On Optimal Communication Cost for Gathering Correlated Data through Wireless Sensor Networks

Junning Liu Micah Adler Don Towsley
Dept. of Computer Science
University of Massachusetts
Amherst, MA, USA
{liujn, micah, towsley}@cs.umass.edu

Chun Zhang
IBM T.J. Watson Research Center
Hawthorne, NY, USA
czhang1@us.ibm.com

ABSTRACT

In many energy-constrained wireless sensor networks, nodes cooperatively forward correlated sensed data to data sinks. In order to reduce the communication cost (e.g. overall energy) used for data collection, previous works have focused on specific coding schemes, such as Slepian-Wolf Code or Explicit Entropy Code. However, the minimum communication cost under arbitrary coding/routing schemes has not yet been characterized. In this paper, we consider the problem of minimizing the total communication cost of a wireless sensor network with a single sink. We prove that the minimum communication cost can be achieved using Slepian-Wolf Code and Commodity Flow Routing when the link communication cost is a convex function of link data rate. Furthermore, we find it useful to introduce a new metric *distance entropy*, a generalization of entropy, to characterize the data collection limit of networked sources. When the energy consumption is proportional to the link data rate (e.g. normally in 802.11), we show that distance entropy provides a lower bound of the communication cost and can be achieved by using a specific rate SWC and shortest path routing. Theoretically, achieving optimality may require global knowledge of the data correlation structure, which may not be available in practice. Therefore, we propose a simple, hierarchical scheme that primarily exploits data correlation between local neighboring nodes. We show that for several correlation structures and topologies, the communication cost achieved by this scheme is within a constant factor of the distance entropy, i.e., it is asymptotically optimal. Finally, we simulate our algorithm using radar reflectivity data as well as traces from Gaussian Markov Fields (GMF). As the network size goes large, for the radar data, we find our algorithm saves two thirds of the communication cost compared to a non-coding approach; as for the GMF data, our algorithm converges to a constant factor ($1.5 \sim 1.8$) of the distance entropy.

MobiCom'06, September 23–26, 2006, Los Angeles, California, USA.
Copyright 2006 ACM 1-59593-286-0/06/0009 ...$5.00.

Categories and Subject Descriptors

C.2.1 [**Computer Communication Networks**]: Network Architecture and Design—*Wireless communication, Network topology*; E.4 [**Data**]: Coding and Information Theory—*Data compaction and compression*; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*Routing and layout*

General Terms

Algorithms, Performance, Design, Theory

Keywords

Distributed Source Coding, Joint Coding and Routing, Communication Cost Minimization, Network Coding

1. INTRODUCTION

In recent years there has been an increasing demand for the use of wireless sensor networks to measure environments (such as temperature, humidity, light, and vibration, etc. [5][12]) and to collect these measurements at a sink. Thus communication is an important task that must be performed in order to collect data from sensors. A common characteristic of these networks, however, is that they are energy-constrained, and thus the communication cost (e.g. overall energy consumption) must be considered in the design of data collection schemes.

Since sensor measurements are often highly correlated, minimizing the overall communication cost is a joint coding/routing problem: routing is required because the source data needs to be shipped through a network to the sink; coding can be used to take advantage of the source correlation and any other known distributional information. Several algorithms based on specific codes have been proposed to minimize the communication cost of wireless sensor networks with a single sink [10][26][14]. When coding is restricted to Explicit Entropy Code (EEC)[1] [10], Cristescu et al. [10] shows that choosing the optimal routes is a NP-hard problem; Pattem et al. [26] proposes a heuristic algorithm to minimize the communication cost assuming a simplified source model. When coding is restricted to

[1] For EEC, a node sends out data with a rate equal to the joint entropy rate of incoming data and its own sensed data.

a Splepian-Wolf Code (SWC) [29][2] and Commodity Flow Routing (CFR) is used, Cristescu et al.'s work [10] [11] finds that Shortest Path Routing (SPR) combined with an optimal rate SWC achieves the minimum communication cost among such schemes. However, in the general case where arbitrary coding/routing operations are allowed, it is still not known what the minimum communication cost is, and how to achieve it. By arbitrary coding/routing operations, we mean that a node can perform arbitrary transformations (functions) on the incoming data and local sensed data.

In this paper, we consider the problem of minimizing the total communication cost over all coding/routing schemes, as well as designing algorithms to achieve it in practice for data communication networks with a single sink (i.e., data collection point). Our work focuses on wireless sensor networks with energy constraints, while the general results apply to wired network (Internet) in which packet delay and bandwidth consumption are typical cost metrics [21] [31].

Theoretically, we prove that, for a wireless sensor network with a single sink, the optimal scheme using only Slepian-Wolf Coding and Commodity Flow Routing is optimal over the class of all possible coding/routing schemes, as long as the energy consumption is a convex function of link data rate. Since this result is based on arbitrary coding/routing schemes that incorporate Network Coding (NC) [3], a corollary of our result is that, for correlated data collection at a single sink, NC can not further improve the minimum communication cost achieved by SWC+CFR. Furthermore, we find it useful to introduce a new metric *distance entropy* to lower bound the minimum communication cost. Distance entropy can be viewed as a generalization of entropy that summarizes a probability distribution while also taking into account the underlying network topology. When the energy consumption is proportional to the link data rate (e.g. normally in 802.11), we show that distance entropy can be achieved by a coding scheme using SWC and Shortest Path Routing (SPR). Last, we extend our results to networks that incorporate broadcast channels. We show that broadcasting does not help in terms of minimizing the total communication cost for the single sink case data collection problems.

For data networks with a single sink, our result shows that the SWC+CFR scheme achieves the minimum communication cost. However, in practice, the minimum communication cost is still difficult to achieve due to several reasons. First, knowledge of the global data correlation structure, which is essential for optimal SWC scheme, is normally unavailable or too costly to learn. Specifically, multidimensional entropy estimation is an extremely costly task due to the curse of dimensionality [4]. Second, SWC [29] is an existential rather than constructive result, even if the correlation knowledge is available through an oracle, hardly any general practical SWC schemes have been developed [31] [1]. In another mostly studied coding scheme, EEC, it was shown in [10] that it is NP-hard to minimize the communication cost, the coding complexity of EEC is also comparable to SWC.

Practically, we design a simple and effective algorithm that achieves (at least asymptotically) the optimal communication cost for several generic and commonly used classes of source models. This algorithm only relies on the source data correlation between local neighboring nodes. The source models include a Hard Continuity Field model, a Linear Variance Continuity Field model and a Gaussian Markov Field model. We provide nontrivial lower bounds on the distance entropy of these source models for a 2D sensor grid. We then propose a simple hierarchical data collection algorithm and demonstrate that it is asymptotically optimal for these source models, i.e., it achieves a communication cost that is within a constant factor of our lower bound on the corresponding distance entropy. We also extend the grid results to corresponding high probability results for randomly deployed sensor networks. We evaluate our algorithm by simulations using 2D radar reflectivity data and a simulated Gaussian Markov Field. We demonstrate that our algorithms reduce communication cost about two thirds compared to a non-coding raw data collecting method (even for medium size network) and within a constant factor around $1.5 \sim 1.8$ of the distance entropy of the GMF data.

The paper is organized as follows: In Section 2, we introduce the background and related work. In Section 3, we formalize the model. In Section 4, we define distance entropy and prove the universal optimal results. In Section 5, we propose the simple hierarchical data collection scheme and prove its asymptotic optimality. In Section 6, we evaluate the performance of our algorithm through simulations. Finally, we conclude and discuss future work in Section 7.

2. BACKGROUND AND RELATED WORK

2.1 Background

There has been considerable interest in applying information theory to data networks recently. By doing so, the traditional routing problems become joint coding/routing optimization problems. In general, Coding consists of Source Coding (SC) and Network Coding. And by routing, we mean the traditional commodity flow routing, where messages can be forwarded, split and merged but not decoded or recoded. With these clarifications, by arbitrarily coding and routing operations we mean any combination of SC, NC and routing[3].

There are two aspects of a joint coding/routing problem, network combinatorics and information theory: as [2] summarizes, Combinatorics is concerned with packing problems (e.g. flows) that are constrained by the graph structure. It grews out of a need to understand the shipment of cargo in transportation networks and does not capture the subtleties of information transmission. On the other hand, information theory provides a deep understanding of complex communication problems over structurally simple channels but does not yet fully extend to arbitrary graph structures. An interesting observation is that when we consider a more general problem by adding the coding elements, the problem often becomes more tractable. Take the maximum multicast throughput problem as an example. If we are restricted to use traditional routing, the problem of maximizing multicast throughput is NP-hard while when network coding is allowed it can be solved using linear programming[19]. For our problem, if the coding part is fixed to be EEC, the routing optimization is related to a multiple travels salesman problem that is NP-hard [10]. However, when arbitrary

[2] SWC is a distributed source coding technique that allows the sensor nodes to encode without explicit communication. Each sensor encodes its data to some rate with the joint rate vector in the achievable Slepian-Wolf region.

[3] Note that these operations could be inseparable.

coding is allowed, combining ideas from both theories of combinatorial optimization and information theory enables us to make significant progress towards understanding the performance limit of such information networks.

2.2 Related work

There has been much research on Distributed Source Coding (DSC) and NC. A thorough review of DSC can be found in [31] where it is claimed that there are still few practical DSC schemes for general source models. [6] proposes a practical SWC scheme based on syndromes. It uses a Hamming distance constraint model so the result can be generalized to a hierarchical scheme applicable to such hard constraint models. There is no spatial or cost consideration in [6]. For NC, if there is just a single sink in addition to independent sources, there is no need for Network Coding. [18] shows that traditional routing where data is treated as commodity flows suffices to solve the data collection problem for such networks. [28] studies the problem of separating SC from NC for collecting data from correlated sources at multiple sinks. They show that the case of 2 sources and 2 sinks is always separable, and give counter-examples for some other cases. Since inseparable NC and SC implies that NC is necessary (not vice versa), we do know that there are cases where NC is needed. Thus further work is needed to determine the utility of NC in our situation. [16] shows that random linear network coding suffices for the network coding of correlated sources. [30] provides a practical low complexity scheme of joint DSC and NC. The scheme is suboptimal and focuses on two sources that are related by a binary symmetric channel. Most of these works on coding apply only to some limited source models, furthermore they all focus on the capacity aspect and ignore costs.

