Collecting Correlated Data Through a Network with Minimum Cost: Distance Entropy and a Practical Asymptotically Optimal Design

Junning Liu, Micah Adler, Don Towsley
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
Email: {liujn, micah, towsley}@cs.umass.edu

Abstract— **We study the communication cost of collecting correlated data at a sink over a network. To do so, we introduce *Distance Entropy*, an intrinsic quantity that characterizes the data gathering limit of networked sources. We demonstrate that, for any network embedded with any set of sources and a cost function [cost]=[data rate]×[link weight], distance entropy is a lower bound on the optimal communication cost. This is true for the most general data collection schemes that allow arbitrary routing and coding operations, including network coding and source coding. This lower bound can be matched using optimal rate Slepian-Wolf encoding plus shortest path routing. For more general communication cost functions, we show that the optimal scheme among schemes using Slepian-Wolf codes is also universally optimal. We then turn to the problem of designing practical and computationally efficient data collection schemes and propose a new, simple, hierarchical data collection scheme that is much more practical than the ones using either Slepian-Wolf Encoding or Explicit Entropy Encoding, another well known technique. We demonstrate for a number of correlation structures, the communication cost required by this scheme is within a constant factor of the distance entropy, and thus its performance is asymptotically optimal. This optimality is shown for two deployment strategies: a 2D grid regular network and a 2D Poisson process random network.**

I. Introduction

A. Problem statement and motivation

Consider the problem of gathering information from a set of correlated data sources. As shown in Fig. 1, each source is located at a solid black node and a set of communication links represented by edges connects the network together. We view this as a graph with the sources as a subset of the nodes and the links[1] as the edges. Each node is capable of sending information and performing coding computations. Each link has an associated weight. Furthermore, each source X_i generates data according to its source distribution. A designated node t acts as a *sink* that must reconstruct the information generated by all the sources. The cost of data transmission over a link is a function of the transmission rate and the link weight. The goal is to minimize the total communication cost. This is a joint source/network coding problem: source coding due to the correlations between the different nodes, as well as any other known distributional information; network coding since we allow the nodes to compute arbitrary functions of the data they receive. In this paper, we study two fundamental questions for this scenario:

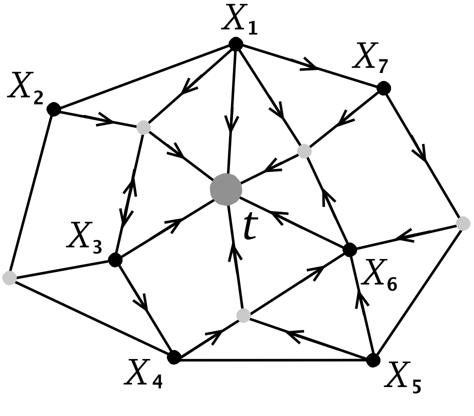


Fig. 1. A Layout of the General Problem of Gathering Correlated Data Through a Network

1) *What is the minimum total communication cost for achieving the data gathering task, possibly under some link capacity constraints?*
2) *What is the tradeoff between communication cost and node complexity? In other words, how complex*

[1] We consider both point to point links and broadcasting links.

do the nodes' functionalities need to be in order to achieve the optimal or close to optimal cost.

These are difficult questions because there are no limitations on what functions a node is allowed to compute for its outgoing messages based on the incoming information and/or its local source. Fig. 1 shows an example data transmission scheme; it uses an arrow on an edge to indicate the actual traffic along the arrow's direction. As we can see, a node can send all its data over one link (e.g. node X_2), can broadcast its data (e.g. X_3), and can send its data to some selected neighbors (e.g. X_6). In addition to this, a node can perform any function on the data and send the output to any neighbor, so long as the sink is able to decode all the source data. With this set up, we are considering the most general scenario that includes all possible schemes. For a limited case where the coding is restricted to only Slepian-Wolf Code, Cristescu etc.'s work [1] [2] finds the optimal rates allocation among the source nodes, and shows that this, combined with shortest path routing, achieves the minimum cost among all such schemes. However, for the general case where arbitrary coding/routing operations are allowed, it is still not known what the optimal cost is and how traditional source coding/network coding techniques can be exploited to achieve it.

This problem is primarily motivated by sensor network applications [3] [4]. 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). Since the battery power is normally quite limited for such cheap sensors and the communication energy cost is a major factor that drains the battery, minimizing the total communication cost is important for such applications. Another example is the collection of correlated data from distributed sources on the Internet, such as images/videos [5] for retrieval or network traces [6] for network management. The cost functions for the Internet are transmission delays and consumption of network resources.

We focus on sensor net applications in this paper while the general results apply to the Internet as well. Consider a sensor network that collects the measurements from the environment to a single sink. We assume an observation at a sensor is a discrete, random variable $\hat{X}$. For continuous field values being measured, quantization techniques are used to convert them to discrete value sensor readings. The N sensors that take measurements generate a vector of discrete random variables $(\hat{X}_1, \ldots, \hat{X}_N)$ that needs to be transmitted to a designated base station, which then decodes the original vector based on the received information. The communication cost per second (*communication power*) for a link is a function of the transmission rate and the link weight between the two nodes. An often used, simplified cost function is a product of a rate term and a separable weight term. We look at both the general and the simplified cost functions. The goal is to solve this data collection problem with reasonable computation/storage complexities for the sensors while minimizing the total communication cost.

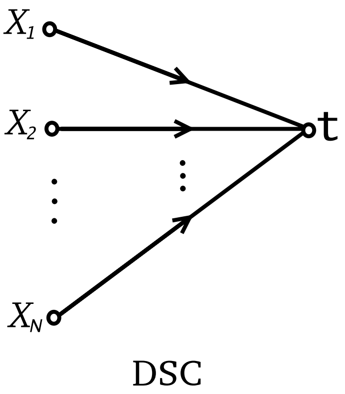


Fig. 2. Distributed Source Coding

In our setting, the node operations are essentially of two basic types: Source Coding (SC) and Network Coding (NC) [7]. When the network contains just N source nodes other than the sink, and only N links connecting each of the source nodes directly to the sink without other links between the source nodes as shown in Fig. 2, the problem reduces to a Distributed Source Coding (DSC) [8] problem; when the sources are independent, the problem reduces to a Network Coding problem, particularly if there is just a single sink, then there is no need for Network Coding: [9] shows that traditional routing where data is treated as commodity flows is sufficient to solve the data collection problem for such networks. On the other hand, sensor information and that of a lot of other applications are often correlated and exhibit large redundancies, e.g. the sensors that measure the temperature or rainfall volumes over a region generate highly correlated readings. [10] studies the problem of separating SC from NC for collecting correlated sources to 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 NC is necessary (not vice versa), we do know that there are cases where NC is mandatory. Thus without sound falsifications it is possible that NC is helpful for achieving the minimum cost or capacity constraints for the case of correlated sources.

There are various possible approaches based on different combinations of SC and NC. The two most studied are based on Slepian-Wolf Coding (SWC) [11] [8] and Explicit Entropy Coding (EEC) [1]. SWC is a distributed source coding technique that allows the sensor nodes to encode without explicit communication. Each sensor encodes its data to some rate with the joint rate vector in the achievable Slepian-Wolf region. For our data collection task, if we are restricted to first apply SWC and then route the encoded bits as incompressible flows through the network, [1] shows that when the cost function is in the form of [cost]=[rate]×[weight] and there is no capacity constraints, the optimal cost is the optimal rate allocation of SWC followed by shortest path routing. For EEC, joint encoding is only possible when side information is available: a node sends out data with a rate equal to the joint entropy rate of incoming data and its own sensed data. [1] shows that optimizing the total cost of EEC is NP-Complete. In general it is still not known whether/how SWC, EEC or any other codes can be exploited to achieve the minimum communication cost of our data collection problem.

Both SWC and EEC have practical limitations. Slepian and Wolf's original work [11] guarantees the existence of an encoding scheme that achieves the joint entropy rate. To ensure the probability of decoding error goes to zero, it requires the size of the block for coding goes to infinity. Thus, their work is an existential rather than constructive result. For designing constructive SWC, most of the progress has been made on highly limited source models [5] [12]. Even for the case of a general constructive SWC, the work required to learn the conditional entropies of N random variables to specify an achievable rate vector is generally too computationally expensive for cheap sensors, especially for large size networks. The long temporal coding block also requires considerable memory on each node. Thus, currently SWC is still a theoretical result that helps us understanding the limits of coding by assuring the existence of an ideal. EEC needs even more training than SWC: it 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.

