The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks

SUNDEEP PATTEM and BHASKAR KRISHNAMACHARI
Dept. of Electrical Engineering-Systems
University of Southern California
and
RAMESH GOVINDAN
Dept. of Computer Science
University of Southern California

The efficacy of data aggregation in sensor networks is a function of the degree of spatial correlation in the sensed phenomenon. The recent literature has examined a variety of schemes that achieve greater data aggregation by routing data with regard to the underlying spatial correlation. A well known conclusion from these papers is that the nature of optimal routing with compression depends on the correlation level. In this work, we show the existence of a simple, practical and static correlation-unaware clustering scheme that satisfies a min-max near-optimality condition. The implication for system design is that a static correlation-unaware scheme can perform as well as sophisticated adaptive schemes for joint routing and compression.

Categories and Subject Descriptors: C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Distributed networks*; I.6 [**Computing Methodologies**]: Simulation and Modeling

General Terms: Design, Performance

Additional Key Words and Phrases: Sensor Networks, Correlated Data Gathering, Analytical Modeling

1. INTRODUCTION

In view of the severe energy constraints of sensor nodes, data aggregation is widely accepted as an essential paradigm for energy-efficient routing in sensor networks. For data-gathering applications in which data originates at multiple correlated sources and is routed to a single sink, aggregation would primarily involve in-network compression of the data. Such compression, and its interaction with routing, has been studied in the literature before; prior work has examined distributed source coding techniques such as Slepian-Wolf coding [Cover and Thomas 1991; Pradhan and Ramchandran 1999], joint source coding and routing techniques [Scaglione and Servetto 2005], and opportunistic compression along the

This work was supported in part by NSF grants numbered 0435505, 0347621, 0430061, 0325875.

shortest path tree [Krishnamachari et al. 2002]. An understanding of various routing schemes across the range of spatial correlations is crucial and this problem has been addressed by several recent papers [Pattem et al. 2004; Cristescu et al. 2004; Enachescu et al. 2004]. Cristescu *et al.* have formalized the correlated data gathering problem and studied the interaction between the correlation in the data measured at nodes in a network and the transmission structure that is used to transport this data to the sink.

In order to understand the space of interactions between routing and compression, we study simplified models of three qualitatively different schemes. In *routing-driven compression* data is routed through shortest paths to the sink, with compression taking place opportunistically wherever these routes happen to overlap [Intanagonwiwat et al. 2002] [Krishnamachari et al. 2002]. In *compression-driven routing* the route is dictated in such a way as to compress the data from all nodes sequentially - not necessarily along a shortest path to the sink. Our analysis of these schemes shows that they each perform well when there is low and high spatial correlation respectively. As an ideal performance bound on joint routing-compression techniques, we consider *distributed source coding* in which perfect source compression is done *a priori* at the sources using complete knowledge of all correlations.

In order to obtain an application-independent abstraction for compression, we use the joint entropy of sources as a measure of the uncorrelated data they generate. An empirical approximation for the joint entropy of sources as a function of the distance between them is developed. A bit-hop metric is used to quantify the total cost of joint routing with compression. Evaluation of the above schemes using these metrics leads naturally to a clustering approach for schemes that perform well over the range of correlations.

We develop a simple scheme based on static, localized clustering that generalizes these techniques. Analysis shows that the nature of optimal routing will depend on the number of nodes, level of correlation and also on where the compression is effected; at the individual nodes or at intermediate aggregation points (cluster heads). Our main contribution is a surprising result that there exists a near-optimal cluster size that performs well over a wide range of spatial correlations. A min-max optimization metric for the near-optimal performance is defined and a rigorous analysis of the solution is presented for both 1-D (line) and 2-D (grid) network topologies. We show further that this near-optimal size is in fact asymptotically optimal in the sense that, for any constant correlation level, the ratio of the energy costs associated with the near-optimal cluster size to those associated with the optimal clustering goes to one as the network size increases. Simulation experiments confirm that the results hold for more general topologies - 2-D random geometric graphs and realistic wireless communication topology with lossy links, and also for a continuous, Gaussian data model for the joint entropy with varying quantization.

From a system-engineering perspective, this is a very desirable result because it eliminates the need for highly sophisticated compression-aware routing algorithms that adapt to changing correlations in the environment (which may even incur additional overhead for adaptation), and therefore simplifies the overall system design.

2. ASSUMPTIONS AND METHODOLOGY

Our focus is on applications which involve continuous data gathering for large scale and distributed physical phenomena using a dense wireless sensor network where joint routing and compression techniques would be useful. An example of this is the collection of data from a field of weather sensors. If the nodes are densely deployed, the readings from nearby nodes are likely to be highly correlated and hence contain redundancies, because of the inherent smoothness or continuity properties of the physical phenomenon.

To compare and evaluate different routing with compression schemes, we will need a common metric. Our focus is on energy expenditure, and we have therefore chosen to use the bit-hop metric. This metric counts the total number of bit transmissions in the network for one round of gathering data from all sources. Formally, let $T = (V, E, \xi_T)$ represent the directed aggregation tree (a subgraph of the communication graph) corresponding to a particular routing scheme with compression, which connects all sources to the sink. Associated with each edge $e = (u, v)$ is the expected number of bits b_e to be transported over that edge in the tree (per cycle). For edges emanating from sources that are leaves on the tree, the bit count is the amount of data generated by a single source. For edges emanating from aggregation points, the outgoing edge may have a smaller bit count than the sum of bits on the incoming edges, due to aggregation. For nodes that are neither sources or aggregation points but act solely as routers, the outgoing edge will contain the same number of bits as the incoming edge. The bit-hop metric ξ_T is simply:

$$\xi_T = \sum_{e \in E} b_e. \tag{1}$$

There are two possible criticisms of this metric that we should address directly. The first is that the total transmissions may not capture the "hot-spot" energy usage of bottleneck nodes, typically near the sink. However, an alternative metric that better captures hot-spot behavior is not necessarily relevant if the *a priori* deployment and energy placement ensure that the bottlenecks are not near the sink or if the sink changes over time. The second possible criticism is that this does not incorporate reception costs explicitly. However, the use of bit-hop metric is justified because it does in-fact implicitly incorporate reception costs. If every bit transmission incurs the same corresponding reception cost in the network, the sum of these transmission and reception costs will be proportional to the total number of bit-hops.

To quantify the bit-hop performance of a particular scheme, therefore, we need to quantify the amount of information generated by sources and by the aggregation points after compression. For this purpose we use the entropy H of a source, which is a measure of the amount of information it originates [Cover and Thomas 1991]. In this paper, we consider only lossless compression of data. In order to characterize correlation in an application-independent manner, we use the joint entropy of multiple sources to measure the total uncorrelated data they originate. Theoretically, using entropy-coding techniques this is the maximum possible lossless compression of the data from these sources. We now attempt to construct a parsimonious model to capture the essential nature of joint entropy of sources as

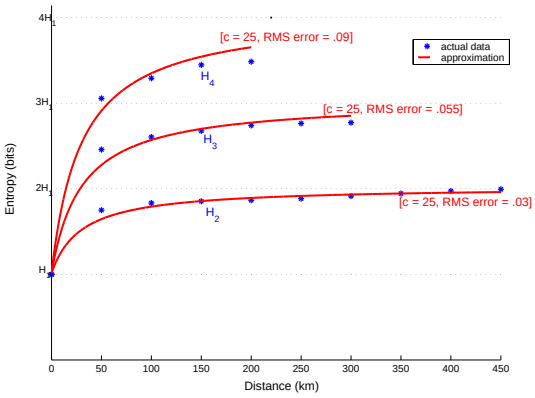


Fig. 1. Empirical data (from the rainfall data-set) and approximation for joint entropy of linearly placed sources separated by different distances

a function of distance. The simplicity of this approximation model enables the analysis presented in Sections 3 and 4.

In general, the extent of correlation in the data from different sources can be expected to be a function of the distance between them. We used an empirical data-set pertaining to rainfall[1] [Widmann and Bretherton 1999] to examine the amount of correlation in the readings of two sources placed at different distances from each other. Since rainfall measurements are a continuous valued random variable and hence would have infinite entropy, we present results obtained from quantization. The range of values was normalized for a maximum value of 100 and all readings 'binned' into intervals of size 10. Fig.1 is a plot of the average joint entropy of multiple sources as a function of inter-source distance.

The figure shows a steeply rising convex curve that reaches saturation quickly. This is expected since the inter-source distance is large (in multiples of 50km). From the empirical curve, a suitable model for the average joint entropy of two sources (H_2) as a function of inter-source distance d is obtained as:

$$H_2(d) = H_1 + [1 - \frac{1}{(\frac{d}{c} + 1)}]H_1. \quad (2)$$

Here c is a constant that characterizes the extent of spatial correlation in the data. It is chosen such that when $d = c$, $H_2 = \frac{3}{2}H_1$. In other words, when inter-source distance $d = c$, the second source generates half the first node's amount in terms of uncorrelated data. In Fig.1, a value of $c = 25$ matches the H_2 curve well.

Finally, this leaves open the question of how to obtain a general expression for the joint entropy of n sources at arbitrary locations. As we shall show later, this is needed in order to study the performance of various strategies for combined routing

[1]This data-set consists of the daily rainfall precipitation for the pacific northwest region over a period of 46 years. The final measurement points in the data-set formed a regular grid of 50km x 50km regions over the entire region under study. Although this is considerably larger-scale than the sensor networks of interest to us, we believe the use of such "real" physical measurements to validate spatial correlation models is important.

and compression. To this end, we now present a constructive technique to calculate approximately the total amount of uncorrelated data generated by a set of n nodes.

From Eqn.2, it appears that on average, each new source contributes an amount of uncorrelated data equal to $[1 - \frac{1}{(\frac{d}{c}+1)}]H_1$, where we take the d as the minimum distance to an existing set of sources. This suggests a constructive iterative technique to calculate approximately the total amount of uncorrelated data generated by a set of n nodes:

(1) initialize a set $S_1 = \{v_1\}$ where v_1 is any node. We will denote by $H(S_i)$ the joint entropy of nodes in set S_i; where $H(S_1) = H_1$. Let V be the set of all nodes.

(2) Iterate the following for $i = 2 : n$
   (a) Update the set by adding a node v_i where $v_i \in V \setminus S_{i-1}$ is the closest (in terms of Euclidean distance) of the nodes not in S_{i-1} to any node in S_{i-1}, i.e. set $S_i = \{S_{i-1}, v_i\}$.
   (b) Let d_i be the shortest distance between v_i and the set of nodes in S_{i-1}. Then calculate the joint entropy as $H(S_i) = H(S_{i-1}) + [1 - \frac{1}{(\frac{d_i}{c}+1)}]H_1$.

(3) The final iteration yields $H(S_n)$ as an approximation of H_n.