Some work has considered network costs; [22] studies the problem of network coding with a cost criterion. For minimum cost correlated data gathering, [14] considers an abstract cost function and a special source model where the joint entropy is a concave function of the number of sources and independent of the source locations. They show that there exists a random approximation of a transmission tree that is universally optimal for all concave cost functions. [26] also studies correlated sensor data collection on a grid. They use a simplified cost function as well as a simplified correlation model that ignores spatial features as in [14]: the joint entropy is a linear function of the number of sources. Thus their discovery of optimal clustering size is consistent with [14]'s general result. Most of these works use simplified abstract source models and assume a given coding algorithm with certain output rates available. Our work, on the other hand, imposes no restrictions on the source correlation model and the coding algorithm.

In order to study the asymptotic behavior of data collecting sensor nets, [24] studies the scaling problem of a large number sensors deployed in a Gaussian Markov Field (GMF) by comparing the per node capacity and node data rate asymptotically. [11] compares SWC and EEC's asymptotic performance on a 1D grid and shows under various conditions that EEC performs asymptotically as well as SWC, which will show to be asymptotically optimal under these conditions. [13] investigates the problem of joint optimization of sensor nodes deployment and data gathering cost in a lossy setting. Most these works assume a Gaussian source distribution. Our practical design targets a more general class of source models that are representative of real spatial data. Of particular interest to us, [26]'s experience equation learned from real rainfall spatial data verifies the validity of the total entropy assumption in our generic source modelling. [17] models spatially correlated sources using real spatial data. Their model also falls within our model framework of LVCF and GMF thus further supports the generality of LVCF and GMF.

3. MODEL FORMULATION

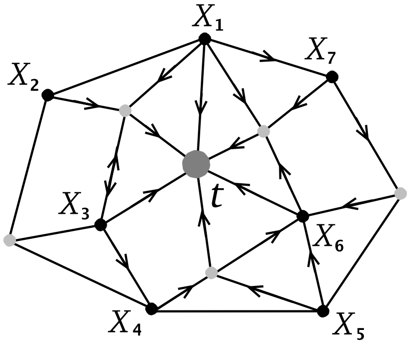

Figure 1: A Layout of the General Problem of Gathering Correlated Data through a Network

We consider a network composed of both source nodes and pure relaying nodes (As shown in Figure 1). For simplicity of representation, we assume all the nodes are source nodes and represent a $N+1$-node network as a graph $G = (V, E)$ (directed or undirected), in which $V = \{v_1, \dots, v_N, t\}$ is the set of nodes, and E is the set of edges. Here t is the sink. All nodes in V are able to code and transmit data. An edge $e = (v_i, v_j) \in E$ iff there is a direct communication link between node v_i and node v_j.

Each node v_i periodically measures a continuous random source X_i and generate a discrete random source $\hat{X}_i$(e.g. quatization). The joint source vector $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_N\}$ is characterized by a joint probability distribution $p(\hat{X}_1 = \hat{x}_1, \dots, \hat{X}_N = \hat{x}_N) = p(\hat{x}_1, \hat{x}_2, ..., \hat{x}_N)$. Let $\{\hat{X}(\tau)\}_{\tau=1}^{\infty}$ be a stationary random process where $\hat{X}(\tau) = (\hat{X}_1(\tau), \dots, \hat{X}_N(\tau))$ is a *field sample* that corresponds to the set of samples gathered from all sources at time-slot τ, $\tau = 1, 2, \dots$ is the time stamp in second. For simplicity of presentation, assume that $\hat{X}(\tau)$ is i.i.d. as we focus on the spatial correlation while our results can be extended to the general case of collecting multiple field samples that are temporally correlated.

Each edge (link) $(v_i, v_j) = e \in E$ has capacity $c_{ij} > 0$ (or c_e), specifying the maximal transmission rate over the link. Link (v_i, v_j) has an associated weight $w_{ij} \geq 0$ (or w_e) that relates to its communication cost. Let r_e be the data rate along edge e in bits per second. Naturally, the communication cost rate (both transmitting and receiving cost per second) along edge e, $g(r_e, w_e)$ is a strictly increasing function of r_e and w_e [11].[4] In practice, if a node uses a fixed transmission power (as the normal mode of 802.11), then

[4] The data rate and the cost rate can be dynamic and changes all the time, we use the average cost rate of the network as the performance metric.

the communication cost rate is a linear function of the data rate. i.e., $g(r_e, w_e) = r_e \cdot w_e$ [11]. For this linear cost function, w_e corresponds to the communication cost per bit. For wireless communication links, $w_{ij} = {l_{ij}}^{\alpha}$ where $2 \leq \alpha \leq 4$ depends on the medium and l_{ij} is the Euclidean distance between nodes v_i and v_j. If the protocol allows nodes to adjust the transmission power, then g is not linear but in general a convex function of data rate. We study both cases of cost functions. Given the edge weights w, we use W_i to denote the sum of the weights of edges on the shortest path from v_i to sink t. We assume that the communication links are implemented as discrete memoryless channels.[5] We fist derive our results based on point to point links (channels)[6] then extend it to include broadcast links. We also omit the negligible communication overhead induced by scheduling and routing control since data can be packed in arbitrarily large packets.

We define *source graph* $G_X = (G, w, c, \hat{X})$ to be the network along with its link costs, capacities and source descriptions. A *Communication Scheme* specifies, for all the nodes, "what to send to whom". It is a set of functions that maps each node's received bits and local generated data (if any) to its output bits and the corresponding selected channels. A *Data Collection Scheme (DCS)* Υ is a communication scheme that allows the network to collect all of the data $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N$ at the sink t *near losslessly* - decode losslessly with zero or an arbitrarily small error probability [8]. A *SWC scheme* Υ_{SWC} is a DCS that only uses Slepian-Wolf source codes at the sources coupled with commodity flow routing. A *SWC-SP scheme* Υ_{SWC-SP} is a SWC scheme that only uses shortest path commodity flow routing. Let Π, Π_{SWC}, Π_{SWC-SP} be the set of all DCSs, the set of all SWC schemes and the set of all SWC-SP schemes, correspondingly.

The *cost rate* for any data collection scheme Υ on a source graph G_X is defined as $W_\Upsilon(G_X) = \sum_{e \in E} g(r_e, w_e)$, or simply denoted as W_Υ. W.l.o.g., we assume the field samples are generated every second, thus W_Υ also equals the cost per field sample. In this paper, our goal is to identify and achieve the minimum communication cost $W_{\Upsilon^*} = \min_{\Upsilon \in \Pi} W_\Upsilon$.

4. OPTIMAL DATA COLLECTION SCHEME

In this section we prove our optimality result. We introduce a new concept, the *Distance Entropy* of a source graph G_X, to characterize the spatial distribution of its source information. Then in Theorem 1, we prove that distance entropy can be achieved by SWC plus shortest path routing. Next, for more general convex cost functions, we prove the universal optimality of the SWC scheme in Theorem 3 based on Theorem 2 and Lemma 1. Finally, we extend the optimality result to networks that include broadcast channels in Theorem 4. W.l.o.g., we assume the nodes $v_1, v_2, \ldots, v_N$ are in a nondecreasing order of shortest path weight to the sink, i.e., $W_1 \leq W_2 \leq \ldots \leq W_N$.

DEFINITION 1. For any source graph G_X, *The Distance Entropy $H_w(G_X)$ is*

$$H_w(G_X) = \sum_{j=1}^{N} W_j \times H(\hat{X}_j | \hat{X}_{j-1}, \ldots, \hat{X}_1)$$

Consider the cost function $g(r_e, w_e) = r_e \cdot w_e$. We have the following theorem describing the total communication cost to collect one field sample.

THEOREM 1. The cost of any DCS Υ on a source graph G_X to collect one field sample is lower bounded by the distance entropy of G_X

$$\min_{\Upsilon \in \Pi} W_\Upsilon(G_X) \geq H_w(G_X).$$

In the absence of capacity constraints, a SWC-SP scheme with an optimal rates allocation $r_j = H(\hat{X}_j | \hat{X}_{j-1}, \ldots, \hat{X}_1)$ achieves the cost of $H_w(G_X)$. Thus

$$\min_{\Upsilon \in \Pi_{SWC-SP}} W_\Upsilon(G_X) = H_w(G_X).$$

PROOF. See technical report [20]. The idea of the proof is to first group nodes into a sequence of sets according to their shortest path weights to the sink, and then investigate the information flow across the cuts between adjacent sets in an equivalent constructed graph. This differs from [10]'s proof and is a more general result. [10] fixes the coding part and shows that for SWC schemes finding the optimal rate (the network combinatorial part) is a Linear Programming problem. Since we have no limitations on coding, our proof is more general applying to arbitrary schemes. □

Theorem 1 shows that distance entropy is a lower bound on the total communication cost. Furthermore, it shows that if there are no capacity constraints (this is often reasonable when the data rates are far less than the capacities), distance entropy is an achievable tight bound and thus the best possible performance for such data collection tasks.

For more general cost functions and networks with or without capacity constraints, we are able to derive a more general result with the help of Han's work, [15]. Han [15] shows the necessary and sufficient condition for the achievable capacity region of a communication network of memoryless channels by exploiting the polymatroidal property of the network capacity function and co-polymatroidal property of the joint conditional entropy functions of the correlated sources. We convert this result to our source graph model and generalize their network topology assumptions as well. [15] models a communication network as a directed graph consisting of a set of sources and a set of relays s.t. there is no incoming edges to any of the source nodes. Replacing min-cut capacity in [15] with cut capacity and because the max-flow min-cut theorem for network flows also applies to an undirected graph, we generalize [15]'s model to any directed/undirected source graph where a source node can have incoming edges.