The source models that have been studied are generally 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 [5], 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.

In summary, first it is important to characterize the optimal communication cost in a general setting; second, given the limited source models studied and the high complexity of current DSC technique, it is desirable to design low complexity data collection schemes for more general source models with close to optimal communication cost.

B. Main contribution

In this paper, we introduce the concept of *distance entropy* as an intrinsic property of a distributed source set to characterize a lower bound on the cost for collecting the distributed information. Distance entropy is a generalization of entropy that, like entropy, measures a probability distribution, but it also takes into account the underlying network topology of the source nodes.

For the case of [cost]=[rate]×[weight] cost functions and no link capacity constraints, we show that the distance entropy equals the cost of the SWC scheme with the optimal rate allocation and shortest path routing. Thus, we prove that the optimal SWC scheme described by [1] is actually optimal over all possible data collection schemes. We also show that for general cost functions with or without capacity constraints, the optimal cost for SWC schemes is also universally optimal for any schemes. A corollary of these results is that for collecting correlated sources to a single sink, network coding is not needed in order to minimize the total communication cost or maximize the achievable capacity.

We then turn to the question of designing simple, practical protocols that achieve (at least asymptotically) the distance entropy. To do so, we first consider various generic and commonly used classes of source models for a two dimensional grid: a Hard Continuity Field, a Linear Variance Continuity field and a Gaussian Markov field. We give nontrivial lower bounds on the distance entropy of these source models for a regular 2D sensor grid. We then propose a simple hierarchical data collection scheme and demonstrate that its communication cost for these source models is within a constant factor of our lower bound on the corresponding distance entropy. This demonstrates the asymptotic optimality of our protocol. Finally we extend the grid results to corresponding high probability results for sensor networks built using randomly deployed nodes.

The paper is organized as follows: In Section II, we formalize the model. In Section III, we define distance entropy and prove the universal optimal results. In Section IV we propose the simple hierarchical data collection scheme and prove its asymptotical optimality. In Section V we introduce the related work. Finally we conclude and discuss future work in Section VI.

II. MODEL FORMULATION

We represent a network as a connected graph (directed or undirected) $G = (V, E, W)$. V is the set of all of the nodes. There is a single sink $t \in V$ corresponding to a central processing point or base station and a *source node set* $\Omega \subseteq V$ corresponding to the set of source nodes that are generating data, $|\Omega| = N$. All nodes in V are able to code and transmit data. E is the edge set, an edge $e = (v_i, v_j) \in E$ iff there is a direct communication link between node v_i and node v_j. The communication links consist of discrete noisy or noiseless memoryless channels.[2] We assume the links (channels) to be independent point to point links since normally there is an underlying MAC layer to solve the wireless contention problem using techniques like TDMA, FDMA, ALOHA etc. We also omit the negligible communication overhead induced by synchronization and routing control since data can be packed in an arbitrarily large packet. There is a weight set W and each edge $(v_i, v_j) \in E$ has an associated weight $w_{ij} \in W, w_{ij} \geq 0$ that relates to the communication cost. There is also possibly an associated positive capacity $c_{ij} \in C$ that specifies the maximal transmission rate over the link. However, if the capacities are much larger than the data rates for all $e \in E$, we can ignore C and treat the network as one without capacity constraints. Define the cost for a path p in G as $W(p) = \sum_{e \in p} w_e$. Denote the set of all the paths from a node $v \in V$ to t as $\wp_v$, the shortest path from v to t as p_v^*. Then $W(p_v^*) = \min_{p \in \wp_v} W(p)$.

Each source node $v_i \in \Omega$ periodically generates samples of a discrete source $\hat{X}_i$. The joint source vector $\hat{X} = \{\hat{X}_1, ..., \hat{X}_N\}$ follows some joint distribution $p(\hat{x}_1, \hat{x}_2, ..., \hat{x}_N)$. Let $\{\hat{X}(\tau)\}_{\tau=1}^{\infty}$ be a stationary random process where $\hat{X}(\tau) = (\hat{X}_1(\tau), \ldots, \hat{X}_N(\tau))$ is a *field sample* that corresponds to the set of samples gathered from all sources at time τ, $\tau = 1, 2, \ldots$. For simplicity of presentation, we analyze the total communication cost of collecting one field sample within one second while our results can be extended to the general case of collecting multiple field samples that are temporally correlated.

A *source graph* G_X consists of a graph $G(V, \Omega, t, E, W)$, a source set $\hat{X}$ and a one to one mapping between $\hat{X}$ and Ω. A *Communication Scheme* specifies for all the nodes "what to send to whom" – a set of functions for the network to map each node's received bits (alphabet) and local generated data (if any) to its output bits (alphabet) and the corresponding selected channels. A *Data Collection Scheme (DCS)* Υ is a communication scheme for the network to collect all of the data at t *near losslessly*–decode losslessly with zero or an arbitrarily small probability of error [8]. A *SWC scheme* Υ_{SWC} is a DCS of particular interest to us that separates source coding from channel coding and separates source coding from routing, more specifically, it only allows Slepian-Wolf source codes and commodity flow routing. A *SWC-SP scheme* Υ_{SWC-SP} is a SWC scheme that only uses the 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.

There is an associated cost for any transmission in G_X. Let r_e be the transmission rate along edge e in bits per second. The cost per second along edge e is given as $g(r_e, w_e)$, a function of r_e and w_e [2]. 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. Denote the optimal cost as $W_{\Upsilon^*} = \min_{\Upsilon \in \Pi} W_\Upsilon$. The cost function g is naturally assumed to be a strictly increasing function of r_e and w_e. The most commonly studied cost function is $g(r_e, w_e) = r_e \cdot w_e$ [2]. Under this form of g, the cost to transmit b bits in τ seconds is $g(b/\tau, w_e) \cdot \tau = b \cdot w_e$, independent of the transmitting period τ. So in this case we can equivalently study the communication cost for collecting one data sample, denoted also as W_Υ. For wireless communication links, $w_e = {l_e}^\alpha$, where $2 \leq \alpha \leq 4$ depending on the medium and l_e is the Euclidean distance between the two nodes connected by e.

[2] A memoryless channel is one that the output is conditionally independent of previous inputs given the current input. The case of noiseless channel reduces the problem to be a pure network source coding problem.

III. UNIVERSAL OPTIMAL COMMUNICATION COST

We introduce a new concept, *Distance Entropy* of a source graph G_X, to characterize its information distribution.

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

$$H_w(G_X) = \sum_{j=1}^{N} W(p^*_{v_j}) \times H(\hat{X}_j | \hat{X}_{j-1}, ..., \hat{X}_0)$$

where $v_i \in \Omega, 1 \leq i \leq N$ is a source node that has a shortest path weight $W(p^*_{v_i})$ satisfying $W(p^*_{v_1}) \leq W(p^*_{v_2}) \leq ... \leq W(p^*_{v_N})$. Also $\hat{X}_0$ denotes $\hat{X}_t$ the source located at sink t which can be null. H is the discrete entropy.

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

Theorem 1: The cost of any data collection scheme 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 $r_j = H(\hat{X}_j | \hat{X}_{j-1}, ..., \hat{X}_1)$ ($v_1, v_2, \ldots, v_N$ is in a nondecreasing order of shortest path weight) achieves the cost of $H_w(G_X)$. Thus

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

Proof: 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 and the lower bound that we derive applies to arbitrary source graphs.

For all the nodes in the vertex set V, order them in a nondecreasing order of the shortest path weight to the sink, i.e. $v_1, v_2, \ldots, v_{|V|}$ satisfies $W(p^*_{v_1}) \leq W(p^*_{v_2}) \leq \ldots \leq W(p^*_{v_{|V|}})$. Particularly the source set $(\hat{X}_t, \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N)$ form a subsequence of it as $W(p^*_{\hat{X}_t}) = 0 < W(p^*_{\hat{X}_1}) \leq W(p^*_{\hat{X}_2}) \leq ... \leq W(p^*_{\hat{X}_N})$. 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 $W(p^*_{\hat{X}_j})$s are distinct values then extend it to the general case.