In the simple case when all nodes are located on a line equally spaced by a distance d, this procedure would yield the expression:

$$H_n(d) = H_1 + (n-1)[1 - \frac{1}{(\frac{d}{c}+1)}]H_1. \quad (3)$$

That this closed-form expression provides a good approximation for a linear scenario is validated by our measurements from the rainfall data set, as seen in Fig.1. The curve for H_3 was obtained by considering all sets of grid points $(p1, p2, p3)$ such that they lie in a straight line with the distance between two adjacent points plotted on the x-axis. The curve for H_4 was similarly obtained using all sets of 4 points.

2.1 Note on Heuristic Approximation

We note that the final approximation $H(S_n)$ is guaranteed to be greater than the true joint entropy $H(v_1, v_2, ..., v_n)$. Thus it does represent a rate achievable by lossless compression. The approximation roughly corresponds to a rate allocation of $H(v_i/\eta_{v_i})$ at every node v_i, where η_{v_i} is the nearest neighbor of v_i. A more precise information-theoretic treatment in terms of the rate allocations at each node is possible, for instance, as in [Cristescu et al. 2004a;2006b]. We relinquish some rigor with the objective of gaining practical insight. This approach makes the problem more tractable and is the basis for analysis in subsequent sections.

Another point of contention is the need for such a heuristic approach instead of using a continuous data model and using analytical expressions for the joint entropy for this model. In this regard, we note that (a) our model matches the standard jointly Gaussian entropy model for low correlation [Appendix A.1.1] and (b) since the standard expression is in covariance form, it cannot be used for high correlation values, necessitating a reasonable approximation.

3. ROUTING SCHEMES

Given this framework, we can now evaluate the performance of different routing schemes across a range of spatial correlations. We choose three qualitatively different routing schemes; these schemes are simplified *models* of schemes that have been proposed in the literature.

(1) Distributed Source Coding (DSC): If the sensor nodes have perfect knowledge about their correlations, they can encode/compress data so as to avoid transmitting redundant information. In this case, each source can send its data to the sink along the shortest path possible without the need for intermediate aggregation. Since we ignore the cost of obtaining this global knowledge, our model for DSC is very idealized and provides a baseline for evaluating the other schemes.

(2) Routing Driven Compression (RDC): In this scheme, the sensor nodes do not have any knowledge about their correlations and send data along the shortest paths to the sink while allowing for opportunistic aggregation wherever the paths overlap. Such shortest path tree aggregation techniques are described, for example, in [Intanagonwiwat et al. 2002] and [Krishnamachari et al. 2002].

(3) Compression Driven Routing (CDR): As in RDC, nodes have no knowledge of the correlations but the data is aggregated close to the sources and initially routed so as to allow for maximum possible aggregation at each hop. Eventually, this leads to the collection of data removed of all redundancy at a central source from where it is sent to the sink along the shortest possible path. This model is motivated by the scheme in [Scaglione and Servetto 2005].

3.1 Comparison of the schemes

Consider the arrangement of sensor nodes in a grid, where only the $2n-1$ nodes in the first column are sources. We assume that there are n_1 hops on the shortest path between the sources and the sink. For each of the three schemes, the paths taken by data and the intermediate aggregation are shown in Fig.2.

In our analysis, we ignore the costs associated for each compressing node to learn the relevant correlations. This cost is particularly high in DSC where each node must learn the correlations with all other source nodes. However the bit-hop cost still provides a useful metric for evaluating the performance of the various schemes and allows us to treat DSC as the optimal policy providing a lower-bound on the bit-hop metric.

Using the approximation formulae for joint entropy and the bit-hop metric for energy, the expressions for the energy expenditure (E) for each scheme are as follows.

For the idealized DSC scheme, each source is able to send exactly the right amount of uncorrelated data, and each source can send the data along the shortest path to the sink, so that:

$$E_{DSC} = n_1 H_{2n-1}. \tag{4}$$

LEMMA 3.1. *E_{DSC} represents a lower bound on bit-hop costs for any possible routing scheme with lossless compression.*

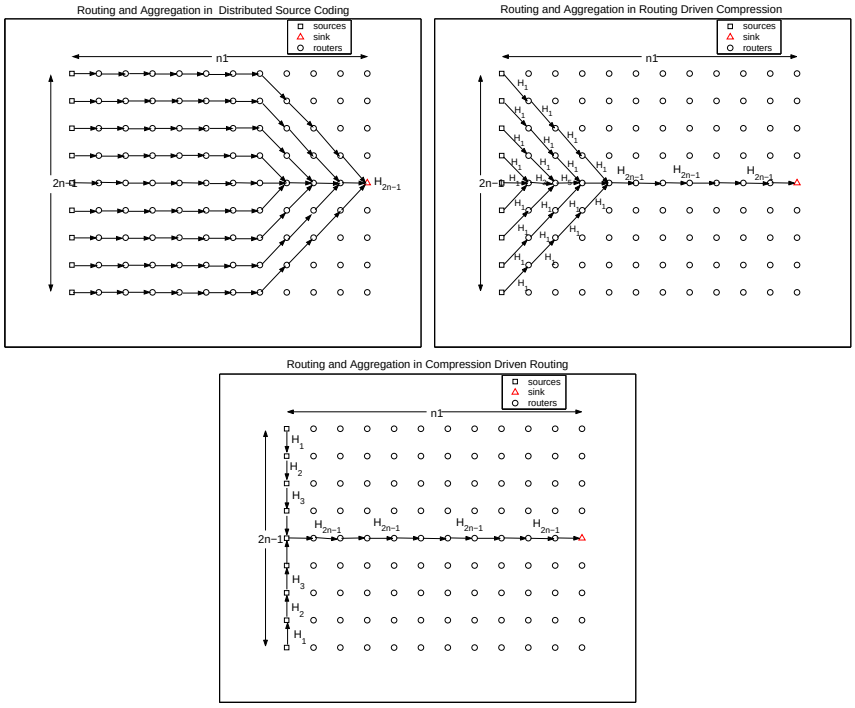


Fig. 2. Illustration of routing for the three schemes: DSC, CDR, and RDC. H_i is the joint entropy of i sources.

PROOF. *The total joint information of all* $(2n-1)$ *sources is* H_{2n-1}. *As discussed before, no lossless compression scheme can reduce the total information transmitted below this level. Each bit of this information must travel at least* n_1 *hops to get from any source to the sink. Thus* $n_1 H_{2n-1}$, *the cost of the idealized DSC scheme, represents a lower bound on all possible routing schemes with lossless compression.* □

In the RDC scheme, the tree is as shown in Fig.2 (middle), with data being compressed along the spine in the middle. It is possible to derive an expression for this scenario:

$$E_{RDC} = (n_1 - n)H_{2n-1} + 2H_1 \sum_{i=1}^{n-1}(i) + \sum_{j=0}^{n-2} H_{2j+1}. \tag{5}$$

For the CDR scheme, the data is compressed along the location of the sources, and then sent together along the middle, as shown in Fig. 2. It can be shown that for this scenario:

$$E_{CDR} = n_1 H_{2n-1} + 2\sum_{i=1}^{n-1} H_i. \tag{6}$$

The above expressions, in conjunction with the expression for H_n presented earlier, allow us to quantify the performance of each scheme. Fig.3 plots the energy expenditure for the DSC, RDC and CDR schemes as a function of the correlation

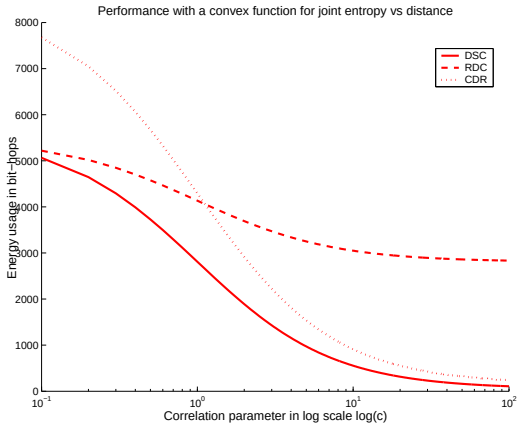


Fig. 3. Comparison of energy expenditures for the RDC, CDR and DSC schemes with respect to the degree of correlation c.

constant c, for different forms of the correlation function. For these calculations, we assumed a grid with $n_1 = n = 53$ and $2n - 1 = 105$ sources. From this figure it is clear that CDR approaches DSC and outperforms RDC for higher values of c (high correlation) while RDC performance matches DSC and outperforms CDR for low c (no correlation). This can be intuitively explained by the tradeoff between compressing close to the sources and transporting information toward the sink. CDR places a greater emphasis on maximizing the amount of compression close to the sources, at the expense of longer routes to the sink, while RDC does the reverse. When there is no correlation in the data (small c), no compression is possible and hence it is RDC that minimizes the total bit-hop metric. When there is high correlation (large c), significant energy gains can be realized by compressing as close to the sources as possible and hence CDR performs better under these conditions.

Interestingly, it appears that neither RDC nor CDR perform well for intermediate correlation values. This suggests that in this range a hybrid scheme may provide energy-efficient performance closer to the DSC curve. CDR and RDC can be viewed as two extremes of a clustering scheme, with CDR having all data sources form a single aggregation cluster before sending data towards the sink while RDC has each source acting as a separate cluster in itself. A hybrid scheme would be one in which sources form small clusters and data is aggregated within them at a cluster head, which then sends data to the sink along a shortest path. This insight leads us to an examination of suitable clustering techniques.

4. A GENERALIZED CLUSTERING SCHEME

The idea behind using clustering for data routing is to achieve a tradeoff between aggregating near the sources and making progress towards the sink. In addition to factors like number of nodes and position of sink, the optimal cluster size will also depend on the amount of correlation in the data originated by the sources (quantified by the value of c). Generally, the amount of correlation in the data is highest for sensor nodes located close to each other and can be expected to decrease

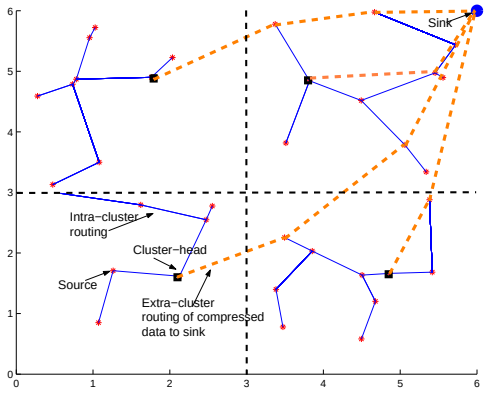


Fig. 4. Illustration of clustering for a two-dimensional field of sensors

as the separation between nodes increases. Once an optimal clustering based on correlations is obtained, aggregation of data is required only for the sources within a cluster, after which data can be routed to the sink without the need for further aggregation. As a consequence, none of the scenarios considered henceforth will resemble RDC exactly.

