

# On the Distance Entropy of a Data Collection Network

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  $[\text{cost}] = [\text{data rate}] \times [\text{link weight}]$ , 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 optimal over all of the possible schemes. Our results imply that for collecting data from correlated sources at a single sink, Network Coding does not help in the sense of lowering the optimal communication cost. We then extend our results to the case that includes broadcast links in the network. We show that the same optimal cost holds even if we allow broadcasting. In other words, neither broadcasting nor Network Coding improves the total cost of collecting data from correlated sources at a single sink.

## 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 connects the network. We view this as a graph with sources forming a subset of the nodes and the links<sup>1</sup> represented as 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. 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 one fundamental question for this scenario:

- 1) *What is the minimum total communication cost for achieving the data gathering task, possibly under some link capacity constraints?*

<sup>1</sup>We consider both point to point links and broadcasting links.

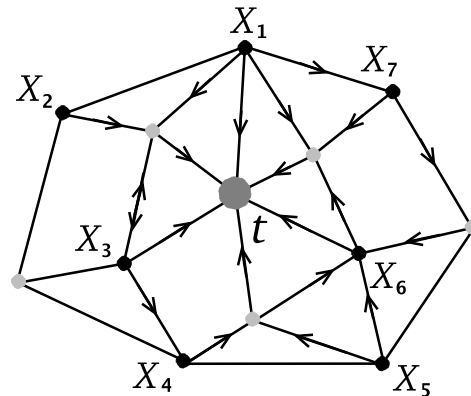


Fig. 1. A Layout of the General Problem of Gathering Correlated Data Through a Network

This is a difficult question because there are no limitations on what functions a node is allowed to apply to the incoming information and/or its local source to generate its outgoing messages. Fig. 1 shows an example data transmission scheme; an arrow on an edge indicates the actual traffic'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 to a subset of neighbors (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 of the source data. With this set up, we consider 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 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 whereas the general results apply to the Internet and any correlated data collection problem as well. Consider a sensor network that collects measurements from the environment at a single sink. We represent an observation at a sensor with 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  $(\hat{X}_1, \dots, \hat{X}_N)$  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*) of a link is a function of the transmission rate and the link weight between the two nodes. An often used, simplified cost function is the product of a rate term and a separable weight term. We consider both this simplified cost function as well as more general cost functions. The goal is to minimize the total communication cost of the whole network for the data collection task.

In our setting, the node operations consist 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 connect each of the source nodes directly to the sink without other links between the source nodes, the problem reduces to a Distributed Source Coding (DSC) [8] problem; when the sources are independent, the problem reduces to a Network Coding problem. If there is just a single sink in addition to independent sources, there is no need for Network Coding; [9] shows that traditional routing where data is treated as commodity flows suffices to solve the data collection problem for such networks. On the other hand, sensor information and data from 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 that NC is necessary (and not vice versa), we do know that there are cases where NC is mandatory. Thus further work is needed to determine the utility of NC in our situation.

There are various possible coding approaches for our data collection problem based on different combinations of SC and NC. The two most studied are based on Slepian-Wolf Coding

(SWC)<sup>2</sup> [11] and Explicit Entropy Coding (EEC)<sup>3</sup> [1]. For our data collection task, [1] shows that when the cost function is  $[\text{cost}] = [\text{rate}] \times [\text{weight}]$  and there are no link capacity constraints, the optimal rate allocation of SWC followed by shortest path commodity flow routing achieves the minimum cost for all schemes using only SWC. For EEC [1] shows that optimizing the total cost 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.

In summary, it is important to characterize the optimal communication cost in a general setting.

## 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, while also accounting for the underlying network topology of the source nodes.

For the case of  $[\text{cost}] = [\text{rate}] \times [\text{weight}]$  cost functions, we show that the distance entropy is a lower bound of the communication cost. For networks without capacity constraints, 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, there exists a SWC scheme that is optimal over the class of all schemes. A corollary of these results is that for collecting correlated sources at a single sink, network coding is not needed in order to minimize the total communication cost or maximize the achievable capacity. We then extend our optimal results to networks that incorporate broadcast channels. We show that the optimal cost is unchanged even if the nodes are allowed to broadcast.

