

On Computing Compression Trees for Data Collection in Wireless Sensor Networks

Jian Li Amol Deshpande Samir Khuller
Department of Computer Science, University of Maryland, College Park
Maryland, USA, 20742
Email: {lijian, amol, samir}@cs.umd.edu

Abstract—We address the problem of efficiently gathering correlated data from a wireless sensor network, with the aim of designing algorithms with provable optimality guarantees, and understanding how close we can get to the known theoretical lower bounds. Our proposed approach is based on finding an optimal or a near-optimal *compression tree* for a given sensor network: a compression tree is a directed tree over the sensor network nodes such that the value of a node is compressed using the value of its parent. We focus on *broadcast communication* model in this paper, but our results are more generally applicable to a unicast communication model as well. We draw connections between the data collection problem and a previously studied graph concept called *weakly connected dominating sets*, and we use this to develop novel approximation algorithms for the problem. We present comparative results on several synthetic and real-world datasets showing that our algorithms construct near-optimal compression trees that yield a significant reduction in the data collection cost.

I. INTRODUCTION

In this paper, we address the problem of designing energy-efficient protocols for collecting all data observed by the sensor nodes in a sensor network at an Internet-connected base station, at a specified frequency. Some of the key challenges in designing an energy-efficient data collection protocol are: (1) effectively exploiting the strong spatio-temporal correlations present in most sensor networks, and (2) optimizing the routing plan for data movement. In most sensor network deployments, especially in environmental monitoring applications, the data generated by the sensor nodes is highly correlated both in time (future values are correlated with current values) and in space (two co-located sensors are strongly correlated). These correlations can usually be captured by constructing predictive models using either prior domain knowledge, or historical data traces. However, the distributed nature of data generation and the resource-constrained nature of the sensor devices make it a challenge to optimally exploit these correlations.

Consider an n -node sensor network, with node i monitoring the value of a variable X_i , and generating a data flow at entropy rate of $H(X_i)$. In the naive protocol, data from each source is simply sent to the base station through the shortest path, rendering a total data transmission cost $\sum_i H(X_i) \cdot d(i, BS)$, where $d(i, BS)$ is the length of a shortest path to the base station. However, because of the strong spatial correlations among the X_i , the joint entropy of the nodes, $H(X_1, \dots, X_n)$, is typically much smaller than the

sum of the individual entropies; the naive protocol ignores these correlations.

A lower bound on the total number of bits that need to be communicated can be computed using the *Distributed Source Coding (DSC) theorem* [1], [2], [3], [4]. In their seminal work, Slepian and Wolf [1] prove that it is theoretically possible to encode the correlated information generated by distributed data sources (in our case, the sensor nodes) at the rate of their joint entropy *even if the data sources do not communicate with each other*. This can be translated into the following lower bound on the total amount of data transmitted for a multi-hop network: $\sum_i d(i, BS) \times H(X_i | X_1, \dots, X_{i-1})$ where X_1, \dots, X_n are sorted in an increasing order by their distances to the base station [5], [4]. With high spatial correlation, this number is expected to be much smaller than the total cost for the naive protocol (i.e., $H(X_i | X_1, \dots, X_{i-1}) \ll H(X_i)$). The DSC result unfortunately is non-constructive, with constructive techniques known for only a few specific distributions [6]; more importantly, DSC requires perfect knowledge of the correlations among the nodes, and may return wrong answers if the observed data values deviate from what is expected.

However, the lower bound does suggest that significant savings in total cost are possible by exploiting the correlations. Pattem et al. [7], Chu et al. [8], Cristescu et al. [9], among others, propose practical data collection protocols that exploit the spatio-temporal correlations while guaranteeing correctness (through *explicit communication* among the sensor nodes). These protocols may exploit only a subset of the correlations, and in many cases, assume uniform entropies and conditional entropies. Further, most of this prior work has not attempted to provide any approximation guarantees on the solutions, nor have they attempted a rigorous analysis of how the performance of the proposed data collection protocol compares with the lower bound suggested by DSC.

We are interested in understanding how to get as close to the DSC lower bound as possible for a given sensor network and a given set of correlations among the sensor nodes. In a recent work, Liu et al. [10] considered a similar problem to ours and developed an algorithm that performs very well compared to the DSC lower bound. However, their results are implicitly based on the assumption that the conditional entropies are quite substantial compared to the base variable entropies (specifically, that $H(X_i | X_1, \dots, X_{i-1})$ is lower bounded). Our results here are complimentary in that, we specifically target

the case when the conditional entropies are close to zero (i.e., the correlations are strong), and we are able to obtain approximation algorithms for that case. We note that we are also able to prove that obtaining better approximation guarantees is NP-hard, so our results are tight for that case. As we will see later, lower bounding conditional entropies enables us to get better approximation results and further exploration of this remains a rich area of future work.

In this paper, we analyze the data collection problem under the restriction that any data collection protocol can directly utilize only *second-order marginal or conditional* probability distributions – in other words, we only directly utilize pairwise correlations between the sensor nodes. There are several reasons for studying this problem. First off, the entropy function typically obeys a strong diminishing returns property in that, utilizing higher-order distributions may not yield significant benefits over using only second-order distributions. Second, learning, and utilizing, second-order distributions is much easier than learning higher-order distributions (which can typically require very high volumes of training data). Finally, we can theoretically analyze the problem of finding the optimal data collection scheme under this restriction, and we are able to develop polynomial-time approximation algorithms for solving it.

The above restriction leads to what we call *compression trees*. Generally speaking, a compression tree is simply a directed spanning tree \mathcal{T} of the communication network, in which, the parents are used to compress the values of the children. More specifically, given a directed edge (u, v) in \mathcal{T} , the value of X_v is compressed using the value of X_u ¹ (i.e., we use the value of $X_u = x_u$ to compute the conditional distribution $p(X_v|X_u = x_u)$ and use this distribution to compress the observed value of X_v (using say Huffman coding)). The compression tree also specifies a data movement scheme, specifying where (at which sensor node) and how the values of X_u and X_v are collected for compression.

The compression tree-based approach can be seen as a special case of the approach presented by one of the authors in prior work [11]. There the authors proposed using *decomposable models* for data collection in wireless sensor networks, of which compression trees can be seen as a special case. However, that work only presented heuristics for solving the problem, and did not present any rigorous analysis or approximation guarantees.

II. PROBLEM DEFINITION

We begin by presenting preliminary background on data compression in sensor networks, discuss the prior approaches, and then introduce the compression tree-based approach.