For each $W(p^*_{\hat{X}_j}) \neq 0$, define a triple partition of the vertex set V as $(M_j^<, M_j, M_j^>)$ where $M_j^< = \{v | W(p^*_v) < W(p^*_{\hat{X}_j})\}$, $M_j^> = \{v | W(p^*_v) > W(p^*_{\hat{X}_j})\}$. Define $M_j = \{v | W(p^*_v) = W(p^*_{\hat{X}_j})\}$ as a boundary set in between. Then we do the following procedure:

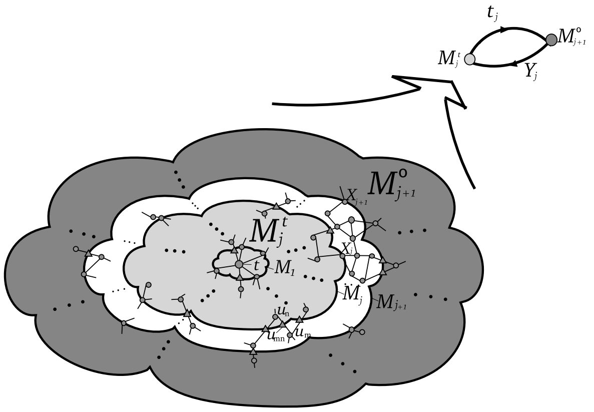


Fig. 3. Construction of the virtual graph

As shown in Fig. 3, from $j = 1$ to N, $\forall (v_m, v_n) \in E$, 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. 3), $G \leftarrow G \cup \{v_{mn}\}$. Replace the edge (v_m, v_n) with two new edges $e_1 = (v_m, v_{mn})$ and $e_2 = (v_{mn}, v_n)$ with weights as $w_{e_1} = W(p^*_{v_m}) - W(p^*_{\hat{X}_j})$, $w_{e_2} = W(p^*_{\hat{X}_j}) - W(p^*_{v_n})$, $E \leftarrow (E \cup \{e_1, e_2\}) \setminus \{(v_m, v_n)\}$. Now $W(p^*_{v_{mn}}) = W(p^*_{\hat{X}_j})$. Update M_j as $M_j \leftarrow M_j \cup \{v_{mn}\}$. Each jth 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 $\Omega \subseteq M_j^t \cup M_{j+1}^o$, M_j^t and M_{j+1}^o partition the source set into two parts, where $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_j$ 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 $\hat{X}_1, \hat{X}_2 \ldots, \hat{X}_j$, map M_{j+1}^o to v_B with all the sources that lie in M_{j+1}^o, namely $\hat{X}_{j+1}, \hat{X}_{j+2} \ldots, \hat{X}_N$. 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 $\hat{X}_{j+1}, \hat{X}_{j+2} \ldots, \hat{X}_N$ 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^o_{j+1} and M^t_j. Consider a further reduced two party distributed source coding scenario where v'_A has side information $\hat{X}_1, \hat{X}_2 \ldots, \hat{X}_j$ and needs to decode $\hat{X}$ near 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^o_{j+1} to M^t_j for any DCS in G_X. Now that $\hat{X}_1, \hat{X}_2 \ldots, \hat{X}_j$ is described perfectly at v'_A, by the results of Slepian-Wolf coding [8] v'_B has to send at least $H(\hat{X}|\hat{X}_1, \hat{X}_2 \ldots, \hat{X}_j) = H(\hat{X}_{j+1}, \hat{X}_{j+2} \ldots, \hat{X}_N|\hat{X}_1, \hat{X}_2 \ldots, \hat{X}_j)$ bits per sample. The same requirement holds for the original traffic from M^o_{j+1} to M^t_j before the reduction. Since there are no capacity constraints here, we actually consider noiseless channels. Thus there has to be $B_j = H(\hat{X}_j, \hat{X}_{j+1} \ldots, \hat{X}_N|\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{j-1})$ bits transmitted from M_j to M_{j-1} for every $2 \leq j \leq N$. Let M^t_0=$\{t\}$, let $M^o_1 = M_1 \cup M^>_1$, using the same argument we get $B_1 = H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N)$. In the case $\hat{X}_1$ is identical as the sink t, $B_1 = H(\hat{X}_2, \hat{X}_3, \ldots, \hat{X}_N|\hat{X}_1)$.

By the new graph's property, a path from any node $v_o \in M^o_{j+1}$ to any node $v_t \in M^t_j$ has to first reach some node $v_m \in M_{j+1}$ then some node $v_n \in M_j$. By definition $W(p^*_{v_m}) = W(p^*_{\hat{X}_{j+1}})$, $W(p^*_{v_n}) = W(p^*_{\hat{X}_j})$. For any path p_{mn} from v_m to v_n, the path's weight has to satisfy the triangle property $W(p_{mn}) \geq W(p^*_{v_m}) - W(p^*_{v_n}) = W(p^*_{\hat{X}_{j+1}}) - W(p^*_{\hat{X}_j})$, otherwise there exists a path from v_m to t with a weight less than $W(p^*_{\hat{X}_{j+1}})$ 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 $E^p = E_1 \cup E_2 \cup \ldots \cup E_N$ where E_j contains all the edges that belong to paths in P_j. It is easy to verify that these E_js are disjoint sets.

We define the communication cost on each edge set E_j as W_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 [18], 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 $W_j \geq W(F_j)$. Since $g = w \cdot r$, from all above we know that for each bit to be transmitted from some $v_o \in M^o_j$ to some $v_t \in M^t_{j-1}$, there has to be $1 \times W(p_{v_{mn}}) \geq (W(p^*_{\hat{X}_j}) - W(p^*_{\hat{X}_{j-1}}))$ cost spent in edge set E_j, so $W(F_j) \geq B_j \times [W(p^*_{\hat{X}_j}) - W(p^*_{\hat{X}_{j-1}})]$.

Thus for any data collection scheme Υ,

$$\begin{aligned}
W_\Upsilon &\geq \sum_{j=1}^{N} W_j \\
&\geq \sum_{j=1}^{N} W(F_j) \\
&\geq \sum_{j=1}^{N} B_j \times \left[W(p^*_{\hat{X}_j}) - W(p^*_{\hat{X}_{j-1}}) \right] \\
&= \sum_{j=1}^{N} H(\hat{X}_j, \ldots, \hat{X}_N|\hat{X}_0, \ldots, \hat{X}_{j-1}) \\
&\quad \times \left[W(p^*_{\hat{X}_j}) - W(p^*_{\hat{X}_{j-1}}) \right] \\
&= \sum_{j=1}^{N} H(\hat{X}_j|\hat{X}_0, ..., \hat{X}_{j-1}) \times W(p^*_{\hat{X}_j}) \\
&= H_w(G_X)
\end{aligned}$$

Since $H_w(G_X)$ is exactly the cost of the SWC scheme that has $r_i = H(\hat{X}_j|\hat{X}_1, ..., \hat{X}_{j-1})$ and routes along shortest paths to the sink, we have $\min_{\Upsilon \in \Pi_{SWC-SP}} W_\Upsilon(G_X) = H_w(G_X)$.

When $W(p^*_{\hat{X}_{i_j}})$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.

■

If we consider SWC scheme a valid scheme, Theorem 1 shows that the minimum communication cost for such a source network is the distance entropy. 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 in 1980 [19]. Han [19] proves 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 their result to our source graph model and generalize their network topology assumptions as well. [19] 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 [19] with cut capacity and also because the max-flow min-cut theorem for network flows also applies to an undirected graph, we generalize [19]'s model to any directed/undirected source graph that has unlimited connections for source nodes. For any graph G, $\forall M \subseteq V, t \in M^c = V \setminus M$ 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 $\hat{X}_L = \{\hat{X}_i | v_i \in L \cap \Omega\}$, $\hat{X}_L^c = \{\hat{X}_i | v_i \in L^c \cap \Omega\}$.

Theorem 2: (Generalized version of Han1980 [19]) *For any source graph G_X (directed or undirected) with an edge capacity set C, there exists a data collection scheme iff*

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

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

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

and there exists a set of flows satisfying the capacity constraints from the source nodes Ω to the sink t with each source node v_i's flow rate magnitude as $f_i = r_i$.

This theorem can be derived by straightforwardly applying the same technique as [19] to our source graph setting. With Theorem 2 we derive a general result on the optimal cost of a source graph, before which we derive a Lemma and introduce some further definitions.

For any source graph G_X and a DCS Υ on it, let the average transmission rate by Υ 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 $r_M = \sum_{v_i \in M, v_j \in M^c} r_{(i,j)}$.