## 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.  $E \subseteq V \times V$  is the edge set,  $e = (v_i, v_j) \in E$  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.<sup>4</sup> We first assume the links

<sup>2</sup>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.

<sup>3</sup>For EEC, a node sends out data with a rate equal to the joint entropy rate of incoming data and its own sensed data.

<sup>4</sup>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.

(channels) are independent point to point links. 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 arbitrarily large packets. 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 maximum 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  $\mathcal{P}_v$  and the shortest path from  $v$  to  $t$  as  $p_v^*$ . Then  $W(p_v^*) = \min_{p \in \mathcal{P}_v} W(p)$ .

Each source node  $v_i \in \Omega$  periodically generates samples from a discrete source  $\hat{X}_i$ . The joint source vector  $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_N\}$  follows some joint distribution  $p(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N)$ . Let  $\{\hat{X}(\tau)\}_{\tau=1}^\infty$  be a stationary random process where  $\hat{X}(\tau) = (\hat{X}_1(\tau), \dots, \hat{X}_N(\tau))$  is a field sample that corresponds to the set of samples gathered from all sources at time  $\tau, \tau = 1, 2, \dots$ . For simplicity of presentation, we analyze the total communication cost per second of collecting i.i.d. field samples with a sample rate of 1 sample/sec while our results can be easily 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  $\Omega$ . 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 and local generated data (if any) to its output bits and the corresponding selected channels. A Data Collection Scheme (DCS)  $\Upsilon$  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  $\Upsilon_{SWC}$  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  $\Upsilon_{SWC-SP}$  is a SWC scheme that only uses the shortest path commodity flow routing. Let  $\Pi, \Pi_{SWC}, \Pi_{SWC-SP}$  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  $\Upsilon$  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  $W_\Upsilon$ . 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  $\tau$  seconds is  $g(b/\tau, w_e) \cdot \tau = b \cdot w_e$ , independent of the transmitting period  $\tau$ . So in this

case we can equivalently study the communication cost for collecting one data sample, denoted also as  $W_\Upsilon$ . 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  $e$ .

### III. 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*

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

where  $W(p_j^*)$  is the shortest path weight of source  $\hat{X}_j$  and satisfies  $W(p_1^*) \leq W(p_2^*) \leq \dots \leq W(p_N^*)$ .

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  $\Upsilon$  on a source graph  $G_X$  to collect one field sample is lower bounded by the distance entropy of  $G_X$*

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

*In the absence of capacity constraints, a SWC-SP scheme with an optimal rates allocation  $r_j = H(\hat{X}_j | \hat{X}_{j-1}, \dots, \hat{X}_1)$  ( $v_1, v_2, \dots, v_N$  is in a nondecreasing order of shortest path weight) achieves the cost of  $H_w(G_X)$ . Thus*

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

The idea of the proof is to first group nodes into a sequence of sets according to their shortest path weights to the sink, then investigate the information flow across the cuts between adjacent sets in a constructed graph equivalently. This is different from [1]’s proof. They show that for SWC schemes finding the optimal rate is a Linear Programming problem, while our proof is for a more general result and thus has to start from a point with less assumptions and exploit the fundamental law of network information theory and graph theory. Due to space limitations, please refer to [12] for details. Theorem 1 shows that the distance entropy is a lower bound of the total communication cost. Furthermore, it shows that if there is no capacity constraints, distance entropy is an achievable tight bound, if we consider SWC scheme to be a valid scheme, then the best performance for such data collection tasks 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 [13]. Han [13] 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. [13] 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 [13] with cut capacity and also because the max-flow min-cut theorem for network flows also applies to an undirected graph, we generalize [13]'s model to any directed/undirected source graph that a source node can connect with any other 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  $\Lambda$ . 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 [13]) For any source graph  $G_X$  (directed or undirected) with an edge capacity set  $C$ , there exists a data collection scheme iff*

