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

Sundeep Pattem
Dept. of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90036, USA
pattem@usc.edu

Bhaskar Krishnamachari
Dept. of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90036, USA
bkrishna@usc.edu

Ramesh Govindan
Dept. of Computer Science
University of Southern California
Los Angeles, CA 90036, USA
ramesh@usc.edu

ABSTRACT

The efficacy of data aggregation in sensor networks is a function of the degree of spatial correlation in the sensed phenomenon. While several data aggregation (*i.e.*, routing with data compression) techniques have been proposed in the literature, an understanding of the performance of various data aggregation schemes across the range of spatial correlations is lacking. We analyze the performance of routing with compression in wireless sensor networks using an application-independent measure of data compression (an empirically obtained approximation for the joint entropy of sources as a function of the distance between them) to quantify the size of compressed information, and a bit-hop metric to quantify the total cost of joint routing with compression. Analytical modeling and simulations reveal that while the nature of optimal routing with compression does depend on the correlation level, surprisingly, there exists a practical static clustering scheme which can provide near-optimal performance for a wide range of spatial correlations. This result is of great practical significance as it shows that a simple cluster-based system design 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

Keywords

Sensor Networks, Analytical Modeling, Simulations

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, have been studied in the literature before; prior work has examined distributed source coding techniques such as Slepian-Wolf coding [10], [6], joint source coding and routing techniques [5], and opportunistic compression along the shortest path tree [4]. What is missing, though, is an understanding of how well these schemes perform across a broad range of spatial source correlations.

We begin our paper by using simplified models of these schemes in order to examine their performance across a wide range of spatial correlations. To do so, we need application-independent abstractions for compression and routing cost. The novel methodology we employ in this paper uses joint-entropy (using an empirically obtained approximation for the joint entropy of sources as a function of the distance between them) to quantify the size of compressed information, and a bit-hop metric to quantify the total cost of joint routing with compression.

Using these measures, we then evaluate three qualitatively different schemes that help us understand the space of interactions between routing and compression. In *routing-driven compression* data is routed through shortest paths to the sink, with compression taking place opportunistically wherever these routes happen to overlap [3] [4]. 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

*This research was funded in part by NSF under grant number 0325875

IPSN'04, April 26–27, 2004, Berkeley, California, USA.

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.

We use insights obtained from this analysis to develop a simpler scheme based on static, localized clustering that generalizes these techniques. Analysis shows that the optimal cluster size depends on the number of sources, sink position and the amount of correlation between sources. However, a surprising result — the principal contribution of this paper — is that for a fixed network topology, there exists a near-optimal cluster size that performs well over a wide range of spatial correlations. The implication, that there exist relatively simple energy-efficient aggregation protocols for correlated sources, has obvious practical importance. From a systems perspective, this is a very desirable result because it implies that we do not need 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.

The rest of the paper is organized as follows. In Section 2, we describe some of the related work to place our work in context. We describe our methodology in Section 3, and then present our evaluation of the three routing schemes in Section 4. Section 5 describes our simplified clustering scheme, and Section 6 summarizes the contributions of this paper.

2. RELATED WORK

Data-centric routing based on attribute naming versus end-to-end address-centric routing has been recognized as an important design principle in sensor networks [1], [4]. These include the opportunistic aggregation scheme that we describe as routing-driven compression (RDC) in this paper. The use of data aggregation operators to optimize the performance of sensor database-type queries is described in [9]. The possibility of adapting the aggregation routing structures to data content and availability in the network has been explored in [7].

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 [10] and the techniques proposed in [6] 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. We find the idea of a compression-driven routing (CDR) scheme such as that described in [5] to be useful for high-correlation scenarios and we therefore explore it in this paper.

The performance of aggregation under an arbitrary, general model is considered in [2]. While [2] takes a more general view of aggregation functions rather than as compression of spatially correlated sources, our finding here 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 [2].

We describe a clustering scheme in this paper that provides near-optimal compression performance across a range of spatial correlations. The motivation of prior studies describing clustering and hierarchical routing in wireless ad-hoc networks [12] has been different, focusing primarily on reduced overhead and scalability. LEACH [11] is an example of sensor network technique that also utilizes static clusters. While the authors of LEACH do describe the utilization of cluster-heads as aggregation points, they do not discuss compression of spatially correlated data and the dependence of optimal cluster-sizes on spatial correlations.