Before we state Theorem 2, we introduce concept of cut capacity. For any graph G, $\forall M \subseteq V, M^c = V \setminus M$ $(t \in M^c)$ defines a cut, denoted as (M, M^c). Define the set for all possible cuts as Λ. Let $C(M, M^c) = \sum_{v_i \in M, v_j \in M^c} c_{ij}$ be the capacity of cut (M, M^c). $\forall L \subseteq V$, let $\hat{X}_L = \{\hat{X}_i | v_i \in L\}$, $\hat{X}_L^c = \{\hat{X}_i | v_i \in L^c\}$. We also define a feasible set of flows as a set of $f_1, f_2, \ldots, f_N$ that maps each source $\hat{X}_i$ to a

[5] A memoryless channel is one that, given the input of current time slot, the output of current time slot is independent of inputs of previous time slot.

[6] Normally there is an underlying MAC layer to solve the wireless contention problem using techniques like TDMA, FDMA, ALOHA, etc.

flow rate f_i such that there exists a set of commodity flows (fractional allowed) from the sources to the sink such that the capacity constraints and flow conservation are satisfied.

THEOREM 2. *(Generalized version of Theorem 3.1 and Lemma 2.3 in Han1980 [15])* For any source graph G_X (directed or undirected) with an edge capacity set C, there exists a data collection scheme iff

$$H(\hat{X}_M|\hat{X}_M^c) \leq C(M, M^c), \quad \forall (M, M^c) \in \Lambda.$$

When this holds, there exists a SWC scheme and a corresponding nonnegative real vector $R = (r_1, r_2, ..., r_N)$ for the SWC's rates such that for any cut (M, M^c)

$$H(\hat{X}_M|\hat{X}_M^c) \leq \sum_{v_i \in M} r_i \leq C(M, M^c).$$

Furthermore, there exists a set of flows from the source nodes $V \setminus \{t\}$ to the sink t with $f_i = r_i$.

This theorem can be derived by applying the same technique as [15] to our source graph setting. Using Theorem 2 we will derive a general result on the optimal cost of a source graph. However, we first derive a Lemma and introduce some further definitions.

For any source graph G_X and a DCS Υ operating on it, let the average transmission rate from v_i to v_j on edge (v_i, v_j) be $r_{(i,j)}$. For any cut (M, M^c), the average bit rate under Υ that crosses the cut is $r_M(\Upsilon) = \sum_{v_i \in M, v_j \in M^c} r_{(i,j)}$.

LEMMA 1. For any source graph G_X with or without capacity constraints and any DCS Υ operating on it, Υ's data rate across any cut $r_M(\Upsilon)$ satisfies

$$r_M(\Upsilon) \geq H(\hat{X}_M|\hat{X}_M^c)$$

PROOF. We prove this by contradiction using Theorem 2. Assume the lemma is not true, then there exists a G_X and DCS Υ that for some cut (M, M^c) of G, $r_M(\Upsilon) < H(\hat{X}_M|\hat{X}_M^c)$.

The total number of edges from M to M^c on which Υ has traffic is finite and we denote it as l_m. Let

$$\epsilon = \frac{H(\hat{X}_M|\hat{X}_M^c) - r_M(\Upsilon)}{2\, l_m}, \tag{1}$$

then $\epsilon > 0$. Construct a directed graph $G'(V, E')$ with the same vertex set as G. Regardless of whether G is undirected or directed, there is a directed edge (v_i, v_j) in G' iff there is traffic routed from node v_i to v_j by Υ. Also the edge has the same weight w_{ij} as in G. Assign each edge in G' a capacity of $c'_{ij} = r_{(i,j)} + \epsilon$. Then for every edge in G', $c'_{ij} > r_{(i,j)}$. Since we also know all rates below the channel capacity are achievable from the Channel Coding Theorem [8], Υ also makes a valid DCS in G'_X. However, the cut capacity of (M, M^c) in G' is $C'(M, M^c) = \sum_{v_i \in M, v_j \in M^c} (r_{(i,j)} + \epsilon) = r_M(\Upsilon) + l_m \cdot \epsilon$. By (1), we have

$$C'(M, M^c) = \frac{H(\hat{X}_M|\hat{X}_M^c) + r_M(\Upsilon)}{2} < H(\hat{X}_M|\hat{X}_M^c).$$

When the cut capacities of G'_X do not satisfy the iff condition of Theorem 2, there exist no DCSs in G'_X. This contradicts the fact that Υ is a DCS in G'_X. So the assumption is incorrect and the lemma is true. □

Any DCS can be thought of as dividing the data on a link into blocks that each has a fixed transmission rate. Thus the traffic generated by Υ on an edge (v_i, v_j) can be characterized as $[(r^1_{(i,j)}, \tau^1_{(i,j)}), (r^2_{(i,j)}, \tau^2_{(i,j)}), \ldots, (r^{K_{ij}}_{(i,j)}, \tau^{K_{ij}}_{(i,j)})]$, where $r^k_{(i,j)} > 0$ is the rate in bits per second for the kth block and $\tau^k_{(i,j)} > 0$ is the corresponding transmission period. Here $K_{ij} \in \{1, 2, \ldots, +\infty\}$. The average rate by Υ along an edge (v_i, v_j) from v_i to v_j is $r_{(i,j)} = \frac{1}{\sum_{k=1}^{K_{i,j}} \tau^k_{(i,j)}} \sum_{k=1}^{K_{i,j}} r^k_{(i,j)} \cdot \tau^k_{(i,j)}$. For edge e, denote $\tau_e = \sum_{k=1}^{K_e} \tau_e^k$ and $\lambda_e^k = \tau_e^k / \tau_e \in (0, 1]$, then $\sum_{k=1}^{K_e} \lambda_e^k = 1$ and $r_e = \sum_{k=1}^{K_e} r_e^k \cdot \lambda_e^k$.

THEOREM 3. Let G_X be an arbitrary source graph with or without capacity constraints. Let the cost function g be nondecreasing in w and r and convex in r, then the optimal SWC scheme is also optimal over the class of all data collection schemes.

$$\min_{\Upsilon \in \Pi} W_\Upsilon(G_X) = \min_{\Upsilon \in \Pi_{SWC}} W_\Upsilon(G_X).$$

PROOF. The proof consists of showing that, for any data collection scheme Υ, there exists at least one SWC scheme that has a communication cost no greater than that of Υ. The trick is to treat the actual transmission rate generated by Υ on each link as a capacity constraint on that link for the SWC scheme.

As in Lemma 1, construct a directed graph $G'(V, E')$ with the same vertex set as G. Regardless of whether G is undirected or directed, there is a directed edge (v_i, v_j) with unchanged weight w_{ij} in G' iff there is traffic routed from node v_i to v_j by Υ. We treat $\{r_{(i,j)}\}$s as capacities of the directed edges in G' i.e. $c'_{ij} = r_{(i,j)} \leq c_{ij}$ for $(v_i, v_j) \in E'$ and $C'(M, M^c) = r_M(\Upsilon) \leq C(M, M^c)$ for any cut (M, M^c); by Lemma 1 we also have $r_M(\Upsilon) \geq H(\hat{X}_M|\hat{X}_M^c)$. So for any cut (M, M^c),

$$H(\hat{X}_M|\hat{X}_M^c) \leq C'(M, M^c) \leq C(M, M^c). \tag{2}$$

(2) matches the iff condition of Theorem 2. Consequently there exists a SWC scheme with a SWC rate vector $R' = (r'_1, r'_2, \ldots, r'_N)$ that satisfies $H(\hat{X}_M|\hat{X}_M^c) \leq \sum_{v_i \in M} r'_i \leq C'(M, M^c)$ for any cut (M, M^c), and there exists a set of flows $F = (f_1, f_2, \ldots, f_N)$ from $V \setminus \{t\}$ to t in G'. For each v_i, the flow magnitude is $f_i = r'_i$. Since R' is in the Slepian-Wolf achievable rate region [7] and the flow magnitudes satisfy the capacity constraints, the set of flows combined with the channel code and SWC defines a SWC scheme in G', which is automatically a SWC scheme in G since the traffic of any DCS Υ' in G' is upper bounded by G''s capacity which is Υ's data rates, which are further bounded by G's capacity, then Υ''s rates are less than G's capacities correspondingly.[7]

The communication cost per second of this SWC scheme is the cost of the flows $W(F) = \sum_{e \in E'} g(\sum_{i=1}^N f_i(e), w_e)$, where $f_i(e)$ is the flow rate of v_i along edge e. With the capacity constraint, we have $\sum_{i=1}^N f_i(e) \leq c'_e$. Since g is

[7] An alternative way of understanding this is to view the channels in G' as the same channels in G with all or part out of all the time divisions usable.

nondecreasing, we conclude

$$W(F) \leq \sum_{e \in E'} g(c'_e, w_e) \tag{3}$$

On the other hand, by the convexity of function g, the average communication cost per second for Υ satisfies

$$\begin{aligned} W_\Upsilon &\geq \sum_{e \in E'} g(r_e, w_e) \\ &= \sum_{e \in E'} g(c'_e, w_e) \end{aligned}$$

Combined with (3) we have $W(F) \leq W_\Upsilon$. □

4.1 Extension to Broadcast Channels

Previously we ignored the multi-access nature of the wireless medium assuming a lower MAC layer to resolve the confliction. Now we consider the case that includes broadcast channels and show that the previous result remains true even if we take advantage of the Multi-Access nature of wireless channels. We use the same source model as before and a slightly modified communication model to incorporate broadcast channels. We first describe the communication model and then show that the minimum communication costs are the same even with broadcast channels, in other words, broadcasting does not help.

