Routing Explicit Side Information for Data Compression in Wireless Sensor Networks

Huiyu Luo and Gregory Pottie

Uiversity of California, Los Angeles, Ca, 90095, USA
huiyu,pottie@ee.ucla.edu

Abstract. Two difficulties in designing data-centric routes [2–5] in wireless sensor networks are the lack of reasonably practical data aggregation models and the high computational complexity resulting from the coupling of routing and in-network data fusion. In this paper, we study combined routing and source coding with explicit side information in wireless sensor networks. Our data aggregation model is built upon the observation that in many physical situations the side information that provides the most coding gain comes from a small number of nearby sensors. Based on this model, we propose a routing strategy that separately routes the explicit side information to achieve data compression and cost minimization. The overall optimization problem is NP hard since it has the minimum Steiner tree as a subproblem. We propose a suboptimal algorithm based on maximum weight branching and the shortest path heuristic for the Steiner tree problem. The worst case and average performances of the algorithm are studied through analysis and simulation.

1 Introduction

The need to lower the communication cost in wireless sensor networks has prompted many researchers to propose data-centric routing schemes that can utilize in-network data fusion to reduce the transmission rate. There are two major difficulties in designing such routes. First, the lack of reasonably practical data aggregation models has led researchers to use overly simplified ones [2–5]. For example, these models generally assume that sensors perform the same aggregation function regardless of the origin of the fused data. As a remedy, [5] suggests looking into models in which data aggregation is not only a function of the number of sources but also the identity of the sources. Second, the resulting optimization problem is often NP hard due to the coupling of routing and in-network data fusion [2, 4]. Hence, algorithms that find exact solutions in polynomial time are unlikely to exist. In this paper, we try to build computationally useful models and devise heuristic algorithms for the combined routing and source coding problem.

Most previous work has considered using trees as the underlying routing structure [2, 4, 5] probably due to the fact that trees are the optimal solution to the shortest path problem and have been pervasive in network routing. However, in data-centric routing, trees are not necessarily optimal. In this paper, a simple strategy, which we call designated side information transmission (DSIT), is proposed. This method results in a non-tree routing structure and tends to distribute the traffic more evenly in the network. To motivate the idea and give a preview of the paper, consider the example depicted in Fig. 1. The edges between adjacent sensors (circles) have the weight $c_e = d$, and the edges connecting a sensor to the fusion center (square) have the weight $c_e = D$. The rate at which each sensor needs to transmit to the fusion center is R without any explicit side information and r if explicit side information from an adjacent sensor is available. We postulate that side information from other sensors can be used to help compress the data only when it is available at both the encoder (the sensor that generated the data) and decoder (fusion center). Assume $r \ll R$ and $d \ll D$. The objective is to minimize the cost $C = \sum_e c_e f_e$ of routing all the data to the fusion center, where f_e is the rate at which data is transmitted across edge e. Consider the following three strategies: (a) the shortest path tree is used; (b) we compress data using explicit side information and optimize over all spanning trees; (c) the data at each sensor is transmitted to an adjacent sensor to be used as explicit side information whenever the coding gain outweighs the transmission cost. This transmission to an adjacent sensor provides only explicit side information and needs not be relayed to the fusion center. The routes corresponding to the three strategies are shown in Fig. 1. Note that at least one sensor has to transmit at rate R to the fusion center so that all the data can be correctly recovered. The costs of the three strategies are:

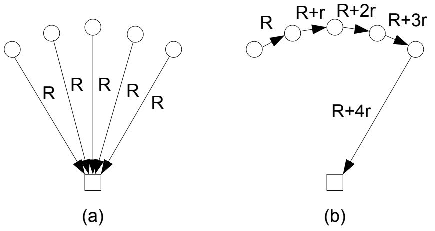

Fig. 1. Three routing strategies: (a) shortest path tree; (b) optimal tree; (c) designated side information transmission.

$$
\begin{aligned}
C_a &= 5RD \\
C_b &= RD + 4rD + 4Rd + 6rd \\
C_c &= RD + 4rD + 4Rd
\end{aligned}
$$

The performance of (b) and (c) are about the same, and both are superior to that of (a). It is also evident that (c) results in more evenly distributed traffic than (b). This is because in DSIT, the communication to the fusion center is separated from the explicit side information transmission, and can be routed through any path.