Lemma 1: For any source graph G_X with or without capacity constraints and any DCS Υ for it. The data rate cross any cut r_M satisfies

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

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

Since the total vertex number is finite, the total number of links from M to M^c on which Υ has traffic is also finite. We denote it as l_m. Let

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

then $\epsilon > 0$. Construct a directed graph $G'(V, E', C', W)$ 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. 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 in [8], Υ also makes a valid DCS in G'_X. However, the cut capacity of (M, M^c) in G' is $C'(M, M^c) = \sum_{v_i \in M, v_j \in M^c} (r_{(i,j)} + \epsilon) = r_M + l_m \cdot \epsilon$. By (1), we have

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

So the cut capacities of G'_X do not satisfy the iff condition of Theorem 2, then there exist no DCSs in G'_X. This contradicts with 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 $[(r_{(i,j)}^1, \tau_{(i,j)}^1), (r_{(i,j)}^2, \tau_{(i,j)}^2), \ldots, (r_{(i,j)}^{K_{ij}}, \tau_{(i,j)}^{K_{ij}})]$, where $r_{(i,j)}^k > 0$ is the rate in bits per second for the kth block and $\tau_{(i,j)}^k > 0$ is the corresponding transmission period. Here $K_{ij} \in \{1, 2, \ldots, +\infty\}$. The average rate by Υ along an edge (v_i, v_j) from v_i to v_j is $r_{(i,j)} = \frac{1}{\sum_{k=1}^{K_{i,j}} \tau_{(i,j)}^k} \sum_{k=1}^{K_{i,j}} r_{(i,j)}^k \cdot \tau_{(i,j)}^k$. 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: For any source graph G_X with or without capacity constraints that its cost function g is nondecreasing in w and r and convex on rate 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: We prove this by 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 by Υ on each link as a capacity constraint on that link for the SWC scheme.

Construct a directed graph $G'(V, E', C', W)$ 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. 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 $C'(M, M^c) = r_M \leq C(M, M^c)$ for any cut (M, M^c); also we have $r_M \geq H(\hat{X}_M|\hat{X}^c_M)$ by Lemma 1. So for any cut (M, M^c),

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

$C'(M, M^c)$ matches the iff condition of Theorem 2, then by Theorem 2 there exists a SWC scheme with a SWC rate vector $R' = (r'_1, r'_2, \ldots, r'_N)$ that satisfies $H(\hat{X}_M|\hat{X}^c_M) \leq \sum_{v_i \in M \cap \Omega} r'_i \leq C'(M, M^c)$ for any cut (M, M^c), and there exists a set of flows $F = (f_1, f_2, \ldots, f_N)$ from Ω to t in G'. For each $v_i \in \Omega$, the flow magnitude is $f_i = r'_i$. Since R' is in the Slepian-Wolf achievable rate region [20] 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 in G' is shadowed by Υ — a DCS in G.[3]

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

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

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

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

By the convexity of function g, we have

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

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. Then our results can be extended to hold if we just replace the $H(\hat{X}_M|\hat{X}^c_M)$ with the entropy rate

$$H_\infty(\hat{X}_M|\hat{X}^c_M) = \lim_{m \to \infty} \frac{1}{m} H(\hat{X}_M(t_1), \cdots, \hat{X}_M(t_m)|\hat{X}^c_M(t_1), \cdots, \hat{X}^c_M(t_m))$$

A. Extension to the case of Broadcast Channels

As mentioned before, previously we ignore the multi-access nature of the wireless medium because of a possible lower MAC layer separation. Now we consider the case that includes broadcast channels and show that the previous result is still true even if we can take advantage of the Multi-Access nature of wireless channels and allow cross-layer optimization. We use the same source model as before and a slightly modified communication model to incorporate broadcast channels. First we describe the communication model then we show the same optimal performance holds even with broadcast channels, in other words, broadcasting does not help.

1) Communication Model: In addition to the independent point to point channels we assumed before, now we also allow the nodes to broadcast: a node sends identical data to multiple receiving nodes simultaneously through a broadcasting channel. Let $NEI(v_i) = \{v_j | (v_i, v_j) \in E\}$ be the neighboring set of node v_i—the set of nodes that v_i can communicate directly via a point to point channel. Broadcasting here means v_i can send the same copy of data simultaneously in a rate r to any subset of its neighbor set $B \subset NEI(v_i)$. The energy cost $g_{i,B}(r)$ of the broadcasting is no less than the cost of sending in 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 the ones using omni-directional antennas for the point to point channels.[4] Also for any $v_j \in B$, r satisfies the capacity constraint $r \leq c_{ij}$ and the broadcasting occupies r of the capacity of the point to point channel from v_i to v_j.

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

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

2) Optimal Result: With the modified communication model we show that any source graph G_X whose nodes are enhanced with this broadcasting capability has the same optimal cost as the one without broadcasting. We give the following theorem.

Theorem 4: For any source graph G_X with or without capacity constraints that its cost function g is nondecreasing in w and r and convex on rate r, then the optimal SWC scheme that does not use broadcasting is also optimal over the class of all data collection schemes using broadcasting or not.

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

Proof: We prove it by showing that for any broadcasting 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 broadcasting.

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, $r_M^B \geq H(\hat{X}_M|\hat{X}_M^c)$. 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 $\hat{X}_M^c$. 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 $r_{s_M} \geq H(\hat{X}_M|\hat{X}_M^c)$ thus $r_M^B \geq H(\hat{X}_M|\hat{X}_M^c)$ for any cut (M, M^c).

Next we 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 in the proof of Theorem 3: G_X^{Υ} contains all vertices in G_X and for all the non-broadcasting traffic of Υ, add a directed edge along the traffic direction with a capacity equal to the traffic rate. For each broadcasting traffic of Υ from node v_i to a set of its neighbors B, we add a pure 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, a directed edge $(v_i, v_{i,B})$ with a capacity equal to the original broadcasting rate r_B and a directed edge from $v_{i,B}$ to each node in B with an infinite capacity. Then because $r_M^B \geq H(\hat{X}_M|\hat{X}_M^c)$ in G_X, it is easy to verify that for any cut (M, M^c) in G_X^{Υ}, the cut capacity satisfies $C'(M, M^c) \geq H(\hat{X}_M|\hat{X}_M^c)$, by Theorem 3 there exists a SWC scheme Υ_{SWC} in G_X^{Υ}. If we copy this Υ_{SWC} to G_X by distributing the flow traffic of $v_i \to v_{i,B} \to v_j$ directly as $v_i \to v_j$, by the construction of G_X^{Υ}, we obtain a non-broadcasting DCS Υ' in G_X. More than that, because g is convex and $g_{i,B}(r_B) \geq \max_{v_j\in B} g(r_B, w_{ij})$ we conclude this DCS Υ' in G_X is also a non-broadcasting DCS with a cost no greater than the broadcasting enhanced DCS Υ. This is true for any broadcasting enhanced DCS Υ in G_X, thus we conclude that including broadcasting in a DCS does not improve the total communication cost for our setting.

■

The theorems in this section show 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 a universally optimal data collection scheme. 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. A key part of the proofs relies on some combinatory geometric properties of submodular and supermodular functions based on Edmonds's result in [21]. In general, there are possible bandwidth benefits applying network coding or broadcasting. While for correlated sources and a single sink, it is first shown here as a corollary of our work that neither network coding nor broadcasting helps either in terms of communication cost or capacity for the most general setting. More than that, our work shows no coding/routing scheme outperforms the SWC schemes. Certainly as we mentioned earlier in Section I 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.

IV. ASYMPTOTICALLY OPTIMAL SIMPLE SCHEME

In this section we describe a simple data collection scheme—*Hierarchical Difference Broadcasting (HDB)* for both regular sensor nets on grid points and random deployed sensor nets. We show that HDB is asymptotically 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. II but with a special spatial deployment strategy. A *sensor grid* is a regular sensor network where sensors are deployed on a two dimensional square grid.[5] There are total of N sensors indexed as $v_{i,j}$, $1 \leq i,j \leq \sqrt{N}$, $i,j = 1,2,\ldots,\sqrt{N}$ are all integers. The location 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 $\hat{X}_{i,j}$ which 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 $\{\hat{X}_{i,j}\}s$ are described by a joint distribution. Denote a sample of $\hat{X}$ as $\hat{x}$, describe the number of bits that $\hat{x}$ is coded into 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 [1], 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 grid edges of unit length. Since there are no capacity constraints, we can equivalently limit the transmissions to be along only such shortest paths without effecting the optimal communication cost. Thus we abstract the sensor network as a grid graph $G(V,E)$, $E = \{(v_{i_1,j_1}, v_{i_2,j_2}) \,|\, |i_1 - i_2| + |j_1 - j_2| = 1\}$. 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 v_{i_1,j_1} as v_{i_2,j_2}'s $\eta_{1,2}$-*hop-neighbor* and vice versa. When $\eta_{1,2} = 1$, we refer to them as each other's *one-hop-neighbor*.

