), \text{ for set } V^1 = \{X_1, X_2\} \\ R_2 &\geq H(X_2|X_1) \end{aligned}$$

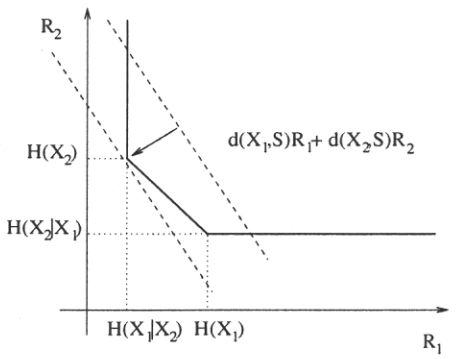



**Fig. 3.** A simple example with two nodes. The total weights $d_1, d_2$, from sources $X_1, X_2$ to the sinks, are respectively $d(X_1, S), d(X_2, S)$, $d(X_1, S) > d(X_2, S)$, in this particular case. In order to achieve the minimization, the cost line $d(X_1, S)R_1 + d(X_2, S)R_2$ has to be tangent to the most interior point of the Slepian-Wolf rate region, given by $(R_1, R_2) = (H(X_1|X_2), H(X_2))$

and respectively,

$$\begin{aligned} R_2 + R_3 &\geq H(X_2, X_3) \\ R_2 &\geq H(X_2|X_3), \text{ for set } V^2 = \{X_2, X_3\} \\ R_3 &\geq H(X_3|X_2). \end{aligned}$$

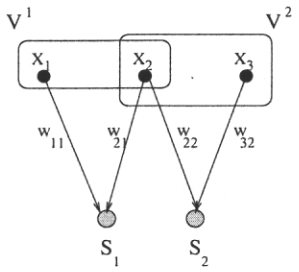



**Fig. 4.** Two sets of sources transmit their correlated data to two sinks

Suppose the weights are such that $w_{11} < w_{21} + w_{22} < w_{32}$. A decentralized algorithm has to use only local information, that is, information only available in a certain local transmission range or cluster neighborhood). We assume that only local Slepian-Wolf constraints are considered in each set $V^j$ for the rate allocation, and no knowledge about the total weights $d_i$ from nodes in the other subsets is available. Then, it readily follows that the optimal rate allocations in

each of the two subsets are:

$$R'_1 = H(X_1)$$
$$R'_2 = H(X_2|X_1), \text{ for set } V^1$$

and respectively,

$$R'_2 = H(X_2)$$
$$R'_3 = H(X_3|X_2), \text{ for set } V^2.$$

If the rate allocation at a source $X_i$ can take different values $R_{ij}$ depending to which sink $j$ source $X_i$ sends its data, then it is straightforward to find this multiple rate assignment. The rate allocation for each cluster will be independent from the rate allocations in the other clusters (see Sect. 4 for the closed-form solution for the rate allocation for each set $V^j$, and a decentralized algorithm for finding the optimal solution).

However, this involves even more additional complexity in coding, so in some situations it might be desirable to assign a unique rate to each node, regardless of which sink the data is sent to. We can see from this simple example that we cannot assign the correct unique optimal rate $R_2$ to source $X_2$, unless node 2 has global knowledge of the whole distance structure from nodes $1, 2, 3$ to the sinks $S_1$ and $S_2$. For a general topology, knowledge in a node from the nodes belonging to other different sets is needed at least at nodes that are at the intersection of sets (in this example, source $X_2$). Even so, it is clear that such global sharing of cluster information over the network is not scalable because the amount of necessary global knowledge grows exponentially.

There are however some important special cases of interest where the problem is tractable, and we treat them in the following two sections.

## 4 Data Gathering: All Sources ($V^j = V$) Sent to One Sink $S = j$

This case has been studied in the context of the *network correlated data gathering problem* [7], and is a particular case of the problem we consider in this paper. An example is shown in Fig. 5.