There has been much recent research activity on data-centric routing. In [7], the interdependence of routing and data compression is addressed from the viewpoint of information theory. Clustering methods have been used by some researchers to aggregate data at the cluster head before transmitting them to the fusion center [8, 9]. Since the cluster head is responsible for data aggregation and relaying, it consumes the most energy. Hence, dynamically electing nodes with more residual power to be cluster heads and evenly distributing energy consumption in network is a major issue in these schemes. In [3], a diffusion type routing paradigm that attaches attribute-value pairs to data packets is proposed to facilitate the in-network data fusion. The correlated data routing problems studied in [2, 4] are closely related to our work. In [2], the authors give a thorough comparison of data-centric and address-centric methods and a overview of recent effort in the field. [4] casts the data-centric routing problem as an optimization problem and seek solutions to it when different source coding schemes are applied. A similar optimization problem is also the subject of [5], where a grossly simplified data model is assumed.

The rest of the paper is organized as follows. In section 2, we present our network flow and data rate models. In section 3, an optimization problem is formulated out of the DSIT strategy, and a heuristic algorithm is proposed. The average performance of the heuristic algorithm is studied through simulations in section 4. Section 5 concludes the paper.

2 Network Models

2.1 Network flows

The sensor network is modelled as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. The node set $\mathcal{N}$ consists of a set $\mathcal{N}_s$ of n sensors and a special node t acting as the fusion center. We call a sensor active if it generates data. Denote by $\mathcal{N}_a$ the set of active sensors. Both active and non-active sensors can be relays. The edge set $\mathcal{E}$ represents m communication links. Here, we assume all the links are bi-directional and symmetric. If they are not, the network can be modelled as a directed graph, and the derivation in this paper will apply similarly. We also assume the network is connected so that data from any sensor can reach t. A weight c_e is associated with each edge $e \in \mathcal{E}$. It represents the cost (e.g. power) of transmitting data at unit rate across e. The flow f_e is defined as the rate at which data is transmitted across edge $e \in \mathcal{E}$. Data generated by node i and terminating at node j are denoted by f_e^{ij}. In particular, we define $f_e^i = f_e^{it}$. Clearly, $f_e = \sum_{i,j \in \mathcal{N}} f_e^{ij}$. Supposing $i, j \in \mathcal{N}_s$, denote by d_{ij} the minimum distance from i to j (i.e. the sum of edge weights along the shortest path from i to j), and d_i the minimum distance from i to t. The objective of our study is to minimize the total cost C while routing all the data from active sensors to the fusion center.

$$C = \sum_{e \in \mathcal{E}} c_e f_e \tag{1}$$

2.2 Source coding with explicit side information

Consider the problem of sampling a distributed field using wireless sensor networks. The measurements at sensors are coded and transmitted back to the fusion center, and used to reconstruct the field under some distortion constraint. There is likely to be a great deal of redundancy in the data collected by different sensors, since they are observing some common physical phenomenon. Denote by X_i the data stream produced by sensor i. (We assume X_i has been quantized and has a discrete alphabet.) Assume X_i satisfies the ergodic condition so that the results of statistical probability theory can be applied. In this paper, we consider source coding with explicit side information. In other words, only when the side information is available at both the encoder (the senor that generates the data) and decoder (fusion center) can it be used to compress the data. Supposing data stream X_k is entropy-coded using $X_1, X_2, \cdots, X_{k-1}$ as side information, we have the following:

$$f^k = H(X_k | X_1, X_2, \cdots, X_{k-1}) \tag{2}$$

Since the data rate f^k to the fusion center depends on the availability of data stream $X_i,\ i = 1, 2, \cdots, k-1$ at k, there are 2^{k-1} possibilities. In an attempt to simplify the data rate model and optimization, we assume that the the rate reduction provided by side information saturates as the number of helpers exceeds one:

$$f^k = \begin{cases} b_0^k & \text{no side information;} \\ \min_j b_1^{kj} & X_j \text{ is available at } k, \text{ and } j \in \mathcal{H}_k \end{cases} \tag{3}$$

in which b_0^k is the rate of coding X_k without any side information; b_1^{kj} ($b_1^{kj} \leq b_0^k$) is the coding rate of X_k when only $X_j, j \in \mathcal{H}_k$ is available at k; $\mathcal{H}_k$ is the set of sensors whose data is correlated with sensor k's observations and can be used as its side information. When side information from more than one sensors is available, the one providing the most coding gain is used. In practice, the information on the set of helping sensors H_k and the rate reduction provided by their data can be obtained using specially designed coding schemes (e.g. [6]). In many physical situations, sensor measurements are highly correlated only in a small neighborhood. In others, although a large number of sensors have similar measurements, the reproduction fidelity constraints often permit thinning the number of active sensors so that again only a small number of sensors have high correlation. As a result, $\mathcal{H}_k$ generally comprises a small number of sensors that are close to sensor k. The quick saturation of coding gain provided by side information indicates when the data stream at a nearby sensor is available as side information, the additional coding gain provided by other sensors' observations is negligible. Note that b_1^{kj} can be about the same as or much less than b_0^k depending on source statistics. This greatly influences the route construction.