A. Notation and Preliminaries

We are given a sensor network modeled as an undirected, edge-weighted graph $\mathcal{G}_C(V = \{1, \dots, n\}, E)$, comprising of n nodes that are continuously monitoring a set of distributed

attributes $\mathcal{X} = \{X_1, \dots, X_n\}$. The edge set E consists of pairs of vertices that are within communication radius of each other, with the edge weights denoting the communication costs. Each attribute, X_i , observed by node i , may be an environmental property being sensed by the node (e.g., *temperature*), or it may be the result of an operation on the sensed values (e.g., in an anomaly-detection application, the sensor node may continuously evaluate a filter such as “*temp* > 100” on the observed values). If the sensed attributes are continuous, we assume that an error threshold of e is provided and the readings are binned into intervals of size $2e$ to discretize them. In this paper, we focus on optimal exploitation of spatial correlations at any given time t ; our approach can be generalized to handle temporal correlations in a straightforward manner.

We are also provided with the entropy rate for each attribute, $H(X_i)$ ($1 \leq i \leq n$) and the conditional entropy rates, $H(X_i|X_j)$ ($1 \leq i, j \leq n$), over all pairs of attributes. More generally, we may be provided with a joint probability distribution, $p(X_1, \dots, X_n)$, over the attributes, using which we can compute the joint entropy rate for any subset of attributes. However accurate computation of such joint entropies for large subsets of attributes is usually not feasible.

We denote the set of neighbors of the node i by $N(i)$ and let $\bar{N}(i) = N(i) \cup \{i\}$ and $\deg(i) = |N(i)|$. We denote by $d(i, j)$ the energy cost of communicating one bit of information along the shortest path between i and j .

We focus on the wireless communication model (WL) in this paper; specifically we assume that when a node transmits a message, all its neighbors can hear the message (*broadcast* model). We further assume that the energy cost of receiving such a broadcast message is negligible, and we only count the cost of transmitting the message. In the extended version of the paper [12], we discuss how our approach generalizes to wired communication networks, and to unicast or multicast models.

B. Prior Approaches

Given the entropy and the joint entropy rates for compressing the sensor network attributes, the key issue with using them for data compression is that the values are generated in a distributed fashion. The naive approach to using *all* the correlations in the data is (a) to gather the sensed values at a central sensor node, and (b) compress them jointly. However, even if the compression itself was computationally feasible, the data gathering cost would typically dwarf any advantages gained by doing joint compression. Prior research in this area has suggested several approaches that utilize a subset of correlations instead. Several of these approaches are illustrated in Figure 1 using a simple 5-node sensor network.

IND: Each node compresses its own value, and sends it to the base station along the shortest path. The total communication cost is given by $\sum_i d(i, BS) \cdot H(X_i)$.

Cluster: In this approach [7], [8], the sensor nodes are grouped into clusters, and the data from the nodes in

¹In the rest of the paper, we denote this by $X_v|X_u$

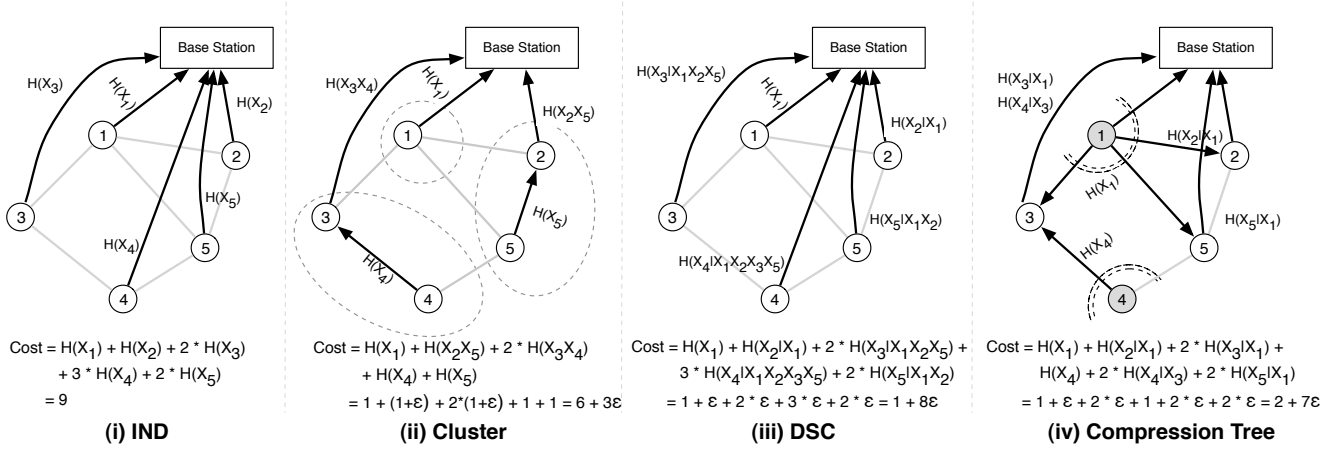


Fig. 1. Illustrating different data collection approaches – costs computed assuming $H(X_i) = 1, H(X_i|X_j) = \epsilon, \forall i, j$: (i) IND: correlations ignored; (ii) Cluster: using 3 clusters $\{X_1\}, \{X_2, X_5\}, \{X_3, X_4\}$; (iii) DSC (theoretical optimal); (iv) Compression tree: with edges $1 \rightarrow 2, 1 \rightarrow 3, 1 \rightarrow 5$ and $3 \rightarrow 4$ (the cost under WN model would have been $5 + 7\epsilon$).

each cluster is gathered at a node (which may be different for different clusters) and is compressed jointly. Figure 1 (ii) shows an example of this using three clusters $\{1\}, \{2, 5\}, \{3, 4\}$. Thus the intra-cluster spatial correlations are exploited during compression; however, the correlations across clusters are not utilized.

Cristescu et al. [9]: The approach proposed by Cristescu et al. is similar to ours, and also only uses second-order distributions. However they only consider the unicast communication model, and further assume that the entropies and conditional entropies are uniform. Rickenbach et al. [13] also present results under similar assumptions.

DSC: Distributed source coding (DSC), although not feasible in this setting for the reasons discussed earlier, can be used to obtain a lower bound on total communication cost as follows [5], [9], [4]. Let the sensor nodes be numbered in increasing order by distances from the base station (i.e., for all $i, d(i, BS) \leq d(i + 1, BS)$). The optimal scheme for using DSC is as follows: X_1 is compressed by itself, and transmitted directly to the sink (incurring a total cost of $d(1, BS) \times H(X_1)$). Then, X_2 is compressed according to the conditional distribution of X_2 given the value of X_1 , resulting in a data flow rate of $H(X_2|X_1)$ (since the sink already has the value of X_1 , it is able to decode according to this distribution). Note that, according to the distributed source coding theorem [1], sensor node 2 does not need to know the actual value of X_1 . Similarly, X_i is compressed according to its conditional distribution given the values of X_1, \dots, X_{i-1} . The total communication cost incurred by this scheme is given by:

$$\sum_{i=1}^n d(i, BS) \times H(X_i|X_1, \dots, X_{i-1})$$

Figure 1 (iii) shows this for our running example (note that 5 is closer to sink than 3 or 4).