4.1.1 Communication Model

In addition to the independent point to point channels assumed before, we allow nodes to broadcast: a node sends identical data to multiple receiving nodes simultaneously through a broadcast channel. Let $\mathcal{N}(v_i) = \{v_j | (v_i, v_j) \in E\}$ be the neighbor set of node v_i—the set of nodes that v_i can communicate directly to via a point to point channel. Broadcasting here means v_i can send the same copy of data simultaneously at a rate r to any subset of its neighbor set $B \subset \mathcal{N}(v_i)$. The energy cost $g_{i,B}(r)$ of broadcasting is no less than the cost of sending at the same rate from v_i to any of the nodes in B through a point to point channel:

$$g_{i,B}(r) \geq \max_{v_j \in B} g(r, w_{ij}).$$

This assumption is valid for both applications using directional antennas and ones using omni-directional antennas for the point to point channels.[8] Now the capacity constraint is not on the independent links but rather on nodes. Each node v_i has a joint capacity constraint c_i for all of its outgoing channels: broadcasting and non-broadcasting ones together. Thus the broadcasting rate $r \leq c_i$ satisfies the capacity constraint and consumes r of the shared capacity of v_i, equivalently as any of the point to point transmissions does.

4.1.2 Optimality Result

With the modified communication model, now we refer to the previously defined DCS that does not use broadcasting as a "unicast scheme" and still use Π to denote the set of all unicast schemes; we refer to a DCS that uses broadcast as a broadcast enabled DCS and denote the set of all broadcast enabled DCSs as Π_B. We show that any source graph G_X whose nodes are enhanced with this broadcast capability has the same optimal cost as the unicast scheme. We state and prove the following theorem.

THEOREM 4. Let G_X be an arbitrary source graph with or without capacity constraints. Let the cost function g be nondecreasing in w and r and convex in rate r, then the optimal SWC unicast scheme is also optimal over the class of all broadcast enabled data collection schemes.

$$\min_{\Upsilon \in \Pi_B} W_\Upsilon(G_X) = \min_{\Upsilon \in \Pi_{SWC}} W_\Upsilon(G_X).$$

PROOF. (*Sketch*) We prove it by showing that for any broadcast enhanced data collection scheme Υ in G_X, there exists a SWC scheme that has a cost that is no greater than Υ and does not use broadcast. We do so by first showing the broadcast reduced rate r_M^B (counting the duplicate data rate only once) across any cut also satisfies the entropy condition of Lemma 1, then constructing a new source graph G_X^Υ based on G_X and Υ. The first part of the construction is similar to the one used in the proof of Theorem 3. The only difference is that for traffic broadcast by Υ from node v_i to a set of its neighbors B, we add a virtual relaying node (has no sources) $v_{i,B}$ and a set of directed edges that bridges together $v_{i,B}$ and nodes in B. Specifically, we add a directed edge $(v_i, v_{i,B})$ with a capacity equal to the original broadcasting rate r_B and a directed edge from $v_{i,B}$ to each node in B with an infinite capacity. Then because $r_M^B \geq H(\hat{X}_M | \hat{X}_M^c)$ in G_X, it is easy to verify that for any cut (M, M^c) in G_X^Υ, the cut capacity satisfies $C'(M, M^c) \geq H(\hat{X}_M | \hat{X}_M^c)$, by Theorem 3 there exists a SWC scheme Υ_{SWC} in G_X^Υ. If we copy this Υ_{SWC} to G_X by distributing the flow traffic of $v_i \to v_{i,B} \to v_j$ directly as $v_i \to v_j$, by the construction of G_X^Υ, we obtain a non-broadcasting DCS Υ' in G_X. More than that, because g is convex and $g_{i,B}(r_B) \geq \max_{v_j \in B} g(r_B, w_{ij})$ we conclude this DCS Υ' in G_X is also a unicast DCS with a cost no greater than the broadcasting enhanced DCS Υ. □

We have established both the achievable capacity region and the minimum communication cost of a source graph. For collecting multiple correlated sources at a single sink, the optimal SWC scheme is also an optimal data collection scheme over all possible DCS. The result is not obvious because the intermediate nodes are allowed to perform any operations that involve arbitrary couplings of network coding and source coding. In general, there are possible bandwidth benefits applying network coding or broadcasting. While for correlated sources and a single sink, it is first shown here as a corollary of our work that neither network coding nor broadcasting helps either in terms of communication cost or capacity for the most general setting. More than that, our work shows no coding/routing scheme outperforms the SWC schemes. Certainly as we mentioned earlier in Section 1 SWC can hardly be considered a practical code and thus SWC scheme is a theoretical scheme that helps us understand the performance limit of the data collection task.

5. ASYMPTOTICALLY OPTIMAL SCHEME

Given the optimality of the SWC scheme, a natural question to ask is how complex do the nodes' functionalities need

[8] For same type of antennas, directional ones consume less energy than omni-directional ones for point to point communications.

to be in order to achieve the optimal or close to optimal cost, and how close can a practical algorithm approach the optimal performance? As mentioned in Sec. 1, both SWC and EEC have practical limitations. Designing practical SWC schemes has been limited to highly constrained source models.

In this section, we tackle the tradeoff between communication cost and node complexity. We describe a simple data collection scheme, *Hierarchical Difference Broadcasting (HDB)*, for both regular sensor nets on grid points and random deployed sensor nets. Given the high dimensional joint compression complexity of SWC and EEC schemes, HDB does not try to exploit the correlations among all sensor data, but rather tries to leverage off the asymptotically dominant part of the total information redundancy through controlled communications. For some naturally constructed generic spatial correlation models, the neighboring correlation actually dominates the total correlation. We show that HDB is asymptotically optimal for three generic source models that are representative of a large class of real spatial data models.

5.1 General Sensor Grid Model

The grid model for our analysis is based on the general model described in Sec. 2 but with a special spatial deployment strategy. A *sensor grid* is a sensor network where sensors are deployed on a two dimensional square grid. There are total of N sensors indexed as $v_{i,j}$, $1 \leq i,j \leq \sqrt{N}$, $i,j = 1,2,\dots,\sqrt{N}$. The location coordinates of sensor $v_{i,j}$ is $\mu = l_0/2 + (i-1)l_0$, $\nu = l_0/2 + (j-1)l_0$, where l_0 is the grid cell size (the minimum distance between neighboring sensors). W.l.o.g. we assume a *unit grid* where $l_0 = 1$. Each sensor $v_{i,j}$ has a reading $\hat{X}_{i,j}$ that is a discrete random variable. The sensor located in the center of the field also serves as the sink and has a reading $\hat{X}_t$. The sensor readings $\{\hat{X}_{i,j}\}$ are described by a joint distribution. Denote a sample of $\hat{X}$ as $\hat{x}$ and describe the number of bits used to encode $\hat{x}$ by $b(\hat{x})$.

Sensors are able to communicate with each other if they are within a certain range. We assume there are no capacity constraints for the communication links. Let $g(r_e, l_e) = ar_e \cdot {l_e}^{\alpha}$ be the communication cost function [10], where l_e is the Euclidean distance of link e, and a and α are constant parameters with $2 \leq \alpha \leq 4$. W.l.o.g. let $a = 1$. Then the energy cost for transmitting b_e bits is $b_e \cdot {l_e}^{\alpha}$. In this section we focus on the total cost of collecting one field sample at the sink. Since $(l_1 + l_2)^{\alpha} \geq {l_1}^{\alpha} + {l_2}^{\alpha}$, the lowest cost path between any two sensors in a grid always consists of only unit length grid edges. Since there are no capacity constraints, we can equivalently limit the transmissions to be along only such shortest paths without affecting the optimal communication cost. Thus we abstract the sensor network as a grid graph $G(V,E)$, $E = \{(v_{i_1,j_1}, v_{i_2,j_2}) \mid |i_1 - i_2| + |j_1 - j_2| = 1\}$. It is easy to see that the *Manhattan distance*, $\eta_{1,2} = |i_1 - i_2| + |j_1 - j_2|$ is the number of hops of any shortest transmission path between two nodes. We will refer to v_{i_1,j_1} as the $\eta_{1,2}$*-hop-neighbor* of v_{i_2,j_2} and vice versa. When $\eta_{1,2} = 1$, they are each other's *one-hop-neighbor*.

5.2 Hierarchical Difference Broadcasting (HDB)

Before describing HDB scheme, we define a set of hierarchical clusters. W.l.o.g. let $N = 3^{2n}, n = 1,2,\dots$ where the sink is node $v_{\frac{3^n+1}{2},\frac{3^n+1}{2}}$. Let $\Omega_0 = \{v_{\frac{3^n+1}{2},\frac{3^n+1}{2}}\}$. Divide the original $3^n \times 3^n$ grid into 9 clusters, each a subgrid of size $3^{n-1} \times 3^{n-1}$, call the set of these subgrids G_1. Let $\Omega_1 = \{v_{i,j} | i = \frac{3^{n-1}+1}{2} + k_1 \cdot 3^{n-1}, j = \frac{3^{n-1}+1}{2} + k_2 \cdot 3^{n-1}, k_1, k_2 \in \{0,1,2\}\}$ be the set of the 9 center nodes of these subgrids. Similarly divide each subgrid in G_1 into nine subclusters, each a $3^{n-2} \times 3^{n-2}$ subgrid. G_2 is the set of all the subgrids at this level. This can be done recursively, producing a set of subgrids G_k at level k with a set of center nodes $\Omega_k = \{v_{i,j} | i = \frac{3^{n-k}+1}{2} + k_1 \cdot 3^{n-k}, j = \frac{3^{n-k}+1}{2} + k_2 \cdot 3^{n-k}, k_1, k_2 \in \{0,1,\dots,3^k - 1\}\}, \dots, k = 0,1,\dots,n-1$. Let $\Omega_n = V \setminus \Omega_{n-1}$. It is easy to see $\Omega_0 \subset \Omega_1 \subset \Omega_2 \ldots \subset \Omega_{n-1}$ and $\bigcup_{i=0}^{n} \Omega_i = V$.