Since $H(X_k) - H(X_k|X_j) = I(X_k, X_j) \leq H(X_j)$, we assume $b_0^k - b_1^{kj} \leq b_0^j$. Therefore, $(b_0^k - b_1^{kj})d_k \leq b_0^j d_k$, and there is no gain in feeding back explicit side information from t to sensors. The total cost of routing data to t can be decomposed as the sum of C_s representing the cost of routing side information and C_t the cost of transmitting data to t.

$$C = \sum_{e \in \mathcal{E}} c_e f_e = C_s + C_t \tag{4}$$

where

$$C_s = \sum_{i,j \in \mathcal{N}_s} \sum_{e \in \mathcal{E}} c_e f_e^{ij}, \qquad C_t = \sum_{i \in \mathcal{N}_s} \sum_{e \in \mathcal{E}} c_e f_e^{i} \tag{5}$$

In applying source coding with explicit side information to sensor networks, we must avoid helping loops. In other words, if X_j's recovery relies on X_i, then X_j cannot be used as the side information for compressing X_i. To formalize this requirement, define a directed network $\mathcal{G}_h$ that consists of all the active sensor nodes. In addition, if X_i is used as side information for coding X_j, a directed edge (i, j) is formed from sensor i to sensor j. Then we have the following proposition:

Proposition 1: No helping loop is formed when using source coding with explicit side information if and only if the directed network $\mathcal{G}_h$ contains no directed cycles.

The proof is straightforward, and hence omitted. It is apparent that if the underlying routing structure is a directed acyclic network (DAG), the above proposition is automatically satisfied. For instance, spanning trees directed toward the fusion center are DAG's, so there will be no helping loops when using trees to route data. However, in this paper the rule of no directed cycles needs to be enforced explicitly.

3 Designated Side Information Transmission

3.1 Problem formulation

In DSIT, we distinguish the data flow from sensor to the fusion center f_e^i and the transmission of explicit side information to other sensors f_e^{ij} $(i, j \in \mathcal{N}_s)$. The side information can only be provided by sensor to sensor transmissions f^{ij} not transmissions to the fusion center f^k. With the network model defined as in section 2, we formulate the following optimization problem.

Designated Side Information Transmission (DSIT)
GIVEN: A graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with weight c_e defined on each edge $e \in \mathcal{E}$, a special node $t \in \mathcal{N}$ acting as the fusion center, a set $\mathcal{H}_k$ of helping sensors and data rate function f^k as in Eq. (3) defined for each sensor $k \in \mathcal{N}_s = \mathcal{N} \setminus \{t\}$.
FIND: A set of routes transmitting the explicit side information among sensors such that the total cost of routing data to the fusion center $C = \sum_{e \in \mathcal{E}} c_e f_e$ is minimized.

We will see in ensuing discussion that the transmission of explicit side information has a Steiner tree problem embedded in it. Hence the overall problem is NP hard. As a result, we will focus on building a heuristic algorithm for the optimization.

3.2 Heuristic algorithm

The total cost can be decomposed into the cost of routing explicit side information C_s and the cost of transmitting data to the fusion center C_t as in Eq. (4). We first consider constructing routes from sensors to the fusion center. These routes affect only C_t. In addition, as $f^k, k \in \mathcal{N}_a$ does not provide any side information, its routing is decoupled from the data aggregation process. Hence, the shortest path should be used to achieve the minimum C_t:

$$C_t = \sum_{k \in \mathcal{N}_a} d_k f^k \tag{6}$$

where d_k is the minimum distance from sensor k to t, and f^k is a function of side information transmission. The design of routes for transmitting f^k must take place before that for side information transmission because the latter depends on the distance information from the former.

Designing routes for side information transmission is more complicated. First, it has the minimum Steiner tree as a subproblem. This is illustrated by the following problem instance. Given network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, we have a subset of the active sensors $\mathcal{S} \subset \mathcal{N}_a$, and there is a sensor $u \in \mathcal{N}_a \setminus \mathcal{S}$. Assume $\mathcal{H}_k = u$ if $k \in \mathcal{S}$, and $\emptyset$ otherwise. In addition, we assume that the rate function and edge weights are defined such that the cost of transmitting side information from u to any sensor in $\mathcal{S}$ using appropriately chosen routes is less than the cost reduction resulting from the coding gain of side information. The optimization problem becomes constructing a subtree that connects u and the sensors in $\mathcal{S}$, which is a minimum Steiner tree. Therefore, the overall optimization problem is NP hard. Second, we need to ensure that no helping loop can be formed while routing the side information. This amounts to avoiding directed cycles in G_h according to Proposition 1.