B. Hierarchical Difference Broadcasting(HDB) Scheme

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

We design the data collection scheme HDB as following:

Step 1: The sink $t \in \Omega_0$ broadcasts its reading $\hat{x}_t$ using a Self-Delimiting Code (SDC) [22] 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 as $\hat{x}_{i,j} \leftarrow \hat{x}_{i,j} - \hat{x}_v$.
} end Do loop
```

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

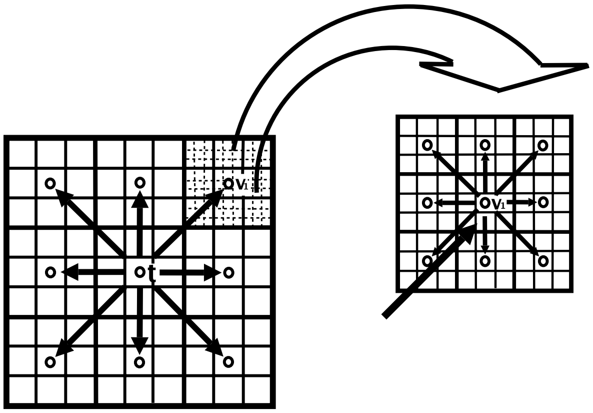


Fig. 4. The Hierarchical Broadcasts of HDB

Fig. 4 shows HDB's hierarchical difference broadcasting. When $N \neq 3^{2n}$, then $N \in (3^{2n}, 3^{2(n+1)})$ 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 $\Omega_0, \Omega_1, \ldots, \Omega_n$ for any N.

C. Asymptotic Optimality of HDB

Coding in HDB is extremely efficient as it relies only on simple subtractions and Self-Delimiting Codes. SDC

[5] Our results can be extended to cases of non-square grid.

is a practical code that encodes $\hat{x}$ into $\Theta(\log \hat{x})$ bits with negligible computation cost [22]. 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.[6] 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, $W(p^*_{\hat{X}_j}) = a\eta_j \cdot l_0^\alpha = \eta_j$ where η_j is the manhattan distance from $\hat{X}_j$ to the sink. By Theorem 1, $H_w(G_{\hat{X}}) = \sum_{j=1}^N W(p^*_{\hat{X}_j}) \times H(\hat{X}_j|\hat{X}_{j-1}, ..., \hat{X}_1) = \sum_{j=1}^N \eta_j \times H(\hat{X}_j|\hat{X}_{j-1}, ..., \hat{X}_1)$ is the optimal communication cost.

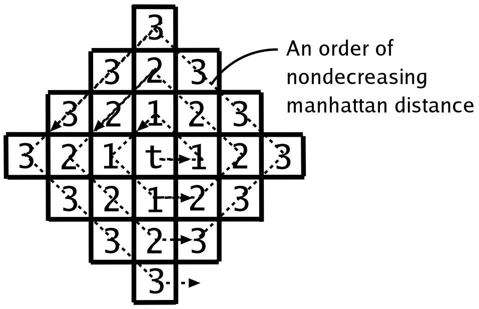


Fig. 5. The sink's k-hop-neighbor set layout on the grid

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. 5). 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. Denote the virtual scheme that collects these $U_0 = N_0 \cdot H_o \leq U$ bits via shortest paths as $\tilde{U}$, it has a cost $W_{\tilde{U}}$.

An optimal SWC scheme Υ^* also has to collect U_0 bits from nodes other than the sink and by Theorem 1 the ith sensor is allocated $H(\hat{X}_i|\hat{X}_{i-1}, \hat{X}_{i-2}, \dots, \hat{X}_1) \leq H_o$ bits, $\forall i > 1$. So for the first N_0 sensors, Υ^* can only allocate to each sensor no more bits than $\tilde{U}$ does. If we order the first U_0 bits collected by Υ^* in the order of nondecreasing manhattan distance, the jth bit of Υ^* has a manhattan distance that is no lower than the distance of the jth bit of $\tilde{U}$. Thus $W_{\Upsilon^*} \geq W_{\tilde{U}}$.

Let k^* be the maximum k that satisfies $\sum_{i=1}^k |S_i| \leq N_0$. Since $|S_i| = 4i$, we get $k^* = \lfloor \frac{\sqrt{2N_0+1}-1}{2} \rfloor$. Let $U_1 = H_o \cdot \sum_{i=1}^{k^*} 4i$, the cost for sending H_o bits from each sensor in $S_1, S_2, \dots, S_{k^*}$ is $W_{\tilde{U}_1} = H_o \cdot \sum_{i=1}^{k^*} 4i \cdot i = \frac{2}{3} H_o k^*(k^*+1)(2k^*+1)$. Replacing with $k^* = \lfloor \frac{\sqrt{2N_0+1}-1}{2} \rfloor$ and $N_0 = \lfloor \frac{U}{H_o} \rfloor$, yields $W_{\tilde{U}_1} = \Theta(U^{\frac{3}{2}}/H_o^{\frac{1}{2}})$. Since $U_1 \leq U_0$, $W_{\tilde{U}_1}$ is just part of the cost of $\tilde{U}$, then $W_{\tilde{U}} \geq W_{\tilde{U}_1}$. So we get $W_{\Upsilon^*} \geq W_{\tilde{U}} \geq W_{\tilde{U}_1} = \Theta(U^{\frac{3}{2}}/H_o^{\frac{1}{2}})$.

■

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 of distant points. This is normally reflected as smaller value difference 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 model the spatial sources using three generic source models that characterize this feature and show that the simple HDB is asymptotically optimal for each of them. Denote the cost of HDB as W_H, then there exists a constant $c > 0$ s.t. $W_H/W_{\Upsilon^*} \leq c$.

1) Hard Continuity Field (HCF):

For HCF, each one of $\hat{X}_{i,j}$ is a discrete random variable that have M different possible values. Without loss of generality, we assume the set for the M values is 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*

[6] For dynamic and non-uniform data resolution requests, we can adaptively adjust HDB's cluster size and hierarchies and form a wavelet type of multi-resolution dynamic scheme.

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

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

$d^{\sqrt{N}} \geq \Theta(M) \Rightarrow H(\hat{X}_t) \leq \log M \leq \Theta(\sqrt{N} \cdot \log d) \Rightarrow H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) - H(\hat{X}_t) \geq \Theta(N \log d) - \Theta(\sqrt{N}\log d) = \Theta(N \log d)$. Let $\hat{X}_1, \hat{X}_2, \ldots, X_N$ be a source sequence in an order of nondecreasing manhattan distances to the sink (as shown in Fig. 5) such that each $\hat{X}_i$ other than the sink has a one-hop-neighbor $\hat{X}_{i_1}$ in the sequence with $i_1 < i$. So $H(\hat{X}_i|\hat{X}_{i-1}, \hat{X}_{i-2}, \ldots, \hat{X}_1) \leq H(\hat{X}_i|\hat{X}_{i_1}) \leq \log(2d+1)$. Applying Lemma 2 with $U = \Theta(N \log d)$ and $H_o = \log(2d+1)$, yields $W(\Upsilon^*) \geq \Theta(N\sqrt{N}\log d)$.

Now we derive a same order upper bound for HDB's cost. $\forall N$, 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 [22]. 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 $\sum_{i=2}^{n} W_B^i \leq \sum_{i=2}^{n}[3^{2(n-i+1)} - 1] \cdot 9^{i-2} \cdot 8[\log(2 \cdot 3^{n-i+1}d) + \log(3^{n-i+1}d) + 2] = \Theta(\tilde{N}\log\tilde{N}) + \Theta(\tilde{N}\log^2\tilde{N}) + \Theta(\tilde{N}\log\tilde{N}\log d) + \Theta(\log\tilde{N}) - \Theta(\tilde{N}) - \Theta(\tilde{N}\log d)$. 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. $B_{n-1} \leq 2(\log 2 \cdot 3^{n-1}d) + \sum_{i=1}^{n-1} 8\frac{3^{2(n-1)}}{9^i}(\log 2 \cdot 3^{i-1}d + \log \cdot 3^{i-1}d + 2) = 2 \cdot 9^{n-1}\log d + (9^{n-1} + 1)\log 2 + 2 \cdot 9^{n-1} + \frac{9^{n-1}-1}{4}\log 3 = \Theta(\tilde{N}\log d) + \Theta(\tilde{N})$ 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 asymptotically 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 $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) = H(\hat{X}_1) + \sum_{i=2}^{N} H(\hat{X}_i|\hat{X}_{i-1}, ..., \hat{X}_1) \geq \log\frac{3d}{2} + (N-1)\log\frac{d}{2} = \Theta(N \cdot \log d)$.

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

Lemma 4: IF a LVCF has a joint entropy of $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) \geq \Theta(N \cdot \log d)$, and $Var(\hat{X}_t) \leq \Theta(d^{\sqrt{N}})$, then HDB's expected communication cost is asymptotically optimal. The optimal cost $W(\Upsilon^*)$ is lower bounded by $\Theta(N\sqrt{N}\log d)$.

