

Collecting Correlated Data through a Network with Minimum Cost: Distance Entropy and a Practical Order Optimal Design

Junning Liu*, Micah Adler*, Don Towsley* and Chun Zhang†

*Dept. of Computer Science, University of Massachusetts, Amherst, MA

Email: {liujn,micah, towsley}@cs.umass.edu

†IBM T.J. Watson Research Center, Hawthorne, NY, USA

Email: czhang1@us.ibm.com

Abstract—We study the communication cost of collecting correlated data at a sink over a network. This is motivated by the correlated data gathering in sensor networks and the internet. In many energy-constrained wireless sensor networks, nodes cooperatively forward correlated sensed data to data sinks. For the internet case, huge amount of correlated network traces need to be collected for management purposes. In order to reduce the communication cost (e.g. overall energy, bandwidth, etc.) used for data collection, previous works have focused on specific coding schemes, such as Slepian-Wolf Code (SWC) or Explicit Entropy Code (EEC). However, the minimum communication cost under arbitrary coding/routing schemes has not yet been characterized. In this article, 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 generic correlation structures, the communication cost achieved by this scheme is within a constant factor of the distance entropy, i.e., it is order optimal. This order optimality is shown for two deployment strategies: a 2D grid regular network and a 2D Poisson process random network. 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 ~ 1.8) of the distance entropy.

Index Terms—istributed Source Coding, Joint Coding and Routing, Communication Cost Minimization, Network Codingistributed Source Coding, Joint Coding and Routing, Communication Cost Minimization, Network Coding

I. INTRODUCTION

With the development of ubiquitous sensing and computing networks, we are approaching to a digital environment where information is generated (e.g. sensing), computed (in-network computation), and gathered everywhere in the network. Correlated data gathering is a fundamental task in such information processing networks. We study the communication cost of collecting correlated data at a sink over a network. As for now, our problem is more directly motivated by two typical applications: Sensor networks and the Internet.

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. [1][2]). Low cost sensors are distributed in a region to collect measurements of field points. Each sensor is capable of sensing, storing, computing and transmitting. The measurements at different sites are usually correlated and all of them need to be reconstructed at a base station or sink for storage or further processing (e.g. inference). A common characteristic of these networks, however, is that they are energy-constrained and the communication energy cost is a dominant factor that drains the battery, thus the communication cost 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 [3][4][5]. When coding is restricted to Explicit Entropy Code (EEC)¹ [3], Cristescu et al. [3] shows that choosing the optimal routes is a NP-hard problem; Pattem et al. [4] proposes a heuristic algorithm to minimize the communication cost assuming a simplified source model. When coding is restricted to a Slepian-Wolf Code (SWC) [6]² and Commodity Flow Routing (CFR) is used, Cristescu et al.'s work [3] [7] 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.

Another typical application of correlated data gathering is to collect the network traces on the internet. Tons of flow statistics need to be transferred to some collecting point for further processing. These data are highly correlated and the same challenge remains as in the sensor network case, only the cost metric here is some network resource consumption such as bandwidth.

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 [8] [9].

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) [10], 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 (the conditional entropies), 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 [11]. Second, SWC [6] 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 [9] [12]; the long temporal coding block also requires considerable memory on each node. In another mostly studied coding scheme, EEC, it was shown in [3] that it is NP-hard to minimize the communication cost. EEC needs to learn and store the conditional distributions for joint encoding. EEC's coding complexity is typically high and in order to reduce coding complexity it requires a larger memory size to store pre-computed values. Another disadvantage of EEC is that the scheduling and coordination of the data flows normally induces large communication/computation costs and delays because coding and routing are not independent. For example, some sensors may need to wait for other sensors' data to do joint encoding. Finally, for both SWC and EEC, when the source model is dynamic and time varying, the cost for retraining is large in terms of delay and resource consumption. In addition to these two schemes, practical schemes that have been proposed generally assumes source models that are limited and not representative enough for general realistic data. For many, the number of sources is limited [13] [14] [15], some are limited to binary sources [15] [16], and some are limited to Gaussian sources [17]. Few adaptive/universal DSC or general coding technique for general source models have been developed [9], especially for low complexity practical ones. Recently, [12] proposes an interactive approach for arbitrary correlation models, their focus is the total number of bits sent from the nodes and there is no notion of network topology or cost function.

Practically, we design a simple and effective algorithm that achieves an order 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

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

²SWC 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.

algorithm and demonstrate that it is order 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 order optimality. In Section 6, we evaluate the performance of our algorithm through simulations. Finally, we conclude and discuss future work in Section 7.

II. BACKGROUND AND RELATED WORK

A. 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 routing³.

There are two aspects of a joint coding/routing problem, network combinatorics and information theory: as [18] summarizes, Combinatorics is concerned with packing problems (e.g. flows) that are constrained by the graph structure. It grows 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 [3]. However, when arbitrary 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.

B. Related work

There has been much research on Distributed Source Coding (DSC) and NC. A thorough review of DSC can be found in [9] where it is claimed that there are still few practical DSC schemes for general source models. [20] 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 [20]. For NC, if there is just a single sink in addition to independent sources, there is no need for Network Coding. [21] shows that traditional routing where data is treated as commodity flows suffices to solve the data collection problem for such networks. [22] 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. [23] shows that random linear network coding suffices for the network coding of correlated sources. [24] 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; [25] studies the problem of network coding with a cost criterion. For minimum cost correlated data gathering, [5] 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. [4] 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 [5]: the joint entropy is a linear function of the number of sources. Thus their discovery of optimal clustering size is consistent with [5]'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, [26] studies the scaling problem of

³Note that these operations could be inseparable.

a large number sensors deployed in a Gaussian Markov Field (GMF) by comparing the per node capacity and node data rate asymptotically. [7] 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 order optimal under these conditions. [27] 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, [4]'s experience equation learned from real rainfall spatial data verifies the validity of the total entropy assumption in our generic source modelling. [28] 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.

III. MODEL FORMULATION

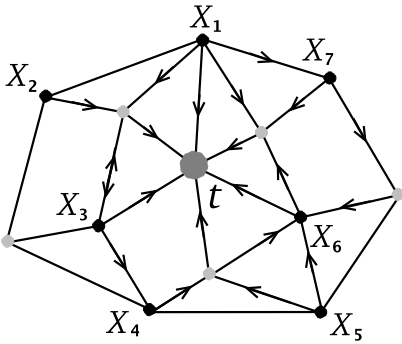


Fig. 1. A layout of the general problem of gathering correlated data through a network

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. quantization). The joint source vector $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_N\}$ is characterized by a joint probability distribution $p(\hat{X}_1 = \hat{x}_1, \dots, \hat{X}_N = \hat{x}_N) = p(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N)$. 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 [7].⁴ In practice, if a node uses a fixed transmission power (as the normal mode of 802.11), then the communication cost rate is a linear function of the data rate. i.e., $g(r_e, w_e) = r_e \cdot w_e$ [7]. 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 first 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 $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N$ at the sink t near losslessly - decode losslessly with zero or an arbitrarily small error probability [29]. 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

⁴The 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.

⁵A 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.

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

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 $W_{\Upsilon^*} = \min_{\Upsilon \in \Pi} W_\Upsilon$.

IV. 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, \dots, v_N are in a nondecreasing order of shortest path weight to the sink, i.e., $W_1 \leq W_2 \leq \dots \leq W_N$.

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

$$H_w(G_X) = \sum_{j=1}^N W_j \times H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_0)$$

Also \hat{X}_0 denotes \hat{X}_t the source located at sink t which can be null.

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 $r_j = H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_1)$ achieves the cost of $H_w(G_X)$. Thus

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

Proof: 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.

Let G_{XC} be a reduced source graph from G_X s.t. there are no capacity constraints in G_{XC} . Then any DCS in G_X will also be a DCS in G_{XC} and a lower bound of minimum DCS cost in G_{XC} is also a lower bound of the one in G_X . Thus below we base our analysis on source graphs without capacity constraints.

Since the nodes $t, 1, 2, \dots, N$ is in a nondecreasing order of the shortest path weight to the sink, i.e. satisfy $W_0 = 0 \leq W_1 \leq W_2 \leq \dots \leq W_N$. Note that \hat{X}_t can be null. For convention we also denote \hat{X}_t as \hat{X}_0 .

We first solve the case when W_j s are distinct values then extend it to the general case.