RDC: Several approaches where data is compressed along the way to the base station (*routing driven compression* [7], [14], [15]) have also been suggested. These however require joint compression and decompression of large numbers of

data sources inside the network, and hence may not be suitable for resource-constrained sensor networks.

Dominating Set-based: Kotidis [16] and Gupta et al. [17], among others, consider approaches based on using a representative set of sensor nodes to approximate the data distribution over the entire network; these approaches however do not solve the problem of exact data collection, and cannot provide correctness guarantees.

As we can see in Figure 1, if the spatial correlation is high, both IND and Cluster incur much higher communication costs than DSC. For example, if $H(X_i) = 1, \forall i$, and if $H(X_i|X_j) = \epsilon \approx 0, \forall i, j$ (i.e., if the spatial correlations are almost perfect), the total communication costs of IND, Cluster (as shown in the figure), and DSC would be 9, 6, and 1 respectively.

C. Compression Trees

As discussed in the introduction, in practice, we are likely to be limited to using only low-order marginal or conditional probability distributions for compression in sensor networks. In this paper, we begin a formal analysis of such algorithms by analyzing the problem of optimally exploiting the spatial correlations under the restriction that we can only use second-order conditional distributions (i.e., two-variable probability distributions). A feasible solution under this restriction is fully specified by a directed spanning tree \mathcal{T} rooted at r (called a *compression tree*) and a data movement scheme according to \mathcal{T} . In particular, the compression tree indicates which of the second-order distributions are to be used, and the data movement scheme specifies an actual plan to implement it.

More formally, let $p(i)$ denote the parent of i in \mathcal{T} . This indicates that both X_i and $X_{p(i)}$ should be gathered together at some common sensor node, and that X_i should be compressed using its conditional probability distribution given the value of $X_{p(i)}$ (i.e., $p(X_i|X_{p(i)}) = x_{p(i)}$). The compressed value is communicated to the base station along the shortest path, resulting in an entropy rate of $H(X_i|X_{p(i)})$. Finally, the root of the tree, r , sends its own value directly to the base station,

resulting in an entropy rate of $H(X_r)$. It is easy to see that the base station can reconstruct all the values. The data movement plan specifies how the values of X_i and $X_{p(i)}$ are collected together for all i .

In this paper, we address the optimization problem of finding the optimal compression tree that minimizes the total communication cost, for a given communication topology and a given probability distribution over the sensor network variables (or the entropy rates for all variables, and the joint entropy rates for all pairs of variables).

We note that the notion of compression trees is quite similar to the so-called *Chow-Liu trees* [18], used for approximating large joint probability distributions.

Example 1: Figure 1 (iv) shows the process of collecting data using a compression tree for our running example, under the broadcast communication model. The compression tree (not explicitly shown) consists of four edges: $1 \rightarrow 2$, $1 \rightarrow 3$, $1 \rightarrow 5$ and $3 \rightarrow 4$. The data collections steps are:

1. Sensor nodes 1 and 4 broadcast their values, using $H(X_1)$ and $H(X_4)$ bits respectively. The Base Station receives the value of X_1 in this step.
2. Sensor nodes 2, 3, and 5 receive the value of X_1 , and compress their own values using the conditional distributions given X_1 . Each of them sends the compressed values to the base station along the shortest path.
3. Sensor node 3 also receives the value of X_4 , and it compresses X_4 using its own value. It sends the compressed value (at an entropy rate of $H(X_4|X_3)$) to the base station along the shortest path.

The total (expected) communication cost is thus given by:

$$H(X_1) + H(X_4) + H(X_2|X_1) + 2 \times H(X_3|X_1) + 2 \times H(X_5|X_1) + 2 \times H(X_4|X_3)$$

If the conditional entropies are very low, as is usually the case, the total cost will be simply $H(X_1) + H(X_4)$.

D. Compression Quality of a Solution

To analyze and compare the quality of the solutions with the DSC approach, we subdivide the total communication cost incurred by a data collection approach into two parts:

Necessary Communication (NC): As discussed above, for practical reasons, data collection schemes typically use a subset of the correlations present in the data (e.g. Cluster only uses intra-cluster correlations, our approach only uses second-order joint distributions). Given the specific set of correlations utilized by an approach, there is a minimum amount of communication that will be incurred during data collection. This cost is obtained by computing the DSC cost assuming only those correlations are present in the data. For a specific compression tree, the NC cost is computed as:

$$H(X_r) \times d(r, BS) + \sum_{i \in V} H(X_i|X_{p(i)}) \times d(i, BS)$$

The NC cost for the Cluster solution shown in Figure 1(ii) is $4 + 5\epsilon$, computed as:

$$H(X_1) + H(X_2) + 2 \cdot H(X_5|X_2) + 2 \cdot H(X_3) + 3 \cdot H(X_4|X_3)$$

In some sense, NC cost measures the penalty of ignoring some of the correlations during compression. For Cluster, this is typically quite high – compare to the NC cost for DSC ($= 1 + 8\epsilon$). On the other hand, the NC cost for the solution in Figure 1 (iv) is $1 + 8\epsilon$ (i.e., it is equal to the NC cost of DSC – we note that this is an artifact of having uniform conditional entropies, and does not always hold).

Intra-source Communication (IC): This measures the cost of explicitly gathering the data together as required for joint compression. By definition, this cost is 0 for DSC. We compute this by subtracting the NC cost from the total cost. For the solutions presented in Figures 1 (ii) and (iv), the IC cost is $2 - 2\epsilon$ and $1 - \epsilon$ respectively. The broadcast communication model significantly helps in reducing this cost for our approach.

The key advantage of our compression tree-based approach is that its NC cost is usually quite close to DSC, whereas the other approaches, such as Cluster, can have very high NC costs because they ignore a large portion of the correlations.

E. Solution Space

In our optimization algorithms, we consider searching among two different classes of compression trees.

- *Subgraphs of \mathcal{G} (SG):* Here we require that the compression tree be a subgraph of the communication graph. In other words, we compress X_i using X_j only if i and j are neighbors.
- *No restrictions (NS):* Here we don't put any restrictions on the compression trees. As expected, searching through this solution space is much harder than SG.