In this case, the problem simplifies: if there is a single sink $S$, then the Steiner tree rooted at $i$ and spanning node $S$ is actually the shortest path, of total weight $d_i$, between the two nodes. The overall optimal transmission structure is thus the superposition of the shortest paths from each node $i$ to the sink $S$. This superposition forms the shortest path tree ($SPT$) rooted in $S$. The $SPT$ can be easily found with a distributed algorithm (e.g. Bellman-Ford).

Let us review in Sect. 4.1 the results contained in [7].

### 4.1 Solution of the LP Problem

The algorithm for finding the optimal rate allocation for this setting is:

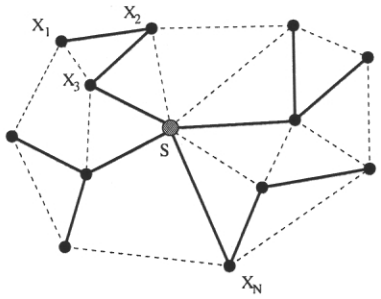



**Fig. 5.** In this example, data from nodes $X_1, X_2, \ldots, X_N$ need to arrive at sink $S$.

**Algorithm 1.** *Data gathering optimal Slepian-Wolf rate allocation:*

- *Find the weights* $d_i = d_{SPT}(i, S)$*, for each node* $i$*, given by the SPT of the graph* $G$*, by running e.g. distributed Bellman-Ford.*
- *Solve the constrained LP:*

$$(R_1^*, \ldots, R_N^*) = \arg\min_{\{R_i\}} \sum_i R_i d_{SPT}(i, S), \quad \text{under constraints:} \quad \sum_{i \in \mathbf{Y}} R_i \geq H(\mathbf{Y}|\mathbf{Y}^C), (\forall)\mathbf{Y} \subseteq V \tag{3}$$

*where* $d_{SPT}(i, S)$ *is the total length of the path in the SPT from node* $i$ *to* $S$*, and* $(R_1^*, \ldots, R_N^*)$ *is the optimal rate allocation.*

As discussed in Sect. 3, we see that in order to express the rate constraints, centralized knowledge of the correlation structure among *all* nodes in the network is needed. Nevertheless, in this case there is a single set of constraints that involves all the nodes, and because of this the solution for the rate allocation can be expressed in a closed-form.

Suppose without loss of generality that nodes are numbered in increasing order of their distance to the sink on the *SPT*: $(X_1, X_2, \ldots, X_N)$ with $d_{SPT}(X_1, S) \leq d_{SPT}(X_2, S) \leq \cdots \leq d_{SPT}(X_N, S)$.

**Proposition 3 (LP solution).** *The solution of the optimization problem in (3) is:*

$$\begin{aligned} R_1^* &= H(X_1), \\ R_2^* &= H(X_2|X_1), \\ &\cdots\cdots\cdots \\ R_N^* &= H(X_N|X_{N-1}, X_{N-2}, \ldots, X_1). \end{aligned} \tag{4}$$

That means that each node codes its data with a rate equal to its respective entropy conditioned on all other nodes which are closer to the sink than itself.

### 4.2 Approximation Algorithm

In the previous subsection, we present the optimal solution of the linear programming rate assignment for the single sink data gathering problem, under Slepian-Wolf constraints. We consider now the problem of designing a distributed approximation algorithm. Even if we can provide the solution in a closed form as (4), the nodes still need local knowledge of the overall structure of the network (distances between nodes and distances to the sink). This local knowledge is needed for:

1. Ordering the distances on the *SPT* from the nodes to the sink: each node needs its index in the ordered sequence of nodes so as to determine on which other nodes to condition when computing its rate assignment.
2. Computation of the rate assignment:

$$\begin{aligned} R_i &= H(X_i|X_{i-1},\ldots,X_1) \\ &= H(X_1,\ldots,X_i) - H(X_1,\ldots,X_{i-1}) \end{aligned}$$

   Note that *all* distances among nodes $(1,\ldots,i)$ are needed locally at node $i$ for computing this rate assignment.