We design the data collection scheme HDB as following:

Step 1: Sink $t \in \Omega_0$ broadcasts its observation $\hat{x}_t$ using a Self-Delimiting Code (SDC) [23] over a minimum spanning tree to all other $N-1$ nodes in the field. Each sensor updates its reading by subtracting the received value, $\hat{x}_{i,j} \leftarrow \hat{x}_{i,j} - \hat{x}_t$.

```
Step 2: Do $i$ from 1 to $n-1$ {
        Each node $v \in \Omega_i \setminus \Omega_{i-1}$ broadcasts its current reading $\hat{x}_v$ in SDC over a minimum spanning tree to all the nodes in the corresponding subgrid of $G_i$. Receiving sensors update the readings, $\hat{x}_{i,j} \leftarrow \hat{x}_{i,j} - \hat{x}_v$.
} end Do loop
```

Step 3: All sensors other than the sink send their remaining readings $\hat{x}_v$ via shortest paths to the sink. The sink first decodes Ω_1's readings by adding the sink's value to the received $\hat{x}_{\Omega_1}$. Then based on the decoded readings the sink sequentially decodes $\Omega_2, \Omega_3, \dots, \Omega_n$ the readings of all sensors.

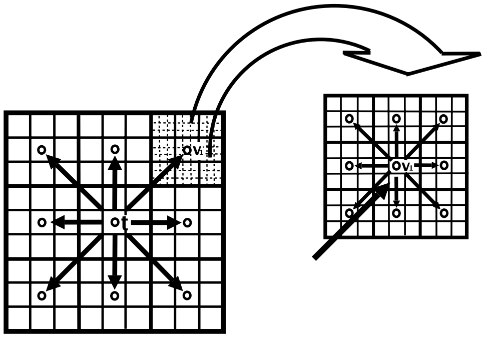

Figure 2: The Hierarchical Broadcasts of HDB

Fig. 2 shows HDB's hierarchical difference broadcasting. When $N \neq 3^{2n}$, $N \in (3^{2n}, 3^{2(n+1)})$ for some n. Expand the grid to size $3^{2(n+1)} \times 3^{2(n+1)}$ with the same center. Divide the expanded grid recursively in the same way, but when a center node of a subgrid is not in the initial grid, choose the closest sensor node from the initial grid. This way we can obtain a sequence of layers $\Omega_0, \Omega_1, \dots, \Omega_n$ for any N.

5.3 Asymptotic Optimality of HDB

Coding in HDB is extremely efficient as it relies only on simple subtractions and Self-Delimiting Codes. SDC is a practical code that encodes $\hat{x}$ into $\Theta(\log \hat{x})$ bits with negligible computation cost [23]. Let the length of the binary representation of $\hat{x}$ be q, SDC sends $q-1$ zeros (q in unary code) followed by the binary representation of $\hat{x}$. For example $\hat{x}=1$ will be coded as '$1'$, $\hat{x}=2$ as '$010'$, 4 as '$00100'$. At the same time, the initialization of HDB is also very simple. Sensors can easily form the series of clusters in a distributed and adaptive fashion. The low coding complexity and high adaptivity of HDB is important for applications of low cost cheap sensors with limited resources.

Lower bound

We apply Theorem 1 to derive a lower bound on the cost of the optimal data collection scheme in a sensor grid network. The result is a lower bound for a general class of correlation models, capturing the topology impact of grid deployment on Distance Entropy.

LEMMA 2. For any sensor grid of size N that has a joint entropy $H(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N) \geq H(\hat{X}_t)+U$, $U>0$, if for some nondecreasing order of the sensor's manhattan distance to the sink $\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N$ ($\hat{X}_1=\hat{X}_t$) we have $H(\hat{X}_i|\hat{X}_{i-1},\hat{X}_{i-2},\ldots,\hat{X}_1) \leq H_o, \forall i>1$ for some $H_o>0$, then the optimal communication cost is lower bounded by $W_{\Upsilon^*} \geq \Theta(U^{\frac{3}{2}}/H_o{}^{\frac{1}{2}})$

PROOF. For a unit grid, $W(p^*_{\hat{X}_j})=\eta_j$ where η_j is the manhattan distance from $\hat{X}_j$ to the sink. By Theorem 1, $H_w(G_{\hat{X}})=\sum_{j=1}^{N} W(p^*_{\hat{X}_j})\times H(\hat{X}_j|\hat{X}_{j-1},...,\hat{X}_1)=\sum_{j=1}^{N}\eta_j\times H(\hat{X}_j|\hat{X}_{j-1},...,\hat{X}_1)$ is the optimal communication cost.

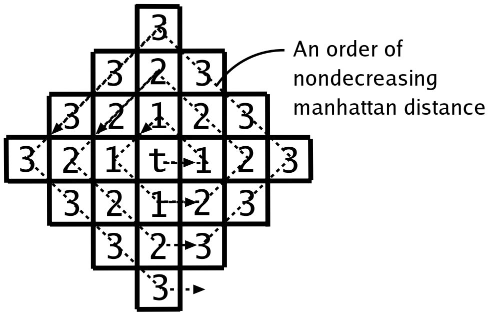

Figure 3: The sink's k-hop-neighbor set layout on the grid

Denote by $S_k=\{v_i|\eta_i=k\}$ the k-hop-neighbor set of the sink. It is easily shown that $|S_k|=4k$ (see Fig. 3). Since we have to collect at least $H(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N)-H(\hat{X}_t)\geq U$ bits at the sink, if we assign H_o bits to each of the sink's neighbors in the order of nondecreasing manhattan distance $(S_1,S_2,\ldots,S_k,\ldots)$ until $N_0=\lfloor U/H_o\rfloor$ sensors are filled, then virtually there is a communication scheme $\tilde{U}$ that collects these $U_0=N_0\cdot H_o\leq U$ bits via shortest paths and it has a cost $W_{\tilde{U}}$.

The optimal SWC scheme Υ^* has to collect $U\geq U_0$ bits from nodes other than the sink and by Theorem 1 the ith sensor is allocated $H(\hat{X}_i|\hat{X}_{i-1},\hat{X}_{i-2},\ldots,\hat{X}_1)\leq H_o$ bits, $\forall i>1$. So for the first N_0 sensors, Υ^* can not allocate to each sensor more bits than $\tilde{U}$ does. If we order the first U_0 bits collected by Υ^* in the order of nondecreasing manhattan distance, the jth bit of Υ^* has a manhattan distance that is no lower than the distance of the jth bit of $\tilde{U}$. Thus $W_{\Upsilon^*}\geq W_{\tilde{U}}$.

Let k^* be the maximum k that satisfies $\sum_{i=1}^{k}|S_i|\leq N_0$. Since $|S_i|=4i$, we get $k^*=\lfloor\frac{\sqrt{2N_0+1}-1}{2}\rfloor$, and $W_{\tilde{U}}\geq H_o\cdot\sum_{i=1}^{k^*}4i\cdot i=\frac{2}{3}H_o k^*(k^*+1)(2k^*+1)$. Applying $k^*=\lfloor\frac{\sqrt{2N_0+1}-1}{2}\rfloor$ and $N_0=\lfloor\frac{U}{H_o}\rfloor$ yields $W_{\tilde{U}}\geq\Theta(U^{\frac{3}{2}}/H_o^{\frac{1}{2}})$. So we get $W_{\Upsilon^*}\geq W_{\tilde{U}}\geq\Theta(U^{\frac{3}{2}}/H_o^{\frac{1}{2}})$. □

Upper bounds

The cost of HDB depends on the spatial correlation among the sensors. In general the correlation exhibits some structure based on the location of the sensors in the graph. For networks in a spatial field, often the correlation structure is a function of its spatial properties. For spatial data, usually the pairwise correlation is a decaying function of the distance. Samples at close by points tend to have higher correlations than those at distant points. This is normally reflected as smaller value differences for closer points, which is especially true for a physical field where the measured phenomena is a result of some micro-scale physical process, e.g. temperature or rainfall distribution. We use three generic source models to model this feature and show that the simple HDB is asymptotically optimal for each of them. Denote the cost of HDB as W_H, then there exists a constant $c>0$ s.t. $W_H/W_{\Upsilon^*}\leq c$.

1) Hard Continuity Field (HCF):

For HCF, $\hat{X}_{i,j}$ is a discrete random variable that has M different possible values. Without loss of generality, we assume the set for the M values is the integer set $\{1,2,\ldots,M\}$. The difference between the samples from any two one-hop-neighbors satisfies a 'hard' continuity constraint as $|\hat{X}_1-\hat{X}_2|\leq d$ for some $d>0$. We assume $d^{\sqrt{N}}\geq\Theta(M)$, this is easy to satisfy when the network scale N is large.

LEMMA 3. If a HCF has a joint entropy $H(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N)\geq\Theta(N\cdot\log d)$, then HDB has an asymptotically optimal communication cost as $\Theta(N\sqrt{N}\log d)$, the same order as the optimal cost $W(\Upsilon^*)$.

PROOF. We first give a lower bound on the optimal cost using Lemma 2 and then demonstrate an upper bound for W_H having the same asymptotic behavior.