In general, we expect to find the optimal solution in the SG solution space; this is because the correlations are likely to be stronger among neighboring sensor nodes than among sensor nodes that are far away from each other.

Finally, we define β as the *bounded conditional entropy parameter*, which bounds the ratio of conditional entropies for any pair of variables that can be used to compress each other. Formally, $\frac{1}{\beta} \leq \frac{H(X_i|X_j)}{H(X_j|X_i)} \leq \beta$ for any nodes i and j and some constant $\beta \geq 1$. For the SG problem, this is taken over pairs of adjacent nodes and for the NS problem, it is taken over all pairs. Moreover, the above property implies that the ratio of entropies between any pair of nodes is also bounded, $\frac{1}{\beta} \leq \frac{H(X_i)}{H(X_j)} \leq \beta$.

We expect β to be quite small (≈ 1) in most cases (especially if we restrict our search space to SG). Note that, if the entropies are uniform ($H(X_i) = H(X_j)$), then $\beta = 1$.

F. Summary of Our Results

We refer to the two problems that we focus on in this paper by WL-SG (where compression trees are restricted to be subgraphs of \mathcal{G}), and WL-NS (no restrictions on compression trees). Below we summarize our key results.

- 1) (Section III-A) We first consider the WL-SG problem under an *uniform entropy and conditional entropy assumption*, i.e., we assume that $H(X_i) = 1 \forall i$ and $H(X_i|X_j) = \epsilon \forall i, j, i \neq j$. We develop a

$\left(\frac{1}{1+2\epsilon(d_{avg}-1/2)}(H_{\Delta} + 1) + 2\right)$ -approximation for this problem, where d_{avg} is the average distance to the base station.

- 2) (Section III-B and III-C) We develop a unified generic greedy framework which can be used for approximating the problem under various communication cost models.
- 3) (Section III-D and III-E) We show that, for wireless communication model, the greedy framework gives a $4\beta^2 H_n$ approximation factor for the SG solution space and an $O(\beta^3 n^{\epsilon} \log n)$ (for any $\epsilon > 0$) factor for the NS solution space.
- 4) (Section IV) We illustrate through an empirical evaluation that our approach usually leads to very good data collection schemes in presence of strong correlations. In many cases, the solution found by our approach performs nearly as well as the theoretical lower bound given by DSC.

III. APPROXIMATION ALGORITHMS

We first present an approximation algorithm for the WL-SG problem under the uniform entropy assumption; this will help us tie the problem with some previously studied graph problems, and will also form the basis for our main algorithms. We then present a generic greedy framework that we use to derive approximation algorithms for the remaining problems.

A. The WL-SG Model: Uniform Entropy and Conditional Entropy Assumption

Without loss of generality, we assume that $H(X_i) = 1$, $\forall i$ and $H(X_i|X_j) = \epsilon$ $\forall i, j$, for all adjacent pairs of nodes (X_i, X_j) . We expect that typically $\epsilon \ll 1$.

For any compression tree that satisfies the SG property, the data movement scheme must have a subset of the sensor nodes locally broadcast their (compressed) values, such that for every edge (u, v) in the compression tree, either u or v (or both) broadcast their values. (If this is not true, then it is not possible to compress X_v using X_u .) Let S denote this subset of nodes. Each of the remaining nodes only transmits ϵ bits of information.

To ensure that the base station can reconstruct all the values, S must further satisfy the following properties: (1) S must form a dominating set of \mathcal{G}_C (any node $\notin S$ must have a neighbor in S). (2) The graph formed by deleting all edges (x, y) where $x, y \in V \setminus S$ is connected. Property (1) implies every node should get at least one of its neighbors' message for compression and property (2) guarantees the connectedness of the compression tree given S broadcast. Graph-theoretically this leads to a slightly different problem than both the classical Dominating Set (DS) and Connected Dominating Set (CDS) problems [19]. Specifically, S must be a Weakly Connected Dominating Set (WCDS) [20] of \mathcal{G}_C .

In the network shown in Figure 2, nodes 4, 3, 9 and 10 form a WCDS, and thus locally broadcasting them can give us a valid compression tree (shown in Figure 2 (ii)). However, note that nodes 4, 9, 10 and 2 form a DS but not a WCDS. As a result, we cannot form a compression tree with these

nodes performing local broadcasts (there would be no way to reconstruct the values of both X_3 and X_2).

The approach for the CDS problem that gives a $2H_{\Delta}$ approximation [19], gives a $H_{\Delta} + 1$ approximation² for WCDS [20]. We use this to prove that:

Theorem 1: Let the average distance to the base station be $d_{avg} = \frac{\sum_j d(j, BS)}{n}$. The approximation for WCSD yields a $\left(\frac{1}{1+2\epsilon(d_{avg}-1/2)}(H_{\Delta} + 1) + 2\right)$ -approximation for WL-SG problem under uniform entropy and conditional entropies assumption.

Proof: The proofs of the theorems and the lemmas are omitted due to space constraints, and can be found in the extended version of the paper [12]. ■

From the above theorem, if ϵ is small enough, say $\epsilon = o(\frac{1}{d_{avg}})$, the approximation ratio is approximately H_{Δ} . On the other hand, if ϵ is large, the approximation ratio becomes better. Specifically, if $\epsilon \approx H_{\Delta}/d_{avg}$, then we get a constant approximation. This matches our intuition that the hardness of approximation comes mainly from the case when the correlations are very strong. We can further formalize this – by a standard reduction from the set cover problem which is hard to approximate within a factor of $(1 - \delta) \ln n$ for any $\delta > 0$ [21], we can prove:

Theorem 2: The WL-SG problem can not be approximated within a factor of $(1 - \delta) \ln n$ for any $\delta > 0$ even with uniform entropy and conditional entropy, unless $NP \subseteq DTIME(n^{\log \log n})$.

B. The Generic Greedy Framework

We next present a generic greedy framework that helps us analyze the rest of the problems.

Suppose node $p(i)$ is the parent of node i in the compression tree \mathcal{T} . Let $I_{i,p(i)}$ denote the node where X_i is compressed using $X_{p(i)}$. We note that this is not required to be i or j , and could be any node in the network. This makes the analysis of the algorithms very hard. Hence we focus on the set of feasible solutions of the following restricted form: $I_{i,j}$ is either node i or j . The following lemma states that the cost of the optimal restricted solution is close to the optimal cost.

Lemma 1: Let the optimal solution be OPT and the optimal restricted solution be \widetilde{OPT} . We have $\text{cost}(\widetilde{OPT}) \leq (2 + \beta)\text{cost}(OPT)$. Furthermore, for WL-SG model, $\text{cost}(\widetilde{OPT}) \leq 2\text{cost}(OPT)$.