Such global knowledge might not be available. Thus, we propose a fully distributed approximation algorithm, which avoids the need for a node to have global knowledge of the network, and which provides solutions very close to the optimum.

Suppose each node $i$ has complete information (distances between nodes and distances to the sink) only about a local vicinity $\mathcal{N}(i)$. This information can be computed by running for example a distributed algorithm for finding the *SPT* (e.g. Bellman-Ford). The approximation algorithm that we propose is based on the observation that nodes that are outside this neighborhood count very little, in terms of rate, in the local entropy conditioning, under the assumption that the correlation decreases with the distance between nodes, which is a natural assumption.

**Algorithm 2.** *Approximated Slepian-Wolf coding:*

- *Find the SPT.*
- *For each node $i$:*
  - *Find in the neighborhood $\mathcal{N}(i)$ the set $C_i$ of nodes that are closer to the sink, on the SPT, than node $i$.*
  - *Transmit at rate $R_i^\dagger = H(X_i|C_i)$.*

This means that data are coded locally at the node with a rate equal to the conditional entropy, where the conditioning is performed *only* on the subset formed by the neighbor nodes which are closer to the sink than the respective node.

The proposed algorithm needs only local information, so it is completely distributed. Still, it will give a solution very close to the optimum since the

neglected conditioning is small in terms of rate for a correlation function that is sufficiently decaying with distance (see Sect. 6 for some numerical simulations).

Similar techniques can be used to derive decentralized approximation algorithms for some of the other particular cases of interests that we discuss in the next section.

## 5 Other Particular Cases

### 5.1 Broadcast of Correlated Data

This case corresponds to the scenario where some sources are sent to all nodes ($S^i = V$). A simple example is shown in Fig. 6. In this example, the traffic matrix has $T_{ij} = 1, (\forall)j$, for some arbitrary $L$ nodes $\{i_1, \ldots, i_L\} \subset V$.

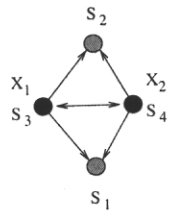



**Fig. 6.** Data from $X_1, X_2$ need to be transmitted to all nodes $S_1, S_2, S_3, S_4$

In this case, for any node $i$, the value $d_i^*$ in (2) is given by the tree of minimum weight which spans $V$; this is the minimum spanning tree (MST), and thus, by definition, it does not depend on $i$.

Note that in this case all weights $\{d_i^*\}_{i=1}^N$ are equal, thus the optimal solution $R_i$ is not unique and therefore we have a degenerate solution case for the LP in (2). There is only one set of constraints, and the cost line is exactly parallel to the diagonal hyper-plane in the Slepian-Wolf region (e.g. in Fig. 3, this happens when the dashed cost line becomes parallel to the diagonal solid line in the boundary). In such a case, not only a corner, but any point on this diagonal hyper-plane of the Slepian-Wolf region is optimal.

Notice that this case includes the typical broadcast scenario where one node transmits its source to all the nodes in the network.

### 5.2 Multiple Sink Data Gathering

This case corresponds to the scenario where all sources ($V^j = V$) are sent to some set $S^a$ of sinks. In this case (see Fig. 7), finding the optimal weights $\{d_i^*\}_{i=1}^N$ is as difficult as in the arbitrary matrix case, presented in Sect. 3. For every $i$, the optimal weight $d_i^*$ is equal to the weight of the minimum Steiner tree rooted at $i$ and spanning the nodes in the set $S^a$.

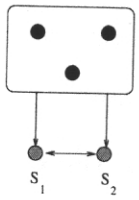



**Fig. 7.** Data from all nodes has to be transmitted to the set of sinks $S^a = \{S_1, S_2\}$. Each sink has to receive data from *all* the sources

However, given the optimal transmission structure, the optimal rate allocation can be easily found in a similar manner as in Sect. 4. First, we order the nodes in increasing order of increasing distance $d_1^* < d_2^* < \cdots < d_N^*$, and then the optimal rate allocation is given as (4).