4.1 Description of the scheme

We now describe a simple, location-based clustering scheme. Given a sensor field and a cluster size, nodes close to each other form clusters. The clusters so formed remain static for the lifetime of the network. Within each cluster, the data from each of the nodes is routed along a shortest path tree (SPT) to a cluster head node. This node then sends the aggregated data from its cluster to the sink along a multi-hop path with no intermediate aggregation. This is illustrated in Fig. 4. The intermediate nodes on the SPT may or may not perform aggregation. Data aggregation in the form of compression is computationally intensive. All nodes in a network might not be capable of performing compression, either because it is too expensive for them to do so or the delays involved are unacceptable. It is conceivable that there will be a few high power nodes or micro-servers [Hu et al. 2004] which will perform the compression. Nodes form clusters around these nodes and route data to them. In this case, data aggregation takes place only at the cluster head.

4.1.1 *Metrics for evaluation of the scheme.* $E_s(c)$ is defined as the energy cost (in bit-hops) for correlation c and cluster size s. The optimal cluster size $s_{opt}(c)$ minimizes the cost for a given c. Let $E^*(c) = E_{s_{opt}}(c)$ represent the optimal energy cost for a given correlation c. For simplifying system design, it is desirable to have a cluster size that performs close to the optimal over the range of c values. We quantify the notion of 'being close to optimal' by defining a ***near-optimal cluster size*** s_{no} as the value of s that minimizes the maximum difference metric, i.e.

$$s_{no} = \arg\min_{s\in[1,n]} \max_{c\in[0,\infty)} \{E_s(c) - E^*(c)\}. \tag{7}$$

In the following sections, we analyze the performance of the clustering scheme for both 1-D and 2-D networks when aggregation is performed

—at intermediate nodes on the SPT, and
—only at the cluster-heads.

4.2 1-D Analysis

We begin with an analysis of the energy costs of clustering for a setup involving a linear array of sources to better understand the tradeoffs. Consider n source nodes linearly placed with unit spacing (i.e. $d = 1$) on one side of a 2-D grid of nodes, with the sink on the other side, and assuming the correlation model, $H_n = H_1(1+\frac{(n-1)}{1+c})$. We consider $\frac{n}{s}$ clusters each consisting of s nodes. Since all sources have the same shortest hop distance to the sink, the position of the cluster head within a cluster has no effect on the results. Within each cluster, the data can either be compressed sequentially on the path to the cluster head or only when it reaches the cluster head. The cluster head then sends the compressed data along a shortest path involving D hops to the sink. The total bit-hop cost for such a routing scheme is therefore

$$E_s(c) = \frac{n}{s}(E_{s,c}^{intra} + E_{s,c}^{extra}), \tag{8}$$

where $E_{s,c}^{intra}$ and $E_{s,c}^{extra}$ are the bit-hop cost within each cluster and the bit-hop cost for each cluster to send the aggregate information to the sink respectively.

4.2.1 *Sequential compression along SPT to cluster head.* At each hop within the cluster, a node receives H_i bits, aggregates them with its own data and transmits H_{i+1} bits. This is done sequentially until the data reaches the cluster head. We have,

$$\begin{aligned} E_{s,c}^{intra} &= \sum_{i=1}^{s-1} H_i = \sum_{i=1}^{s-1} \Big(1 + \frac{i-1}{1+c}\Big) H_1 \\ &= \Big(s - 1 + \frac{(s-2)(s-1)}{2(1+c)}\Big) H_1. \end{aligned}$$

Since the cluster heads get aggregated data from s sources and send it to the sink using a shortest path of D hops,

$$\begin{aligned} E_{s,c}^{extra} &= H_s \cdot D = \Big(1 + \frac{s-1}{1+c}\Big) H_1 \cdot D \\ \Rightarrow E_s(c) &= nH_1\Big(\frac{s-1}{s} + \frac{(s-2)(s-1)}{2s(1+c)} + \frac{D}{s} + \frac{(s-1)D}{s(1+c)}\Big). \end{aligned} \tag{9}$$

The optimum value of the cluster size s_{opt} can be determined by setting the derivative of the above expression equal to zero. It can be shown that

$$\begin{aligned} s_{opt} &= 1, && \text{if } c \le \frac{1}{2(D-1)} \\ &= \sqrt{2c(D-1)}, && \text{if } \frac{1}{2(D-1)} < c < \frac{n^2}{2(D-1)} \\ &= n, && \text{if } c \ge \frac{n^2}{2(D-1)}. \end{aligned}$$

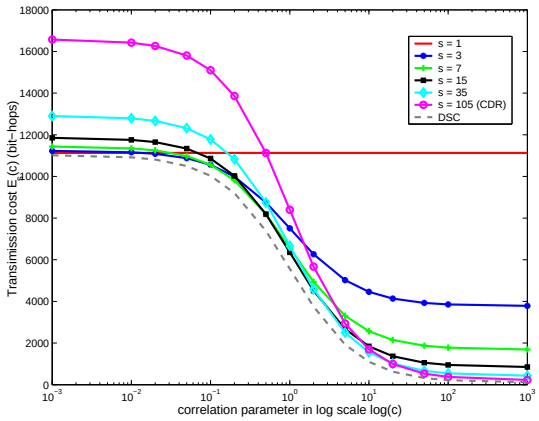


Fig. 5. Comparison of the performance of different cluster-sizes for linear array of sources($n = D = 105$) with compression performed sequentially along the path to cluster heads. The optimal cluster size is a function of correlation parameter c. Also, cluster size $s = 15$ performs close to optimal over the range of c

Note that s_{opt} depends on the distance from the sources to the sink[2] and the degree of correlation c.

Fig.5 shows (based on the analysis) how different cluster sizes perform across a range of correlation levels, based on the analysis presented above for a set of 105 linearly placed nodes. As expected the small cluster sizes and large cluster sizes perform well at low and high correlations respectively. However, it appears that an intermediate cluster size near 15 would perform well across the whole range of correlation values. The curve with $s = 105$ corresponds to CDR and the DSC curve is also plotted for reference.

THEOREM 4.1. *For $E_s(c)$ given by Equation.9, the near-optimal cluster size s_{no} defined by Equation.7 exists, and is given by*

$$s_{no} \;=\; \Theta(\min(\sqrt{D}, n)).$$

Proof is in Appendix A.2.2

This is illustrated in Fig.6, in which the costs are plotted with respect to the cluster sizes for a few different values of the spatial correlation. The figure shows clearly that although the optimal cluster size does increase with correlation level, the near-optimal static cluster size performs very well across a range of correlation values. In this figure, $D = n = 105$ and the near-optimal cluster size obtained from Theorem.4.1, $s_{no} = 14$ is indicated by the vertical line in the plot. Intersections of the dotted lines and the nearest c curve with this vertical line show the difference in energy cost between the near-optimal and optimal solutions.

[2]It is, however, assumed that $D \geq n$, so there is an implicit dependence on n.

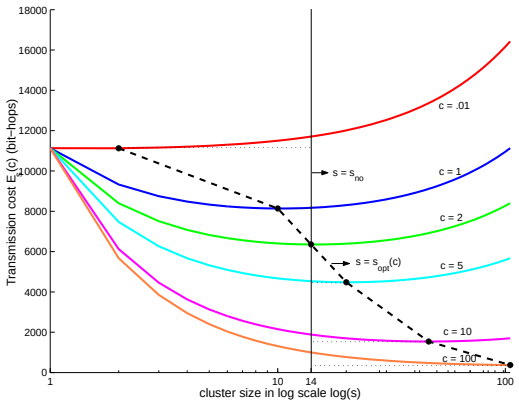


Fig. 6. Illustration of the existence of a static cluster for near-optimal performance across a range of correlations. The sources are in a linear array and data is sequentially compressed along the path to cluster heads.

4.2.2 *Compression at cluster head only.* In this case, each source within a cluster sends data to the cluster head using a shortest path. There is no aggregation before reaching the cluster head. We have,

$$\begin{aligned} E_{s,c}^{intra} &= \sum_{i=1}^{s-1} i \cdot H_1 = \frac{s(s-1)}{2} H_1 \\ E_{s,c}^{extra} &= \Big(1 + \frac{s-1}{1+c}\Big) H_1 \cdot D \\ \Rightarrow E_s(c) &= nH_1 \Big(\frac{s-1}{2} + \frac{D}{s} + \frac{(s-1)D}{(s)(1+c)}\Big). \end{aligned} \tag{10}$$

It can be shown that

$$\begin{aligned} s_{opt} &= 1, && \text{if } c \leq \frac{1}{2D-1} \\ &= n, && \text{if } c > \frac{n^2}{2D-n^2}, 2D > n^2 \\ &= \sqrt{\frac{2Dc}{c+1}}, && \text{else .} \end{aligned}$$

Fig.7 shows that for a linear array of sources (with $n = D = 105$), the performance for cluster sizes $s = 5, 7$ are close to optimal over the range of c. The DSC curve is plotted for reference.

THEOREM 4.2. *For $E_s(c)$ given by Equation.10, the near-optimal cluster size s_{no} defined by Equation.7 exists, and is given by*

$$s_{no} = \Theta(\min(\sqrt{D}, n))$$

.

Proof is in Appendix A.2.4

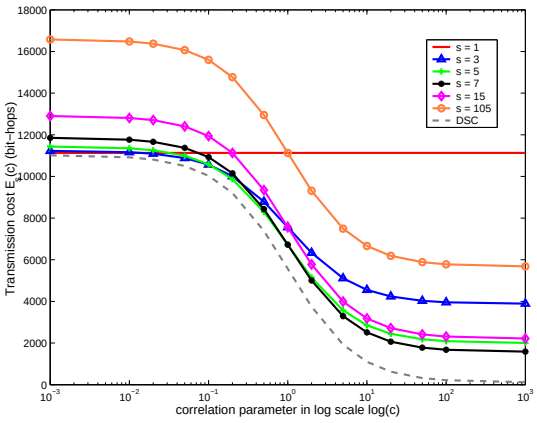


Fig. 7. Performance with compression only at cluster head with nodes in a linear array($n = D = 105$). Cluster sizes $s = 5, 7$ are close to optimal over the range of c

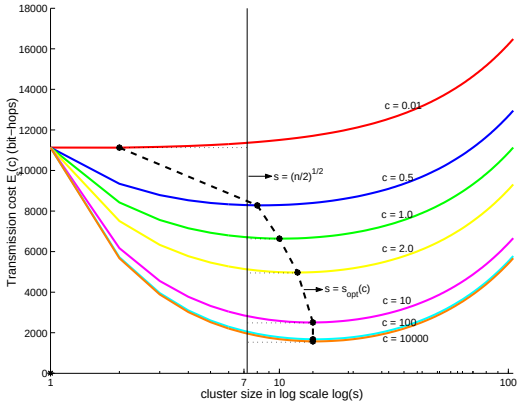


Fig. 8. Illustration of the near-optimal cluster size with compression only at cluster head with nodes in a linear array. The performance of cluster sizes near $s = 7(\approx \sqrt{\frac{105}{2}})$ is close to optimal over the range of c values

The existence of a near-optimal cluster size is illustrated in Fig. 8. The performance of cluster sizes near $s = 7$ is close to optimal over the range of c values.