Our algorithm finds what we call an *extended compression tree*, which in a final step is converted to a compression tree. An *extended compression tree* $\widetilde{\mathcal{T}}$ corresponding to a compression tree \mathcal{T} has the same underlying tree structure, but each edge $e(i, j) \in \mathcal{T}$ is associated with an *orientation* specifying the raw data movement. Basically, an extended compression tree naturally suggests a restricted solution in which an edge from i to j in $\widetilde{\mathcal{T}}$ implies that i ships its raw data to j and the corresponding compression is carried out at j . We note that the direction of the edges in $\widetilde{\mathcal{T}}$ may not be

² Δ is the maximum degree and H_n is the n th harmonic number, i.e., $H_n = \sum_{i=1}^n \frac{1}{i}$.

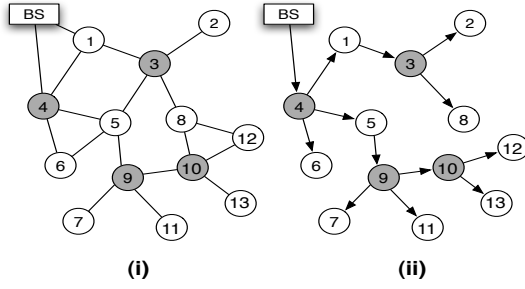


Fig. 2. (i) A weakly connected dominating set of the sensor network is indicated by the shaded nodes, which locally broadcast their values; (ii) The corresponding compression tree (e.g. Node 3 is compressed using the value of Node 1 at Node 1, whereas Node 5 is compressed using the value of Node 4 at Node 5).

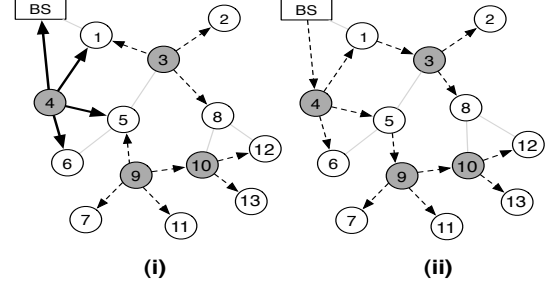


Fig. 3. Illustrating the Treestar algorithm: First the treestars centered at nodes 10, 9 and 3 are chosen, and finally the treestar centered at node 4 is chosen. This causes the parents of nodes 1 and 5 to be re-defined as node 4, the parent of node 9 to be defined as node 5, and the parent of node 3 to be defined as node 1. (i) also shows an extended compression tree.

the same as in \mathcal{T} where edges are always oriented from the root to the leaves, irrespective of the data movement. In the following, we refer *the parent of node i* to be the parent in \mathcal{T} , i.e., the node one hop closer to the root, denoted by $p(i)$.

The main algorithm greedily constructs an extended compression tree by greedily choosing subtrees to merge in iterations. We start with a empty graph \mathcal{F}_1 that consists of only isolated nodes. During the execution, we maintain a forest in which each edge is directed. In each iteration, we combine some trees together into a new larger tree by choosing the most (or approximately) cost-effective *treestar* (defined later). Let the forest at the start of the i th iteration be \mathcal{F}_i . A treestar \mathcal{TS} is specified by k trees in \mathcal{F}_i , say T_1, \dots, T_k , a node $r \notin T_j (1 \leq j \leq k)$ and k directed edges $e_j = (r, v_j) (v_j \in T_j, 1 \leq j \leq k)$. We call r the center, T_1, \dots, T_k the *leaf-trees*, e_j the *leaf-edges*. The treestar \mathcal{TS} is a specification of the data movement of X_r , which we will explain in detail shortly. Once a treestar is chosen, the corresponding data movement is added to our solution. The algorithm terminates when only one tree is left which will be our extended compression tree $\tilde{\mathcal{T}}$.

Let r be the center of \mathcal{TS} and S be the subset indices of leaf-trees. We define the cost of \mathcal{TS} ($\text{cost}(\mathcal{TS})$) to be
$$\min_{v_j \in T_j, j \in S} (c(r, \{v_j\}_{j \in S})H(X_r) + \sum_{j \in S} H(X_{v_j}|X_r)d(v_j, BS))$$
 where $c(r, \{v_j\}_{j \in S})$ is the minimum cost for sending X_r from r to all v_j 's. Essentially, the first term corresponds roughly to the cost of intra-source communication (raw data movement of X_r), denoted $IC(\mathcal{TS})$ and the second roughly to the necessary communication (conditional data movement), denoted $NC(\mathcal{TS})$. We say that the corresponding data movement is an *implementation* of the treestar. The cost function $c()$ differs for different cost models of the problem; we will specify its concrete form later.

We define the *cost effectiveness* of the treestar \mathcal{TS} to be $\text{ceff}(\mathcal{TS}) = \frac{\text{cost}(\mathcal{TS})}{k+1}$ where k is the number of leaf-trees in \mathcal{TS} . In each iteration, we will try to find the most cost effective treestar. Let $\text{Mce-Treestar}(\mathcal{F}_i)$ be the procedure for finding the most (or approximately) cost effective treestar on \mathcal{F}_i . The actual implementation of the procedure Mce-Treestar

will be described in detail in the discussion of each cost model. In some cases, finding the most cost-effective treestar is NP-hard and we can only approximate it.

We now discuss the final data movement scheme and how the cost of the final solution has been properly accounted in the treestars that were chosen. Suppose in some iteration, a treestar \mathcal{TS} is chosen in which the center node r sends its raw information to each $v_j (v_j \in T_j, j \in S)$ (S is the set of indices of leaf-trees in \mathcal{TS}). The definition of the cost function suggests that X_{v_j} is compressed using X_r at v_j , and the result is sent from v_j to BS. However, this may not be consistent with the extended compression tree $\tilde{\mathcal{T}}$. In other words, some v_j may later become the parent of r , due to latter treestars being chosen, in $\tilde{\mathcal{T}}$ which implies that r should be compressed using v_j instead of the other way around. Suppose some leaf $v_p (v_p \in T_p, p \in S)$ is the parent of r in $\tilde{\mathcal{T}}$. The actual data movement scheme is determined as follows. We keep the raw data movement induced by \mathcal{TS} unchanged, i.e., r still sends X_r to each $v_j (j \in S)$. But now, $X_r|X_{v_p}$ instead of $X_{v_p}|X_r$ is computed on node v_p and sent to the base station. Other leaves $v_j (j \neq p)$ still compute and send $X_{v_j}|X_r$. It is easy to check this data movement scheme actually implements the extended compression tree $\tilde{\mathcal{T}}$.