For a moment, we ignore the Steiner tree part, and use the shortest path to route all the side information. This leads us to construct a network $\mathcal{G}_a$ as follows. $\mathcal{G}_a$ includes the set of active sensors $\mathcal{N}_a$. In addition, for each ordered pair of nodes $(i, j) \in \mathcal{N}_a$, create a directed edge (i, j) from sensor i to j and assign the weight w_{ij} to represent the net coding gain resulting from routing side information from i to j.

$$w_{ij} = \begin{cases} (b_0^j - b_1^{ji})d_j - d_{ij}b_0^i & i \in \mathcal{H}_j \\ -d_{ij}b_0^i & \text{otherwise} \end{cases} \tag{7}$$

where b_0^j, b_1^{ji} and $\mathcal{H}_j$ are the data rates and set of helping sensors as in Eq. (3); d_j and d_{ij} are the minimum distances defined in section 2. Denote by $\mathcal{A}_a$ the set of directed edges with $w_{ij} > 0$. A branching on the directed graph $\mathcal{G}_a = (\mathcal{N}_a, \mathcal{A}_a)$ is a set of directed edges $\mathcal{B} \subseteq \mathcal{A}_a$ satisfying the conditions that no two edges in $\mathcal{B}$ enter the same node, and $\mathcal{B}$ has no directed cycle. It is evident that a branching on $\mathcal{G}_a$ represents a feasible set of routes for side information transmission. No two directed edges in $\mathcal{B}$ entering the same node ensures that a sensor uses side information from at most one helper and no directed cycle avoids the helping loop. The problem of minimizing the total cost is equivalent to maximizing the weight sum of the set of directed edges $\mathcal{B}$ that is a branching on $\mathcal{G}_a$, which is the so called maximum weight branching problem.

Maximum Weight Branching (MWB)
GIVEN: A directed graph $\mathcal{G}_a = (\mathcal{N}_a, \mathcal{A}_a)$ with weight w_e defined on each directed edge $e \in \mathcal{A}_a$.
FIND: A branching $\mathcal{B} \subseteq \mathcal{A}_a$ that maximizes $\sum_{e\in\mathcal{B}} w_e$.

It has been shown that this problem can be solved efficiently [10]. Once the optimal branching $\mathcal{B}$ is determined, we revert to using Steiner trees. Define $\mathcal{S}_k$ as the set of sensors that receive side information from k based on the optimal branching $\mathcal{B}$. We use the shortest path heuristic proposed by [11] to construct the subtree that connects k and $\mathcal{S}_k$. This method has a worst case performance ratio of 2. Our heuristic algorithm is a combination of the maximum weight branching and the Steiner tree approximation. We state it as follows:

Designated Side Information Transmission Heuristic (DSIT Heuristic)

Given a network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ with edge weights and rate function properly defined, carry out the following steps.

1. Find the shortest path from each active sensor to the fusion center. These are the routes for transmitting data to the fusion center.
2. Construct a directed graph $\mathcal{G}_a = (\mathcal{N}_a, \mathcal{A}_a)$. $\mathcal{N}_a \subseteq \mathcal{N}$ consists of all active sensors, and $\mathcal{A}_a$ is the set of directed edges from i to j ($i, j \in \mathcal{N}_a$ and $i \neq j$) whose weight w_{ij} defined as in Eq. (7) is greater than zero.
3. Find the maximum weight branching on $\mathcal{G}_a$. Based on the optimal branching $\mathcal{B}$, determine the set of sensors $\mathcal{S}_k$ that each active sensor $k \in \mathcal{N}_a$ transmits side information to.
4. Run a shortest path heuristic for the Steiner tree problem to find the subtree that connects k and the sensors in $\mathcal{S}_k$.

3.3 Performance analysis

Finding the maximum weight branching takes $O(m_a \log n_a)$ time, where $m_a = |\mathcal{A}_a|$ and $n_a = |\mathcal{N}_a|$. ($|\mathcal{S}|$ is the number of elements in finite set $\mathcal{S}$.) The shortest path heuristic for a Steiner tree requires $O(n_a n^2)$ time. The actual running time of the shortest path heuristic is in general much less because the number of nodes involved in constructing the shortest path is often a lot fewer than n. Regarding the performance of our heuristic algorithm compared to that of the optimal solution, we prove the following proposition.