3. 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 would be the collection of data from a field of weather sensors. If the nodes are densely deployed, the readings from nearby nodes is likely to be highly correlated and hence contain redundancies, because of the inherent smoothness/continuity properties of the physical phenomenon.

To compare and evaluate different routing-plus-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 (per cycle). Formally, let $T = (V_T, E_T)$ 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 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 all other intermediate nodes on the tree, the outgoing edge will contain the same number of bits as the incoming edge. The bit-hop metric E_T is simply:

$$E_T = \sum_{e \in T} b_e \tag{1}$$

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 the bit-hop metric does not explicitly incorporate reception costs. 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 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 [10]. In this paper, we consider only lossless com-

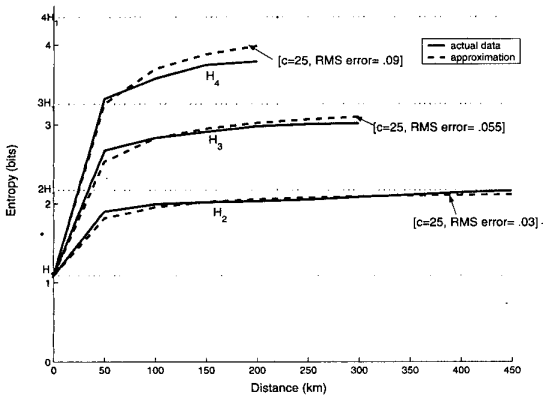


Figure 1: Empirical data (from the rainfall data-set [13]) and approximation for joint entropy of linearly placed sources separated by different distances

pression 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. 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] [13] 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. Figure 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 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 $c = d$, $H_2 = \frac{3}{2}H_1$. In other words, when $c = d$, the second source generates half the first node's amount in terms of uncorrelated data.

Finally, this leaves open the question of how to obtain a general expression for the joint entropy of n sources at arbitrary locations. This is an extremely hard problem. However, as we shall show later, this is precisely what we need in order to study the performance of various strategies for combined routing 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 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 $v_i \in V$ 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

We should 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. 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 [2] provides a good approximation for a linear scenario is validated by our measurements from the rainfall data set, as seen in figure 1.

4. 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.

[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.

[2]In addition to this convex curve, as a precaution against incorrect generalization, we also used some linear and concave models for the joint entropy as a function of inter-source distance and the correlation parameter c. Both analytical and simulation results from these models were found to provide similar results to the convex model and are therefore not included here, due to space considerations.

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 [3] and [4].

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 [5].

4.1 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 figure 2.

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: E_{DSC} represents a lower bound on bit-hop costs for any possible routing scheme with lossless compression.
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 figure 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 figure 2. It can be shown that for this scenario:

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

The above expressions, in conjunction with the expression for H_n presented earlier, allow us to quantify the performance of each scheme. Figure 3 plots the energy expenditure for the DSC, RDC and CDR schemes as a function of the correlation constant c, for different forms of the correlation function. For these calculations, we assumed a grid with $n_1 = n = 53$.