For each $W_j \neq 0$, define a triple partition of the vertex set V as $(M_j^<, M_j, M_j^>)$ where $M_j^< = \{i | W_i < W_j\}$, $M_j^> = \{i | W_i > W_j\}$. Define $M_j = \{i | W_i = W_j\}$ as a boundary set in between. Then we do the following procedure:

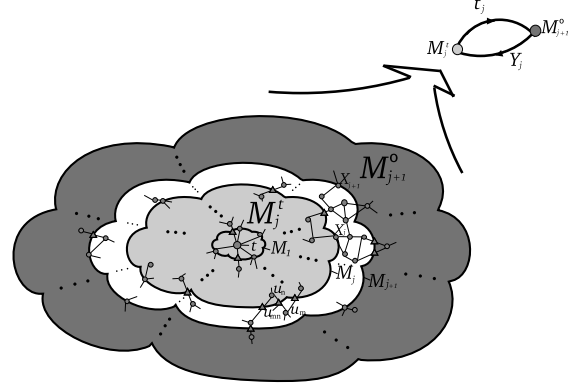


Fig. 2. Construction of the virtual graph

As shown in Fig. 2, from $j = 1$ to N , for any edge $e = (v_m, v_n) \in E$, if $v_m \in M_j^>$ and $v_n \in M_j^<$, we create a virtual node v_{mn} (identified by a triangle node in Fig. 2), $V \leftarrow V \cup \{v_{mn}\}$. Replace the edge $e = (v_m, v_n)$ with two new edges $e_1 = (v_{mn}, v_n)$ and $e_2 = (v_m, v_{mn})$ with weights as $w_{e_1} = W_j - W(v_n)$, $w_{e_2} = w_e - w_{e_1}$, $E \leftarrow (E \cup \{e_1, e_2\}) \setminus \{e\}$, where $W(v_n)$ is the shortest path weight of node v_n . The resulting w_{e_2} is guaranteed to be positive otherwise contradicted with $v_m \in M_j^>$. Now $W(v_{mn}) = W_j$. Update M_j as $M_j \leftarrow M_j \cup \{v_{mn}\}$. Each j th set of such updates on G satisfies that for any data collection scheme in the original graph before the updates, there exists a corresponding data collection scheme in the resulting graph with the same communication cost, and vice versa. Thus we can just evaluate the communication cost of data collection schemes on the resulting graph.

The resulting graph has another property: There are no edges going from $M_j^>$ to $M_j^<$ for any $1 \leq j \leq N$. In other words, for any node $v \in M_j^>$, any path connects v with a node in $M_j^<$ have to reach some node in M_j first.

Then from $j = 1$ to $N - 1$, we define subsets of V as $M_j^t = M_j^< \cup M_j$ and $M_{j+1}^o = M_{j+1} \cup M_{j+1}^>$. Since the sink is in M_j^t and $V \subseteq M_j^t \cup M_{j+1}^o$, M_j^t and M_{j+1}^o partition the source set into two parts, where $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_j$ are in M_j^t and the rest are in M_{j+1}^o . The transmissions between M_{j+1}^o and M_j^t can be viewed as a two party Alice/Bob communications. We map M_j^t to a sink v_A with all the sources that lie in M_j^t , namely $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_j$, map M_{j+1}^o to v_B with all the sources that lie in M_{j+1}^o , namely $\hat{X}_{j+1}, \hat{X}_{j+2}, \dots, \hat{X}_N$. There are two directed edges between v_A and v_B . v_A sends the bits from M_j^t to M_{j+1}^o to v_B while v_B sends to v_A all the bits from M_{j+1}^o to M_j^t . v_A needs to decode $\hat{X}_{j+1}, \hat{X}_{j+2}, \dots, \hat{X}_N$ near losslessly. It is easy to see that any DCS in the original source graph has a

corresponding DCS in this simplified source graph where the traffics between v_B and v_A are the same as those between M_{j+1}^o and M_j^t . Consider a further reduced two party distributed source coding scenario where v'_A has side information $\hat{X}_1, \hat{X}_2 \dots, \hat{X}_j$ and needs to decode \hat{X} losslessly, v'_B has the whole source vector \hat{X} . Now v'_A does not send anything to v'_B but v'_B sends all the bits $v_B \rightarrow v_A$ to v'_A . Since the bits sent from v_A to v_B are ultimately functions of \hat{X} , we know any DCS in the old source graph corresponds to a DCS in this new source graph G'_X . Thus if v'_B has to send to v'_A at least b bits for any DCS in G'_X , then there has to be at least b bits transmitted from M_{j+1}^o to M_j^t for any DCS in G_X . Now that $\hat{X}_1, \hat{X}_2 \dots, \hat{X}_j$ is described perfectly at v'_A , by the results of Slepian-Wolf coding [29] v'_B has to send at least $H(\hat{X}|\hat{X}_1, \hat{X}_2 \dots, \hat{X}_j) = H(\hat{X}_{j+1}, \hat{X}_{j+2} \dots, \hat{X}_N|\hat{X}_1, \hat{X}_2 \dots, \hat{X}_j)$ bits per sample. The same requirement holds for the original traffic from M_{j+1}^o to M_j^t before the reduction. Since there are no capacity constraints here, we actually consider noiseless channels. Thus there has to be $B_j = H(\hat{X}_j, \hat{X}_{j+1} \dots, \hat{X}_N|\hat{X}_1, \hat{X}_2, \dots, \hat{X}_{j-1})$ bits transmitted from M_j to M_{j-1} for every $2 \leq j \leq N$. Let $M_0^t = \{t\}$, let $M_1^o = M_1 \cup M_1^>$, using the same argument we get $B_1 = H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N)$. In the case that there is a source \hat{X}_0 located at the sink t , $B_1 = H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N|\hat{X}_0)$.

By the new graph's property, a path from any node $v_o \in M_{j+1}^o$ to any node $v_t \in M_j^t$ has to first reach some node $v_m \in M_{j+1}$ then some node $v_n \in M_j$. By definition $W(v_m) = W_{j+1}$, $W(v_n) = W_j$. For any path p_{mn} from v_m to v_n , the path's weight has to satisfy the triangle property $W(p_{mn}) \geq W(v_m) - W(v_n) = W_{j+1} - W_j$, otherwise there exists a path from v_m to t with a weight less than W_{j+1} and contradict with M_{j+1} 's definition.

For $2 \leq j \leq N$, define P_j as the set of all the paths that go from a node in M_j to a node in M_{j-1} and intersect with M_j only the starting node with M_{j-1} only the ending node. Define P_1 as all the paths that go from M_1 to the sink t and only intersect with them once. Then any bit that goes from M_j to M_{j-1} has to be transmitted along some path in P_j , even it may goes along some path that intersects with M_j and/or M_{j-1} more than once, it has to contain a sub-path that belong to P_j .

Then we form a subset of the edge set E as $E^p = E_1 \cup E_2 \cup \dots \cup E_N$ where E_j contains all the edges that belong to paths in P_j . It is easy to verify that these E_j 's are disjoint sets.

We define the communication cost on each edge set E_j as $W(E_j)$. Since B_j bits have to be crossed for any cut in between M_j and M_{j-1} , by the Max-flow Min-cut theorem [30], there exist a set of flows F_j from M_j to M_{j-1} . Each flow of F_j has to go along some path in P_j at some part of its trajectory. Let $W(F_j)$ be the part of F_j 's cost that is consumed in E_j , then $W(F_j) \leq W(E_j)$. Since $g = w \cdot r$, from all above we know that for each bit to be

transmitted from some $v_o \in M_j^o$ to some $v_t \in M_{j-1}^t$, there has to be $1 \times W(v_{mn}) \geq W_j - W_{j-1}$ cost spent in edge set E_j , so $W(F_j) \geq B_j \times (W_j - W_{j-1})$.

Thus for any data collection scheme Υ ,

$$\begin{aligned} W_\Upsilon &\geq \sum_{j=1}^N W(E_j) \\ &\geq \sum_{j=1}^N W(F_j) \\ &\geq \sum_{j=1}^N B_j \times (W_j - W_{j-1}) \\ &= \sum_{j=1}^N H(\hat{X}_j, \dots, \hat{X}_N|\hat{X}_0, \dots, \hat{X}_{j-1}) \\ &\quad \times (W_j - W_{j-1}) \\ &= \sum_{j=1}^N H(\hat{X}_j|\hat{X}_0, \dots, \hat{X}_{j-1}) \times W_j \\ &= H_w(G_X) \end{aligned}$$

Since $H_w(G_X)$ is exactly the cost of the SWC scheme that has $r_i = H(\hat{X}_j|\hat{X}_0, \dots, \hat{X}_{j-1})$ and routes along shortest paths to the sink, we have $\min_{r \in \Pi_{SWC-SP}} W_\Upsilon(G_X) = H_w(G_X)$.

When $W(p_{\hat{X}_i}^*)$'s are not distinct values, the subscript of M_j enumerates from 1 to the number of distinct shortest path weights. Now the order between the sources with the same shortest path weight does not matter and we get the same formula as before. ■

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. This differs from [3]'s proof and is a more general result. [3] 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.