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

First by [8]

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

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

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

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

Next we derive the upper bound for W_H. First $\mathbf{E}[(\hat{X}_i - \hat{X}_{i_1})^2] \leq d^2 \Rightarrow \mathbf{E}|\hat{X}_i - \hat{X}_{i_1}| \leq d$. Applying the triangle inequality of an absolute function, any two readings satisfy $\mathbf{E}|\hat{X}_i - \hat{X}_j| \leq \eta_{i,j} \cdot d$. Since Self-delimiting code can compress any $\hat{x}$ into $b(\hat{x}) = 2(\lfloor\log|\hat{x}|\rfloor + 1)$ bits [22] and $\log(x)$ is a concave function, by Jensen's inequality [8],

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

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

■

3) Gaussian-Markov field (GMF):

Multivariate Normal (MVN) is an often used model for multivariate distributions. It is a good approximation of many applications yet mathematically tractable. Gaussian-Markov Field (GMF) [23] is one common MVN model to model spatial fields exhibiting the close-points-high-correlation property. Let $X_1, X_2, \ldots, X_N$ be N continuous random values being measured at N different points of a GMF, they follow a joint MVN distribution: $N(\mu, \mathbf{\Sigma})$. Without loss of generality we assume the sources have the same mean $\mu = 0$. $\mathbf{\Sigma} = (\sigma_{\mathbf{i},\mathbf{j}})_{N\times N}$ is the covariance matrix with $\sigma_{i,j} = \sigma^2 \cdot e^{-c\eta_{i,j}l_0}$, where $c > 0$ is a constant and σ^2 is the unconditional variance of a source. So 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, our simulation suggests that the joint entropy ratio between GMF with manhattan distance and GMF with Euclidean distance is bounded in a range close to 1 as shown in Fig. 6.

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

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

Lemma 5: For any GMF on a k-dimensional hypercube grid of $N = m^k$ nodes, the field's joint entropy $H(\hat{X}_1, \ldots, \hat{X}_N) =$ $\frac{1}{2}\log\left((2\pi e)^N\sigma^{2N}(1-\gamma^2)^{km^{k-1}(m-1)}\right) - N\log\Delta$

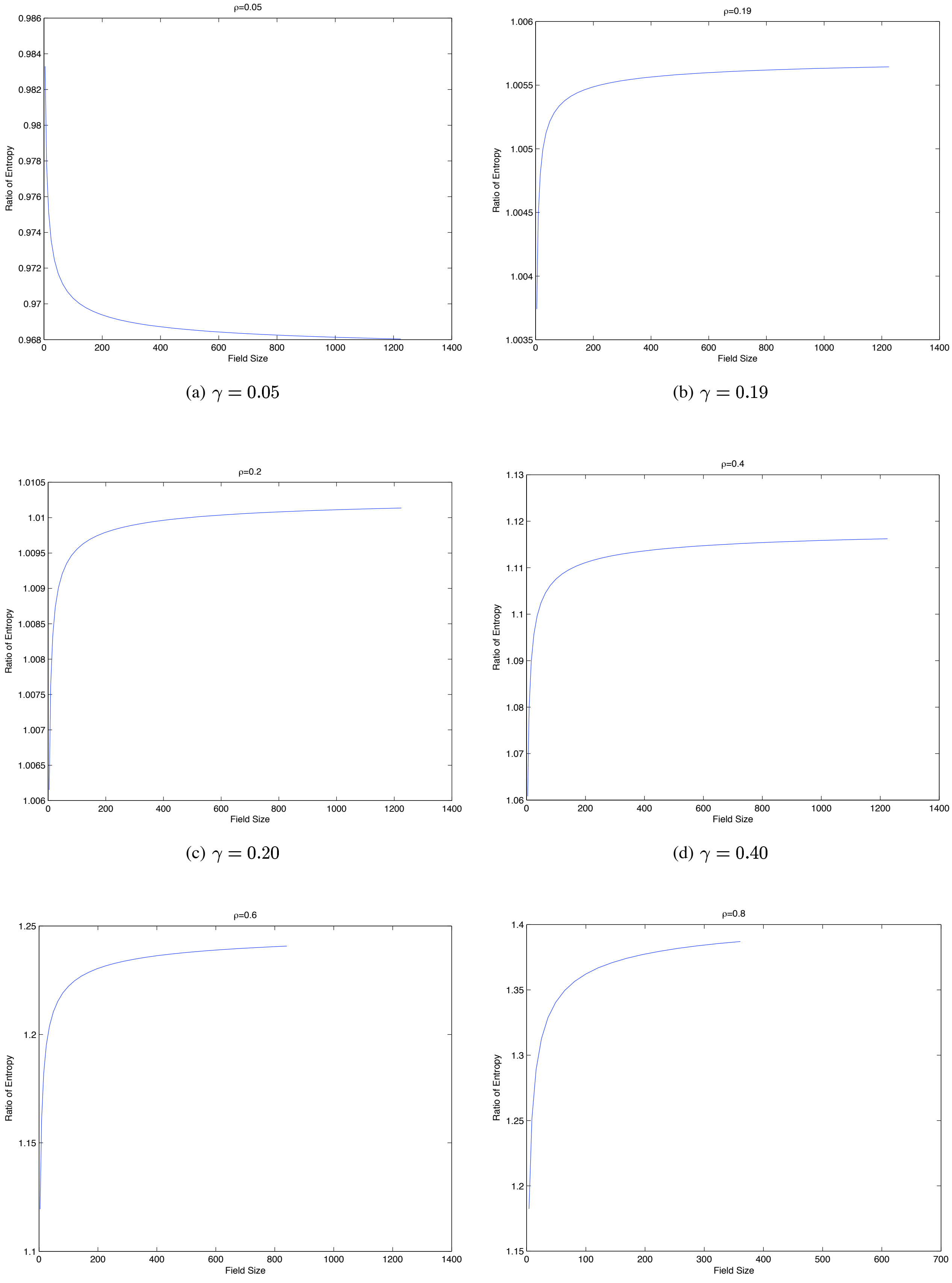
Proof: By [8],

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

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

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

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



(a) $\gamma = 0.05$

(b) $\gamma = 0.19$

(c) $\gamma = 0.20$

(d) $\gamma = 0.40$

(e) $\gamma = 0.60$

(f) $\gamma = 0.80$

Fig. 6. Joint entropy ratio of Manhattan distance GMF to Euclidean distance GMF

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 $q^k_{i,j} = \gamma^{|i-j|} Q_{k-1}$.

Q_1 is a Toeplitz matrix and $det(Q_1) = (1-\gamma^2)^{m-1}$ [24]. 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

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

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 [2]'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 (7), get $det(Q_k) \geq (1-\gamma^2)^{kN}$, by Lemma 5 we prove the corollary. Note that particularly for a 2D grid we have $det(\Sigma_2) = \sigma^{2N} det(Q_2) \geq \sigma^{2N}(1-\gamma^2)^{2N}$. ■

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

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

By Corollary 1,

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

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 bits of SDC code is larger than the corresponding entropy: $H(\hat{X}_i - \hat{X}_{i_1}) \leq \mathbf{E}[b(\hat{X}_i - \hat{X}_{i_1})]$. By (5),

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

By [25], $\mathbf{E}[(X_i - X_j)^2] = 2\sigma^2(1-\gamma_{i,j})$, then $\mathbf{E}|X_i - X_j| \leq \sqrt{2\sigma^2(1-\gamma_{i,j})} \Rightarrow \mathbf{E}|\hat{X}_i - \hat{X}_j| \leq \mathbf{E}|X_i - X_j|/\Delta + 1 \leq \sqrt{2\sigma^2(1-\gamma_{i,j})}/\Delta + 1$, by (6), we have $\mathbf{E}b[(\hat{X}_i - \hat{X}_j)] \leq 2[\log(\sqrt{2\sigma^2(1-\gamma_{i,j})}/\Delta + 1) + 1]$. Also $\Delta \ll \sqrt{\sigma^2(1-\gamma)}$ (high resolution quantizer), we get

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

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

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

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

we have $\mathbf{E}[b(\hat{X}_i - \hat{X}_j)] < 2H_o + \log \eta_{i,j}$. Then apply the same counting technique as in Lemma 3, we have HDB's cost is upper bounded by $W_H < \Theta(N\sqrt{N}H_o) + \Theta(N\sqrt{N}) = \Theta(N\sqrt{N}H_o)$. ■

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

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 $H(\hat{X}_t) \leq \Theta(\sqrt{N}\log d)$, the expected total communication cost of HDB is asymptotically 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. Nevertheless, it turns out that we are able to 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. As justified in [26], this can be due to the method of deployment, such as air-dropping in an unknown environment. 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 $N(A)$, it follows a Poisson distribution of 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:
Another difference of the model from the grid case is instead of using a one hop continuity constraint, we have to define a continuity constraint depending on the distance continuously because now the one hop distance is not a fixed value as in grid case. The constraint is modeled appropriately according to the sensed field being HCF, LVCF or GMF. Here we use LVCF as an example and it is easy to adjust for the other two. Assume for any two sensors v_i and v_j that has a Euclidean distance $l_{i,j}$, their reading difference satisfies $\mathbf{E}[(\hat{X}_i - \hat{X}_j)^2] \leq f(l_{i,j}) > 0$ where f is any nondecreasing function that maps the distance between two sensors to an upper bound of their reading differences. We call this model a Poisson LVCF field or $PLVCF$.

4) Protocol–RHDB:
We refer to the modified HDB 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. Stop dividing a subregion when it is of size $3c_\epsilon l_o \times 3c_\epsilon l_o$($c_\epsilon$ is some constant) or there are no sensors in it. Choose the sensor closest to the geometric center of the subregion as its cluster head. Then we have the following Theorem.

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

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