4.3 2-D analysis

Consider a 2-D network in which $N = n^2$ nodes are placed on a $n \times n$ unit grid and are divided into clusters of size $s \times s$. We assume that each node can communicate directly only with its 8 immediate neighbors. The routing pattern within a cluster and from the cluster-heads to the sink is similar and is illustrated in Fig.9. Note that using the iterative approximation described in section 3, the joint entropy of k adjacent[3] nodes on a grid is the same as the joint entropy of k sensors lying on

[3]nodes forming a contiguous set

a straight line. Fig.9(a) illustrates this along the diagonal.

The results for the linear array of sources do not extend directly to a two-dimensional arrangement where every node is both a source and a router. In the 1-D case, the optimal aggregation tree is different from the shortest path tree (except for the case with zero correlation). This is because moving towards the sources allows greater compression than moving towards the sink. In the 2-D case however, there are opportunities for compression in all directions. Hence, it is always possible to achieve compression while making progress towards the sink.

4.3.1 *Opportunistic compression along SPT to cluster head.* According to the approximation we have been using for the joint entropy, the contribution of a node v is $H(v/\eta_v)$, where η_v is the nearest neighbor of v. If we assume that $H(v/\eta_v)$ is the fixed rate allocation for every node v, it follows[4] that a network-wide SPT is the optimal routing structure. In other words, the optimal cluster size $s = n$ for all values of correlation parameter c. There is no incentive for data to deviate from a shortest path to the sink. The result is established more precisely in the following lemma.

LEMMA 4.3. *For a 2-D grid with opportunistic compression along an SPT to cluster head, the optimal cluster size is $s = n$ for any value of correlation parameter $c \in [0, \infty]$.*

Proof is in Appendix A.2.

It should be noted that the optimality of a network-wide SPT obtained above is contingent on two of our assumptions:

—a grid topology.

—routing within clusters is along an SPT.

Results for general graph topologies are presented in [von Rickenbach and Wattenhofer 2004] and [Cristescu et al. 2004]. These are discussed in the related work section.

4.3.2 *Compression at cluster head only.* When compression is possible only at cluster heads, there is a definite tradeoff in progress towards the sink and compression at intermediate points. Since there is no compression before reaching and after leaving the cluster-heads, shortest-path routing is optimal within clusters and from cluster-heads to sink (Fig.9(b), (c)). Let $E_s(c)$ be the total cost for a network with cluster size $s \times s$ and correlation parameter c. E_s^{intra} and E_s^{extra} are defined as the combined intra-cluster costs and the overall cost for routing from cluster heads to the sink respectively. From Fig.9, a node at (i, j) will take $\max\{i, j\}$ hops to reach the cluster head at $(0, 0)$. Since there are $(\frac{n}{s})^2$ clusters, we have

$$E_{s,c}^{intra} = \left(\frac{n}{s}\right)^2 \sum_{i=0}^{s-1}\sum_{j=0}^{s-1} max\{i,j\}H_1 = \left(\frac{n}{s}\right)^2 \Big(\sum_{i=0}^{s-1}\sum_{j=0}^{i} i + \sum_{i=0}^{s-1}\sum_{j=i+1}^{s-1} j\Big)H_1$$

[4]see [Cristescu et al. 2004] for a formal proof

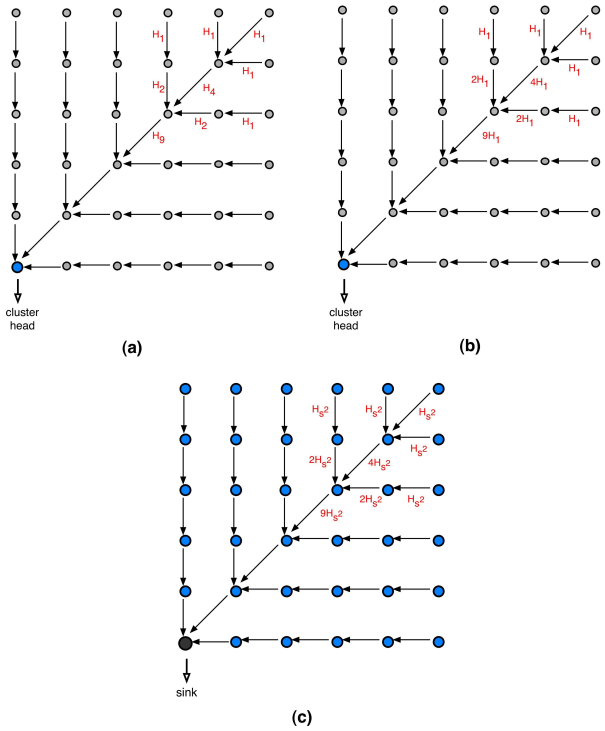


Fig. 9. Routing in a 2-D grid arrangement. (a) Calculation of joint entropy. Using the iterative approximation joint entropy of k nodes forming a contiguous set is the same as the joint entropy of k sensors lying on a straight line. This is illustrated along the diagonal. (b) Intra-cluster, shortest path from source to cluster head routing with compression only at cluster head. (c) Inter-cluster, shortest path routing from cluster heads to sink. There is no compression enroute to sink

$$
\begin{aligned}
&= \left(\frac{n}{s}\right)^2 \Big(\sum_{i=0}^{s-1} i(i+1) + \sum_{i=0}^{s-1} \big((i+1)+(i+2)+...+(s-1)\big) \Big) H_1 \\
&= \left(\frac{n}{s}\right)^2 \Big(\sum_{i=0}^{s-1} i(i+1) + \sum_{i=0}^{s-1} \Big(\frac{(s-1)s}{2} - \frac{i(i+1)}{2} \Big) \Big) H_1 \\
&= \frac{n^2}{6s}(s-1)(4s+1)H_1. \qquad (11)
\end{aligned}
$$

Now, the shortest route between adjacent cluster-heads is s hops. Hence,

$$
\begin{aligned}
E_{s,c}^{extra} &= \sum_{i=0}^{\frac{n}{s}-1} \sum_{j=0}^{\frac{n}{s}-1} max\{s\cdot i, s\cdot j\} H_{s^2} = s \sum_{i=0}^{\frac{n}{s}-1} \sum_{j=0}^{\frac{n}{s}-1} max\{i,j\} \Big(1 + \frac{s^2-1}{1+c}\Big) H_1 \\
&= \frac{n}{6}\Big(\frac{n}{s}-1\Big)\Big(\frac{4n}{s}+1\Big)\Big(1+\frac{s^2-1}{1+c}\Big)H_1. \qquad (12) \\
&\quad \text{[using the expression for } \textstyle\sum\sum max\{i,j\} \text{ from Eqn.11]}
\end{aligned}
$$

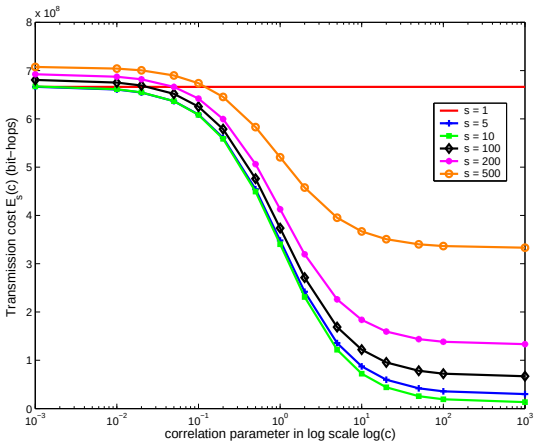


Fig. 10. Comparison of the performance of various cluster sizes for a network with 10^6 nodes on a 1000x1000 grid when compression is possible only at cluster heads. The performance for $s = 5, 10$ is observed to be close to optimal over the range of c values.

$$\begin{aligned} E_s(c) &= E_{s,c}^{intra} + E_{s,c}^{extra} \\ &= \Big[\frac{n^2}{6s}(s-1)(4s+1) + \frac{n}{6}\Big(\frac{n}{s}-1\Big)\Big(\frac{4n}{s}+1\Big)\Big(1+\frac{s^2-1}{1+c}\Big)\Big]H_1. \quad (13) \end{aligned}$$

Fig.10 shows the performance of the scheme for various cluster sizes for a 1000×1000 network. While the optimal cluster size depends on the value of c, we again find that there are certain intermediate cluster sizes (s =5, 10, 25) that perform near optimally over a wide range of spatial correlations.

It can be shown (derivation in Appendix A.3.2) that

$$s_{opt}(c) = \Big(\frac{8c}{4c+1}n\Big)^{\frac{1}{3}} + o(n^{\frac{1}{3}}).$$

THEOREM 4.4. *For $E_s(c)$ given by Equation.13, the near-optimal cluster size*

$$s_{no} = \Theta(n^{\frac{1}{3}})(\approx 0.6847 n^{\frac{1}{3}}).$$

Proof is in Appendix A.3.4.

The number of nodes in the near-optimal cluster is $N_{no} = \Theta(n^{\frac{2}{3}}) = \Theta(N^{\frac{1}{3}})$.

Fig.11 illustrates the existence of the near-optimal cluster size for a network of 10^6 nodes on a 1000×1000 grid. Clearly, the transmission cost with cluster side values near $s = 7(= \lceil .6487 n^{\frac{1}{3}} \rceil)$ is quite close to the optimal for a large range of correlation coefficient c values.

5. SIMULATION RESULTS

The analysis in Section 4 is based on simple and restricted communication, topology and joint entropy models. To verify the robustness of the conclusions from analysis,

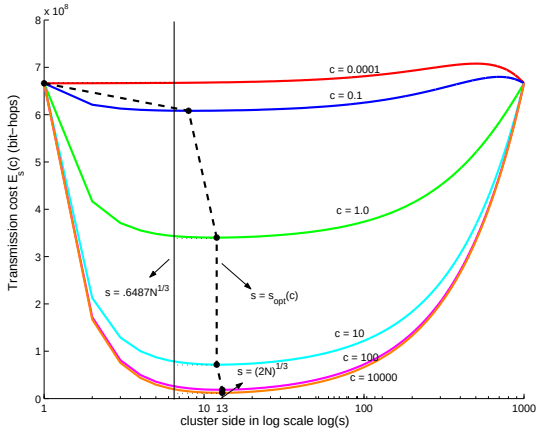


Fig. 11. Illustration of the existence of a near-optimal cluster size for a 2D network. The network size is $n \times n = 1000 \times 1000$ and compression is possible only at cluster heads. The performance of cluster side values near $s = .6487n^{\frac{1}{3}}$ is quite close to optimal for all values of c ranging from 0.0001 to 10000

we present results from extensive simulation experiments with more general models. As before, the network is deployed in a $N \times N$ area which is partitioned into grids of size $s \times s$, for $s \in [1, N]$. All nodes which are located within the same grid form a cluster.