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, [31]. Han [31] 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. [31] 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 [31] with cut capacity and because the max-flow min-cut theorem for network flows also applies to an undirected graph, we generalize [31]'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, \dots, f_N that maps each source \hat{X}_i to a 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 [31]) *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 $R = (r_1, r_2, \dots, r_N)$ 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 [31] 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

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

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 [29], 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 $[(r_{(i,j)}^1, \tau_{(i,j)}^1), (r_{(i,j)}^2, \tau_{(i,j)}^2), \dots, (r_{(i,j)}^{K_{ij}}, \tau_{(i,j)}^{K_{ij}})]$, where $r_{(i,j)}^k > 0$ is the rate in bits per second for the k th block and $\tau_{(i,j)}^k > 0$ is the corresponding transmission period. Here $K_{ij} \in \{1, 2, \dots, +\infty\}$. The average rate by Υ along an edge (v_i, v_j) from v_i to v_j is $r_{(i,j)} = \frac{1}{\sum_{k=1}^{K_{ij}} \tau_{(i,j)}^k} \sum_{k=1}^{K_{ij}} r_{(i,j)}^k \cdot \tau_{(i,j)}^k$. 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) ,

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

(2) matches the iff condition of Theorem 2. Consequently there exists a SWC scheme with a SWC rate vector $R' = (r'_1, r'_2, \dots, r'_N)$ 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 $F = (f_1, f_2, \dots, f_N)$ 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 [32] 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 nondecreasing, we conclude

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

On the other hand, the average communication cost per second for Υ is

$$\begin{aligned} W_\Upsilon &= \sum_{e \in E'} \frac{1}{\tau_e} \sum_{k=1}^{K_e} g(r_e^k, w_e) * \tau_e^k \\ &= \sum_{e \in E'} \left[\sum_{k=1}^{K_e} g(r_e^k, w_e) * \lambda_e^k \right] \end{aligned}$$

By the convexity of function g , we have

$$\begin{aligned} W_\Upsilon &\geq \sum_{e \in E'} g\left(\sum_{k=1}^{K_e} r_e^k * \lambda_e^k, w_e\right) \\ &= \sum_{e \in E'} g(r_e, w_e) \\ &= \sum_{e \in E'} g(c'_e, w_e) \end{aligned}$$

Combined with (3) we have $W(F) \leq W_\Upsilon$. Thus for any data collection scheme Υ there exists a SWC scheme with a communication cost no bigger than Υ . As a result, the optimal SWC scheme is also optimal among all the possible data collection schemes. ■

When samples are temporally correlated, we group and encode them in temporal blocks. Our results can be

⁷An 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.

extended to hold by replacing the $H(\hat{X}_M | \hat{X}_M^c)$ with the entropy rate

$$\begin{aligned} H_\infty(\hat{X}_M | \hat{X}_M^c) &= \\ \lim_{m \rightarrow \infty} \frac{1}{m} H(\hat{X}_M(\tau_1), \dots, \hat{X}_M(\tau_m) | \hat{X}_M^c(\tau_1), \dots, \hat{X}_M^c(\tau_m)) \end{aligned}$$

A. 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.

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 $r \leq c_i$ also follows the capacity constraint and consumes r of the shared capacity of v_i , equivalently as any of the point to point transmissions does.

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

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

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: 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.

For any broadcasting enhanced DCS Υ for G_X , we define r_M^B as the broadcasting reduced data rate across any cut (M, M^c) , that is, if a broadcasting sender is in M , and there is at least one receiver across the cut in M^c , the data rate across the cut of this broadcasting will only be counted as the broadcasting rate r without double counting the multiple receiving rates.

We first show that for any cut (M, M^c) in G_X , $r_M^B \geq H(\hat{X}_M | \hat{X}_{M^c}^c)$. We do this by shrinking G_X to a simple source graph of two nodes, s_M and t_M , where node s_M has source \hat{X}_M and t_M has source $\hat{X}_{M^c}^c$. A pair of infinite capacity channels connect s_M and t_M . We emulate all the traffic between M and M^c under Υ now between s_M and t_M in the new source graph except that for the broadcasting traffic we only emulate one copy of it between s_M and t_M . Since all coding/routing operations within M or M^c under Υ are now all achievable internal coding operations inside s_M or t_M in the new source graph, any DCS Υ in G_X corresponds to a DCS in the new source graph with a rate from s_M to t_M as $r_{s_M} = r_M^B$. By Lemma 1 $r_{s_M} \geq H(\hat{X}_M | \hat{X}_{M^c}^c)$ thus $r_M^B \geq H(\hat{X}_M | \hat{X}_{M^c}^c)$ for any cut (M, M^c) .

We then construct 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 $r_M^B \geq H(\hat{X}_M | \hat{X}_{M^c}^c)$ in G_X , it is easy to verify that for any cut (M, M^c) in G_X^Υ , the cut capacity satisfies $C'(M, M^c) \geq H(\hat{X}_M | \hat{X}_{M^c}^c)$, 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 $v_i \rightarrow v_{i,B} \rightarrow v_j$ directly as $v_i \rightarrow v_j$, 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 [33] [34] [35]. 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.

V. ORDER OPTIMAL SCHEME

Given the optimality of the SWC scheme, a natural question to ask is how complex do the nodes' functionalities need 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 order optimal for three generic source models that are representative of a large class of real spatial data models.

A. 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 [3], 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*.

B. 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 $v_{\frac{3^{2n+1}}{2}, \frac{3^{2n+1}}{2}}$. Let $\Omega_0 = \{v_{\frac{3^{2n+1}}{2}, \frac{3^{2n+1}}{2}}\}$. 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 $\Omega_1 = \{v_{i,j} \mid i = \frac{3^{n-1}+1}{2} + k_1 \cdot 3^{n-1}, j = \frac{3^{n-1}+1}{2} + k_2 \cdot 3^{n-1}, k_1, k_2 \in \{0, 1, 2\}\}$ 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 $\Omega_k = \{v_{i,j} \mid i = \frac{3^{n-k}+1}{2} + k_1 \cdot 3^{n-k}, j = \frac{3^{n-k}+1}{2} + k_2 \cdot 3^{n-k}, k_1, k_2 \in \{0, 1, \dots, 3^k - 1\}\}, \dots, k = 0, 1, \dots, n-1$. Let $\Omega_n = V \setminus \Omega_{n-1}$. It is easy to see $\Omega_0 \subset \Omega_1 \subset \Omega_2 \dots \subset \Omega_{n-1}$ 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) [36] 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}_{Ω_1} . Then based on the decoded readings the sink sequentially decodes $\Omega_2, \Omega_3, \dots, \Omega_n$ the readings of all sensors.

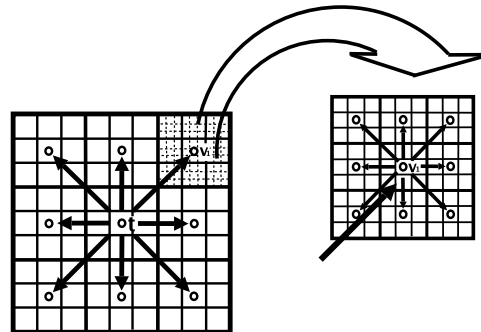


Fig. 3. The hierarchical broadcasts of HDB

Fig. 3 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 .

C. order 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 [36]. 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 $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq H(\hat{X}_t) + U$, $U > 0$, if for some nondecreasing order of the sensor's manhattan distance to the sink $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N$ ($\hat{X}_1 = \hat{X}_t$) we have $H(\hat{X}_i | \hat{X}_{i-1}, \hat{X}_{i-2}, \dots, \hat{X}_1) \leq H_o, \forall i > 1$ 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, $H_w(G_{\hat{X}}) = \sum_{j=1}^N W(p_{\hat{X}_j}^*) \times H(\hat{X}_j|\hat{X}_{j-1}, \dots, \hat{X}_1) = \sum_{j=1}^N \eta_j \times H(\hat{X}_j|\hat{X}_{j-1}, \dots, \hat{X}_1)$ is the optimal communication cost.

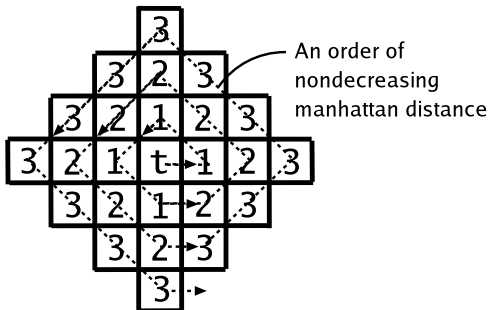


Fig. 4. The sink's k -hop-neighbor set layout on the grid

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. 4). Since we have to collect at least $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) - H(\hat{X}_t) \geq U$ bits at the sink, if we assign H_o bits to each of the sink's neighbors in the order of nondecreasing manhattan distance ($S_1, S_2, \dots, S_k, \dots$) 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 i th sensor is allocated $H(\hat{X}_i|\hat{X}_{i-1}, \hat{X}_{i-2}, \dots, \hat{X}_1) \leq H_o$ 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 j th bit of Υ^* has a manhattan distance that is no lower than the distance of the j th 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 order 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 $\{1, 2, \dots, M\}$. 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 $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq \Theta(N \cdot \log d)$, then HDB has an order 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.