For instance, in Figure 3, node 3 is initially the parent of node 1, but later node 4 becomes the parent of node 1, and in fact node 1 ships $X_1|X_4$ to the base station (and not $X_1|X_3$). Node 1 now being the parent of node 3 also compresses X_3 and sends $X_3|X_1$ to BS. Due to the fact that $\frac{1}{\beta} \leq \frac{H(X|Y)}{H(Y|X)} \leq \beta$, the actual data movement cost is at most β times the sum of the treestar costs. Thus every part of the communication cost incurred is counted in some treestar. We formalize the above observations as the following lemma:

Lemma 2: Let \mathcal{TS}_i be the treestars we choose in iteration i for $1 \leq i \leq \ell$. Then: $\text{cost}(\mathcal{T}) \leq \beta \sum_{i=1}^{\ell} \text{cost}(\mathcal{TS}_i)$.

The pseudocode for constructing $\tilde{\mathcal{T}}$ and the corresponding communication scheme is given in Algorithm 1.

C. The Generic Analysis Framework

OPT is defined as the optimal restricted solution. Let \mathcal{TS}_i be the treestar computed in iteration i . After ℓ iterations (it is

Algorithm 1: The Generic Greedy Framework

```

 $\mathcal{F}_1 = \bigcup_{i=1}^n \{\{X_i\}\};$ 
 $i \rightarrow 1;$ 
while  $\mathcal{F}_i$  is not a spanning tree do
   $\mathcal{TS}_i = Mce - Treestar(\mathcal{F}_i);$ 
  Let  $E(\mathcal{TS}_i)$  be leaf-edges of  $\mathcal{TS}_i$  and  $r$  is the center
  of  $\mathcal{TS}_i;$ 
   $\mathcal{F}_{i+1} \leftarrow \mathcal{F}_i + E(\mathcal{TS}_i);$ 
   $T_r \leftarrow T_r + IC(\mathcal{TS}_i);$ 
   $i = i + 1;$ 
 $\vec{T} = \mathcal{F}_i;$ 
for each directed edge  $e(i, j) \in E(\vec{T})$  do
  if  $i$  is the parent of  $j$  then
    | Compute  $X_j|X_i$  at  $j$  and send it to  $BS;$ 
  else
    | Compute  $X_i|X_j$  at  $j$  and send it to  $BS;$ 

```

easy to see ℓ must be smaller than n), the algorithm terminates. We assume in each iteration, *Mce-Treestar* is guaranteed to find an α -approximate most cost-effective treestar. We assume further the bounded conditional entropy parameter is β . Given these, we can prove that:

Theorem 3: We can find a $2\alpha\beta^2H_n$ approximate restricted solution \mathcal{T} in polynomial time, i.e., $\text{cost}(\mathcal{T}) \leq \beta \sum_{i=1}^{\ell} \text{cost}(\mathcal{TS}_i) \leq 2\alpha\beta^2H_n \text{cost}(\text{OPT})$.

D. The WL-SG Model

We first specify the cost function $c(r, \{v_j\}_{j \in S})$ in the wireless sensor network model where we require the compression tree to be a subgraph of the communication graph and then give a polynomial time algorithm for finding the most cost-effective treestar.

Recall $c(r, \{v_j\}_{j \in S})$ is cost of sending X_r from r to all v_j 's. It is easy to see $c(r, \{v_j\}_{j \in S}) = H(X_r)$ since we require v_j to be adjacent to r and a single broadcast of X_r from r can accomplish the communication. The most cost-effective treestar can be computed as follows: We fix a node r as the center to which all leaf-edges will connect. Assume T_1, T_2, \dots are sorted in a non-increasing order of $h(r, T_j) = \min_{v \in T_j \cap N(r)} H(X_v|X_r)d(v, BS)$. $h(r, T_j)$ captures the minimum cost of sending the data of some node in $T_j \cap N(r)$ conditioned on X_r to the base station. The most cost-effective treestar is determined simply by

$$\min_k \left\{ \frac{H(X_r) + \sum_{j=1}^k h(r, T_j)}{k+1} \right\}.$$

In each iteration, for each candidate center r , sorting $h(r, T_j)$ s needs $O(\deg(r) \log \deg(r))$ time. So, the most-effective treestar can be found in $O(|E| \log n)$ time. Therefore, the total running time is $O(n|E| \log n)$. In each iteration, for each candidate center r , sorting $h(r, T_j)$ s needs $O(\deg(r) \log \deg(r))$ time. Using Lemma 1 and Theorem 3, we obtain the following.

Theorem 4: We can compute a $4\beta^2H_n$ -approximation for the WL-SG model in $O(n|E| \log n)$ time.

E. The WL-NS Model

Here we don't put any restrictions on the compression trees. Thus, a source node is able to send the message to a set of nodes through a Steiner tree and the cost for sending one bit is the sum of the weights of all inner nodes of the Steiner tree (due to the broadcasting nature of wireless networks). In graph theoretic terminology, it is the cost of the connected dominating set that includes the source node and dominates all terminals. Formally, the cost of the treestar \mathcal{TS} with node r as the center and S be the set of indices of the leaf-trees is defined to be:

$$\min_{v_j \in T_j} \left(Cds(r, \{v_j\}_{j \in S})H(X_r) + \sum_{j \in S} H(X_{v_j}|X_r)d(v_j, BS) \right)$$

where $Cds()$ is the minimum connected set dominating all nodes in its argument.

Next, we discuss how to find the most effective treestar. We reduce the problem to the following version of the directed steiner tree problem [22].

Definition 1: Given a weighted directed graph G , a specified root $r \in V(G)$, an integer k and a set $X \subseteq V$ of terminals, the *D-Steiner*(k, r, X) problem asks for a minimum weight directed tree rooted at r that can reach any k terminals in X . It has been shown that the D-Steiner(k, r, X) problem can be approximated within a factor of $O(n^\epsilon)$ for any fixed $\epsilon > 0$ within time $O(n^{O(\frac{1}{\epsilon})})$ [22].

The reduction is as follows. We first fix the center r . Then, we create a undirected node-weighted graph D . The weight of each node is $H(X_r)$. For each node v , we create a copy v' with weight $w(v') = H(X_v|X_r)d(v, BS)$ and add an edge (u, v') for each $u \in \bar{N}(v)$. For each tree component T_j , we create a group $g_j = \{v'|v \in T_j\}$. Then, we construct the directed edge-weighted graph. We replace each undirected edge with two directed edges of opposite directions. For each group g_i , we add one node t_i and edges (v, t_i) for all $v \in g_i$. The following standard trick will transfer the weight on nodes to directed edges. For each vertex $v \in V(D)$, we replace it with a directed edge (v', v'') with the same weight as $w(v)$ such that v' absorbs all incoming edges of v and v'' takes all outgoing edges of v . We let all t_i'' 's be the terminals we want to connect. It is easy to see a directed steiner tree connecting k terminals in the new directed graph corresponds exactly to a treestar with k leaf-trees.