5.1 Communication and Topology models

We consider more general communication and topology models, while using the same entropy model as in the analysis. Nodes are deployed uniformly at random within the network area. Each node is assumed to transmit 1 bit of data. The joint entropy of nodes within the cluster are calculated using the iterative, approximation technique described in Section 2.

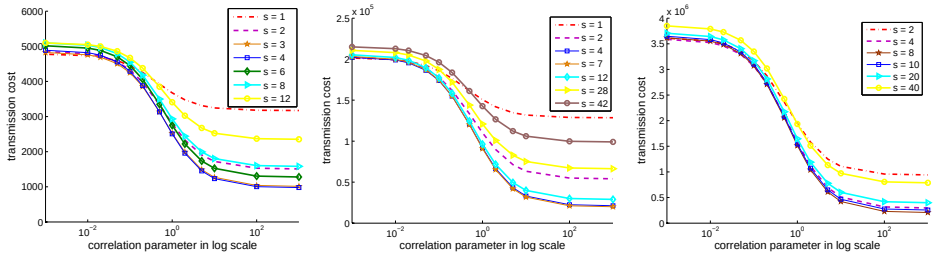
5.1.1 *Random geometric graphs.* In this model, all nodes that are within the communication radius can communicate with each other over ideal, lossless links. Since each link has a unit cost, the routing cost is calculated as:

intra-cluster cost = $\sum_{\text{all nodes in cluster}}$ (node depth in cluster SPT)
extra-cluster cost =
$\sum_{\text{all clusters in network}}$ (cluster-head depth in network SPT)· (cluster joint entropy)
total cost = intra-cluster cost + extra-cluster cost.

The simulation parameters are as follows:

—network sizes 24mx24m, 84mx84m, 240mx240m
—density of deployment = 1 node/m^2
—communication radius = $3m$

Figures 12 (a), (b), (c) show performance of clustering for the network sizes considered. As predicted by the analysis, for a network of N nodes, $N^{\frac{1}{3}}$ is a good



(a) (b) (c)

Fig. 12. Random geometric graph topology. Performance of clustering with density = 1 $node/m^2$, communication radius = 3m for network of size (a) 24x24 (b) 84x84 (c) 200x200. Near-optimal cluster sizes are (a) 3,4 (b) 4,7 (c) 8,10.

estimate of the near-optimal cluster size.

5.1.2 *Realistic Wireless Communication model.* We consider the model for lossy, low power wireless links proposed in [Zuniga and Krishnamachari 2004a]. Link costs are the average number of transmissions required for a successful transfer and these are used as weights for obtaining the shortest-path tree. The routing cost is calculated as:

intra-cluster cost = $\sum_{\text{all nodes in cluster}}$ (node cost in cluster SPT)
extra-cluster =
$\sum_{\text{all clusters in network}}$(cluster head cost in network SPT) · (cluster joint entropy)

The authors have made code available online for a topology generator based on the model [Zuniga and Krishnamachari 2004b]. The parameters used in the simulations are as follows:

—network size = 48mx48m , density of deployment = .25 nodes/m^2

—random node placement

—NCSFK modulation, Manchester encoding

—PREAMBLE_LENGTH = 2, FRAME_LENGTH = 50,

—NOISE_FLOOR = -105.0; Power levels: -3dB, -7dB and -10dB.

Figures 13 (a), (b) (c) show performance of clustering for the different power values. For lower power, there is an increase in the routing cost since links become more lossy. However, since proximity relationships between nodes do not change drastically, the relative routing costs for different cluster sizes remain similar.

5.2 Joint entropy models

We now consider more general models for the joint entropy of sources while using the realistic lossy link model from Section 5.1.2. The routing cost is calculated using the same equations and simulations are performed with power level of -3dB, all other parameters remaining the same.

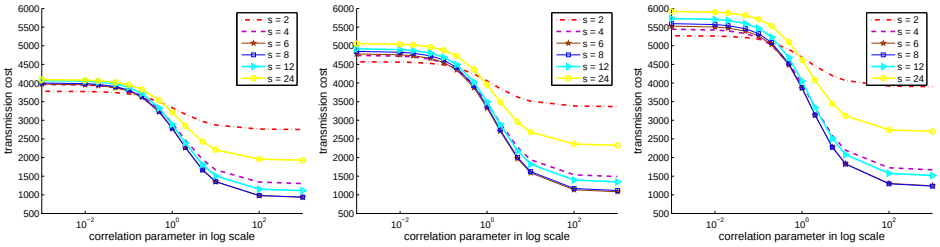


Fig. 13. Realistic wireless communication topology. Performance of clustering in 48mx48m network with density = .25 $nodes/m^2$ for power level (a) -3dB (b) -7dB (c) -10dB. Cluster sizes 6, 8 are near-optimal.

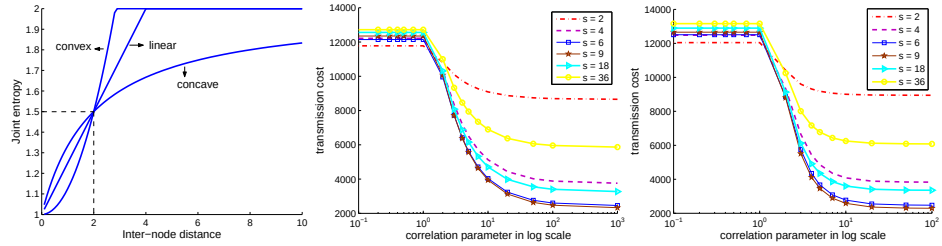


Fig. 14. (a) Example forms of joint entropy functions for 2 sources. The entropy of each source is normalized to 1 unit. The convex and linear curves are clipped when the joint entropy equals the sum of individual entropies. The curves shown are for correlation parameter $c = 2$. Performance of clustering in $72m \times 72m$ network with density = .25 $nodes/m^2$ for (b) linear model (c) convex model of joint topology. Cluster size 6 is near-optimal.

5.2.1 *Linear and convex functions of distance.* In the empirically obtained model, the joint entropy is a concave function of the distance between sources. We also look at a linear function, for which

$$H_2(d) = H_1 + min(1, \frac{d}{c}) \cdot H_1$$

and a convex function, for which

$$H_2(d) = H_1 + min(1, \frac{d^2}{c^2}) \cdot H_1$$

.

Fig 14 (a) illustrates the three forms of joint entropy functions for 2 sources. The entropy of each source is normalized to 1 unit. The convex and linear curves are clipped when the joint entropy equals the sum of individual entropies. Figures 14 (b) and (c) show performance of clustering.

5.2.2 *Continuous, Gaussian data model.* In order to verify that the results from analysis and all earlier simulations is not an artifact of the simple approximation models for joint entropy, we now consider a continuous, jointly Gaussian data model

and use its entropy as the metric for uncorrelated data in the routing cost calculations. The data is assumed to have a zero-mean jointly Gaussian distribution $X \sim N^N(0, K)$, with unit variances $\sigma_{ii} = 1$:

$$f(X) = \frac{1}{\sqrt{(2\pi)|K|^{\frac{1}{2}}}} e^{-\frac{1}{2}(X)^T K^{-1}(X)}.$$

, where K is the covariance matrix of X, with elements depending on the distance between the corresponding nodes and the degree of correlation, $K_{ij} = e^{-\frac{d_{ij}}{c}}$, where d_{ij} is the distance between nodes i and j and c is the correlation parameter. For this distribution and with quantization step size δ, entropy of a single source is [Cover and Thomas 1991]:

$$H_1 = \frac{1}{2} log_2(2\pi e) - log_2(\delta)$$

and joint entropy of n sources is:

$$H_n = \frac{1}{2} log_2((2\pi e)^n |K|) - n log_2(\delta).$$

Since $|K|$ becomes singular for large c values, we clip H_n by using

$$\max\left\{\frac{1}{2} log_2(2\pi e), \frac{1}{2} log_2((2\pi e)^n |K|)\right\} - n log_2(\delta)$$

.

Figures 15 (a), (b) and (c) show performance of clustering for quantization step $\delta = 1$, 0.5 and .05. The cluster sizes $s = 6, 8$ are near-optimal. In Figures 15 (d), (e) and (f) , the same curves are presented but the transmission cost is normalized to make the highest value equal to 1. For lower values of δ, the quantization cost dominates and the gains from removing inter-source correlations in data are diminished. Accordingly, the relative gains from optimizing cluster size are also reduced.

5.3 Summary of simulation results

Overall, the results presented in this section show that the basic conclusions from the analysis hold even when the limiting assumptions of the analysis regarding node placement, communication link quality, exact form of the correlation model, quantization, are relaxed. In all cases, we observe the existence of small cluster-sizes that provide near-optimal performance over a wide range of correlation settings.