$d\sqrt{N} \geq \Theta(M) \Rightarrow H(\hat{X}_t) \leq \log M \leq \Theta(\sqrt{N} \cdot \log d) \Rightarrow H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) - H(\hat{X}_t) \geq \Theta(N \log d) - \Theta(\sqrt{N} \log d) = \Theta(N \log d)$. Let $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N$ be a source sequence in an order of nondecreasing manhattan distances to the sink (as shown in Fig. 4) 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 $H(\hat{X}_i|\hat{X}_{i-1}, \hat{X}_{i-2}, \dots, \hat{X}_1) \leq H(\hat{X}_i|\hat{X}_{i_1}) \leq \log(2d+1)$. Applying Lemma 2 with $U = \Theta(N \log d)$ and $H_o = \log(2d+1)$ yields $W(\Upsilon^*) \geq \Theta(N\sqrt{N} \log d)$.

Now we derive a same order upper bound for HDB's cost using simple counting techniques. $\forall N$, let $\tilde{N} = \min_{3^{2n} \geq N} 3^{2n}$, n is a positive integer. $N \leq \tilde{N} < 9N$, so HDB's energy cost for \tilde{N} sensors under the same model and continuity constraint is at least as large as the cost for N sensors, $W_H(\tilde{N}) \geq W_H(N)$. We next derive an upper bound for $W_H(\tilde{N})$.

W_H consists of two parts, the broadcast cost W_B and the collection cost W_C . There are n broadcast rounds, $W_B(\tilde{N}) = \sum_{i=1}^n W_B^i$. The first round broadcasts the sink's reading throughout the network. We code \hat{x}_t into $\Theta(\log \hat{x}_t)$ bits, $\hat{x}_t \leq M \Rightarrow \log \hat{x}_t \leq \log M \leq \Theta(\sqrt{N} \cdot \log d)$. Also because the broadcast needs exactly one hop transmission to cover each sensor in the minimum spanning tree, $W_B^1 \leq \Theta[\sqrt{N} \log d \cdot (\tilde{N} - 1)]$. The second round is to broadcast the readings of sensors in $\Omega_1 \setminus \Omega_0$ within G_1 . From $|a+b| \leq |a| + |b|$ we know $|\hat{x}_i - \hat{x}_j| \leq \eta_{i,j} \cdot d$, the reading difference between any two sensors is bounded by their manhattan distance times the one hop difference bound. So after the first round's reading updates, four

of the sensors in $\Omega_1 \setminus \Omega_0$ have $|\hat{x}| \leq 2 \cdot 3^{n-1} \cdot d$, the other four have $|\hat{x}| \leq 3^{n-1} \cdot d$. Using a self-delimiting code, we code any integer $K > 0$ into $b(K) = 2 \log K + 1$ bits [36]. So any $|\hat{x}| \leq K$, $b(\hat{x}) \leq 2(\log K + 1)$ bits and $b(\hat{x}_i - \hat{x}_j) \leq 2(\log(\eta_{i,j} \cdot d) + 1)$ bits. Then four readings of sensors in $\Omega_1 \setminus \Omega_0$ are coded into no more than $2(\log(2 \cdot 3^{n-1} \cdot d) + 1)$ bits each, the other four readings of $\Omega_1 \setminus \Omega_0$ are coded into no more than $2(\log(3^{n-1} \cdot d) + 1)$ bits each. $W_B^2 \leq [3^{2(n-1)} - 1]8[\log(2 \cdot 3^{n-1} \cdot d) + \log(3^{n-1} \cdot d) + 2]$. Similarly the 3rd round broadcasts the readings of sensors in $\Omega_2 \setminus \Omega_1$ within G_2 and $W_B^3 \leq [3^{2(n-2)} - 1]9 * 8[\log(2 \cdot 3^{n-2} \cdot d) + \log(3^{n-2} \cdot d) + 2]$. In general, $W_B^i \leq [3^{2(n-i+1)} - 1]9^{i-2} * 8[\log(2 \cdot 3^{n-i+1} \cdot d) + \log(3^{n-i+1} \cdot d) + 2]$, $W_B^n \leq [3^{2(n-n+1)} - 1]9^{n-2} * 8[\log(2 \cdot 3^{n-n+1} \cdot d) + \log(3^{n-n+1} \cdot d) + 2]$.

So $\sum_{i=2}^n W_B^i \leq \sum_{i=2}^n [3^{2(n-i+1)} - 1] \cdot 9^{i-2} \cdot 8[\log(2 \cdot 3^{n-i+1} \cdot d) + \log(3^{n-i+1} \cdot d) + 2] = \Theta(\tilde{N} \log \tilde{N}) + \Theta(\tilde{N} \log^2 \tilde{N}) + \Theta(\tilde{N} \log \tilde{N} \log d) + \Theta(\log \tilde{N}) - \Theta(\tilde{N}) - \Theta(\tilde{N} \log d)$. Combined with W_B^1 we have $W_B(\tilde{N}) \leq \Theta[\tilde{N} \sqrt{\tilde{N}} \log d]$

The data transmission cost after n rounds of broadcasting is composed of nine parts—collecting the nine subgrids of G_1 . The four corner subgrids have a larger bound for cost than the other four. Let the cost for collecting the upper-left corner subgrid be W_C^u , then $W_C(\tilde{N}) \leq 9W_C^u$. Let the total number of bits sent to the sink from the upper-left subgrid be B_{n-1} . Note that the bits distribution in the subgrid is symmetric to the subgrid center $v_{\frac{3^{n-1}+1}{2}, \frac{3^{n-1}+1}{2}}$: the center's i -hop-neighbors have the same upper bound for the remaining bits. $B_{n-1} \leq 2(\log 2 \cdot 3^{n-1} \cdot d) + \sum_{i=1}^{n-1} 8 \frac{3^{2(n-1)}}{9^i} (\log 2 \cdot 3^{i-1} \cdot d + \log \cdot 3^{i-1} \cdot d + 2) = 2 \cdot 9^{n-1} \log d + (9^{n-1} + 1) \log 2 + 2 \cdot 9^{n-1} + \frac{9^{n-1}-1}{4} \log 3 = \Theta(\tilde{N} \log d) + \Theta(\tilde{N})$. By symmetry $W_C^u = (\Theta(\tilde{N} \log d) + \Theta(\tilde{N})) \cdot \eta_a$. The manhattan distance from the center to the sink is $\eta_a = 2 \cdot 3^{n-1} = \Theta(\sqrt{\tilde{N}})$, thus $W_C(\tilde{N}) \leq 9W_C^u \leq \Theta(\sqrt{\tilde{N}}) \cdot \Theta(\tilde{N} \log d) = \Theta(\tilde{N} \sqrt{\tilde{N}} \log d)$.

From above and $\tilde{N} < 9N$ we have $W_H(\tilde{N}) = W_B(\tilde{N}) + W_C(\tilde{N}) \leq \Theta(\tilde{N} \sqrt{\tilde{N}} \log d) = \Theta[N \sqrt{N} \log d]$, $\Rightarrow W_H(N) \leq \Theta[N \sqrt{N} \log d]$. Thus compared with $W(\Upsilon^*)$ we know HDB is order optimal for such HCF models. ■

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 $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) = H(\hat{X}_1) + \sum_{i=2}^N H(\hat{X}_i | \hat{X}_{i-1}, \dots, \hat{X}_1) \geq \log \frac{3d}{2} + (N-1) \log \frac{d}{2} = \Theta(N \cdot \log d)$.

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 $\mathbb{E}[(\hat{X}_1 - \hat{X}_2)^2] \leq d^2$, $d > 0$.

Lemma 4: *IF a LVCF has a joint entropy of at least $\Theta(N \cdot \log d)$, and $\text{Var}(\hat{X}_t) \leq \Theta(d\sqrt{N})$, then HDB’s expected communication cost is order 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 [29]

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

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

$$\begin{aligned} H(\hat{X}_i | \hat{X}_{i-1}, \hat{X}_{i-2}, \dots, \hat{X}_1) &\leq H(\hat{X}_i | \hat{X}_{i_1}) \\ &= H(\hat{X}_i - \hat{X}_{i_1} | \hat{X}_{i_1}) \\ &\leq H(\hat{X}_i - \hat{X}_{i_1}) \end{aligned} \quad (5)$$

Also by [29], $H(\hat{X}) \leq \frac{1}{2} \log [(2\pi e) (\text{Var}(\hat{X}) + \frac{1}{12})]$, since the variance $\text{Var}(\hat{X}_i - \hat{X}_{i_1}) \leq \mathbb{E}[(\hat{X}_i - \hat{X}_{i_1})^2] \leq d^2$, 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 $\mathbb{E}[(\hat{X}_i - \hat{X}_{i_1})^2] \leq d^2 \Rightarrow \mathbb{E}|\hat{X}_i - \hat{X}_{i_1}| \leq d$. Applying the triangle inequality of an absolute function, any two readings satisfy $\mathbb{E}|\hat{X}_i - \hat{X}_j| \leq \eta_{i,j} \cdot d$. Since a self-delimiting code can compress any \hat{x} into $b(\hat{x}) = 2(\lceil \log |\hat{x}| \rceil + 1)$ bits [36] and $\log(x)$ is a concave function, by Jensen’s inequality [29],

$$\mathbb{E}(b(\hat{X})) = 2(\mathbb{E}[\log |\hat{X}|] + 1) \leq 2(\log \mathbb{E}|\hat{X}| + 1) \quad (6)$$