### 5.3 Localized Data Gathering

This case corresponds to the scenario where disjoint sets $\{V^1, V^2, \ldots, V^L\}$ are sent to some sinks $\{S_1, S_2, \ldots, S_L\}$. In this case, for each $i$, the solution for the optimal weight $d_i^*$ is again the corresponding Steiner tree rooted at $i$ and that spans $S^i$. If $d_i^*$ can be found, then the rate allocation can be approximated by a decentralized algorithm for each set $\{V^j\}_{j=1}^L$, in the same way as in Sect. 4, that is, we solve $L$ LP programs independently (decentralization up to cluster level).

**Algorithm 3.** *Disjoint sets.*

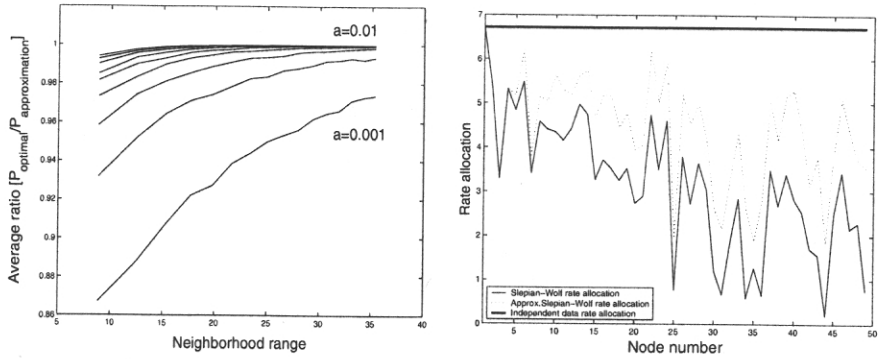
- *For each set $V^j$, order nodes $\{i, i \in V^j\}$ as a function of the total weight $d_i$.*
- *Assign rates in each $V^j$ as in (4), taking into account this order.*

## 6 Numerical Simulations

We present numerical simulations that show the performance of the approximation algorithm introduced in Sect. 4, for the case of data gathering. We consider a stochastic data model given by a multi-variate Gaussian random field, and a correlation model where the inter-node correlation decays exponentially with the distance between the nodes. In this case, the joint entropy of the data measured at a set of nodes is essentially given by the logarithm of the determinant of the corresponding covariance matrix.

Then, the performance of our approximation algorithm will be close to optimal even if we consider a small neighborhood $\mathcal{N}(i)$ for each sensor $i$. We use an exponential model of the covariance $K_{ij} = \exp(-ad_{i,j}^2)$, for varying neighborhood range radius and several values for the correlation exponent $a$. The weight of an edge $(i, j)$ is $w_{i,j} = d_{i,j}^2$ and the total cost is given by expression (3). Figure 8(a) presents the average ratio between the approximated solution and the

optimal one. In Fig. 8(b) we show a comparison of the rate allocations with our different approaches for rate allocation, as a function of the distances from the nodes to the sink.




(a) Slepian-Wolf coding.

(b) Typical rate allocation.

**Fig. 8.** (a) Average value of the ratio between the optimal and the approximated solution, in terms of total cost, vs. the neighborhood range. The network instances have 50 nodes uniformly distributed on a square area of size $100 \times 100$, and the correlation exponent varies from $a = 0.001$ (high correlation) to $a = 0.01$ (low correlation). The average has been computed over 20 instances for each $(a, \text{radius})$ pair. (b) Typical rate allocation for a network instance of 50 nodes, and correlation exponent $a = 0.0005$. On the x-axis, nodes are numbered in order as the distance from $S$ increases, on the corresponding spanning tree

## 7 Conclusions and Future Work

We addressed in this paper the problem of joint rate allocation and transmission structure optimization for sensor networks, when the flow cost metric [rate] $\times$ [link weight] is considered. We showed that if the cost function is separable, then the tasks of optimal rate allocation and transmission structure optimization separates. We assess the difficulty of the problem, namely we showed that for an arbitrary transfer matrix the problem of finding the optimal transmission structure is NP-complete. The problem of optimal rate allocation can be posed as a linear programming (LP) problem, but it is difficult in general to find decentralized algorithms that use only local information for this task. We also studied some particular cases of interest where the problem becomes easier and a closed form solution can be found and where efficient approximation algorithms can be derived.