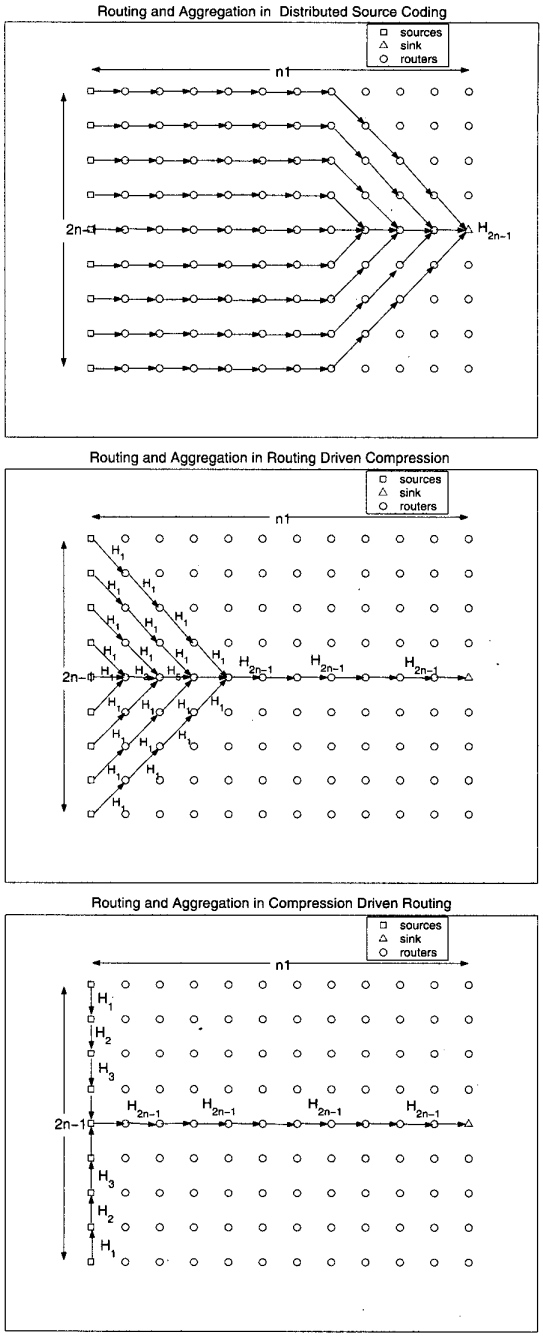


Figure 2: Illustration of routing for the three schemes: DSC, CDR, and RDC

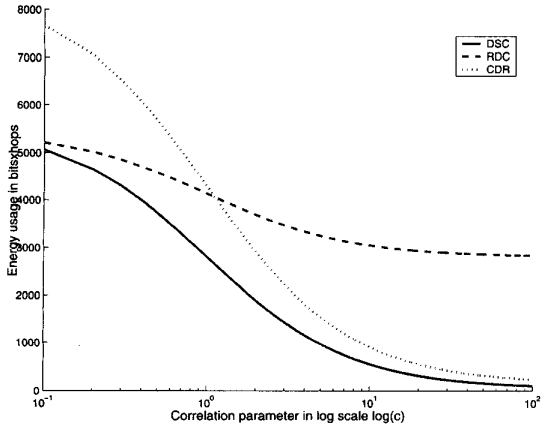


Figure 3: Comparison of energy expenditures for the RDC, CDR and DSC schemes with respect to the degree of correlation c.

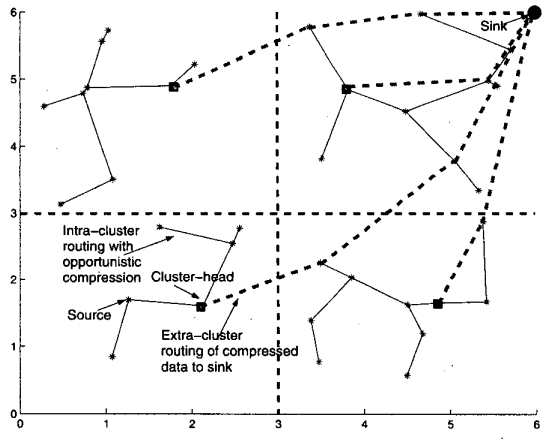


Figure 4: Illustration of clustering for a two-dimensional field of sensors

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, thus our model for DSC is very idealized). However this 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. 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.

What is interesting in these figures is that 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 (CDR), which then sends data to the sink along a shortest path (RDC). This insight leads us to an examination of suitable clustering techniques.

5. 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 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 the aggregated data can be routed to the sink without the need for further aggregation.

5.1 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. Data aggregation takes place at each of the intermediate nodes along the SPT. The cluster head then sends the aggregated data from its cluster to the sink along a multi-hop path with no intermediate aggregation. This is illustrated in Figure 4.

We analyze the performance of the clustering scheme for a one-dimensional array of sensors first, and then provide simulation results for both 1D and 2D scenarios.