So $\mathbb{E}[b(\hat{X}_i - \hat{X}_j)] \leq 2(\log(\eta_{i,j} \cdot d) + 1)$. Also from $\text{Var}(\hat{X}_t) \leq \Theta(d\sqrt{N})$ we have $\mathbb{E}\hat{X}_t \leq \Theta(d\sqrt{N})$, by (6) $\mathbb{E}(b(\hat{X}_t)) \leq \Theta(\sqrt{N} \log d)$ 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 $\mathbb{E}[W_H(N)] \leq \Theta[N\sqrt{N} \log d]$. Compared with $W(\Upsilon^*)$ we know that HDB is order 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) [37] is a common MVN model to model spatial fields exhibiting the close-points-high-correlation property. Let X_1, X_2, \dots, X_N be N continuous random values being measured at N different points of a GMF, they follow a joint MVN distribution: $N(\mu, \Sigma)$. Without loss of generality we assume the sources have the same mean $\mu = 0$. $\Sigma = (\sigma_{i,j})_{N \times N}$ is the covariance matrix with $\sigma_{i,j} = \sigma^2 \cdot e^{-cn_{i,j}l_0}$, 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 suggests⁹.

Let $\gamma = e^{-cl_0}$, $\gamma_{i,j} = \gamma^{n_{i,j}}$, then $\gamma_{i,j}$ is the correlation coefficient between sensor i and j and the covariance matrix can be written as $\Sigma = \sigma^2 \cdot (\gamma_{i,j})_{N \times N}$. 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 [29], 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, $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_k) = h(X_1, X_2, \dots, X_k) - k \log \Delta$.

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

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

Proof: By [29],

$$H(\hat{X}_1, \dots, \hat{X}_N) = \frac{1}{2} \log \left((2\pi e)^N \det \Sigma_k \right) - N \log \Delta$$

where Σ_k is the k dimensional grid's covariance matrix. Order the N sensors in a dimension-recursive enumerating order. For example, the 1D order is sequentially enumerating the nodes; the 2D order is enumerate the nodes line by line, and use the 1D order within each line:

⁹The joint entropy ratio between GMF with manhattan distance and GMF with Euclidean distance is bounded in a range close to 1.

$v_{1,1}, \dots, v_{1,m}, v_{2,1}, \dots, v_{2,m}, \dots, v_{m,1}, \dots, v_{m,m}$. Then define matrix Q_k as the correlation coefficient matrix.

$$Q_1 = (q_{i,j}^1)_{m \times m} \begin{pmatrix} 1 & \gamma & \gamma^2 & \dots & \gamma^{m-1} \\ \gamma & 1 & \gamma & \dots & \gamma^{m-2} \\ \dots & \dots & \dots & \dots & \dots \\ \gamma^{m-2} & \gamma^{m-1} & \dots & 1 & \gamma \\ \gamma^{m-1} & \gamma^{m-2} & \dots & \gamma & 1 \end{pmatrix}$$

with the entry as a $m^{k-1} \times m^{k-1}$ submatrix $q_{i,j}^1 = \gamma^{|i-j|}$. Inductively,

$$Q_k = \begin{pmatrix} Q_{k-1} & \gamma Q_{k-1} & \gamma^2 Q_{k-1} & \dots & \gamma^{m-1} Q_{k-1} \\ \gamma Q_{k-1} & Q_{k-1} & \gamma Q_{k-1} & \dots & \gamma^{m-2} Q_{k-1} \\ \dots & \dots & \dots & \dots & \dots \\ \gamma^{m-2} Q_{k-1} & \gamma^{m-1} Q_{k-1} & \dots & Q_{k-1} & \gamma Q_{k-1} \\ \gamma^{m-1} Q_{k-1} & \gamma^{m-2} Q_{k-1} & \dots & \gamma Q_{k-1} & Q_{k-1} \end{pmatrix}$$

is a partitioned matrix with the entry as $q_{i,j}^k = \gamma^{|i-j|} Q_{k-1}$.

Q_1 is a Toeplitz matrix and $\det(Q_1) = (1-\gamma^2)^{m-1}$ [38]. For Q_k , from top to bottom, each row subtracts the next row times γ , we can obtain a lower triangular matrix and thus get

$$\det(Q_k) = [(1-\gamma^2)^{m^{k-1}} \det(Q_{k-1})]^{m-1} \cdot \det(Q_{k-1})$$

Inductively, we prove the

$$\det(Q_k) = (1-\gamma^2)^{km^{k-1}(m-1)} \quad (7)$$

Combined with $\Sigma_k = \sigma^2 \cdot Q_k$ we get the entropy result. ■ To the best of our knowledge, Lemma 5 is the first characterization of the joint entropy of a general grid GMF. The closest work is [7]'s 1D grid result. Also (7) is the first equation for the determinant of this general type of matrices.

Corollary 1:

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

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 order 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,

$$H(\hat{X}_1, \dots, \hat{X}_N) \geq N H_o$$

also

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

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

W.l.o.g. let $1, 2, \dots, N$ 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: $H(\hat{X}_i - \hat{X}_{i_1}) \leq \mathbb{E}[b(\hat{X}_i - \hat{X}_{i_1})]$. By (5),

$$H(\hat{X}_i | \hat{X}_{i-1}, \dots, \hat{X}_1) \leq \mathbb{E}[b(\hat{X}_i - \hat{X}_{i_1})] \quad (8)$$

By [39], $\mathbb{E}[(X_i - X_j)^2] = 2\sigma^2(1 - \gamma_{i,j})$, then $\mathbb{E}|X_i - X_j| \leq \sqrt{2\sigma^2(1 - \gamma_{i,j})} \Rightarrow \mathbb{E}|\hat{X}_i - \hat{X}_j| \leq \mathbb{E}|X_i - X_j|/\Delta + 1 \leq \sqrt{2\sigma^2(1 - \gamma_{i,j})}/\Delta + 1$, by (6), we have $\mathbb{E}b[(\hat{X}_i - \hat{X}_j)] \leq 2[\log(\sqrt{2\sigma^2(1 - \gamma_{i,j})}/\Delta + 1) + 1]$. Also since $\Delta \ll \sqrt{\sigma^2(1 - \gamma)}$ (high resolution quantizer), we get

$$\mathbb{E}[b(\hat{X}_i - \hat{X}_j)] \leq (1 + \epsilon) \log[8\sigma^2(1 - \gamma_{i,j})/\Delta^2] \quad (9)$$

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

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

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

we have $\mathbb{E}[b(\hat{X}_i - \hat{X}_j)] < 2H_o + \log \eta_{i,j}$. 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 order optimal. This is intuitively right because as the correlation coefficient $\gamma \rightarrow 1$ (either $c \rightarrow 0$ or $l_0 \rightarrow 0$), 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 order optimal: $\gamma \leq 0.86539$ as opposed to the full possible range of $(0, 1)$. Applying Theorem 5 and the same technique, HDB's order 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.

Non-Square Grid

All the results for square grids can be extended to non-square-shape regions as long as the region *weight center* (equally weighted average location of all sensors) has a distance $\eta_G = \Theta(\sqrt{N})$ to the sink. HDB still uniformly and hierarchically divides the region into cluster series of geometrically decreasing sizes.

Corollary 2: For a unit grid of arbitrary shapes with $\eta_G = \Theta(\sqrt{N})$, if $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq \Theta(N \cdot \log d)$ and $H(\hat{X}_i) \leq \Theta(\sqrt{N} \log d)$, the expected total communication

cost of HDB is order optimal. And the optimal cost $W(\Upsilon^*)$ is lower bounded by $\Theta(N\sqrt{N} \log d)$.

We give a sketch of the proof that uses the same techniques as previous results. For the lower bound, we construct a infinite large virtual grid that is of the same l_0 and contains the original sensor grid as a subgraph. When we fill in the sensors each with H_o bits as close to the sink as possible, the virtual grid is used instead of the original grid. This gives the same lower bound as before.

Since the sensors with the same bits upper bound are uniformly distributed in the field, the average distance from them to the sink is the same as the region weight center's distance to the sink, thus the expected data collecting cost is upper bounded by the sum of all the bits upper bounds times the weight center's distance to the sink $\Theta(\sqrt{N})$, which gives the same order upper bound as square region, also because the broadcasting cost is independent of the region shape, we have the same order upper bound as in the square region case.