Our future research efforts include the derivation of efficient distributed approximation algorithms for both finding the optimal transmission structure and the optimal distribution of rates among the various subsets of sources for more general cases of transmission matrices. Moreover, an interesting research issue is

to find tight bounds for the approximation ratios, in terms of power costs, for these distributed algorithms. Also, we consider more general network problems where for each node $i$, there is a source vector $\overrightarrow{X}_i = (X_{i1}, \ldots, X_{im})$ and any subvector of this vector has to be transmitted to some set of sinks.

## References


1. [2002] Aaron, A., Girod, B.: Compression with side information using turbo codes, Data Compression Conference 2002.
2. [2002], Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks, IEEE Communications Magazine, vol. 40, nr. 8, pp:102–116, 2002.
3. [1996] Arora, S.: Polynomial time approximation scheme for Euclidean TSP and other geometric problems, In Proc. 37th Ann. IEEE Symp. on Foundations of Comput. Sci., IEEE Computer Society, 2–11, 1996.
4. [1989] Bern, M., Plassmann, P.: The Steiner problem with edge lengths 1 and 2, Inform. Process. Lett. 32, 171–176, 1989.
5. [1998] Bertsekas, D.: Network optimization: continuous and discrete models, Athena Scientific, 1998.
6. [1991] Cover, T.M., Thomas, J.A: Elements of information theory, John Wiley and Sons, Inc., 1991.
7. [2003] Cristescu, R., Beferull-Lozano, B., Vetterli, M.: On network correlated data gathering, submitted to Infocom 2004.
8. [2003] Goel, A., Estrin, D.: Simultaneous optimization for concave costs: single sink aggregation or single source buy-at-bulk, ACM-SIAM Symposium on Discrete Algorithms, 2003.
9. [2001] Lindsey, S., Raghavendra, C.S., Sivalingam, K.: Data gathering in sensor networks using the energy*delay metric, Proc. of IPDPS Workshop on Issues in Wireless Networks and Mobile Computing, April 2001.
10. [1984] Luenberger, D.: Linear and nonlinear programming, Addison-Wesley, 1984.
11. [2003] Marco, D., Duarte-Melo, E., Liu, M., Neuhoff, D.L.: On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data, IPSN 2003.
12. [2000] Pottie, G.J., Kaiser, W.J.: Wireless integrated sensor networks, Communications of the ACM, nr. 43, pp:51–58, 2000.
13. [2001] Pradhan, S.: Distributed Source Coding Using Syndromes (DISCUS), Ph.D. thesis, U.C. Berkeley, 2001.
14. [1999] Pradhan, S., Ramchandran, K.:Distributed Source Coding Using Syndromes (DISCUS): Design and construction, in Proc. IEEE DCC, March, 1999.
15. [2000] Rabaey, J., Ammer, M.J., da Silva, J.L., Patel, D., Roundy, S.: PicoRadio supports ad-hoc ultra-low power wireless networking, IEEE Computer, vol. 33, nr. 7, pp:42–48.
16. [2000] Rabiner-Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks, in Proc. of the 33rd International Conference on System Sciences (HICSS '00), January, 2000.
17. [2000] Robins, G., Zelikovsky, A.: Improved steiner tree approximation in graphs, in Proc. 10th Ann. ACM-SIAM Symp. on Discrete Algorithms, 770–779, 2000.
18. [2002] Scaglione, A., Servetto, S.D.: On the interdependence of routing and data compression, in multi-hop sensor networks, in Proc. ACM MOBICOM, 2002.
19. [1973] Slepian, D., Wolf, J.K.: Noiseless coding of correlated information sources. IEEE Trans. Information Theory, IT-19, 1973, pp. 471–480.