5.2 Analytical Results

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 same correlation model $H_n = H_1(1 + \frac{(n-1)}{1+c})$ that we have been using. We consider n/s clusters each consisting of s nodes. The cluster head for each cluster is located at the end of each cluster. Within each cluster, the data is compressed sequentially from the one end to the cluster-head end. The cluster head then sends the compressed data along a shortest path involving D hops to the sink. The total bit-hop cost for this routing scheme is therefore

$$E_s(c) = \frac{n}{s}(E_{intra} + E_{extra}) \quad (7)$$

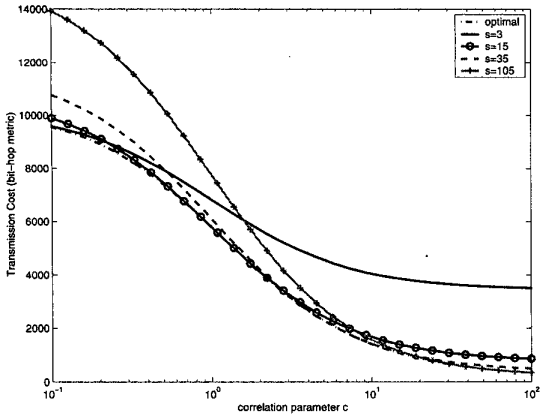


Figure 5: Analytical comparison of the performance of different cluster-sizes for linear array of sources

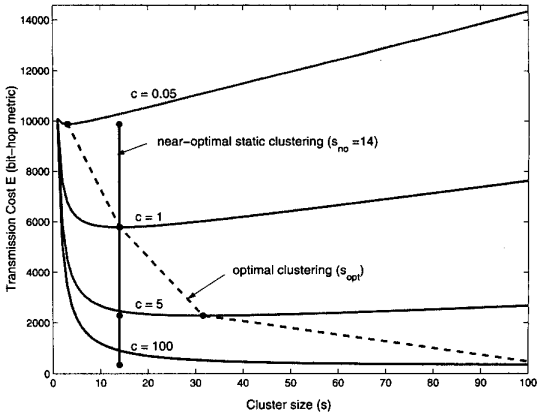


Figure 6: Analytical curves for a linear array of sources showing the possible use of a static cluster for near-optimal performance across a range of correlations

where E_{intra} and E_{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. We can obtain expressions for each of these

$$E_{intra} = \sum_{i=1}^{s} H_i = \sum_{i=1}^{s}(1+\frac{i-1}{1+c})H_1$$
$$= (s+\frac{(s-1)s}{2(1+c)})H_1 \quad (8)$$
$$E_{extra} = (1+\frac{s-1}{1+c})H_1 D \quad (9)$$
$$\Rightarrow E_s(c) = nH_1[1+\frac{(s-1)}{2(1+c)}+$$
$$\frac{D}{s}+\frac{(s-1)D}{(s)(1+c)}] \quad (10)$$

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 this

$$s_{opt} = \sqrt{2Dc} \quad (11)$$

Note that s_{opt} depends on the distance from the sources to the sink[3] and the degree of correlation c. This expression makes it clear why RDC (which corresponds to $s = 0$) performs better than CDR (which corresponds to $s = n$) when the correlation is low and vice versa.

Figure 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 very well across the whole range of correlation values. We now try to quantify this notion of a "near-optimal" static cluster size.

Let $E^*(c) = E_{s_{opt}}(c)$ represent the optimal energy cost for a given correlation c. We can formalize the notion of near-optimal as finding a cluster size $s = s_{no}$ that minimizes the maximum difference metric, i.e.

$$\min_{s} \max_{c\in[0,\infty)} |E_s(c) - E^*(c)| \quad (12)$$

It can be shown that for any arbitrary s, this difference is maximum at one of the two extremes (i.e. at $c = 0$ and $c \to \infty$). To minimize the above metric it suffices to find $s = s_{no}$ such that

$$E_{s_{no}}(0) - E^*(0) = E_{s_{no}}(\infty) - E^*(\infty) \quad (13)$$

From Equation (10), we can derive the following expressions for energy costs of clustering schemes for the two extreme correlation values:

$$E_s(0) = nH_1(1+\frac{s-1}{2}+D) \quad (14)$$
$$E^*(0) = nH_1(1+D) \quad (15)$$
$$E_s(\infty) = nH_1(1+\frac{D}{s}) \quad (16)$$
$$E^*(\infty) = nH_1(1+\frac{D}{n}) \quad (17)$$

Therefore to satisfy condition (13), we have that

$$(1+\frac{s-1}{2}+D-(1+D)) = (1+\frac{D}{s}-(1+\frac{D}{n}))$$
$$\Rightarrow \frac{s-1}{2} = \frac{D}{s}-\frac{D}{n} \quad (18)$$