D. 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 $\mathbb{E}[(\hat{X}_i - \hat{X}_j)^2] \leq f(l_{i,j}) > 0$ 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 sub-regions, 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 $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq \Theta[N \cdot \log f(c_\epsilon l_o)]$,

$$\text{Var}(\hat{X}_t) \leq \Theta \left[\left(f(c_\epsilon l_o) \right)^{\sqrt{N}} \right] \text{ where } c_\epsilon = \sqrt{\frac{(1+\epsilon)}{(1-\delta)(\frac{2\pi}{3} - \frac{\sqrt{3}}{2})}}$$

also $\log f(x)$ is a concave function and $\eta_G = \Theta(\sqrt{N}l_o)$, then RHDB is order 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 $\Theta[N\sqrt{N}l_o^\alpha \log f(c_\epsilon l_o)]$.

Proof: a) The lower bound

$H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq \theta[N \cdot \log f(c_\epsilon l_o)]$, so there exists a constant c , for N large enough, we have $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq cN \cdot \log f(c_\epsilon l_o)$

Call a disk centered at t with radius r as $\varphi_t(r)$. Let $r_0 = \sqrt{\frac{c}{2\pi}}\sqrt{N}l_o$. Call the region of $\varphi_t(r_0)$ as G_0 . Let $\hat{X}_{G_0} = \{\hat{X}_k | l_k \leq r_0\}$ be the set of readings of sensors in G_0 . Let $\partial_0 = G \setminus G_0$, denote the readings in ∂_0 as $H(\hat{X}_{\partial_0})$. Denote the number of sensors in G_0 as N_0 . N_0 is a random variable. We first prove that w.h.p. $N_0 \in ((1-\delta)cN/2, (1+\delta)cN/2)$.

N_0 can be viewed as the sum of N independent identical Poisson trials: $N_0 = \sum_{k=1}^N Y_k$, each Y is either 1 or 0 and has the same probability distribution as $\text{Pr}(Y_k = 1) = \frac{|G_0|}{|G|}$, corresponding to the probability for the k th sensor being deployed in region $\varphi_t(r_0)$. So $\mathbb{E}(N_0) = N \frac{|G_0|}{|G|} = \pi r_0^2 \rho = cN/2$. Using Chernoff bound [40], $\text{Pr}[N_0 \notin ((1-\delta)\mathbb{E}(N_0), (1+\delta)\mathbb{E}(N_0))] < (\frac{e^\delta}{(1+\delta)^{(1+\delta)}})^{\mathbb{E}(N_0)} + \exp(-\mathbb{E}(N_0)\delta^2/2)$. Denote this probability as P_e^0 , Easy to see $P_e^0 = O(2^{-\theta(N)})$. Let ϕ_0 be the event of $N_0 \in ((1-\delta)\mathbb{E}(N_0), (1+\delta)\mathbb{E}(N_0))$. $\text{Pr}[\phi_0] = 1 - P_e^0$, so ϕ_0

occurs w.h.p. ($1 - O(2^{-\theta(N)})$).

Order all the sensors as v_1, v_2, \dots, v_N , in nondecreasing Euclidean distance to the sink $l_1 \leq l_2 \leq \dots \leq l_N$. Particularly, $l_1 = 0, v_1 = t$. By Theorem 1, the optimal communication cost

$$\begin{aligned} H_w(G_{\hat{X}}) &= \sum_{j=1}^N W(p_{\hat{X}_j}^*) \times H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_1) \\ &= l_o^{\alpha-1} \sum_{j=1}^N l_j \times H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_1) \\ &\geq l_o^{\alpha-1} r_0 \times H(\hat{X}_{\partial_0} | \hat{X}_{G_0}) \end{aligned} \quad (10)$$

So for the lower bound it is sufficient to show $H(\hat{X}_{\partial_0} | \hat{X}_{G_0}) \geq \theta[N \cdot \log f(c_\epsilon l_o)]$ w.h.p. Since $H(\hat{X}_{\partial_0} | \hat{X}_{G_0}) = H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) - H(\hat{X}_{G_0} | \hat{X}_t) - H(\hat{X}_t)$, we evaluate $H(\hat{X}_{G_0} | \hat{X}_t)$ first.

Among sensors closer to the sink than v_j , let v_k be the closest one to v_j , that is,

$$l_{k,j} = \min_{i < j} l_{i,j}$$

Let $d_j = l_{k,j}$ be the Euclidean distance between sensor j and k . Then by $H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_1) \leq H(\hat{X}_j | \hat{X}_k)$ and the same argument as in Lemma 4, $H(\hat{X}_j | \hat{X}_k) \leq \frac{1}{2} \log [(2\pi e)(f(d_j) + \frac{1}{12})]$. So

$$\begin{aligned} H(\hat{X}_{G_0} | \hat{X}_t) &= \sum_{v_k \in G_0 \setminus \{t\}} H(\hat{X}_k | \hat{X}_{k-1}, \hat{X}_{k-2}, \dots, \hat{X}_1) \\ &\leq \sum_{v_k \in G_0 \setminus \{t\}} \frac{1}{2} \log [(2\pi e)(f(d_k) + \frac{1}{12})] \end{aligned} \quad (11)$$

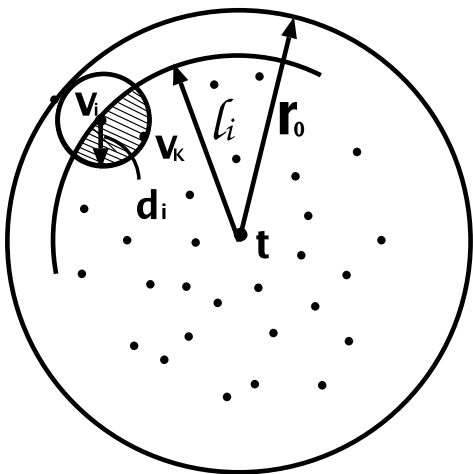
Let $\bar{d}_0 = \frac{1}{N_0-1} \sum_{v_k \in G_0 \setminus \{t\}} d_k$. Since $\log f()$ is a concave function,

$$\begin{aligned} &\sum_{v_k \in G_0 \setminus \{t\}} \frac{1}{2} \log [(2\pi e)(f(d_k) + \frac{1}{12})] \\ &\leq (N_0 - 1) \cdot \frac{1}{2} \log [(2\pi e)(f(\bar{d}_0) + \frac{1}{12})] \end{aligned} \quad (12)$$

Define $U_0 = (N_0 - 1) \cdot \frac{1}{2} \log [(2\pi e)(f(\bar{d}_0) + \frac{1}{12})]$, then

$$H(\hat{X}_{G_0} | \hat{X}_t) \leq U_0 \quad (13)$$

Next we study the distribution of d_i for any node v_i in the field. We do this indirectly with an area distribution related to d_i . Define ω_i as the area of the intersection of two disks $\varphi_t(l_i)$ and $\varphi_{v_i}(d_i)$ (See Fig. 5), $\omega_i = |\varphi_t(l_i) \cap \varphi_{v_i}(d_i)| = \beta(d_i)\pi d_i^2$, where β is a decreasing function of d_i and takes the minimum value of $\beta_{min} = \frac{2}{3} - \frac{\sqrt{3}}{2\pi}$ as d_i takes its max value $d_{i,max} = l_i$ when there are no sensors in between the sink and v_i . The sensor deployment satisfies a 2D Poisson process, so a ω_i follows an exponential distribution with mean $\mathbb{E}\omega_i = \frac{1}{\rho} = l_o^2$ variance $\text{Var}\omega_i = l_o^4$. It is easy to see that ω_i 's distribution for sensors in G_0

Fig. 5. Statistics of d_i

is not independent of N_0 . The conditional distribution is exponential distribution with mean $\mathbf{E}(\omega_i|N_0) = \frac{1}{\rho_0} = |G_0|/N_0 = \pi r_0^2/N_0 = \frac{c}{2}l_o^2 \frac{N}{N_0}$ and variance $\mathbf{Var}(\omega_i|N_0) = \mathbf{E}(\omega_i|N_0) = \frac{1}{\rho_0^2}$. Let $\bar{\omega}_0 = \frac{1}{N_0-1} \sum_{v_j \in G_0 \setminus \{t\}} \omega_j$. Then

$$\mathbf{E}(\bar{\omega}_0|N_0) = \mathbf{E}(\omega_i|N_0) = \frac{c}{2}l_o^2 \frac{N}{N_0} \quad (14)$$

$$\begin{aligned} \mathbf{Var}(\bar{\omega}_0|N_0) &= \frac{1}{(N_0-1)^2} \left[\sum_{v_i \in G_0 \setminus \{t\}} \mathbf{Var}(\omega_i|N_0) \right. \\ &\quad \left. + \sum_{v_i, v_j \in G_0 \setminus \{t\}} 2\mathbf{Cov}(\omega_i, \omega_j|N_0) \right] \end{aligned}$$

For any pair of $v_i, v_j \in G_0 \setminus \{t\}$, we next show ω_i, ω_j are negatively associated conditioned on N_0 , or

$$\mathbf{Cov}(\omega_i, \omega_j|N_0) = \mathbf{E}\left[\left(\omega_i - \frac{c}{2}l_o^2 \frac{N}{N_0}\right)\left(\omega_j - \frac{c}{2}l_o^2 \frac{N}{N_0}\right)\right] \leq 0 \quad (15)$$