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 $\hat{X}_{G_0} = \{\hat{X}_k | l_k \le r_0\}$ 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 $Pr(Y_k = 1) = \frac{|G_0|}{|G|}$, corresponding to the probability for the kth sensor being deployed in region $\varphi_t(r_0)$. So $\mathbf{E}(N_0) = N\frac{|G_0|}{N/\rho} = \pi {r_0}^2 \rho = cN/2$. Using Chernoff bound [27], $Pr[N_0 \notin ((1-\delta)\mathbf{E}(N_0), (1+\delta)\mathbf{E}(N_0))] < (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{\mathbf{E}(N_0)} + \exp(-\mathbf{E}(N_0)\delta^2/2)$.Denote this probability as P_e^0, Easy to see $P_e^0 = O(2^{-\theta(N)})$. Let ϕ_0 be the event of $N_0 \in (\ (1-\delta)\mathbf{E}(N_0), (1+\delta)\mathbf{E}(N_0)\ )$. $Pr[\phi_0] = 1 - P_e^0$, so ϕ_0 occurs w.h.p. $(1 - O(2^{-\theta(N)}))$.

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

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

So for the lower bound it is sufficient to show $H(\hat{X}_{\partial_0}|\hat{X}_{G_0}) \ge \theta[N \cdot \log f(c_\epsilon l_o)]$ w.h.p. Since $H(\hat{X}_{\partial_0}|\hat{X}_{G_0}) = H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) - H(\hat{X}_{G_0}|\hat{X}_t) - H(\hat{X}_t)$, 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 $H(\hat{X}_j|\hat{X}_{j-1}, ..., \hat{X}_1) \le H(\hat{X}_j|\hat{X}_k)$ and the same argument as in Lemma 4, $H(\hat{X}_j|\hat{X}_k) \le \frac{1}{2}\log[(2\pi e)(f(d_j) + \frac{1}{12})]$. So

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

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,

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

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

$$H(\hat{X}_{G_0}|\hat{X}_t) \le U_0 \tag{13}$$

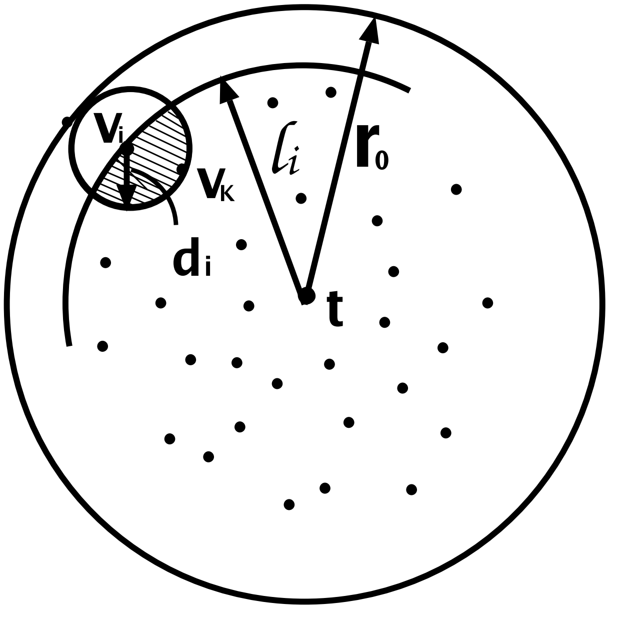


Fig. 7. Statistics of d_i

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. 7), $\omega_i = |\varphi_t(l_i) \cap \varphi_{v_i}(d_i)| = \beta(d_i)\pi {d_i}^2$, 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 $\mathbf{E}\omega_i = \frac{1}{\rho} = l_o^2$ variance $\mathbf{Var}\omega_i = l_o^4$. It is easy to see that ω_i's distribution for sensors in G_0 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

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

$$\mathbf{Var}(\bar{\omega}_{_0}|N_0) = \frac{1}{(N_0-1)^2}\Bigg[\sum_{v_i\in G_0\setminus\{t\}} \mathbf{Var}(\omega_{_i}|N_0) + \sum_{v_i,v_j\in G_0\setminus\{t\}} 2\mathbf{Cov}(\omega_{_i},\omega_{_j}|N_0)\Bigg]$$

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

$$\mathbf{Cov}(\omega_{_i},\omega_{_j}|N_0) = \mathbf{E}[(\omega_{_i} - \frac{c}{2}{l_o}^2\frac{N}{N_0})(\omega_{_j} - \frac{c}{2}{l_o}^2\frac{N}{N_0})] \leq 0 \tag{15}$$

This is because when we know $\omega_{_i}$'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, $\omega_{_i}$ 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:

$$Pr(N_{G_\varepsilon} = k) = C^k_{N_0-1}(\frac{|G_\varepsilon|}{|G_i|})^k(\frac{|G_i| - |G_\varepsilon|}{|G_i|})^{N_0-1-k}$$

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 $\omega_{_i}$, $\omega_{_j}$ 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

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

By Chebyshev's Inequality [27],

$$Pr\bigg(\bar{\omega}_{_0} > (1+\epsilon)\mathbf{E}(\bar{\omega}_{_0}|N_0)\bigg|N_0\bigg) < \frac{1}{(N_0-1)\epsilon^2} \tag{17}$$

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

$$Pr\bigg(\bar{\omega}_{_0} > \frac{1+\epsilon}{1-\delta}{l_o}^2\bigg|\phi_0\bigg) < \frac{1}{\Big[\frac{(1-\delta)cN}{2} - 1\Big]\epsilon^2} \tag{18}$$

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

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

Apply it to (18) we have

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

Or

$$Pr\Bigg(\bar{d}_0 > c_\epsilon l_o \;\Bigg|\; \phi_0\Bigg) < \frac{1}{\Big[\frac{(1-\delta)cN}{2} - 1\Big]\epsilon^2} \tag{19}$$

Let ϕ_1 be the event of $\bar{d}_0 \leq c_\epsilon l_o$. Then $P^1_e = Pr(\phi^c_1) = Pr(\phi^c_1\wedge\phi^c_0) + Pr(\phi^c_1\wedge\phi_0) \leq P^0_e + Pr(\phi^c_1|\phi_0)\cdot Pr(\phi_0) \leq P^0_e + Pr[\bar{d}_0 > c_\epsilon l_o|\phi_0]$. So $P^1_e \leq O(2^{-\theta(N)}) + \theta(\frac{1}{N})$, 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 $Pr(\phi^c) \leq Pr(\phi^c_0) + Pr(\phi^c_1) \leq O(2^{-\theta(N)}) + \theta(\frac{1}{N})$. So ϕ is true w.h.p. Apply this to (12),(13), also f is nondecreasing, we have w.h.p.

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

From $H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) \geq cN \cdot \log f(c_\epsilon l_o)$, get $H(\hat{X}_{\partial_0}|\hat{X}_{G_0}) = H(\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_N) - H(\hat{X}_{G_0}|\hat{X}_t) - H(\hat{X}_t) \geq \frac{1-\delta}{2}cN\cdot\log f(c_\epsilon l_o) - \theta(N) - \theta[\sqrt{N}\log f(c_\epsilon l_o)]$ 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 $H_w(G_{\hat{X}}) \geq \theta[N\sqrt{N}l^\alpha_o\log f(c_\epsilon l_o)]$ 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 asymptotically optimal w.h.p. for PLVCF.

■

For HCF, GMF fields with PPP deployment process, similar extensions of previous results can be derived applying the same technique.

V. Related work

There have been a huge amount of research in the field of distributed source coding and network coding. A full description of the current state is beyond the scope of this paper, we here briefly introduce some most related ones and refer the reader to more thorough surveys.

In [7], the authors introduce the concept of network coding, state and prove the Max-flow Min-cut theorem for network information flow. [5] gives a thorough review on DSC. [28] shows that random linear network coding suffices for the network coding of correlated sources. [29] first gives 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. [30] finds a similar entropy-capacity iff condition as [19] for a sensor incast problem: reproduce the whole sensor field at any one of the sensors. [31] studies the problem of network coding with a cost criterion. For minimal cost correlated data gathering, [32] 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. A universal random approximation on optimal transmission tree is given for all concave functions of the source model.