Proposition 2: The ratio of the cost C_H resulting from our DSIT heuristic algorithm and the minimum cost C_{MIN} using the DSIT strategy is bounded by:

$$\frac{C_H}{C_{MIN}} \leq M \tag{8}$$

where $M = \max\{1, \max_{k\in\mathcal{N}_a} |\mathcal{S}_k^{opt}|\}$, the greater of one and the maximum number of sensors one sensor needs to transmit side information to in the optimal solution. The bound is tight in the sense that there is a network that attains the worst performance ratio.

Proof: First, we note that $\mathcal{S}_k^{opt}$ is in general not the same as the $\mathcal{S}_k$ in our heuristic algorithm. Consider the structure of an optimal solution. It can be given by the set of sensors $\mathcal{S}_k^{opt}$ that each $k \in \mathcal{N}_a$ sends explicit side information to. The side information circulated within the group of k sensors in $\mathcal{S}_k^{opt}$ is routed using the minimum Steiner tree. Denote by C_k^{ST} the sum of edge costs of these Steiner trees ($C_k^{ST} = 0$ if $\mathcal{S}_k^{opt} = \emptyset$). The data is transmitted to the fusion center using the shortest path tree. Hence we can write the minimum cost as:

$$C_{MIN} = \sum_{k\in\mathcal{N}_a} f^k d_k + \sum_{k\in\mathcal{N}_a} b_0^k C_k^{ST} \tag{9}$$

Instead of the Steiner tree, consider relying on a shortest path tree to route the side information from k to the sensors in $\mathcal{S}_k^{opt}$. Denote by C_k^{SPT} the sum of edge costs of such shortest path trees. The corresponding cost C' will be:

$$C' = \sum_{k\in\mathcal{N}_a} f^k d_k + \sum_{k\in\mathcal{N}_a} b_0^k C_k^{SPT} \tag{10}$$

Since $M_k C_k^{ST} \geq C_k^{SPT}$ [11], where $M_k = |\mathcal{S}_k^{opt}|$, we have

$$\frac{C'}{C_{MIN}} \leq \max\{1, \max_{k\in\mathcal{N}_a} M_k\} = M$$

On the other hand, C_H is the optimal result of using the shortest path tree to route the side information. Therefore, $C_H \leq C'$. This gives rise to the bound in Eq. (8). To show the bound is tight, we look at the example in Fig. 2. The

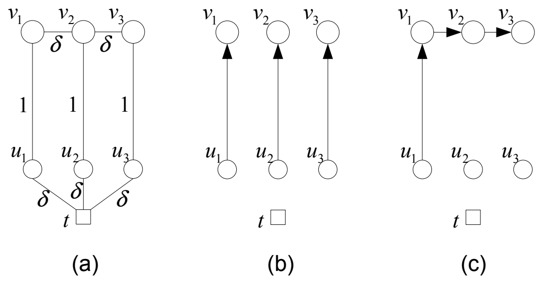

Fig. 2. A problem instance that approaches the worst performance ratio: (a) sensor network setup; (b) routes of side information transmission using DSIT heuristic; (c) routes of side information transmission in optimal solution.

network setup is given in (a). The edge weights between sensors v_k and u_k

$(k = 1, 2, 3)$ is 1. Other edges have weight $\delta \ll 1$. All the sensors are active with data rate R without side information and 0 when side information is available. Denote $\mathcal{U} = \{u_1, u_2, u_3\}$, and $\mathcal{V} = \{v_1, v_2, v_3\}$. We assume $\mathcal{H}_k = \mathcal{U}$ when $k \in \mathcal{V}$ and $\emptyset$ when $k \in \mathcal{U}$. In Fig. 2, (b) and (c) illustrate how side information is transmitted in DSIT heuristic and optimal solutions. Therefore, $C_H = 3R + 3R\delta$ and $C_{MIN} = R + 5R\delta$. When $\delta \to 0$, the ratio C_H/C_{MIN} approaches $M = 3$ asymptotically. In a similar fashion, problem instances with arbitrary values of M can be devised. Q.E.D.

The worst case scenario in the proof can be avoided by changing the heuristic algorithm to run multiple maximum weight branching and shortest path heuristic iterations. At each iteration, only one sensor is added to $\mathcal{S}_k, k \in \mathcal{N}_a$. However, this greatly increases the computational cost. Moreover, the pathological case in the proof rarely occurs in our assumed data rate model. The value of M is expected to be small as one's data helps mostly nearby sensors. Also since side information is often circulated within one's neighborhood, using shortest paths to approximate a Steiner tree introduces a moderate amount of error. What we are more interested in is the average behavior of the algorithm, which is examined through simulations in the next section.