This is because when we know ω_i 's value, the point process in G_0 is not any more the Poisson process without this information. First, N_0 points are deployed in a region $G_i = G_0 \setminus (\varphi_t(l_i) \cap \varphi_i(d_i))$ with an area of $|G_i| = |G_0| - \omega_i$; second, ω_i as a finite value implies d_i is finite value that can not be arbitrarily small, this means $N_0 - 2$ sensors are independently distributed in G_i , sensor v_i and another sensor are deployed independently of these $N_0 - 2$ sensors but not independently of each other, they have to be d_i distance away. All N_0 sensors are still uniformly deployed in G_i . If we look at an arbitrary small region G_ε adjacent to v_j , then v_i and another sensor can not simultaneously reside in G_ε . Thus the number of sensors in G_ε satisfy a equivalent binomial distribution of $N_0 - 1$ sensors independently and uniformly deployed in G_i :

$$\Pr(N_{G_\varepsilon} = k) = C_{N_0-1}^k \left(\frac{|G_\varepsilon|}{|G_i|}\right)^k \left(\frac{|G_i| - |G_\varepsilon|}{|G_i|}\right)^{N_0-1-k}$$

When the number of sensors is large, the point process in G_i approaches an equivalent Poisson process with rate $\lambda_i = \rho_i = \frac{N_0-1}{|G_0|-\omega_i}$ per unit area. So condition on ω_i, ω_j follows an exponential distribution with mean $\mathbf{E}(\omega_j|\omega_i, N_0) = \frac{1}{\lambda_i} = \frac{|G_0|-\omega_i}{N_0-1}$. Also from $|G_0| = \mathbf{E}(\omega_i|N_0) \cdot N_0$ we know

$$\begin{aligned} \mathbf{E}(\omega_j|\omega_i, N_0) &= \mathbf{E}(\omega_j|N_0) - \frac{\omega_i - \mathbf{E}(\omega_i|N_0)}{N_0-1} \\ &= \frac{c}{2}l_o^2 \frac{N}{N_0} - \frac{\omega_i - \frac{c}{2}l_o^2 \frac{N}{N_0}}{N_0-1} \end{aligned}$$

From this we get the negative association result of (15).

Thus

$$\begin{aligned} \mathbf{Var}(\bar{\omega}_0|N_0) &\leq \frac{1}{(N_0-1)^2} \sum_{v_i \in G_0 \setminus \{t\}} \mathbf{Var}(\omega_i|N_0) \\ &= \frac{1}{(N_0-1)^2} \frac{N_0-1}{\rho_0^2} \\ &= \frac{[\mathbf{E}(\bar{\omega}_0|N_0)]^2}{N_0-1} \end{aligned} \quad (16)$$

By Chebyshev's Inequality [40],

$$\Pr\left(\bar{\omega}_0 > (1+\epsilon)\mathbf{E}(\bar{\omega}_0|N_0) \mid N_0\right) < \frac{1}{(N_0-1)\epsilon^2} \quad (17)$$

Now assume ϕ_0 is true, $N_0 \in ((1-\delta)cN/2, (1+\delta)cN/2)$. Then by (14) $\mathbf{E}(\bar{\omega}_0|N_0) \in (\frac{l_o^2}{1+\delta}, \frac{l_o^2}{1-\delta})$. Apply (16), (17) we have

$$\Pr\left(\bar{\omega}_0 > \frac{1+\epsilon}{1-\delta}l_o^2 \mid \phi_0\right) < \frac{1}{\left[\frac{(1-\delta)cN}{2} - 1\right]\epsilon^2} \quad (18)$$

At the same time, $\bar{\omega}_0 = \frac{1}{N_0-1} \sum_{j \in G_0 \setminus \{t\}} \beta(d_j)\pi d_j^2 \geq \beta_{\min}\pi \left[\frac{1}{N_0-1} \sum_{j \in G_0 \setminus \{t\}} d_j^2\right]$. Since x^2 is a convex function,

$$\bar{\omega}_0 \geq \beta_{\min}\pi \bar{d}_0^2$$

Apply it to (18) we have

$$\Pr\left(\bar{d}_0 > \sqrt{\frac{1+\epsilon}{(1-\delta)\beta_{\min}\pi}}l_o \mid \phi_0\right) < \frac{1}{\left[\frac{(1-\delta)cN}{2} - 1\right]\epsilon^2}$$

Or

$$\Pr\left(\bar{d}_0 > c_\epsilon l_o \mid \phi_0\right) < \frac{1}{\left[\frac{(1-\delta)cN}{2} - 1\right]\epsilon^2} \quad (19)$$

Let ϕ_1 be the event of $\bar{d}_0 \leq c_\epsilon l_o$. Then $P_e^1 = \Pr(\phi_1^c) = \Pr(\phi_1^c \cap \phi_0^c) + \Pr(\phi_1^c \cap \phi_0) \leq P_e^0 + \Pr(\phi_1^c | \phi_0) \cdot \Pr(\phi_0) \leq P_e^0 + \Pr[\bar{d}_0 > c_\epsilon l_o | \phi_0]$. So $P_e^1 \leq O(2^{-\theta(N)}) + \theta(\frac{1}{N})$, ϕ_1 is true w.h.p. $(1 - O(2^{-\theta(N)}) + \theta(\frac{1}{N}))$

Let $\phi = \phi_0 \wedge \phi_1$, easy to see $Pr(\phi^c) \leq Pr(\phi_0^c) + Pr(\phi_1^c) \leq O(2^{-\theta(N)}) + \theta(\frac{1}{N})$. So ϕ is true w.h.p. Apply this to (12),(13), also f is nondecreasing, we have w.h.p.

$$H(\hat{X}_{G_0}|\hat{X}_t) \leq ((1 + \delta)cN/2 - 1)\frac{1}{2} \log [(2\pi e)(f(c_\epsilon l_o) + \frac{1}{12})]$$

From $H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) \geq cN \cdot \log f(c_\epsilon l_o)$, get $H(\hat{X}_{\partial_0}|\hat{X}_{G_0}) = H(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_N) - H(\hat{X}_{G_0}|\hat{X}_t) - H(\hat{X}_t) \geq \frac{1-\delta}{2}cN \cdot \log f(c_\epsilon l_o) - \theta(N) - \theta[\sqrt{N} \log f(c_\epsilon l_o)]$ w.h.p., or $H(\hat{X}_{\partial_0}|\hat{X}_{G_0}) = \theta(N \cdot \log f(c_\epsilon l_o))$ w.h.p. Then by (10) we have $H_w(G_{\hat{X}}) \geq \theta[N\sqrt{N}l_o^\alpha \log f(c_\epsilon l_o)]$ with high probability.

b)upper bound

Now that RHDB's stopping subregion size is modified as $3c_\epsilon l_o \times 3c_\epsilon l_o$, follow the same technique as before, we can derive an upper bound of the same order.

Since the upper bound matches the lower bound, we prove RHDB is order optimal w.h.p. for PLVCF. ■

VI. 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 [41], 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 6 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 [11], we are unable to compare HDB's cost to the distance entropy (optimal). Nevertheless, our simulation demonstrates HDB's improvement and trend vs. RAW.

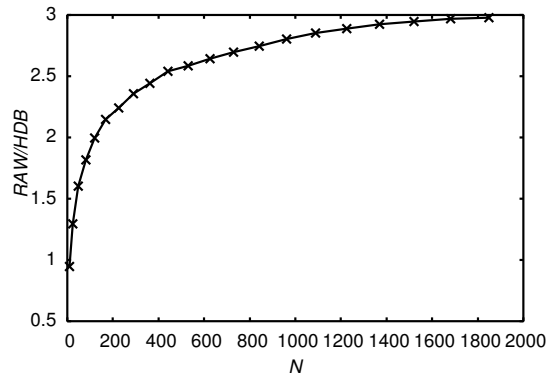


Fig. 6. The communication cost ratio of RAW to HDB for radar data set

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 $N \rightarrow \infty$ for two asymptotic topologies: fixed density network topology in which the cell size is constant (we set the size to one) and fixed area increasing density topology. 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.

Fixed Density, Expanded Network Topology:

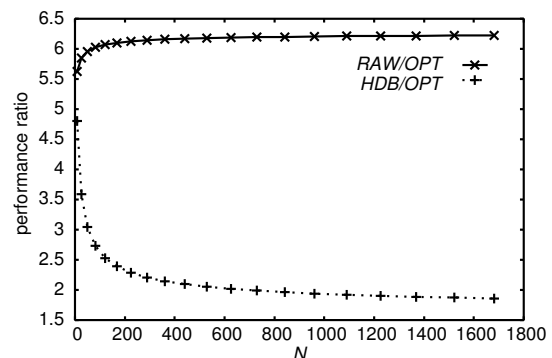


Fig. 7. The communication cost ratio of RAW to OPT, and HDB to OPT for data set generated by Gaussian Markov Field.