Theorem 5: We develop an $O(\beta^3 n^\epsilon \log n)$ -approximation for the WL-NS model for any fixed constant $\epsilon > 0$ in $O(n^{O(\frac{1}{\epsilon})})$ time.

IV. EXPERIMENTAL EVALUATION

We conducted a comprehensive simulation study over several datasets comparing the performance of several approaches for data collection. Our results illustrate that our algorithms can exploit the spatial correlations in the data effectively, and perform comparably to the DSC lower bound. Below we present results over a few representative settings.

Comparison systems:

We compare the following data collection methods.

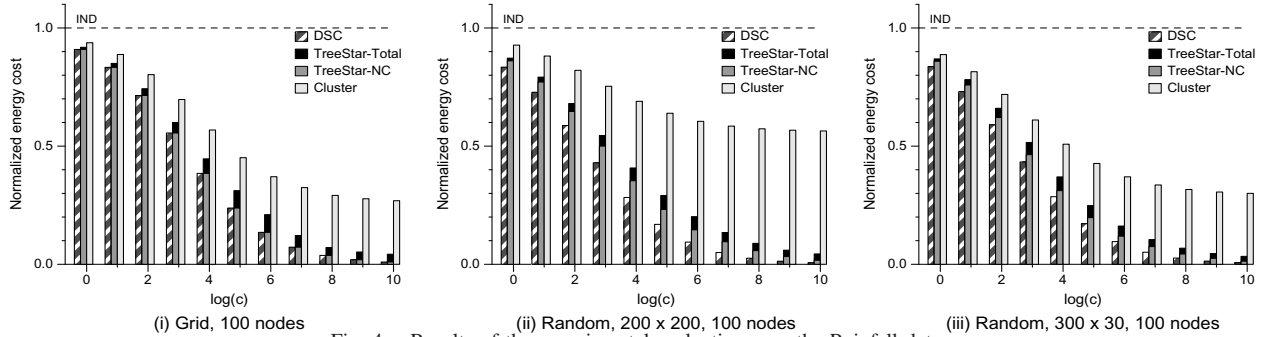


Fig. 4. Results of the experimental evaluation over the Rainfall data

- IND (Sec. II-B): Each node compresses its data independently of the others.
- Cluster (Sec. II-B): The clusters are chosen using the greedy algorithm presented in Chu et al. [8] – we start with each node being in its own cluster, and combine clusters greedily, till no improvement is observed.
- DSC: the theoretical lower bound is plotted (Sec. II-B).
- TreeStar: Our algorithm, presented in Sec. III-D, augmented with a greedy local improvement step³.

For the TreeStar algorithm, we also show the NC cost (which measures how well the compression tree chosen by TreeStar approximates the original distribution). This cost is lower bounded by the cost of DSC (which uses the best possible compression tree).

Rainfall Data:

For our first set of experiments, we use an analytical expression of the entropy that was derived by Pattem et al. [7] for a data set containing *precipitation* data collected in the states of Washington and Oregon during 1949-1994⁴. All the nodes have uniform entropy ($H(X_i) = h$), and the conditional entropies are given by: $H(X_i|X_j) = (1 - \frac{c}{c+dist(i,j)})h$, where $dist(i, j)$ is the Euclidean distance between the sensors i and j . The parameter c controls the correlation. For small values of c , $H(X_i|X_j) \approx h$ (indicating independence), but as c increases, the conditional entropy approaches 0.

Figure 4 shows the results for 3 synthetically generated sensor networks. We plot the total communication cost for each of the above approaches normalized by the cost of IND. The first plot shows the results for a 100-node network where the sensor nodes are arranged in a uniform grid. Since the conditional entropies depend only on the distance, for any two adjacent nodes i, j , $H(X_i|X_j)$ is constant. Because of this, TreeStar-NC is always equal to DSC in this case. As we can see, the extra cost (of local broadcasts) is quite small, and overall TreeStar performs much better than either Cluster or IND, and performs nearly as well as DSC.

We then ran experiments on randomly generated sensor networks, both containing 100 nodes each. The nodes were randomly placed in either a 200x200 square or a 300x30 rectangle, and communication links were added between nodes

³After the TreeStar algorithm finds a feasible solution, adding a few redundant local broadcasts can cause significant reduction in the NC cost. We greedily add such local broadcasts till the solution stops improving.

⁴http://www.jisao.washington.edu/data_sets/widmann

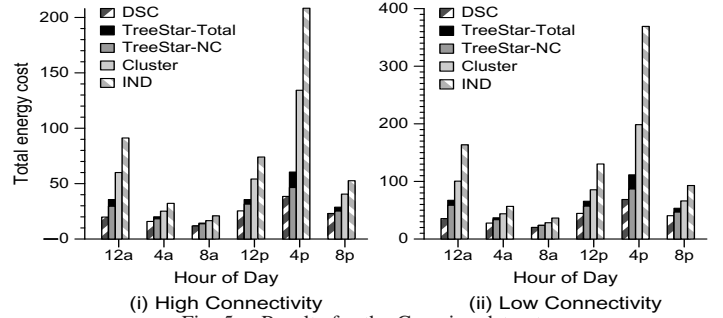


Fig. 5. Results for the Gaussian dataset

that were sufficiently close to each other ($distance < 30$). For each plotted data point, we ran the algorithms on 10 randomly chosen networks, and averaged the results. As we can see in Figures 4 (ii) and (iii), the relative performance of the algorithms is quite similar to the first experiment. Note that, because the conditional entropies are not uniform, TreeStar-NC cost was typically somewhat higher than DSC. The cost of local broadcasts for TreeStar was again relatively low.

Gaussian approximation to the Intel Lab Data:

For our second set of experiments, we used *multivariate Gaussian* models learned over the *temperature* data collected at an indoor, 49-node deployment at the Intel Research Lab, Berkeley⁵. Separate models were learned for each hour of day [23] and we show results for 6 of those. After learning the Gaussian model, we use the *differential entropy* of these Gaussians for comparing the data collection costs. We use the aggregated connectivity data available with the dataset to simulate different connectivity behavior: in one case, we put communication links between nodes where the success probability was $> .35$, resulting in somewhat sparse network, whereas in the other case, we used a threshold of $.20$.