$d^{\sqrt{N}}\geq\Theta(M)\Rightarrow H(\hat{X}_t)\leq\log M\leq\Theta(\sqrt{N}\cdot\log d)\Rightarrow H(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N)-H(\hat{X}_t)\geq\Theta(N\log d)-\Theta(\sqrt{N}\log d)=\Theta(N\log d)$. Let $\hat{X}_1,\hat{X}_2,\ldots,X_N$ be a source sequence in an order of nondecreasing manhattan distances to the sink (as shown in Fig. 3) such that each $\hat{X}_i$ other than the sink has a one-hop-neighbor $\hat{X}_{i_1}$ in the sequence with $i_1<i$. So $H(\hat{X}_i|\hat{X}_{i-1},\hat{X}_{i-2},\ldots,\hat{X}_1)\leq H(\hat{X}_i|\hat{X}_{i_1})\leq\log(2d+1)$. Applying Lemma 2 with $U=\Theta(N\log d)$ and $H_o=\log(2d+1)$ yields $W(\Upsilon^*)\geq\Theta(N\sqrt{N}\log d)$.

The same order upper bound can be derived by simple counting technique, please refer to [20]. □

The joint entropy assumption of Lemma 3 is a natural assumption. Here is an example to demonstrate that there exist HCFs with a $\Theta(N\cdot\log d)$ order joint entropy. Consider a case that $M=\frac{3d}{2}$ and a sensor has uniform conditional distribution based on its neighbor readings, then $H(\hat{X}_1,\hat{X}_2,\ldots,\hat{X}_N)=H(\hat{X}_1)+\sum_{i=2}^{N}H(\hat{X}_i|\hat{X}_{i-1},...,\hat{X}_1)\geq\log\frac{3d}{2}+(N-1)\log\frac{d}{2}=\Theta(N\cdot\log d)$.

2) Linear Variance Continuity Field (LVCF): For real sensor data, it is more reasonable to assume a 'soft' continuity constraint rather than the 'hard' one as in HCF. Using the same setting as HCF, a Linear Variance Continuity Field (LVCF) is one where data continuity is modeled as a constraint on the expected data values. We replace the hard continuity constraint with a 'soft' one: any two one-hop-neighbors' reading difference satisfies $\mathbf{E}[(\hat{X}_1 - \hat{X}_2)^2] \leq d^2$, $d > 0$.

LEMMA 4. IF a LVCF has a joint entropy of at least $\Theta(N \cdot \log d)$, and $Var(\hat{X}_t) \leq \Theta(d^{\sqrt{N}})$, then HDB's expected communication cost is asymptotically optimal. The optimal cost $W(\Upsilon^*)$ is lower bounded by $\Theta(N\sqrt{N}\log d)$.

PROOF. We use the same method as Lemma 3 to prove this lemma. The only difference is that here we work with the expected number of bits and apply information theory inequalities.

First by [8]

$$H(\hat{X}) \leq \frac{1}{2}\log\left[(2\pi e)(Var(\hat{X}) + \frac{1}{12})\right] \qquad (4)$$

We have $H(\hat{X}_t) \leq \Theta(\sqrt{N}\log d)$ thus $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) - H(\hat{X}_t) \geq \Theta(N \cdot \log d)$. For the same sequence of nondecreasing manhattan distance to the sink as in Lemma 3,

$$\begin{array}{rcl} H(\hat{X}_i|\hat{X}_{i-1}, \hat{X}_{i-2}, \ldots, \hat{X}_1) & \leq & H(\hat{X}_i|\hat{X}_{i_1}) \\ & = & H(\hat{X}_i - \hat{X}_{i_1}|\hat{X}_{i_1}) \\ & \leq & H(\hat{X}_i - \hat{X}_{i_1}) \qquad (5) \end{array}$$

Also by [8], $H(\hat{X}) \leq \frac{1}{2}\log[(2\pi e)(Var(\hat{X}) + \frac{1}{12})]$, since the variance $Var(\hat{X}_i - \hat{X}_{i_1}) \leq \mathbf{E}[(\hat{X}_i - \hat{X}_{i_1})^2] \leq d^2$, we have $H(\hat{X}_i - \hat{X}_{i_1}) \leq \frac{1}{2}\log[(2\pi e)(d^2 + \frac{1}{12})]$. Applying Lemma 2 with $U = \Theta(N \log d)$ and $H_o = \frac{1}{2}\log[(2\pi e)(d^2 + \frac{1}{12})] = \Theta(\log d)$, get $W(\Upsilon^*) \geq \Theta(N\sqrt{N}\log d)$.

Next we derive the upper bound for W_H. First $\mathbf{E}[(\hat{X}_i - \hat{X}_{i_1})^2] \leq d^2 \Rightarrow \mathbf{E}|\hat{X}_i - \hat{X}_{i_1}| \leq d$. Applying the triangle inequality of an absolute function, any two readings satisfy $\mathbf{E}|\hat{X}_i - \hat{X}_j| \leq \eta_{i,j} \cdot d$. Since a self-delimiting code can compress any $\hat{x}$ into $b(\hat{x}) = 2(\lfloor\log|\hat{x}|\rfloor + 1)$ bits [23] and $\log(x)$ is a concave function, by Jensen's inequality [8],

$$\mathbf{E}(b(\hat{X})) = 2(\mathbf{E}\lfloor\log|\hat{X}|\rfloor + 1) \leq 2(\log \mathbf{E}|\hat{X}| + 1) \qquad (6)$$

So $\mathbf{E}[b(\hat{X}_i - \hat{X}_j)] \leq 2(\log(\eta_{i,j} \cdot d) + 1)$. Also from $Var(\hat{X}_t) \leq \Theta(d^{\sqrt{N}})$ we have $\mathbf{E}\hat{X}_t \leq \Theta(d^{\sqrt{N}})$, by (6) $\mathbf{E}(b(\hat{X}_t)) \leq \Theta(\sqrt{N}\log d)$ and thus replacing the hard bound for the coded bits $b(\hat{x})$ of Lemma 3 with a bound on its expected value and applying the same counting technique, we show $\mathbf{E}[W_H(N)] \leq \Theta[N\sqrt{N}\log d]$. Compared with $W(\Upsilon^*)$ we know that HDB is asymptotically optimal for such LVCF models. □

3) Gaussian-Markov field (GMF): Multivariate Normal (MVN) is an often used model for multivariate distributions. Actually MVN is a subset of the LVCF source model. However, since LVCF is a very general model and HDB is not optimal for any LVCF model, it is worth analyzing the optimality conditions of HDB on MVN. Furthermore, MVN is a good approximation of many applications while being mathematically tractable. Among all the possible spatial correlation gaussian structures, Gaussian-Markov Field (GMF) [9] is a common MVN model to model spatial fields exhibiting the close-points-high-correlation property. Let $X_1, X_2, \ldots, X_N$ be N continuous random values being measured at N different points of a GMF, they follow a joint MVN distribution: $\boldsymbol{N}(\mu, \Sigma)$. Without loss of generality we assume the sources have the same mean $\mu = 0$. $\Sigma = (\sigma_{\mathbf{i},\mathbf{j}})_{N\times N}$ is the covariance matrix with $\sigma_{i,j} = \sigma^2 \cdot e^{-c\eta_{i,j}l_0}$, where $c > 0$ is a constant and σ^2 is the unconditional variance of a source. The correlation between sensors decays exponentially as the distance between them goes up. We use Manhattan distance instead of Euclidean distance because the former is much more tractable yet is a good approximation of the latter as our simulation suggests.

Let $\gamma = e^{-c \cdot l_0}$, $\gamma_{i,j} = \gamma^{\eta_{i,j}}$, then $\gamma_{i,j}$ is the correlation coefficient between sensor i and j and the covariance matrix can be written as $\Sigma = \sigma^2 \cdot (\gamma_{\mathbf{i},\mathbf{j}})_{N\times N}$. Notice $0 < \gamma_{i,j} < 1$ for any $i \neq j$ and $\gamma_{i,i} = 1$ for any i. This avoids the trivial case of $\gamma_{i,j} \equiv 1$ when all readings are fully dependent on each other, in which case the sink's reading is exactly the same as that of any other sensor and there is no need for communication. The other trivial case is when we have independent readings, $\gamma_{i,j} = 0$ for all $i \neq j$, then the problem reduces to a single source coding problem with no need for distributed coding.

Each sensor's reading $\hat{X}_i$ is a quantized version of X_i where each sensor uses the same type uniform scalar quantizer. When the quantization precision is high and thus the step size Δ is small, by [8], the entropy of $\hat{X}$ is approximately the differential entropy of X minus $\log\Delta$. We assume a high resolution quantizer is used and $H(\hat{X}_j) = h(X_j) - \log\Delta$, where $h(X)$ is the differential entropy of X. For any k sources, $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_k) = h(X_1, X_2, \ldots, X_k) - k\log\Delta$.

LEMMA 5. For any GMF on a k-dimensional hyper-cube grid of $N = m^k$ nodes, the field's joint entropy

$$H(\hat{X}_1, \ldots, \hat{X}_N) = \frac{1}{2}\log\left((2\pi e)^N\sigma^{2N}(1-\gamma^2)^{km^{k-1}(m-1)}\right) - N\log\Delta$$

The proof can be found in [20].

COROLLARY 1.

$$H(\hat{X}_1, \ldots, \hat{X}_N) \geq \frac{1}{2}\log\left((2\pi e)^N\sigma^{2N}(1-\gamma^2)^{kN}\right) - N\log\Delta$$

PROOF. Just apply the fact $\gamma \in (0,1)$ to Lemma (5), get $det(Q_k) \geq (1-\gamma^2)^{kN}$, by Lemma 5 we prove the corollary. Note that particularly for a 2D grid we have $det(\Sigma_2) = \sigma^{2N}det(Q_2) \geq \sigma^{2N}(1-\gamma^2)^{2N}$. □

THEOREM 5. For any two dimensional GMF that has $\gamma \leq 0.86539$ and $\frac{1}{2}\log\frac{2\pi e\sigma^2}{\Delta} \leq \sqrt{N}H_o$, where $H_o = \log\frac{\sqrt{2\pi e\sigma^2}(1-\gamma^2)}{\Delta}$. The expected communication cost of HDB is asymptotically optimal. The optimal cost $W(\Upsilon^*)$ is lower bounded by $\Theta(N\sqrt{N}H_o)$.