Figure 7 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 order 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 7, 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 $N \rightarrow \infty$, 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 results show that by exploiting local data correlation, HDB substantially reduces the communication cost, and achieves the order optimality with small constant factor.

Fixed Area, Increasing Density Topology: The above fixed density expanded network is only one way to look at the asymptotic behavior. Actually, a even more popular and interesting perspective for evaluating such networks' asymptotic behavior is to fix the area and increase the nodes number and thus increase the density [26]. We also evaluate the RAW vs. HDB vs. OPT under this kind of networks. With the density increasing as the nodes number increases, the correlation between neighboring nodes also increase. This is equivalent to increase the node number and correlation coefficient simultaneously with $\gamma \propto \Theta(c \frac{1}{\sqrt{N}})$ where c is a constant. As Fig. 8 shows, in this case, the RAW/OPT ratio increases with a unbounded trend, while the HDB/OPT ratio decreases faster than the fixed density case. This is because that as the node number increases, the correlation also increases, thus the HDB's benefit and the RAW's inefficiency shows up faster than the previous case. However, when N goes to very large, in the end the data sources will be totally dependent thus knowing the sink's value is sufficient to derive the whole field's values, thus the OPT cost will goes to zero and the RAW/OPT ratio goes to infinity. At the same time, if HDB does not use any block code and has to send at least one bit even for nothing, and also if we require HDB to be robust enough to handle different source models with the same operations, then there will be a minimum cost for HDB that does not diminish away as N increases beyond this point. Since the OPT cost goes to zero ultimately, as N goes beyond certain point, the HDB/OPT ratio will reach a minimum value then start to increase and ultimately go to

infinity. If however, we allow HDB to use some block code techniques and learning algorithms to detect the source model dynamics and adapt accordingly, then HDB's cost can also be made to go to zero as N goes to infinity under this fixed area increased density network. How fast it goes to zero depends on how much we want to trade the coding complexity for energy efficiency.

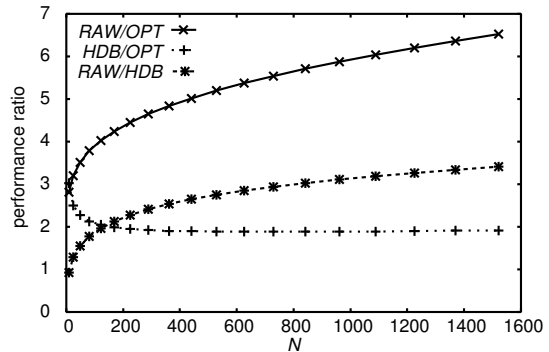


Fig. 8. The communication cost ratio RAW/OPT, HDB/OPT and RAW/HDB in a fixed area increased density network of a Gaussian Markov Field.

VII. CONCLUDING REMARKS

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 order 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 order optimality with small constant factor.

Our work is a first step in solving the much larger problem of general in-network computation as proposed by Giridhar and Kumar in [42]. In general, we want to minimize the communication cost of multi-sink lossy in-network computation task, where we have multiple sinks each needs to estimate some general function of the correlated data. In this case, we would expect that network coding can help and needs to cooperate with distributed source coding techniques to improve the performance, finding the minimum cost and identifying the optimal scheme under this general setup seems to be a quite challenging problem.

VIII. 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.

REFERENCES

- [1] C. Chong and S. Kumar, "Sensor networks: Evolution, opportunities, and challenges," in *IEEE Symposium on Foundations of Computer Science*, 2003, pp. 1247–1256.
- [2] D. Estrin, D. Culler, K. Pister, and G. Sukhatme, "Connecting the physical world with pervasive networks," *IEEE Pervasive Computing*, vol. 1, no. 1, pp. 59–69, 2002.
- [3] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On network correlated data gathering," in *IEEE INFOCOM*, 2004.
- [4] S. Patten, B. Krishnamachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *IPSN*, 2004.
- [5] A. Goel and D. Estrin, "Simultaneous optimization for concave costs: Single sink aggregation or single source buy-at-bulk," in *SODA*, 2003.
- [6] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.
- [7] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Networked Slepian-Wolf: Theory, Algorithms and Scaling Laws," *IEEE Transactions on Information Theory*, 2005.
- [8] Y. Liu, D. Towsley, J. Weng, and D. Goeckel, "An information theoretic approach to network trace compression," University of Massachusetts, Amherst, Tech. Rep. CS TR05-03, 2005.
- [9] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," in *IEEE Signal Processing Magazine*, September 2004, pp. 522–533.
- [10] R. Ahlswede, N. Cai, S. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.
- [11] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, "The complexity of approximating entropy," 2002.
- [12] M. Adler, "Collecting correlated information from a sensor network," in *SODA*, 2005.
- [13] D. Schongberg, K. Ramchandran, and S. Pradhan, "Distributed code constructions for the entire slepian-wolf rate region for arbitrarily correlated sources," in *Proc. DCC'04*, March 2004, pp. 292–301.
- [14] A. Liveris, Z. Xiong, and C. Georghiades, "Distributed compression of binary sources using conventional parallel and serial concatenated convolutional codes," in *Proc. DCC'04*, March 2004, pp. 292–301.
- [15] —, "Compression of binary sources with side information at the decoder using ldpc codes," *IEEE Communication Letters*, vol. 6, pp. 440–442, Oct. 2002.
- [16] V. Stanković, A. Liveris, Z. Xiong, and C. Georghiades, "Design of slepian-wolf codes by channel code partitioning," in *Proc. DCC'04*, March 2004, pp. 302–311.
- [17] S. Cheng and Z. Xiong, "Successive refinement for the wyner-ziv problem and layered code design," in *Proc. DCC'04*, March 2004, p. 531.
- [18] M. Adler, N. Harvey, K. Jain, R. Kleinberg, and A. Lehman, "On the capacity of information networks," in *SODA*, 2006.
- [19] Z. Li, B. Li, D. Jiang, and L. Lau, "On achieving optimal throughput with network coding," in *INFOCOM*, 2005.
- [20] J. Chou, D. Petrovic, and K. Ramchandran, "A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks," in *INFOCOM*, 2003.
- [21] A. Lehman and E. Lehman, "Complexity classification of network information flow problems," in *SODA*, 2004.
- [22] A. Ramamoorthy, K. Jain, P. Chou, and M. Effros, "Separating distributed source coding from network coding," in *Allerton*, 2004.
- [23] T. Ho, M. Médard, M. Effros, and R. Koetter, "Network coding for correlated sources," in *CISS*, 2004.
- [24] Y. Wu, V. Stankovic, Z. Xiong, and S. Kung, "On practical design for joint distributed source and network coding," in *NETCOD*, 2005.
- [25] D. Lun, Médard, T. Ho, and R. Koetter, "Network coding with a cost criterion," MIT, Tech. Rep. P-2584, 2004.
- [26] 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.
- [27] D. Ganesan, R. Cristescu, and B. Beferull-Lozano, "Power-efficient sensor placement and transmission structure for data gathering under distortion constraints," in *IPSN*, 2004.
- [28] A. Jindal and K. Psounis, "Modeling spatially-correlated sensor network data," in *SECON*, 2004.
- [29] T. Cover and J. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York, NY, USA: John Wiley & Sons, 1991.
- [30] N. Megiddo, "Optimal flows in networks with multiple sources and sinks," *Math. Programming*, vol. 7, pp. 97–107, 1974.
- [31] T. Han, "Slepian-wolf-cover theorem for networks of channels," *Information and Control*, vol. 47, no. 1, pp. 67–83, 1980.
- [32] T. Cover, "A proof of the data compression theorem of slepian and wolf for ergodic sources," *IEEE Transactions on Information Theory*, vol. 22, pp. 226–228, March 1975.
- [33] S. Katti, D. Katabi, W. Hu, and R. Hariharan, "The importance of being opportunistic: Practical network coding for wireless environments," in *Allerton*, 2005.
- [34] J. Liu, D. Goeckel, and D. Towsley, "The throughput order of ad hoc networks employing network coding and broadcasting," in *to appear, MILCOM*, 2006. [Online]. Available: <http://www.cs.umass.edu/liujn/research/milcom06.pdf>
- [35] —, "Bounds on the gain of network coding and broadcasting in wireless networks," in *Submitted to Infocom*, 2007. [Online]. Available: <http://www.cs.umass.edu/liujn/research/1568996949.pdf>
- [36] D. Mackay, *Information theory, inference, and learning algorithms*. John Wiley & Sons, 2004.
- [37] N. Cressie, *Statistics for Spatial Data*. John Wiley & Sons, 1993.
- [38] D. Bernstein, *Matrix Mathematics*. Princeton University Press, 2005.
- [39] D. W. R. Johnson, *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.
- [40] R. Motwani and P. Raghavan, *Randomized algorithms*. Cambridge University Press, 1995.

- [41] 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, vol. 76, 2001.
- [42] A. Giridhar and P. R. Kumar, "Towards a theory of in-network computation in wireless sensor networks," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 98–107, 2006.