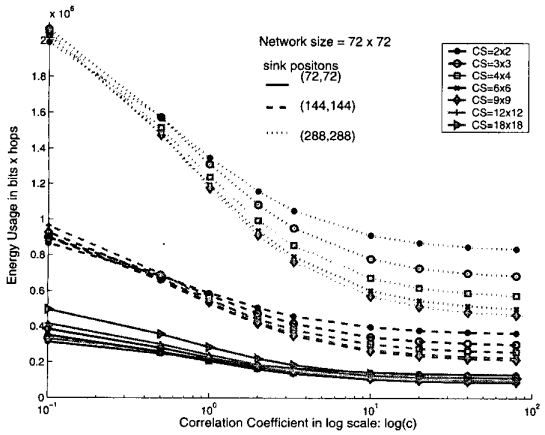
Solving the quadratic expression that results from the above equation, and simplifying by letting $D = n$, we get that

$$s_{no} = \frac{\sqrt{8n+1}-1}{2} \quad (19)$$

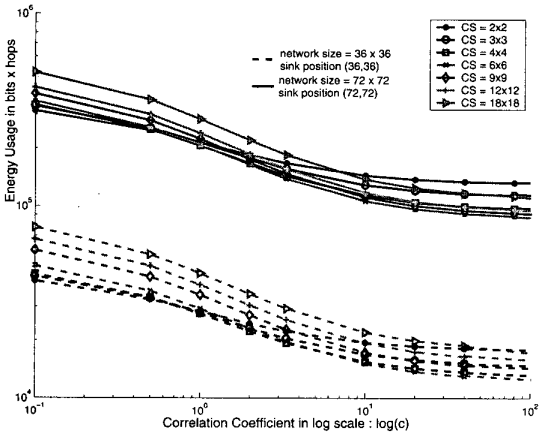
From Equation (19), we can also determine the worst case maximum difference between this near-optimal solution and the optimal solution (which occurs at the two extremes, $c = 0$ and $c = \infty$):

$$\max_{c\in[0,\infty)} E_{s_{no}}(c) - E^*(c) = E_{s_{no}}(0) - E^*(0)$$
$$= \frac{nH_1(\sqrt{8n+1}-1)}{4} \quad (20)$$

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



(a)



(b)

Figure 7: Comparisons of the performance of different cluster-sizes for a two-dimensional field of randomly placed sources, showing the impact of both (a) sink position and (b) network size.

This is illustrated in Figure 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 the near optimal cluster size is $s_{no} = 14$ (assuming $D = n = 105$) and is indicated by the vertical line in the plot — the marks on the vertical line show the energy cost of the optimal solution corresponding to the nearest c curve for comparison.[4].

[4]We should note that our difference metric for near-optimality, i.e. $E - E^*$ is different from the ratio metric E/E^* that is treated in [2]. For the latter metric, it turns out that $s = n$ is the best choice and offers a constant approximation ratio of 2 (lower than the $O(logn)$ ratio described in [2], due to the restricted topology we consider here).

5.3 Simulation Results

We now consider a two-dimensional arrangement where each node is both a data source and router. This removes the rather artificial constraint of having only a linear array of sources which we considered earlier for ease of analysis. The simulated networks consist of sensor nodes placed at random positions with a uniform density of one node per unit area within a $N \times N$ unit square region, which is then divided into smaller grids of size $g \times g$ units. The nodes within each of the smaller grids form a cluster. In our results, the cluster size denotes the area of each cluster, i.e, a $g \times g$ unit grid is said to have cluster size g^2 (with a uniformly random distribution, this is equal to the average number of sensor nodes in the cluster).

Figure 7 shows the energy performance of the scheme for various cluster sizes. The trends are similar to those observed in the mathematically-analyzed case of the linear array of sources. As expected, a large cluster size is optimal for low correlation (small c) and a small cluster size performs optimally for high correlation (large c). While the optimal cluster size depends on the value of c, we again find that there are certain intermediate cluster sizes that perform near optimally over a wide range of spatial correlations. This near-optimal cluster size depends only on network topology (sink position) and network size (total number of nodes).