PROOF. The proof uses the same type of technique as the case for HCF and LVCF, only now we work on the entropy of gaussian variables.

By Corollary 1,

$$H(\hat{X}_1, \ldots, \hat{X}_N) \geq NH_o$$

also

$$H(\hat{X}_t) = \frac{1}{2}\log\frac{2\pi e\sigma^2}{\Delta} \leq \sqrt{N}H_o$$

so $U{=}H(\hat{X}_1,\ldots,\hat{X}_N){-}H(\hat{X}_t){\geq}\Theta(NH_o)$.

W.l.o.g. let $1,2,\ldots,N$ be the same type of nondecreasing manhattan distance order as in the proofs for HCF and LVCF, since entropy is a lower bound for any codes, the expected coded bits of SDC is larger than the corresponding entropy: $H(\hat{X}_i - \hat{X}_{i_1}) \leq \mathbf{E}[b(\hat{X}_i - \hat{X}_{i_1})]$. By (5),

$$H(\hat{X}_i|\hat{X}_{i-1},\ldots,\hat{X}_1) \leq \ \mathbf{E}[b(\hat{X}_i - \hat{X}_{i_1})] \tag{7}$$

By [27], $\mathbf{E}[(X_i - X_j)^2] = 2\sigma^2(1-\gamma_{i,j})$, then $\mathbf{E}|X_i - X_j| \leq \sqrt{2\sigma^2(1-\gamma_{i,j})} \Rightarrow \mathbf{E}|\hat{X}_i - \hat{X}_j| \leq \mathbf{E}|X_i - X_j|/\Delta + 1 \leq \sqrt{2\sigma^2(1-\gamma_{i,j})}/\Delta + 1$, by (6), we have $\mathbf{E}b[(\hat{X}_i - \hat{X}_j)] \leq 2[\log(\sqrt{2\sigma^2(1-\gamma_{i,j})}/\Delta{+}1){+}1]$. Also since $\Delta \ll \sqrt{\sigma^2(1-\gamma)}$ (high resolution quantizer), we get

$$\mathbf{E}[b(\hat{X}_i - \hat{X}_j)] \leq (1+\epsilon)\log[8\sigma^2(1-\gamma_{i,j})/\Delta^2] \tag{8}$$

$\epsilon > 0$ is a small constant. Particularly $\mathbf{E}[b(\hat{X}_i - \hat{X}_{i_1})] \leq (1+\epsilon)\log[8\sigma^2(1-\gamma)/\Delta^2]$. When $\gamma \leq 0.86539$, $\log[8\sigma^2(1-\gamma)/\Delta^2] < 2H_o$, thus combined with (7), we have $2(1+\epsilon)H_o > H(\hat{X}_i|\hat{X}_{i-1},\ldots,\hat{X}_1)$. Applying U and H_o to Lemma 2 yields $W(\Upsilon^*) > \Theta(N\sqrt{N}H_o)$.

At the same time, it follows that $\mathbf{E}[b(\hat{X}_i{-}\hat{X}_j)] \leq \log[8\sigma^2(1{-}\gamma_{i,j})]$. Since for any $\gamma \in (0,1)$,

$$(1-\gamma_{i,j}) = (1-\gamma^{\eta_{i,j}}) \leq \eta_{i,j}(1-\gamma)$$

we have $\mathbf{E}[b(\hat{X}_i - \hat{X}_j)] < 2H_o + \log\eta_{i,j}$. Applying the same counting technique as in Lemma 3 yields the following upper bound on HDB cost $W_H < \Theta(N\sqrt{N}H_o) + \Theta(N\sqrt{N}) = \Theta(N\sqrt{N}H_o)$. □

From Theorem 5 we conclude that for large portion of a GMF grids without too high correlations between the nodes, HDB is asymptotically optimal. This is intuitively right because as the correlation coefficient $\gamma \to 1$ (either $c \to 0$ or $l_0 \to 0$), the field approaches the trivial case of completely dependent with no need for communications. However, as long as the field is not anywhere close to this, for a large range HDB remains asymptotically optimal:$\gamma \leq 0.86539$ as opposed to the full possible range of $(0,1)$. Applying Theorem 5 and the same technique, HDB's asymptotic optimality can be generalized to high dimensional GMF grid as well as *Gaussian Uniform Field(GUF)* which is a multivariate gaussian field with $\gamma_{i,j} = \gamma$ for any two nodes. Due to space limitations, we do not present the details here.

5.4 Non-grid Models

Grid deployment is a good approximation for a large class of sensor applications where sensors can be deployed in a regular manner. We also extend the techniques and insights developed from the grid case to the random deployment case.

1) Deployment Model:
Assume N sensors are uniformly and independently distributed in a two-dimensional geographical region G. Under this assumption, for large N the sensor locations can be approximated or modelled as a two-dimensional Poisson Point Process (PPP). Let the average sensor density be $\rho = N/|G|$(number of sensors per unit area, $|\cdot|$ is the area function). Let the number of sensors in a region A be $N(A)$; $N(A)$ follows a Poisson distribution with parameter $\rho|A|$,

$$P(N(A) = k) = \frac{e^{-\rho|A|}(\rho|A|)^k}{k!}$$

The rate of the Poisson process λ is just the density $\lambda = \rho$.

There is a single sink in the region to collect all the readings. Each sensor v_i's Euclidean distance to the sink is l_i. Let $l_G = \frac{1}{N}\sum_{i=1}^{N} l_i$ be the field's average distance to the sink.

2) Communication Cost Model:
We use the same linearly separable communication cost function $g = l_e^\alpha \cdot b_e$ as the grid case. Let $l_o = \sqrt{\frac{|G|}{N}} = \frac{1}{\sqrt{\rho}}$ be the average neighbor distance of the sensors. Assume the minimum communication cost per bit from a sensor v_i to the sink t is $W(p_i^*) = \frac{l_i}{l_o} \cdot {l_o}^\alpha = l_i \cdot l_o^{\alpha-1}$. The minimum per bit cost between any two sensors v_i, v_j are $W(p_{i,j}^*) = l_{i,j}l_o^{\alpha-1}$ This is a close approximation when N is large, the majority of the sensors are many hops away from the sink.

3) Source Model:
For the random deployment case, instead of having a one hop continuity constraint, we have to define a continuity constraint depending on the distance continuously because now the one hop distance is not a fixed value as in grid case. The constraint is modeled appropriately according to the sensed field being HCF, LVCF or GMF. Here we use LVCF as an example and it is easy to adjust for the other two. Assume for any two sensors v_i and v_j that has a Euclidean distance $l_{i,j}$, their reading difference satisfies $\mathbf{E}[(\hat{X}_i - \hat{X}_j)^2] \leq f(l_{i,j}) > 0$ where f is any nondecreasing function that maps the distance between two sensors to an upper bound of their reading differences. We call this model a Poisson LVCF field or $PLVCF$.

4) Protocol–RHDB:
We refer to the modified HDB scheme as Random deployed HDB (RHDB). The modifications are simple: Instead of dividing the sensors into clusters directly, now we divide the geometric region uniformly into nine square shape subregions, sensors in the same square are clustered together, then further divide each cluster into sub-regions of $\frac{1}{9}$size. Division stops when it is the size of a region $3c_\epsilon l_o \times 3c_\epsilon l_o$ (c_ϵ is some constant) or there are no sensors in it. Choose the sensor closest to the geometric center of the subregion as its cluster head. Then we have the following Theorem.

THEOREM 6. For a PLVCF field, if there exists a pair of constants $\epsilon > 0$ and $0 < \delta < 1$ such that the field has a joint entropy $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) \geq \Theta[N{\cdot}\log f(c_\epsilon l_o)]$, $Var(\hat{X}_t) \leq \Theta\left[\left(f(c_\epsilon l_o)\right)^{\sqrt{N}}\right]$ where $c_\epsilon = \sqrt{\frac{(1+\epsilon)}{(1-\delta)(\frac{2\pi}{3}-\frac{\sqrt{3}}{2})}}$, also $\log f(x)$ is a concave function and $\eta_G = \Theta(\sqrt{N}l_o)$, then RHDB is asymptotically optimal for the expected total communication cost w.h.p.(with high probability). And w.h.p. the optimal cost $W(\Upsilon^*)$ is lower bounded by $\Theta[N\sqrt{N}l_o^\alpha \log f(c_\epsilon l_o)]$.

The proof of Theorem (6) uses several randomized techniques including Chernoff bound, Chebyshev's Inequality [25] and is quite involved, we refer the reader to [20].

6. PERFORMANCE EVALUATION

In this section we evaluate our HDB scheme using two data sets: a 2D radar reflectivity data set generated by a weather simulating/forecasting tool ARPS [32], and a synthetic data set generated by a Gaussian Markov Field model. We ignore the radar data's temporal correlations and focus on HDB's performance on reducing its spatial redundancy. Nevertheless, we point out that it is not hard to generalize HDB to deal with temporal correlation as well. For both data sets, we use a 2D square grid and place the sink at the center.

The 2D radar data set is formatted as a 43 × 43 grid covering a region of about 41 square kilometers. We evaluate our HDB algorithm assuming the data is collected by a corresponding sensor grid network with unit cell size (the absolute cell length does not influence the ratio of one hop cost per bit for different schemes). Since we expect most spatial data to share some form of the continuity feature described by LVCF defined earlier, HDB's performance trend based on radar data should apply to spatial data collected in other ad-hoc sensor networks. Since we are interested in the large network performance of HDB, we also ignore the asymptotically diminishing routing&scheduling overhead to focus on the asymptotically dominating part of the cost.