[33] studies the scaling problem of large number sensors deployed in a GMF by comparing the per node capacity and node data rate asymptotically. [2] compares SWC and EEC's asymptotic performance on a 1D grid and shows under various conditions EEC performs asymptotically as well as SWC, which we now know should be asymptotically optimal under these conditions based on our work. [34] propose a practical SWC scheme based on syndromes. They use a hamming distance constraint model so their result can be generalized to a hierarchical scheme similar to HDB and applicable to HCF but not LVCF or GMF. There is also no spatial or cost consideration in [34]. [35] investigates the problem of joint optimization of sensor nodes deployment and data gathering cost in a lossy setting. [36] also studies correlated sensor data collection on a grid. They use a simplified cost function for which the cost over the diagonal hops of longer distance is considered to be the same as the cost over the shorter vertical/horizontal hops, they also use a simplified correlation model that ignores spatial features as in [32]: the joint entropy is a linear function of the number of sources. Thus their discovery of optimal clustering size is consistent with [32]'s general result. Of particular interest to us, [36]'s experience equation learned from real rainfall spatial data verifies the validity of our generic source model LVCF. [37] models spatially correlated sources using real sensor data. Their model also falls within our model framework of LVCF and GMF thus further supports the generality of LVCF and GMF.

VI. Conclusion and Future Work

We introduce the concept of distance entropy and prove that it is a lower bound of the minimum communication cost for gathering distributed information. We introduce several generic classes of source models, and design one simple data collection scheme that is asymptotically optimal for all of them. The broad asymptotical success of HDB suggests a bolder conjecture that a dominant portion of most models' data redundancies can be removed by truly simple distributed computations.

Our work has many possibilities for future directions. For the optimal cost, there are work to do to incorporate considerations of channel coding, capacity constraints, multi-sinks, and lossy data collection. For asymptotically optimal schemes and general source models, fully characterize HDB's power by finding more network source models for which HDB is asymptotically optimal, particularly, evaluate HDB on other variants of gaussian models. Finally, considering the wireless medium's multiaccess property and thus a lower cost for broadcasting, both the theoretical optimal communication cost and HDB's cost need to be reevaluated under this case.

References

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

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

[3] 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.

[4] C. Chong and S. Kumar, "Sensor networks: Evolution, opportunities, and challenges," in *IEEE Symposium on Foundations of Computer Science*, pp. 1247–1256, 2003.

[5] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," in *IEEE Signal Processing Magazine*, pp. 522–533, September 2004.

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

[7] R. Ahlswede, N. Cai, S. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, pp. 1204–1216, July 2000.

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

[9] A. Lehman and E. Lehman, "Complexity classification of network information flow problems," in *ACM-SIAM SODA*, (Philadelphia, PA, USA), pp. 142–150, Society for Industrial and Applied Mathematics, 2004.

[10] A. Ramamoorthy, K. Jain, P. Chou, and M. Effros, “Separating distributed source coding from network coding,” in *Allerton Conference on Communication, Control, and Computing*, Oct. 2004.

[11] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Transactions on Information Theory*, vol. IT-19, pp. 471–480, July 1973.

[12] M. Adler, “Collecting correlated information from a sensor network,” in *Proceedings of ACM-SIAM 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*, pp. 292–301, March 2004.

[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*, pp. 292–301, March 2004.

[15] A. Liveris, Z. Xiong, and C. Georghiades, “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*, pp. 302–311, March 2004.

[17] S. Cheng and Z. Xiong, “Successive refinement for the wyner-ziv problem and layered code design,” in *Proc. DCC’04*, p. 531, March 2004.

[18] N. Megiddo, “Optimal flows in networks with multiple sources and sinks,” *Math. Programming*, vol. 7, pp. 97–107, 1974.

[19] T. Han, “Slepian-wolf-cover theorem for networks of channels,” *Information and Control*, vol. 47, no. 1, pp. 67–83, 1980.

[20] T. Cover, “A proof of the data compression theorem of slepian and wolf for ergodic sources,” *IEEE Transactions on Information Theory*, vol. IT-22, pp. 226–228, March 1975.

[21] J. Edmonds, “Submodular functions, matroids, and certain polyhedra.,” in *Combinatorial Optimization*, pp. 11–26, 2001.

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

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

[24] D. Bernstein, *Matrix Mathematics*. Pinceton University Press, 2005.

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

[26] B. Liu, P. Brass, O. Dousse, P. Nain, and D. Towsley, “Mobility improves coverage of sensor network,” in *ACM MobiHoc*, 2005.

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

[28] T. Ho, M. Médard, M. Effros, and R. Koetter, “Network coding for correlated sources,” in *CISS*, 2004.

[29] Y. Wu, V. Stankovic, Z. Xiong, and S. Kung, “On practical design for joint distributed source and network coding,” in *Proc. of the First Workshop on Network Coding, Theory and Applications*, 2005.

[30] J. Barros and S. D. Servetto, “Network information flow with correlated sources,” *IEEE Transactions on Information Theory*.

[31] D. Lun, M. edard, T. Ho, and R. Koetter, “Network coding with a cost criterion,” Tech. Rep. P-2584, MIT, 2004.

[32] A. Goel and D. Estrin, “Simultaneous optimization for concave costs: Single sink aggregation or single source buy-at-bulk,” in *Proceedings of ACM-SIAM SODA*, 2003.

[33] 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,” April 2003.

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

[35] D. Ganesan, R. Cristescu, and B. Beferull-Lozano, “Power-efficient sensor placement and transmission structure for data gathering under distortion constraints,” in *Proceedings of the third international symposium on Information processing in sensor networks (IPSN-04)*, (New York), pp. 142–150, ACM Press, Apr. 26–27 2004.

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

[37] A. Jindal and K. Psounis, “Modeling spatially-correlated sensor network data,” in *SECON’04*, 2004.