Figure 7(a) compares the performance of the scheme for different sink positions. It shows gradual increase in the near optimal cluster size from 4x4 for sink position (72,72) to 9x9 for sink position (288,288). As the sink moves farther from the sources, it is useful to spend more energy trying to aggregate close to the sources and to have a smaller number of clusters(larger cluster size) in order to minimize the number of long-haul data routes to the sink. Figure 7 (b) shows the performance of the scheme for two different network sizes. It should be noted that the average distance to the sink also increases with network size. For the 72x72 network, the near optimal cluster size is about 6x6 while for the 36x36 network it is about 4x4. In general, the performance of larger cluster sizes improves for larger networks, as expected.

Other experiments we have conducted, for which we do not present results here due to space limitations, show that the results are also robust to the form of the joint entropy function (i.e. whether it is linear, convex or concave with respect to inter-node distances).

An important feature of our clustering scheme is that it is static (non-adaptive) and the memberships of a cluster do not change over time. We performed extensive simulations to examine the sensitivity of the scheme to various factors. We also found that the results do not vary with the choice of cluster head within each cluster. Hence, rotating the cluster head (as discussed, for example, in [11]) would ensure a longer network lifetime. This also allows the clustering to be static over time. This cluster size is relatively small compared to the total network size and hence clustering is simple to implement and does not involve much communication overhead for setup and maintenance.

6. CONCLUSION AND FUTURE WORK

In this paper, we have argued that, for a given network size, there exists a simple, static clustering scheme that is near-optimal (in terms of energy efficiency) across a wide range of spatial correlations. We have also sketched a simple

implementation of this clustering that leverages geographic routing techniques. Our result has important consequences—it obviates the need for sophisticated adaptive routing and compression schemes.

There are several promising avenues of research that this work can lead to. We have not considered temporal correlations and temporal compression in this work; whether our conclusions hold up under these circumstances remains to be seen. Similarly, our work has ignored lossy compression; it is conceivable that our clustering scheme can provide bounded distortion under lossy compression.

7. REFERENCES

[1] C. Intanagonwiwat, R. Govindan, D. Estrin, J. S. Heidemann, F. Silva, "Directed diffusion for wireless sensor networking," IEEE/ACM Transactions on Networking vol. 11, no. 1, p. 2-16, 2003.

[2] Ashish Goel, Deborah Estrin, "Simultaneous optimization for concave costs: single sink aggregation or single source buy-at-bulk," *SODA 2003*, p. 499-505.

[3] C. Intanagonwiwat, D. Estrin, R. Govindan, J. Heidemann, "Impact of Network Density on Data Aggregation in Wireless Sensor Networks," *ICDCS 2002.*

[4] B. Krishnamachari, D. Estrin, S. B. Wicker, "The Impact of Data Aggregation in Wireless Sensor Networks," *ICDCS Workshop on Distributed Event-based Systems (DEBS)*, 2002.

[5] A. Scaglione, S. D. Servetto, "On the Interdependence of Routing and Data Compression in Multi-Hop Sensor Networks," *8th ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2002.

[6] S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," *Proc. IEEE Data Compression Conference (DCC)*, 1999.

[7] B. Bonfils and P. Bonnet, "Adaptive and Decentralized Operator Placement for In-Network Query Processing," *Workshop on Information Processing in Sensor Networks (IPSN)*, April 2003.

[8] P. Bonnet, J. Gehrke, P. Seshadri, " Querying the Physical World," *IEEE Personal Communications Special Issue on Networking the Physical World*, October 2000.

[9] S. R. Madden, R. Szewczyk, M. J. Franklin and D. Culler, "Supporting Aggregate Queries Over Ad-Hoc Wireless Sensor Networks," *Workshop on Mobile Computing and Systems Applications*, 2002.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory* John Wiley, 1991.

[11] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," *33rd Hawaii International Conference on System Sciences (HICSS '00)*, 2000.

[12] M. Steenstrup, "Cluster Based Networks," in *Ad Hoc Networking, Ch. 4*, Ed. Perkins, C.E., Addison Wesley, 2001.

[13] M. Widmann and C. Bretherton, "50 km resolution daily preciptation for the Pacific Northwest, 1949-94," Cimate Data Archive, Joint Institute for the Study of the Atmosphere and the Ocean, 1999. Online data-set located at <http://www.jisao.washington.edu/data_sets/widmann>