6. RELATED WORK

Estrin *et al.* first discussed the ideas of data-centric routing and in-network aggregation for scalable and efficient designs for sensor networks [Estrin et al. 1999]. LEACH [Heinzelman et al. 2000] was an early proposal for a randomized clustering protocol that demonstrated some of the gains of in-network compression and its relation to routing. Other early work developed models and presented analysis of simple aggregation (duplicate suppression, min, max) [Krishnamachari et al. 2002] and greedy aggregation based on directed diffusion [Intanagonwiwat et al. 2002, 2003], and explored the use of data aggregation operators to optimize the performance of sensor database-type queries [Madden et al. 2002] and the possibility of

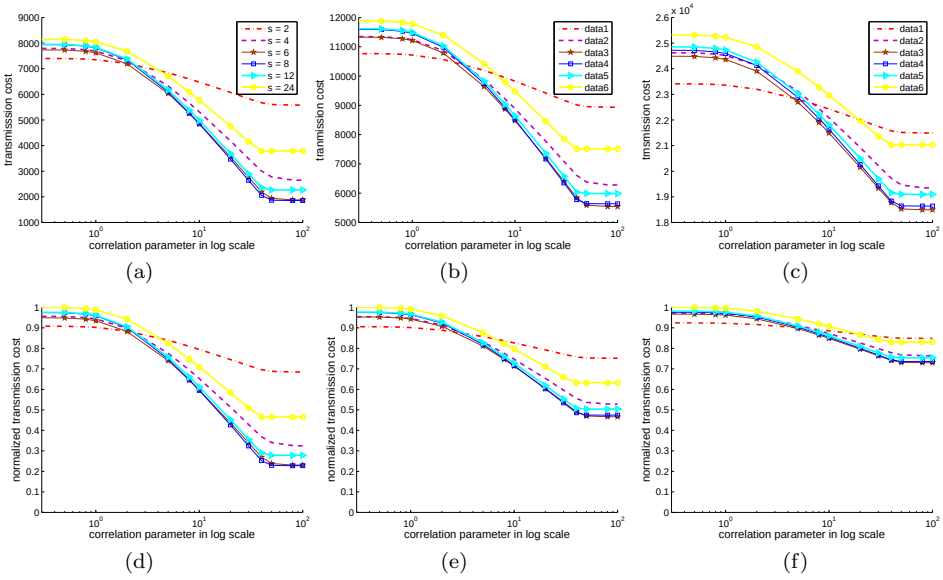


Fig. 15. Performance of clustering in $48m \times 48m$ network with density = .25 $nodes/m^2$ with a continuous, jointly Gaussian data model and quatization step (a) $\delta = 1$ (b) $\delta = 0.5$ (c) $\delta = 0.05$. Cluster size 6, 8 are near-optimal. (d) (e) (f) show normalized curves corresponding to (a) (b) and (c) respectively. For lower values of δ, quantization costs dominate, reducing the gains from optimizing for removal of correlations.

adapting the aggregation routing structures to data content and availability in the network [Bonfils and Bonnet 2003]. Krishnamachari *et al.* studied the effects of network topology and the nature of optimal routing for simple aggregation. The scheme described as routing-driven compression (RDC) in our analysis is inspired by this work.

In this paper, we consider compression of correlated sources as the principal form of data aggregation employed in the network. This is the approach taken by several works with an information-theoretic perspective. Distributed source coding (which we refer to as DSC) such as Slepian-Wolf coding [Cover and Thomas 1991] and DISCUS [Pradhan and Ramchandran 1999] suggest mechanisms to compress the content at the original sources in a distributed manner without explicit routing-based aggregation. However the implementation of DSC in a practical setting is still an open problem and likely to incur significant additional costs since it requires the complete knowledge of all source correlations *a priori* at each source. Work by Scaglione and Servetto was the first to explicitly consider the problem of joint routing and compression [Scaglione and Servetto 2002, 2005]. Using the joint entropy of sources as the data metric, the network broadcast problem in multi-hop networks is claimed to be feasible by adapting routing for compression within localized partitions (or clusters), regardless of network size. This result is disputed by work which showed that for the same problem the per sensor capacity asymptotically goes to zero [Duarte-Melo and Liu 2003, Marco et al. 2003]. Along with approaching the problem in different ways, a fundamental contrast is that while

Marco *et al.* account for wireless interference, Scaglione and Servetto ignore it. Our work assumes that data rates are well below the network capacity and the essential conclusions are shown to hold for large but finite sized networks. We explore the idea of a compression-driven routing (CDR) scheme, described by Scaglione and Servetto to be useful for high-correlation scenarios.

Our analysis of the representative routing schemes is based on using an empirically motivated model for the joint entropy as a function of inter-source distances [Pattem et al. 2004] and shows that there exist efficient correlation independent routing structures. Cristescu *et al.* formalized the correlated data gathering problem and the need for jointly optimizing the coding rate at nodes and routing structure [Cristescu et al. 2004]. The authors provide analysis of two strategies: (a) the Slepian-Wolf or DSC model, for which the optimal coding is complex (needs global knowledge of correlations) and optimal routing is simple (always along a shortest path tree) and (b) a joint entropy coding model with explicit communication for which coding is simple and optimizing routing structure is difficult. For the Slepian-Wolf model, a closed form solution is derived while for the explicit communication case it is shown that the optimization problem is NP-complete and approximation algorithms are presented. Our approach is to simplify the optimization for the explicit communication case by using the empirical model for joint entropy. The optimal routing structure is then analyzed under this approximation. The analysis demonstrates that the optimal routing structure also depends on where the actual data compression is performed; at each individual node or at "micro-servers" acting as intermediate data collection points. Von Rickenbach *et al.* differentiate "self-coding" and "foreign-coding" [von Rickenbach and Wattenhofer 2004]. In self-coding, a node uses data from other nodes to compress its own data, while in foreign-coding a node can also compress data from other nodes. With foreign-coding, it is shown that energy-optimal data gathering involves building a directed minimum spanning tree (DMST). Self-coding corresponds to the explicit communication model described by Cristescu *et al.*, for which the optimal solution is NP-complete. Good solutions are expected to be tradeoffs between a shortest path tree (SPT) and a traveling salesman path (TSP). Both these works assume that the data is compressed only once, after which it is decompressed at the sink. Recently proposed techniques [Ciancio and Ortega 2005; Ciancio et al. 2006] allow compression at several hops, potentially leading to greater reductions in transported data. Non-homogeneous networks [Hu et al. 2004] might allow routing with compression while extending the network lifetime. With some highly capable nodes acting as intermediate cluster-heads, sensor nodes do not need to expend their energy on compression. Adapting and optimizing Slepian-Wolf coding for a clustered network has been studied recently [Wang et al. 2007].

In closely related work, [Enachescu et al. 2004] presents a randomized algorithm which is a constant factor approximation (in expectation) to the optimum aggregation tree simultaneously for all correlation parameters. A notion of correlation is introduced in which the information gathered by a sensor is proportional to the area it covers and the aggregate information generated by a set of sensors is the total area they cover. The performance of aggregation under an arbitrary, general model is considered by Goel and Estrin [Goel and Estrin 2003]. Zhu *et al.* have

shown that under many network scenarios, a shortest path tree has performance that is comparable to an optimal correlation aware routing structure [Zhu et al. 2005]. While Goel and Estrin take a more general view of aggregation functions rather than as compression of spatially correlated sources and Zhu *et al.* use a different model of compression, our finding that there exists a near-optimal clustering scheme that performs well for a wide range of correlations is in keeping with the results presented in these works.

While most existing work assumes that nodes that are closest to each other have the most correlated data, Dang *et al.* have recently proposed compression over a logical mapping of nodes based on their data content, independent of locations [Dang et al. 2007]. All the work described above, including this paper, does not consider the practical details of how compression is achieved and the accompanying cost for the operations required. Ciancio and Ortega have developed a distributed scheme for removing spatial correlations using wavelet transforms via lifting steps [Ciancio and Ortega 2005]. Follow-up work has studied optimization of the choice of wavelet decomposition levels at nodes in conjunction with the routing [Ciancio et al. 2006]. Results show how practical compression schemes have to adapt to the routing and that network topology is a deciding factor in the choice of routing scheme. An improved transform, better suited for 2D topologies, has also been developed [Shen and Ortega 2008a, 2008b]. Further work is needed on developing practical compression schemes for sensor networks and evaluating them on testbed implementations [Lee et al. 2007].

7. CONCLUSION

We study the correlated data gathering problem in sensor networks using an empirically obtained approximation for the joint entropy of sources. We present analysis of the optimal routing structure under this approximation. This analysis leads naturally to a clustering approach for schemes that perform well(in terms of energy-efficiency) over the range of correlations. The optimal clustering depends on the level of correlation and also on where the actual data compression is performed; at each individual node or at intermediate data collection points or cluster heads. Remarkably, however, there exists a static, near-optimal cluster size which performs well over the range of correlations. The notion of near-optimality is formulated as a min-max optimization problem and rigorous analysis of the solution is presented for both 1-D and 2-D network topologies. For a linear arrangement of N sources, the near-optimal cluster size is $\Theta(\sqrt{D})$ irrespective of where compression occurs, where $D(\geq N, O(N^2))$ is the shortest hop distance of each source to the sink. For a 2-D grid deployment, with N sources and unit density, a network-wide shortest path tree is optimal if every node compresses its data using side information from its neighbors. If compression is possible only at cluster-heads, a $\Theta(N^{\frac{1}{6}})$ cluster size is shown to be near-optimal. The robustness of the conclusions from analysis is established using extensive simulations with more general communication and entropy models.

The practical implication of these results for sensor network data gathering is that a simple, static cluster-based system design can perform as well as sophisticated adaptive schemes for joint routing and compression.

APPENDIX

A.1 Continuous Gaussian Model

A.1.1 *Assuming non-singular covariance matrix.* Assume a zero-mean jointly Gaussian bi-variate distribution $X \sim N^2(0, K)$, with unit variances $\sigma_{ii} = 1$:

$$f(X) = \frac{1}{\sqrt{(2\pi)|K|^{\frac{1}{2}}}} e^{-\frac{1}{2}(X)^T K^{-1}(X)}.$$

, where K is the covariance matrix of X, with $K_{ij} = \sqrt{\frac{c}{c+d_{ij}}}$, where d_{ij} is the distance between nodes i and j, c is the correlation parameter. Entropy of each source is given by:

$$H_1 = \frac{1}{2} log_2(2\pi e)$$

The joint entropy is given by:

$$\begin{aligned} H_2 &= \frac{1}{2} log_2((2\pi e)^2|K|) = \frac{1}{2} log_2\Big((2\pi e)^2\big(1 - \frac{c}{c+d}\big)\Big) \\ &= \frac{1}{2} log_2(2\pi e)^2 + \frac{1}{2} log_2\Big(1 - \frac{c}{c+d}\Big) \\ &\approx log_2(2\pi e) + \frac{1}{2} log_2 e \cdot \Big(-\frac{c}{c+d}\Big) \quad \text{for small } c \\ &= 2H_1 - \frac{log_2 e}{2}\Big(\frac{c}{c+d}\Big) \\ &= H_1 + H_1\Big(1 - \frac{\alpha}{\frac{d}{c}+1}\Big) \quad \text{for } \alpha = \frac{log_2(e)}{log_2(2\pi e)} \end{aligned}$$

A.1.2 *With singular covariance matrix.* For an N-dimensional Gaussian process with a singular covariance matrix K having rank $\psi(K) < N$, the joint density of the samples can be expressed as the product of the densities of an auxiliary set of independent Gaussian random variables with variance equal to the non-zero eigenvalues of (also called principal components) whose number is equal to $\psi(K)$, and a set of N - $\psi(K)$ Dirac delta functions [Scaglione and Servetto 2005]. Consequently, if we denote by $|K|^+$ the product of the non-zero eigenvalues and by $\psi(K)$ the rank of K, the joint entropy of a Gaussian multivariate density can be, in general, written as:

$$H_N = \frac{1}{2} log_2((2\pi e)^{\psi(K)}|K|^+)$$

With $K_{ij} = \sqrt{\frac{c}{c+d_{ij}}}$, K becomes singular for $c \to \infty$, with rank 1 and the single eigenvalue = 1, to give:

$$H_N = \frac{1}{2} log_2(2\pi e) = H_1$$

While it is clear that the joint entropy will converge to that of a single source for high correlation, this does not give us a tractable expression for H_N in the high correlation region.