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

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

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

*and there exists a set of flows satisfying the capacity constraints from the source nodes  $\Omega$  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 [13] 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  $\Upsilon$  on it, let the average transmission rate by  $\Upsilon$  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  $\Upsilon$  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  $\Upsilon$  on it,  $\Upsilon$ 's data rate cross any cut  $r_M(\Upsilon)$  satisfies*

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

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

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

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

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

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

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  $\Upsilon$  is a DCS in  $G'_X$ . So the assumption is incorrect and the lemma is true.  $\blacksquare$

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

For edge  $e$ , denote  $\tau_e = \sum_{k=1}^{K_e} \tau_e^k$  and  $\lambda_e^k = \tau_e^k / \tau_e \in (0, 1]$ , then  $\sum_{k=1}^{K_e} \lambda_e^k = 1$  and  $r_e = \sum_{k=1}^{K_e} r_e^k \cdot \lambda_e^k$ .

*Theorem 3: Let  $G_X$  be an arbitrary source graph with or without capacity constraints. Let the cost function  $g$  be nondecreasing in  $w$  and  $r$  and convex in 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 the Theorem by showing that for any data collection scheme  $\Upsilon$ , there exists at least one SWC scheme that has a communication cost no greater than that of  $\Upsilon$ . The trick is to treat the actual transmission rate generated by  $\Upsilon$  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  $\Upsilon$ . We treat  $\{r_{(i,j)}\}$ s as capacities of the directed edges in  $G'$  i.e.  $c'_{ij} = r_{(i,j)} \leq c_{ij}$  for  $(v_i, v_j) \in E'$  and  $C'(M, M^c) = r_M(\Upsilon) \leq C(M, M^c)$  for any cut  $(M, M^c)$ ; also we have  $r_M(\Upsilon) \geq H(\hat{X}_M | \hat{X}_M^c)$  by Lemma 1. So for any cut  $(M, M^c)$ ,

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

$C'(M, M^c)$  matches the iff condition of Theorem 2, then by Theorem 2 there exists a SWC scheme with a SWC rate vector  $R' = (r'_1, r'_2, \dots, r'_N)$  that satisfies  $H(\hat{X}_M | \hat{X}_M^c) \leq \sum_{v_i \in M \cap \Omega} r'_i \leq C'(M, M^c)$  for any cut  $(M, M^c)$ , and there exists a set of flows  $F = (f_1, f_2, \dots, f_N)$  from  $\Omega$  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 [14] 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  $\Upsilon$  — a DCS in  $G$ .<sup>5</sup>

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

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

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

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

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

#### 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. We use the same source model as before and a slightly modified communication model to incorporate broadcast channels. First we describe the communication model and then show that 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 broadcast channel. Let  $\mathcal{N}(v_i) = \{v_j | (v_i, v_j) \in E\}$  be the neighbor set of node  $v_i$ —the set of nodes that  $v_i$  can communicate directly via a point to point channel. Broadcasting here means  $v_i$  can send the same copy of data simultaneously at a rate  $r$  to any subset of its neighbor set  $B \subset \mathcal{N}(v_i)$ . The energy cost  $g_{i,B}(r)$  of broadcasting is no less than the cost of sending at the same rate from  $v_i$  to any of the nodes in  $B$  through a point to point channel:

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

This assumption is valid for both applications using directional antennas and ones using omni-directional antennas for the point to point channels.<sup>6</sup> Also for any  $v_j \in B$ ,  $r$  satisfies the capacity constraint  $r \leq c_{ij}$  and broadcasting consumes  $r$  of the capacity of the point to point channel from  $v_i$  to  $v_j$ .

<sup>5</sup>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.

<sup>6</sup>For same type of antennas, directional ones consume less energy than omni-directional ones for point to point communications.

2) *Optimality Result*: With the modified communication model, now we refer to the previously defined DCSs that do not use broadcasting as none broadcast schemes and still use  $\Pi$  to denote the set of all none broadcast schemes; we refer to the DCSs using broadcasting or not as broadcast enabled DCSs and denote the set of all broadcast enabled DCSs as  $\Pi_B$ . We show that any source graph  $G_X$  whose nodes are enhanced with this broadcast capability has the same optimal cost as the none broadcast scheme. We state and prove the following theorem.