4 Simulations

In our simulations, we place $(n + 1)$ nodes including the fusion center and n sensors in an $n_d \times n_d$ square, where $n_d = \lceil \sqrt{n+1} \rceil$. (Denote by $\lceil z \rceil$ the smallest integer such that $\lceil z \rceil \leq z$, and $\lfloor z \rfloor$ the largest integer such that $\lfloor z \rfloor \geq z$.) Supposing $\tilde{x}_i$ and $\tilde{y}_i, i = 1, \cdots, n+1$, are random variables that are uniformly distributed in $[0, 1]$, the coordinates of node i is given by:

$$\begin{aligned} x_i &= [(i \bmod n_d) - 1] + \tilde{x}_i \\ y_i &= \lfloor (i-1)/n_d \rfloor + \tilde{y}_i \end{aligned}$$

We define a transmission radius r_c. If two nodes are no more than r_c away from each other, direct communication between the two nodes is allowed. Otherwise, a relay has to be used. Denote by d_e the length of edge e. When $d_e \leq r_c$, the edge weight c_e is proportional to d_e^α, where $\alpha = 2$ is the path loss factor. When the number of sensors increases, the network covers a larger area while maintaining the communication range and sensor to sensor spacing. A typical 100 node network constructed in this manner is depicted in Fig. 3. The node (in lower left corner) with a letter ”t” next to it is the fusion center.

In our simulation, we assume that all the sensors are active. The helping set H_i of sensor i is defined as follows. Any pair of sensors that are no more than r_d away from one another has a probability of 0.5 to be in the helping sets of one another. Fig. 4 shows the resulting data correlation in the network. For simplicity, we assume the data rate function is the same for all the sensors:

$$f^k = \begin{cases} b_0 & \text{no side information} \\ \beta b_0 & \text{with side information} \end{cases} \tag{11}$$

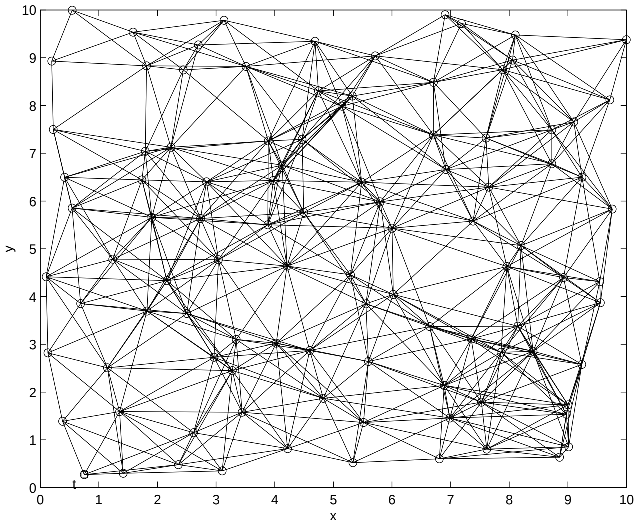

Fig. 3. A network of 100 nodes with $r_c = \sqrt{5}$. Two nodes are connected if direct transmission is allowed between the two.

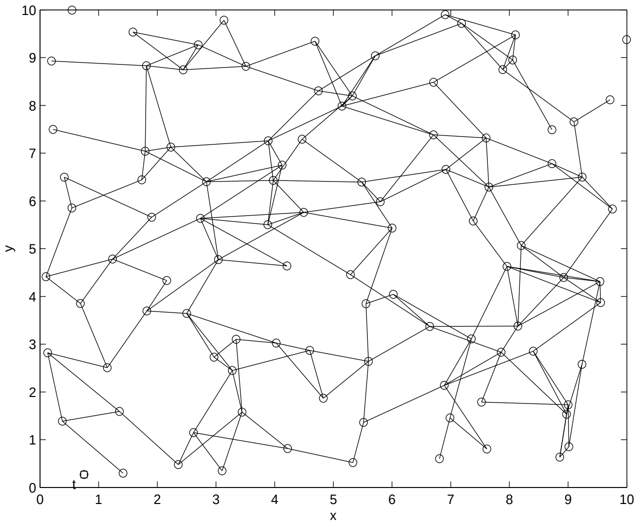

Fig. 4. This graph shows the correlation of data at different sensors when $r_d = 1.8$. Two nodes are connected by an edge if they are in the helping set H_i of one another.

where $i \in \mathcal{N}_a$ and $0 \leq \beta \leq 1$. Fig. 5 shows the maximum weight branching on the network described in Fig. 3 and 4. The graph is a forest, and hence acyclic. The root of each tree is indicated by a circle with a cross inside, from which