A.2 1D Analysis

A.2.1 . The following lemma is required for proving Theorem 4.1.

LEMMA A.1. *To solve the optimization problem in Eqn.7 for $E_s(c)$ given by Eqn.9 it suffices to find $s = s_{no}$ such that*

$$E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty). \tag{14}$$

PROOF. We first show that for any arbitrary s, this difference is maximum at one of the two extremes (i.e. at $c = 0$ and $c \to \infty$). Let

$$\begin{aligned}
E_s^d(c) &= E_s(c) - E^*(c) = E_s(c) - E_{s_{opt}}(c) \\
&= nH_1 \frac{(s - s_{opt})\big(s \cdot s_{opt} - 2c(D-1)\big)}{2s \cdot s_{opt}(1+c)} \\
\frac{\partial E_s^d(c)}{\partial c} &= -nH_1 \frac{(s-1)\big(s + 2(D-1)\big)}{2s(1+c)^2}, && \text{if } c \le \frac{1}{2(D-1)} \\
&= -nH_1 \frac{\Big(s - \sqrt{2c(D-1)}\Big)\Big(s + \sqrt{\frac{2(D-1)}{c}}\Big)}{2s(1+c)^2}, && \text{if } \frac{1}{2(D-1)} < c < \frac{n^2}{2(D-1)} \\
&= -nH_1 \frac{(s-n)\big(s \cdot n + 2(D-1)\big)}{2s \cdot n(1+c)^2}, && \text{if } c \ge \frac{n^2}{2(D-1)}.
\end{aligned}$$

$E_s^d(c)$ and its derivative vanish for the same values of c and since $E_s^d(c)$ is non-negative, the minimum is achieved at these values of c.

The derivative is continuous for all $s \in [1, n]$, and

—for a particular value of $s \in (1, n)$, it is zero only for one value of c.
—for $s = 1$, it is zero only for $c \in [0, \frac{1}{2(D-1)}]$.
—for $s = n$, it is zero only for $c \in [\frac{n^2}{2(D-1)}, \infty)$.

From the non-negativity of $E_s^d(c)$ and the above properties of its derivative, we can conclude that:

—for $s \in (1, n)$, $E_s^d(c)$ is convex
—for $s = 1$, it is monotonously increasing
—for $s = n$, it is monotonously decreasing.

This implies that $E_s^d(c)$ is maximum either for $c = 0$ or $c = \infty$ and Eqn.(7) reduces to

$$\min_{s \in [1,n]} \max(E_s(0) - E^*(0), E_s(\infty) - E^*(\infty)). \tag{15}$$

From Eqn. (9), we can derive the following expressions for energy costs of clustering schemes for the two extreme correlation values:

$$\begin{aligned}
E_s(0) &= nH_1(\frac{s-1}{2} + D) \\
E^*(0) &= nH_1 D
\end{aligned}$$

$$E_s(\infty) = nH_1(1+\frac{D-1}{s})$$
$$E^*(\infty) = nH_1(1+\frac{D-1}{n}). \quad (16)$$

Substituting Eqn. (16) in Eqn. (15) and disregarding common factors, we obtain:

$$\min_{s\in[1,n]} \max(\frac{s-1}{2}, \frac{D-1}{s} - \frac{D-1}{n}). \quad (17)$$

Let $f_1(s) = \frac{s-1}{2}, f_2(s) = \frac{D-1}{s} - \frac{D-1}{n}$. We have

$$\max_{s=1}(f_1, f_2) = f_2(1)$$
$$\max_{s=n}(f_1, f_2) = f_1(n).$$

For $s \in (1, n)$, f_1, f_2 are continuous, f_1 is increasing and f_2 is decreasing. Therefore, $max(f_1, f_2)$ achieves its minimum for $s = s_{no}$ such that

$$f_1(s_{no}) = f_2(s_{no})$$
$$i.e. \qquad E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty). \quad \square$$

A.2.2

Proof of Theorem 4.1. Solving for $f_1(s_{no}) = f_2(s_{no})$, we get

$$\frac{s_{no}-1}{2} = \frac{D-1}{s_{no}} - \frac{D-1}{n} \Rightarrow s_{no}^2 + (\frac{2(D-1)}{n} - 1)s_{no} - 2(D-1) = 0$$
$$\Rightarrow s_{no} = \sqrt{2(D-1) + (\frac{D-1}{n} - \frac{1}{2})^2} - (\frac{D-1}{n} - \frac{1}{2})$$
$$= \Theta(\min(\sqrt{D}, n)). \quad \square$$

A.2.3 . The following lemma is required for proving Theorem 4.2.

Lemma A.2. *The near-optimal cluster size $s = s_{no}$ for $E_s(c)$ given by Eqn.10 satisfies the condition*

$$E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty).$$

Proof. The proof is similar to proof of Lemma A.1 with

$$f_1(s) = \frac{E_s(0) - E^*(0)}{nH_1} = \frac{s-1}{2}, \text{and}$$

$$f_2(s) = \frac{E_s(\infty) - E^*(\infty)}{nH_1} = \frac{s}{2} + \frac{D}{s} - \sqrt{2D} \quad \text{if } 2D \le n^2$$
$$= \frac{s-n}{2} + \frac{D}{s} - \frac{D}{n} \quad \text{else.}$$

$\square$

A.2.4

Proof of Theorem 4.2. Using Lemma A.2 and solving

$$E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty)$$

for $E_s(c)$ given by Eqn.10, we get

$$\begin{aligned} s_{no} &= \frac{2D}{2\sqrt{2D}-1}(\approx \sqrt{\frac{D}{2}}) \quad \text{if } 2D < n^2 \\ &= \frac{2Dn}{2D+n(n-1)} \qquad\qquad \text{else.} \end{aligned}$$

It can be verified that

$$\begin{aligned} s_{no} &= \Theta(\sqrt{D}) \quad \text{if } D = o(n^2) \\ &= n \qquad\quad \text{if } D = \Omega(n^2). \quad \square \end{aligned}$$

A.3 2D Analysis

A.3.1

Proof of Lemma 4.3. Consider a cluster of size sxs. The routing within the cluster is as shown in Fig. 9a and routing from cluster head to sink is as shown in Fig. 9b. The routing costs are obtained as follows:

$$\begin{aligned} E_{s,c}^{intra} &= \Big(\frac{n}{s}\Big)^2 \sum_{i=1}^{s-1} \big(2(s-i)H_i + H_{i^2}\big) \\ &= \Big(\frac{n}{s}\Big)^2 \sum_{i=1}^{s-1} \Big((2(s-i)\big(1+\frac{i-1}{1+c}\big)H_1 + \big(1+\frac{i^2-1}{1+c}\big)H_1\Big) \\ &= \Big(\frac{n}{s}\Big)^2 (s-1)\Big(s+1+\frac{(s-2)(4s+3)}{6(1+c)}\Big)H_1 \\ E_{s,c}^{extra} &= \sum_{i=0}^{\frac{n}{s}-1}\sum_{j=0}^{\frac{n}{s}-1} max\{s\cdot i, s\cdot j\}H_{s^2} \\ &= s(\sum_{i=0}^{\frac{n}{s}-1}\sum_{j=0}^{i} i + \sum_{i=0}^{\frac{n}{s}-1}\sum_{j=i+1}^{\frac{n}{s}-1} j)\Big(1+\frac{s^2-1}{1+c}\Big)H_1 \\ &= \frac{n}{6}(\frac{n}{s}-1)\Big(\frac{4n}{s}+1\Big)\Big(1+\frac{s^2-1}{1+c}\Big)H_1. \end{aligned}$$

The total cost is

$$E_s(c) = E_{s,c}^{intra} + E_{s,c}^{extra}$$

The routing cost for a network-wide SPT i.e. with $s = n$ is

$$E_n(c) = E_{n,c}^{intra} + 0 = (n-1)\Big(n+1+\frac{(n-2)(4n+3)}{6(1+c)}\Big)H_1.$$

now for any $s < n$ and any value of c consider the difference

$$\begin{aligned} &E_s(c) - E_n(c) \\ &= \frac{n}{6(1+c)}\Big(\big(ns - \frac{n}{s} - s^2 + 1\big) + \frac{c}{s^2}\big(4n^2 - 3ns - s^2 - 6n + \frac{6s^2}{n}\big)\Big). \end{aligned} \quad (18)$$

It can be verified that the two terms

$$ns - \frac{n}{s} - s^2 + 1 \text{ and } 4n^2 - 3ns - s^2 - 6n + \frac{6s^2}{n}$$

are positive for any value of $s < n$. Hence the difference in Eqn. 18 is always positive. This implies that for all values of $c \in [0, \infty]$, $E_s(c)$ is minimum for $s = n$. □

A.3.2 *Derivation of optimal cluster size.* Setting the partial derivative of $E_s(c)$ w.r.t s to zero,

$$\begin{aligned}
&\frac{\partial E_s(c)}{\partial s} = \frac{n}{6(c+1)}\Big(-2s + (4c+1)n + (c-2)\frac{n}{s^2} - 8c\frac{n^2}{s^3}\Big)H_1 = 0 \\
&\Rightarrow -2s^3 + ns^2 + n = 0, \text{ if } c = 0 \\
&\Rightarrow -2s^4 + (4c+1)ns^3 + (c-2)ns - 8cn^2 = 0, \text{ if } c \neq 0. \qquad (19)
\end{aligned}$$

Differentiating again w.r.t s

$$\begin{aligned}
\frac{\partial E^2{}_s(c)}{\partial^2 s} &= -\Big(\frac{2n}{s^2} + 2\Big)H_1, \text{ if } c = 0 \qquad (20) \\
&= \frac{n}{3(c+1)s^4}(12cn^2 - s^4 - (c-2)ns)H_1, \text{ if } c \neq 0. \qquad (21)
\end{aligned}$$

If $c = 0$, the second derivative in Eqn.20 is always negative and hence the minimum is achieved at the two extremities $s = 1$ and $s = n$. Therefore,

$$s_{opt}(0) = \{1, n\}. \qquad (22)$$

—If $c > 0$, for $s = o(n^{\frac{1}{2}}), s^4 = o(n^2)$ and $(c-2)ns = o(n^2)$. Solving Eqn.19 with this constraint,

$$\begin{aligned}
&(4c+1)ns^3 - 8cn^2 + o(n^2) = 0 \\
&\Rightarrow s_{opt}(c) = \Big(\frac{8c}{4c+1}n\Big)^{\frac{1}{3}} + o(n^{\frac{1}{3}}). \qquad (23)
\end{aligned}$$

It can be verified that a minimum is achieved since the second derivative in Eqn.21 is positive for this value of s.