*Theorem 4*: Let  $G_X$  be an arbitrary source graph with or without capacity constraints. Let the cost function  $g$  be nondecreasing in  $w$  and  $r$  and convex in rate  $r$ , then the optimal SWC scheme that does not use broadcasting is also optimal over the class of all broadcast enabled data collection schemes.

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

*Proof*: We prove the theorem by showing that for any broadcast enhanced data collection scheme  $\Upsilon$  in  $G_X$ , there exists a SWC scheme that has a cost that is no greater than  $\Upsilon$  and does not use broadcasting.

For any broadcasting enhanced DCS  $\Upsilon$  for  $G_X$ , we define  $r_M^B(\Upsilon)$  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  $G_X$ ,  $r_M^B(\Upsilon) \geq H(\hat{X}_M | \hat{X}_M^c)$ . We do this by shrinking  $G_X$  to a simple source graph consisting 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  $\Upsilon$  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  $\Upsilon$  are now all achievable internal coding operations inside  $s_M$  or  $t_M$  in the new source graph, any DCS  $\Upsilon$  in  $G_X$  corresponds to a DCS in the new source graph with a rate from  $s_M$  to  $t_M$  as  $r_{s_M} = r_M^B(\Upsilon)$ . By Lemma 1  $r_{s_M} \geq H(\hat{X}_M | \hat{X}_M^c)$  thus  $r_M^B(\Upsilon) \geq H(\hat{X}_M | \hat{X}_M^c)$  for any cut  $(M, M^c)$ .

Next we construct a new source graph  $G_X^\Upsilon$  based on  $G_X$  and  $\Upsilon$ . The first part of the construction is similar to the one in the proof of Theorem 3:  $G_X^\Upsilon$  contains all vertices in  $G_X$  and for all the non-broadcast traffic of  $\Upsilon$ , add a directed edge along the traffic direction with a capacity equal to the traffic rate. For each broadcasting traffic of  $\Upsilon$  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  $r_M^B(\Upsilon) \geq H(\hat{X}_M | \hat{X}_M^c)$  in  $G_X$ , it is easy to verify that for any cut  $(M, M^c)$  in  $G_X^\Upsilon$ , 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  $\Upsilon_{SWC}$  in  $G_X^\Upsilon$ . If we copy this  $\Upsilon_{SWC}$  to  $G_X$  by distributing the flow traffic of  $v_i \rightarrow v_{i,B} \rightarrow v_j$  directly as  $v_i \rightarrow v_j$ , by the construction of  $G_X^\Upsilon$ , we obtain a non-broadcast DCS  $\Upsilon'$  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  $\Upsilon'$  in  $G_X$  is also a non-broadcast DCS with a cost no greater than the broadcasting enhanced DCS  $\Upsilon$ . This is true for any broadcasting enhanced DCS  $\Upsilon$  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 optimal over the set of all data collection schemes. The result is not obvious because the intermediate nodes are allowed to perform operations that involve arbitrary couplings of network coding and source coding. A key part of the proofs relies on some combinatorial geometric properties of submodular and supermodular functions based on Edmonds's result in [15]. 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 network coding does not help 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. We should note that SWC can hardly be considered a practical code. Nevertheless, SWC scheme is a useful theoretical scheme that helps us understand the performance limit of the data collection task.

#### IV. CONCLUSION AND FUTURE WORK

We introduced the concept of distance entropy and proved that it is a lower bound of the minimum communication cost for gathering distributed information. We exploited both the combinatory optimization and information theory to achieve the minimum data gathering cost. The combinatory part involves with analyzing packing problems (e.g. flows) that are constrained by the graph structure. This theory grew out of a need to understand the shipment of cargo in transportation networks and does not capture the subtleties of information transmission. On the other hand, information theory provides a deep understanding of complex communication problems over structurally simple channels but does not yet fully extend to arbitrary graph structures. Combining ideas from both of these theories allow us to make significant progress on understanding the performance limit of information networks.

Our future work is to study the optimal cost for the more general case of collecting correlated sources to multiple sinks, incorporating considerations of capacity constraints, broadcasting channels and lossy data collection.

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

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

[14] 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.

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