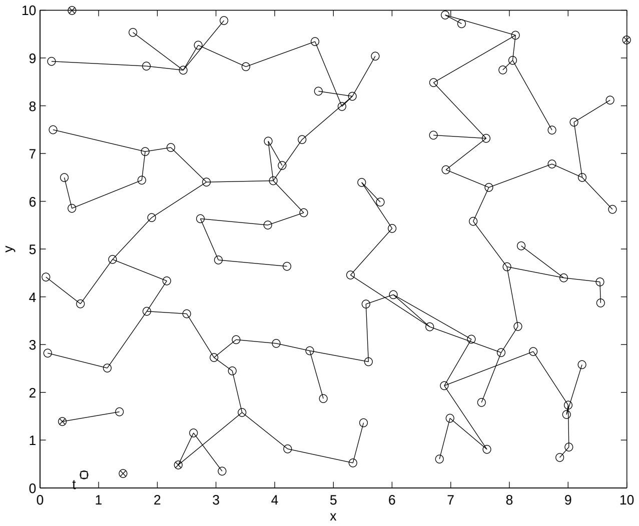

Fig. 5. A maximum weight branching on the network described in Fig. 3 and 4.

there is a simple path to any other member of the tree. Based on this rule, the helping set S_k of each sensor k can be easily determined.

We simulate for different network sizes and change the value of β. The performance of DSIT heuristic is compared to that of the shortest path tree, in which the same compression scheme based on explicit side information is used. Define the cost ratio $\mu = C_{DSIT}/C_{SPT}$. In Fig. 6, we plot μ against the number of nodes in the network. DSIT heuristic outperforms the shortest path tree in all cases. In addition, we observe that as the coding gain decreases (i.e. β increases), μ drops. This is expected considering that DSIT becomes the shortest path tree, which is also the optimum solution, when coding gain is zero. It is also noticed that μ increases as the number of nodes increases. This is explained by looking at a shortest path tree solution plotted in Fig. 7. The leaf nodes of a shortest path tree are generally far away from the fusion center while the source coding at these nodes receives no side information from other sensors. In contrast, in DSIT (Fig. 5), the nodes that receives zero side information (the roots of the subtrees) are mostly near the fusion center. As the network size increases, the leaf nodes become farther and farther away from the fusion center. Consequently, C_{SPT} rises faster than C_{DSIT}.

5 Discussion and Conclusion

The DSIT strategy relies heavily on our assumed network model, in particular, the assumptions that data streams are highly correlated only when they are from a small group of sensors close to one another, and thus the coding gain saturates when the number of helpers exceeds one. Therefore, this scheme may not be

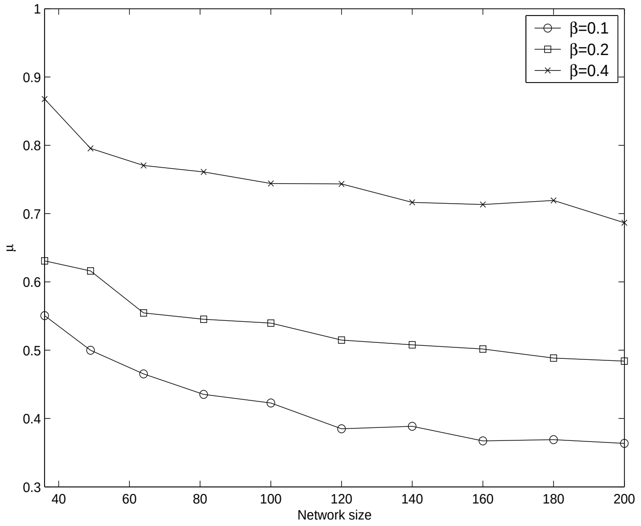

Fig. 6. Plot $\mu = C_{DSIT}/C_{SPT}$ against network size for $\beta = 0.1, 0.2$, and 0.4.

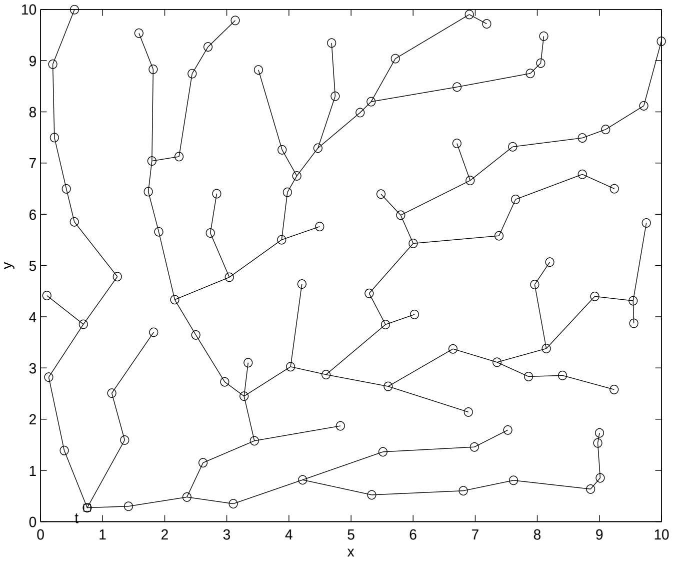

Fig. 7. The shortest path tree solution for the network described in Fig. 3 and 4.

as effective in cases that deviate from these assumptions. Nonetheless, there are practical reasons to consider this case. First, identifying the set of sensors that provides most coding gain and then processing side information incurs cost. The gain of an additional helper may not be enough to outweigh these costs.