—If $c > 0$, for $s = \Omega(n^{\frac{1}{2}})$, it can be verified that Eqn.19 has no solution for $s \leq n$.

A.3.3

LEMMA A.3. *The near-optimal cluster size $s = s_{no}$ for $E_s(c)$ given by Eqn.13 satisfies the condition*

$$E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty).$$

The proof is similar to proof of Lemma A.1 with

$$\begin{aligned}
f_1(s) &= \frac{E_s(0) - E^*(0)}{\frac{n}{6}H_1} - \frac{n}{s}(s-1)(4s+1) \\
&= -s^2 - 3ns + 3n + 1, and \\
f_2(s) &= \frac{E_s(\infty) - E^*(\infty)}{\frac{n}{6}H_1} - \frac{n}{s}(s-1)(4s+1)
\end{aligned}$$

$$= -\frac{4n^2}{s^2} - \frac{3n}{s} - 6 \cdot 2^{\frac{1}{3}} n^{\frac{4}{3}} + 3n + 2 \cdot 2^{\frac{2}{3}} n^{\frac{2}{3}}.$$

A.3.4

PROOF OF THEOREM 4.4. From Eqns. 22 and 23, $s_{opt}(0) = 1, n$ and $s_{opt}(\infty) \rightarrow (2n)^{\frac{1}{3}}$.
Using Lemma A.3, the near-optimal cluster size $s = s_{no}$ satisfies:

$$E_s(0) - E^*(0) = E_s(\infty) - E^*(\infty)$$

$$\begin{aligned}
\Rightarrow & \Big[\frac{n^2}{6s}(s-1)(4s+1) + \frac{n}{6}\Big(\frac{n}{s}-1\Big)\Big(\frac{4n}{s}+1\Big)s^2\Big] - \Big[\frac{n}{6}(n-1)(4n+1)\Big] \\
= & \Big[\frac{n^2}{6s}(s-1)(4s+1) + \frac{n}{6}\Big(\frac{n}{s}-1\Big)\Big(\frac{4n}{s}+1\Big)\Big] \\
& -\Big[\frac{n^2}{6(2n)^{\frac{1}{3}}}\big((2n)^{\frac{1}{3}}-1\big)\big(4(2n)^{\frac{1}{3}}+1\big) + \frac{n}{6}\Big(\frac{n}{(2n)^{\frac{1}{3}}}-1\Big)\Big(\frac{4n}{(2n)^{\frac{1}{3}}}+1\Big)\Big]. \quad (24)
\end{aligned}$$

Rearranging Eqn.24 and factoring out $\frac{n}{6s^2}$, we get the condition:

$$s^4 + 3ns^3 - (6 \cdot 2^{\frac{1}{3}} n^{\frac{4}{3}} + 3n + 2)s^2 - 3ns + 4n^2 + o(n^2) = 0. \quad (25)$$

Since $s^4 = o(ns^3), ns = o(n^2)$, by factoring out n, Eqn.25 reduces to

$$3s^3 - 6 \cdot 2^{\frac{1}{3}} n^{\frac{1}{3}} s^2 + 4n + o(s^3) + o(n) = 0. \quad (26)$$

It can be verified that Eqn.26 has only one non negative solution,

$$s_{no} = 0.6487 n^{\frac{1}{3}} + o(n^{\frac{1}{3}}). \quad \square$$

REFERENCES

BONFILS, B. AND BONNET, P. 2003. Adaptive and decentralized operator placement for in-network query processing. In *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN 2003), Palo Alto, CA, USA*. Springer-Verlag, 47–62.

CIANCIO, A. AND ORTEGA, A. 2005. A distributed wavelet compression algorithm for wireless multihop sensor networks using lifting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA*. IEEE.

CIANCIO, A., PATTEM, S., ORTEGA, A., AND KRISHNAMACHARI, B. 2006. Energy-efficient data representation and routing for wireless sensor networks based on a distributed wavelet compression algorithm. In *Proceedings of the ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN), Palo Alto, CA, USA*. Springer Verlag.

COVER, T. M. AND THOMAS, J. A. 1991. *Elements of Information Theory*. John Wiley, New York, N.Y., USA.

CRISTESCU, R., BEFERULL-LOZANO, B., AND VETTERLI, M. 2004. On network correlated data gathering. In *Proceedings of the 23rd Conference of the IEEE Communications Society (INFOCOM 2004), Hong Kong*. IEEE Communications Society.

CRISTESCU, R., BEFERULL-LOZANO, B., VETTERLI, M., AND WATTENHOFER, R. 2006. Network correlated data gathering with explicit communication: Np-completeness and algorithms. *IEEE/ACM Transactions on Networking 14,* 1 (Feb.), 41–54.

DANG, T., BULUSU, N., AND FENG, W. 2007. Rida: A robust information-driven data compression architecture for irregular wireless sensor networks. In *Proceedings of the 4th European Workshop on Sensor Networks (EWSN), Delft, Netherlands*. IEEE.

DUARTE-MELO, E. J. AND LIU, M. 2003. Data-gathering wireless sensor networks: organization and capacity. *Computer Networks (COMNET) Special Issue on Wireless Sensor Networks 43,* 4 (Nov.), 519–537.

Enachescu, M., Goel, A., Govindan, R., and Motwani, R. 2004. Scale-free aggregation in sensor networks. In *1st International Workshop on Algorithmic Aspects of Wireless Sensor Networks (AlgoSensors 2004), Turku, Finland*. Springer-Verlag, 71–84.

Estrin, D., Heidemann, J., Govindan, R., and Kumar, S. 1999. Next century challenges: Scalable coordination in sensor networks. In *Proceedings of The 5th ACM International Conference on Mobile Computing and Networking (Mobicom), Seattle, Washington, USA*. ACM, 263–270.

Goel, A. and Estrin, D. 2003. Simultaneous optimization for concave costs: single sink aggregation or single source buy-at-bulk. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2003), Baltimore, Maryland, USA*. ACM/SIAM, 499–505.

Heinzelman, W., Chandrakasan, A., and Balakrishnan, H. 2000. Energy-efficient communication protocol for wireless microsensor networks. In *33rd Hawaii International Conference on System Sciences (HICSS 2000), Maui, Hawaai, USA*. IEEE Computer Society, 8020.

Hu, W., Chou, C., Jha, S., and Bulusu, N. 2004. Deploying long-lived and cost-effective hybrid sensor networks. In *The 1st Workshop on Broadband Advanced Sensor Networks(BaseNets 2004), San Jose, CA, USA*. IEEE Communications Society.

Intanagonwiwat, C., Estrin, D., Govindan, R., and Heidemann, J. 2002. Impact of network density on data aggregation in wireless sensor networks. In *Proceedings of The 22nd International Conference on Distributed Computing Systems (ICDCS 2002), Vienna, Austria*. IEEE Computer Society, 457–458.

Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J., and Silva, F. 2003. Directed diffusion for wireless sensor networking. *IEEE/ACM Transactions on Networking 11,* 1 (Jan.), 2–16.

Krishnamachari, B., Estrin, D., and Wicker, S. 2002. The impact of data aggregation in wireless sensor networks. In *Proceedings of the 22nd International Conference on Distributed Computing Systems, Workshops (ICDCSW 2002), Vienna, Austria*. IEEE Computer Society, 575–578.

Lee, S., Pattem, S., Shen, G., Tu, A., Krishnamachari, B., Ortega, A., Cheng, M., Dolinar, S., Kiely, A., and Xie, H. 2007. A distributed wavelet approach for efficient information representation and data gathering in sensor webs. In *NASA Science Technology Conference (NSTC), College Park, Maryland, USA*. NASA.

Madden, S., Szewczyk, R., Franklin, M., and Culler, D. 2002. Supporting aggregate queries over ad-hoc wireless sensor networks. In *4th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2002), Callicoon, NY, USA*. IEEE Computer Society, 49–58.

Marco, D., Duarte-Melo, E., Liu, M., and Neuhoff, D. L. 2003. On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data. In *Proceedings of International Workshop on Information Processing in Sensor Networks (IPSN)*. IEEE/ACM.

Pattem, S., Krishnamachari, B., and Govindan, R. 2004. The impact of spatial correlation on routing with compression in wireless sensor networks. In *ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN 2004), Palo Alto, CA, USA*. Springer-Verlag, 28–35.

Pradhan, S. and Ramchandran, K. 1999. Distributed source coding using syndromes (discus): Design and construction. In *IEEE Data Compression Conference (DCC 1999), Snowbird, Utah, USA*. IEEE Computer Society, 158–167.

Scaglione, A. and Servetto, S. 2002. On the interdependence of routing and data compression in multi-hop sensor networks,. In *Proceedings of The 8th ACM International Conference on Mobile Computing and Networking (MobiCom 2002)*. ACM, 140–147.

Scaglione, A. and Servetto, S. 2005. On the interdependence of routing and data compression in multi-hop sensor networks. In *Wireless Networks, Volume 11, Number 1-2*. ACM, 149–160.

Shen, G. and Ortega, A. Joint routing and 2d transform optimization for irregular sensor network grids using wavelet lifting. In *Proceedings of the ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN), St. Louis, MO, USA*.

Shen, G. and Ortega, A. Optimized distributed 2d transforms for irregularly sampled sensor network grids using wavelet lifting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, NV, USA*.

von Rickenbach, P. and Wattenhofer, R. 2004. Gathering correlated data in sensor networks. In *Proceedings of the DIALM-POMC Joint Workshop on Foundations of Mobile Computing, Philadelphia, PA, USA*. ACM, 60–66.

Wang, P., Li, C., and Zheng, J. 2007. Distributed data aggregation using clustered slepian-wolf coding in wireless sensor networks. In *Proceedings of the IEEE International Conference Communication (ICC)*. IEEE, 3616 – 3622.

Widmann, M. and Bretherton, C. 1999. 50 km resolution daily preciptation for the pacific northwest, 1949-94. Online data-set located at <http://www.jisao.washington.edu/data_sets/widmann>.

Zhu, Y., Sundaresan, K., and Sivakumar, R. 2005. Practical limits on achievable energy improvements and useable delay tolerance in correlation aware data gathering in wireless sensor networks. In *Proceedings of the IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON), Santa Clara, California, USA*. IEEE.

Zuniga, M. and Krishnamachari, B. 2004a. Analyzing the transitional region in low power wireless links. In *Proceedings of the First IEEE International Conference on Sensor and Ad hoc Communications and Networks (SECON), Santa Clara, CA, USA*. IEEE.

Zuniga, M. and Krishnamachari, B. 2004b. Realistic wireless link quality model and generator. Available online for download at <http://ceng.usc.edu/ anrg/downloads.html>.

...