For the radar data, Figure 4 shows that the communication cost ratio between our HDB scheme and a non-coding raw scheme (RAW). RAW sends some fixed number bits from each node to the sink following a shortest path. Compared to RAW, HDB saves approximately 2/3 of the communication cost. 316 different snapshots of the field are used to estimate the average cost for HDB and RAW. Since there are no good spatial models for the radar data and non-parametric multidimensional entropy estimation is difficult [4], we are unable to compare HDB's cost to the distance entropy (optimal). Nevertheless, our simulation demonstrates HDB's improvement and trend vs. RAW.

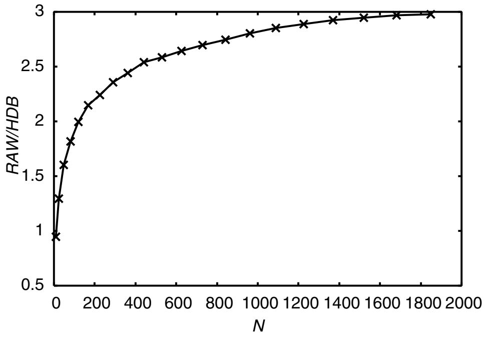

Figure 4: The communication cost ratio of RAW to HDB for radar data set

In order to further evaluate the performance of HDB, by comparing it to RAW and distance entropy (OPT), we turn to a synthetic data set that is generated by Gaussian Markov Field model. We set the mean of the field as 250 and variance $\sigma^2 = 5625$. The 2D grid is composed of N sensors ($\sqrt{N} \times \sqrt{N}$) with grid cell size one. $\gamma_1 = 0.9999$ is set by the correlation coefficient between one hop neighbor nodes estimated from the radar data.

We are interested in the asymptotic behavior as $N \to \infty$ for fixed density network topology in which the cell size is constant (we set the size to one). Note that in a fixed density topology, since the cell size is fixed, the correlation coefficient between one hop neighbors γ_1 is fixed as well.

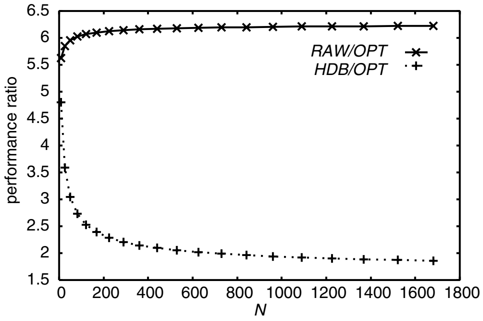

Figure 5: The communication cost ratio of RAW to OPT, and HDB to OPT for data set generated by Gaussian Markov Field.

Figure 5 shows the performance ratios of HDB/OPT and RAW/OPT for a fixed density network as N increases. We see that the ratios HDB/OPT and RAW/OPT both approach a constant as N increases. This result matches well the asymptotically optimal result given by Theorem 5. Note that in the simulation we use Euclidean distance instead of Manhattan distance for the correlation decaying which is more realistic. When the curves converge, the constant ratio between RAW, OPT and HDB can be explained as follows: RAW sends all source data without any coding; OPT reduces data using full knowledge of data correlation; HDB reduces data based on only local data correlation. The constant ratio between RAW, OPT and HDB shows the ratio of communication cost when using no correlation information, complete correlation information, and local correlation information. In this example, by exploiting local neighboring correlation, HDB approaches about one third of the communication cost of RAW, and achieves about 1.5 times of the OPT communication cost. Note in the radar case, HDB also approaches one third of RAW's cost and both cases have the same one hop correlation coefficient.

From Figure 5, we also see that HDB/OPT decreases as N increases. This is because in HDB, the communication cost (step 1 and step 2 of HDB) to obtain the side information from neighboring nodes is asymptotically dominated by the communication cost of sending the remaining data after compression ($\Theta(N^{3/2})$) (step 3 of HDB): as $N \to \infty$, the communication cost for obtaining neighborhood side information becomes more and more negligible compared to that for sending the data. Thus, the ratio of HDB/OPT decreases accordingly.

For fixed density topology, our about results show that by exploiting local data correlation, our HDB algorithm substantially reduces the communication cost, and achieves the asymptotically optimality with small constant factor.

7. CONCLUSION AND FUTURE WORK

Our main contributions are summarized as follows:

- We show that, for a single sink data network, the Slepian-Wolf Code and Commodity Flow Routing can achieve the minimum communication cost even if arbitrary coding/routing scheme is allowed.

- We propose a new metric *distance entropy*, a generalization of entropy, to characterize the "spatial" information distribution in a weighted graph (abstraction of the communication network).

- We design a simple and effective algorithm that exploits local data correlation to achieve an asymptotically optimal performance for some generic classes of source models.

- We further evaluate our algorithm with 2D radar reflectivity data and a simulated Gaussian Markov Field model. Result demonstrates that our HDB algorithm substantially reduces the communication cost and also achieves the asymptotically optimality with small constant factor.

Future research can be pursued in the following directions: further evaluating HDB algorithm on more data correlation models; implementing our distributed algorithm in an energy constrained wireless sensor network; solving the general multi-sink lossy correlated data collection problem, this is quite challenging.

8. ACKNOWLEDGMENTS

We thank Jim Kurose for suggesting CASA's radar data and his helpful comments, thank Eric Lyons and Ming Xue for providing the radar data and explaining the data format. This material is based upon work supported by the Engineering Research Centers Program under award EEC-0313747 001, and the National Science Foundation under NSF Faculty Early Award CCR-0133664, NSF Award ITR-0325726, and NSF Research Infrastructure Award EIA-0080119. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

9. REFERENCES

[1] M. Adler. Collecting correlated information from a sensor network. In *SODA*, 2005.

[2] M. Adler, N. Harvey, K. Jain, R. Kleinberg, and A. Lehman. On the capacity of information networks. In *SODA*, 2006.

[3] R. Ahlswede, N. Cai, S. Li, and R. Yeung. Network information flow. *IEEE Transactions on Information Theory*, 46(4):1204–1216, July 2000.

[4] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy, 2002.

[5] C. Chong and S. Kumar. Sensor networks: Evolution, opportunities, and challenges. In *IEEE Symposium on Foundations of Computer Science*, pages 1247–1256, 2003.

[6] J. Chou, D. Petrovic, and K. Ramchandran. A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks. In *INFOCOM*, 2003.

[7] T. Cover. A proof of the data compression theorem of slepian and wolf for ergodic sources. *IEEE Transactions on Information Theory*, 22:226–228, March 1975.

[8] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[9] N. Cressie. *Statistics for Spatial Data.* John Wiley & Sons, 1993.

[10] R. Cristescu, B. Beferull-Lozano, and M. Vetterli. On network correlated data gathering. In *IEEE INFOCOM*, 2004.

[11] R. Cristescu, B. Beferull-Lozano, and M. Vetterli. Networked Slepian-Wolf: Theory, Algorithms and Scaling Laws. *IEEE Transactions on Information Theory*, 2005.

[12] D. Estrin, D. Culler, K. Pister, and G. Sukhatme. Connecting the physical world with pervasive networks. *IEEE Pervasive Computing*, 1(1):59–69, 2002.

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

[14] A. Goel and D. Estrin. Simultaneous optimization for concave costs: Single sink aggregation or single source buy-at-bulk. In *SODA*, 2003.

[15] T. Han. Slepian-wolf-cover theorem for networks of channels. *Information and Control*, 47(1):67–83, 1980.

[16] T. Ho, M. Médard, M. Effros, and R. Koetter. Network coding for correlated sources. In *CISS*, 2004.

[17] A. Jindal and K. Psounis. Modeling spatially-correlated sensor network data. In *SECON*, 2004.

[18] A. Lehman and E. Lehman. Complexity classification of network information flow problems. In *SODA*, 2004.

[19] Z. Li, B. Li, D. Jiang, and L. Lau. On achieving optimal throughput with network coding. In *INFOCOM*, 2005.

[20] J. Liu, M. Adler, and D. Towsley. Collecting correlated data through a network with minimum cost: Distance entropy and a practical asymptotically optimal design. Technical Report CS TR05-64, University of Massachusetts, Amherst, 2005.

[21] Y. Liu, D. Towsley, J. Weng, and D. Goeckel. An information theoretic approach to network trace compression. Technical Report CS TR05-03, University of Massachusetts, Amherst, 2005.

[22] D. Lun, Médard, T. Ho, and R. Koetter. Network coding with a cost criterion. Technical Report P-2584, MIT, 2004.

[23] D. Mackay. *Information theory, inferrence, and learning algorithms.* John Wiley & Sons, 2004.

[24] D. Marco, E. Duarte-Melo, M. Liu, and D. L. Neuhoff. On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data, 2003.

[25] R. Motwani and P. Raghavan. *Randomized algorithms.* Cambridge University Press, 1995.

[26] S. Pattem, B. Krishnamachari, and R. Govindan. The impact of spatial correlation on routing with compression in wireless sensor networks. In *IPSN*, 2004.

[27] D. W. R. Johnson. *Applied Multivariate Statistical Analysis.* Prentice Hall, 2002.

[28] A. Ramamoorthy, K. Jain, P. Chou, and M. Effros. Separating distributed source coding from network coding. In *Allerton*, 2004.

[29] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, IT-19(4):471–480, July 1973.

[30] Y. Wu, V. Stankovic, Z. Xiong, and S. Kung. On practical design for joint distributed source and network coding. In *NETCOD*, 2005.

[31] Z. Xiong, A. Liveris, and S. Cheng. Distributed source coding for sensor networks. In *IEEE Signal Processing Magazine*, pages 522–533, September 2004.

[32] M. Xue, K. K. Droegemeier, V. Wong, A. Shapiro, K. Brewster, F. Carr, D. Weber, Y. Liu, and D. Wang. The advanced regional prediction system (arps)a multi-scale nonhydrostatic atmospheric simulation and prediction tool. part ii: Model physics and applications. *Meteorol. Atmos. Phys. 76*, 76, 2001.