Second, using more than one helper increases the complexity of the model and the optimization. For example, it is possible to jointly code two helpers' data if they are to are be used as side information at the same sensor. If more than one helper has to be considered, we speculate that the problem can be approached in a multiple-step procedure. At each step, the number of helpers is restricted to at most one, and an algorithm similar to our heuristic scheme is used. This is an area that needs further research.

The decoupling of route design for f^k from side information transmission offers greater flexibilities than strategies based on trees. Unlike a tree structure that bundles the network flows, the transmission of f^k can virtually be routed to t through any path. As a result, traditional address-centric routing schemes that evenly distribute the traffic load and maximize node lifetime [12, 13] can be applied. If the optimization objective is to maximize the network lifetime, we surmise the DSIT can take into account both the energy reserve of sensor nodes and source correlation. In contrast, address-centric routing focuses on node energy, and data-centric routing concentrates on source correlation only.

As we discussed in section 2, the number of sensors with highly correlated data can be brought down in a process of thinning the number of active sensors based on the reconstruction requirement. This pre-routing step makes in practice our procedure a two-phase operation. First determine the set of sensors that will participate in the fusion, then design the routes for transmitting the data to the fusion center. Currently, the first step is generally approached from a sampling point of view [14] trying to meet the distortion constraint, while route design attempts to minimize the energy consumption. It is of interest to ask whether a combined approach will yield better results. [15] is an interesting preliminary effort on that direction.

Acknowledgement

This material is based upon work supported by National Science Foundation (NSF) Grant #CCR-0121778.

References

[1] G. Pottie and W. Kaiser, "Wireless sensor networks," *Comm. ACM,* vol. 43, no. 5, pp. 51-58, May 2000.

[2] B. Krishnamachari, D. Estrin, and S. Wicker, "Modelling data-centric routing in wireless sensor network," *USC Computer Engineering Technical Report CENG 02-14.*

[3] C. Intanagonwiwat, et al. "Directed diffusion for wireless sensor networking," *IEEE/ACM Trans. Networking,* vol. 11, no. 1, pp. 2-16, Feb 2003.

[4] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On network correlated data gathering," *IEEE Infocom,* Hongkong, March 7-11, 2004.

[5] A. Goel and D. Estrin, "Simultaneous optimization for concave costs: single sink aggregation or single source buy-at-bulk," *ACM/SIMA Symposium on Discrete Algorithms,* 2003.

[6] H. Luo, Y. Tong, and G. Pottie, "A two-stage DPCM scheme for wireless sensor networks," to appear in *ICASSP 2005,* Philadelphia, USA.

[7] A. Scaglione and S. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," *ACM/IEEE Mobicom,* 2002.

[8] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," *IEEE Infocom,* 2003.

[9] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication procotol for wireless microsensor networks," *Hawaii Inter. Conf. System Sciences,* Jan. 4-7, 2000.

[10] R. E. Tarjan, "Finding optimum branchings," *Networks,* pp 25-35, vol. 7, 1977.

[11] H. Takahashi and A. Matsuyama, "An approximate solution for the steiner problem in graphs," *Math. Japonica,* 24, No. 6, pp 573-577, 1980.

[12] J. H. Chang, L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," *IEEE Infocom,* 2000.

[13] S. Singh, M. Woo, C. S. Raghavendra, "Power-aware routing in mobile ad-hoc networks," *ACM/IEEE Mobicom, 1998,* Dallas, Texas.

[14] R. Willett, A. Martin, and R. Nowak, "Backcasting: adaptive sampling for sensor networks," *IPSN,* 2004, Berkeley, Ca, USA.

[15] D. Ganesan, R. Cristescu, and B. Beferull-Lozano, "Power-efficient sensor placement and transmission structure for data gathering under distortion constraints," *IPSN,* 2004, Berkeley, Ca, USA.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* John Wiley & Sons, 1991.

[17] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* W. H. Freeman and Company, 1979.

[18] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem,* North-Holland, 1992.