Figure 5 shows the comparative results for this dataset. The dataset does not exhibit very strong spatial correlations: as we can see, optimal exploitation of the spatial correlations (using DSC) can only result in at best a factor of 4 or 5 improvement over IND (which ignores the correlations). However, TreeStar still performs very well compared to the lower bound on the data collection cost, and much better than the Cluster approach. Different connectivity behavior does not affect the relative performance of the algorithms much, with the low-connectivity network consistently incurring about twice as much energy cost compared to the high-connectivity network.

⁵<http://db.csail.mit.edu/labdata/labdata.html>

V. RELATED WORK

Wireless sensor networks have been a very active area of research in recent years (see [24] for a survey). Due to space constraints, we only discuss some of the most closely related work on data collection in sensor networks here. Directed diffusion [25], TinyDB [26], LEACH [27] are some of the general purpose data collection mechanisms that have been proposed in the literature. The focus of that work has been on designing protocols and/or declarative interfaces to collect data, and not on optimizing continuous data collection. Aside from the works discussed earlier in the paper [7], [8], [9], the BBQ system [23] also uses a predictive modeling-based approach to collect data from a sensor network. However, the BBQ system only provides probabilistic, approximate answers to queries, without any guarantees on the correctness. Scaglione and Servetto [14] also consider the interdependence of routing and data compression, but the problem they focus on (getting all data to all nodes) is different from the problem we address. In seminal work, Gupta and Kumar [28] proved that the transport capacity of a random wireless network scales only as $O(\sqrt{n})$, where n is the number of sensor nodes. Although this seriously limits the scalability of sensor networks in some domains, in the kinds of applications we are looking at, the *bandwidth* or the *rate* is rarely the limiting factor; to be able to last a long time, the sensor nodes are typically almost always in sleep mode.

Several approaches not based on predictive modeling have also been proposed for data collection in sensor networks or distributed environments. For example, constraint chaining [29] is a suppression-based exact data collection approach that monitors a minimal set of node and edge constraints to ensure correct recovery of the values at the base station.

VI. CONCLUSIONS

Designing practical data collection protocols that can optimally exploit the strong spatial correlations typically observed in a given sensor network remains an open problem. In this paper, we considered this problem with the restriction that the data collection protocol can only utilize second-order marginal or conditional distributions. We analyzed the problem, and drew strong connections to the previously studied weakly-connected dominating set problem. This enabled us to develop a greedy framework for approximating this problem under various different communication model or solution space settings. Although we are not able to obtain constant factor approximations, our empirical study showed that our approach performs very well compared to the DSC lower bound. We observe that the worst case for the problem appears to be when the conditional entropies are close to zero, and that we can get better approximation bounds if we lower-bound the conditional entropies. Future research directions include generalizing our approach to consider higher-order marginal and conditional distributions, improving the approximation bounds by incorporating lower bounds on the conditional entropy values, and also understanding how to apply such

approximation algorithms in practice in presence of node and communication link failures.

REFERENCES

- [1] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, no. 4, 1973.
- [2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Info. Theory*, 1976.
- [3] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, 2004.
- [4] X. Su, "A combinatorial algorithmic approach to energy efficient information collection in wireless sensor networks," *ACM Trans. Sen. Netw.*, vol. 3, no. 1, p. 6, 2007.
- [5] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Networked slepian-wolf: Theory, algorithms, and scaling laws," *IEEE Trans. Info. Theory*, vol. 51(12), pp. 4057–4073, 2005.
- [6] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *IEEE Trans. Information Theory*, 2003.
- [7] S. Pattem, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *IPSN*, 2004.
- [8] D. Chu, A. Deshpande, J. Hellerstein, W. Hong, "Approximate data collection in sensor networks using probabilistic models," *ICDE*, 2006.
- [9] R. Cristescu, B. Beferull-Lozano, M. Vetterli, and R. Wattenhofer, "Network correlated data gathering with explicit communication: Np-completeness and algorithms," *IEEE/ACM Trans. Netw.*, 2006.
- [10] J. Liu, M. Adler, D. Towsley, and C. Zhang, "On optimal communication cost for gathering correlated data through wireless sensor networks," in *Proceedings of ACM Mobicom*, 2006.
- [11] L. Wang and A. Deshpande, "Predictive modeling-based data collection in wireless sensor networks," in *EWSN*, 2008.
- [12] J. Li, A. Deshpande, and S. Khuller, "On computing compression trees for data collection in sensor networks," 2009, <http://arxiv.org/abs/0907.5442>.
- [13] P. von Rickenbach and R. Wattenhofer, "Gathering correlated data in sensor networks," in *DIALM-POMC*, 2004.
- [14] A. Scaglione and S. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," in *Mobicom*, 2002.
- [15] A. Goel and D. Estrin, "Simultaneous optimization for concave costs: Single sink aggregation or single source buy-at-bulk," in *SODA*, 2003.
- [16] Y. Kotidis, "Snapshot queries: Towards data-centric sensor networks," in *ICDE*, 2005.
- [17] H. Gupta, V. Navda, S. Das, and V. Chowdhary, "Efficient gathering of correlated data in sensor networks," in *MobiHoc*, 2005.
- [18] C. Chow and C. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Trans. on Information Theory*, 1968.
- [19] S. Guha and S. Khuller, "Approximation algorithms for connected dominating sets," *Algorithmica*, vol. 20, no. 4, 1998.
- [20] Y. Chen and A. L. Liestman, "Approximating minimum size weakly-connected dominating sets for clustering mobile ad hoc networks," in *Mobihoc*, 2002, pp. 165–172.
- [21] U. Feige, "A threshold of $\ln n$ for approximating set cover," *J. ACM*, vol. 45, no. 4, pp. 634–652, 1998.
- [22] M. Charikar, C. Chekuri, T. Cheung, Z. Dai, A. Goel, and M. Li, "Approximation algorithm for directed Steiner problem," *J. Algo.*, vol. 33, no. 1, pp. 73–91, 1999.
- [23] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *VLDB*, 2004.
- [24] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, 2002.
- [25] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in *ACM Mobicom*, 2000.
- [26] S. Madden, W. Hong, J. M. Hellerstein, and M. Franklin, "TinyDB web page," <http://telegraph.cs.berkeley.edu/tinydb>.
- [27] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *HICSS*, 2000.
- [28] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. on Information Theory*, vol. 46, 2000.
- [29] A. Silberstein, R. Braynard, and J. Yang, "Constraint-chaining: Energy-efficient continuous monitoring in sensor networks," in *SIGMOD*, 2006.