This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2005 proceedings.

Clustering-Based Correlation Aware Data Aggregation for Distributed Sensor Networks

Ramanan Subramanian, Hossein Pishro-Nik and Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0250
Emails: {ramanan, hossein, fekri}@ece.gatech.edu

***Abstract*— Temporal and spatial correlation in the sensed data in Wireless Distributed Sensor Networks gives room for better energy efficiency in the network. Several data aggregation schemes have been suggested in the literature. However a clear-cut solution which quantitatively describes most energy-efficient routing scheme is still lacking. In this paper, we propose a novel, generalized clustering-based aggregation scheme, called "Annular Slicing-based Clustering (ASC)" and show that by varying the cluster size and the distribution of clusters in the deployment area, one can approach the most energy-efficient aggregation scheme. Analytical expressions for the optimal cluster size and distribution have been arrived at, for a specific correlation model and a cost function based on the Euclidean distance traversed by the transmitted data. With the help of numerical simulation, it has been found that the proposed aggregation technique can achieve optimality over a wide range of correlation.**

I. Introduction

Recent advances in wireless communication, networking as well as in hardware technology such as nanotechnology and mems have opened up the potential for large-scale applications of Distributed Sensor Networks in fields such as defense, environment sensing/tracking, power system monitoring etc.,. wherein the sensor nodes co-ordinate to accomplish a specified sensing tasks. The foremost task of the Sensor Network in any of its applications is to jointly collect and transmit back the sensed data in response to query requests by the sink(s). However, any application involving Sensor Networks should also take into account the node constraints, namely the limited availability of energy, computational power and memory, to be practically viable. Hence, a lot of emphasis is recently being laid on the design of efficient routing protocols for such networks.

By nature, physical phenomena are spatially and temporally correlated. This spatial correlation results in redundancy in the data transmitted back to the sink node. This gives room for improving the energy-efficiency of the network by compressing the incoming data at key "junction" nodes (thus eliminating all redundancy) before further transmission towards the sink node. This technique is termed *Data Aggregation*, *Data Fusion* or *Routing with Compression*. In general, this is more efficient compared to locally optimal techniques such as shortest path routing at individual nodes. The primary focus of this paper is to provide an energy-efficient clustering-based aggregation scheme for a random deployment of sensor nodes, to route sensed data to the sink node. We consider a large number of sensor nodes randomly deployed in a circular field. We then divide the region into several layers, and each layer is divided into sectors, thus defining the cluster boundaries. Using this scheme, we quantify the "cost" of transmission by a metric based on the number of bits required for each hop and the Euclidean distance involved in the hop. Hence, our problem of finding the optimum transmission structure results in a non-linear optimization problem with certain constraints. The rest of the paper is organized as follows: In section II we present a brief introduction to some of the work related to our own. In section III, we discuss the assumptions on the network we work with and the analytical models we use for the energy-metric cost function and the correlation model. In section IV, we formulate the problem mathematically and show how we arrive at a non-linear optimization problem. In section V we describe a scheme from [1], and relate it with our setup. In section VI we propose our clustering scheme for the problem. We then discuss in section VII the performance of our ASC scheme in various settings and discuss the results from numerical simulation. We then show analytically that our method improves upon this scheme. Section VIII is a summary of our conclusions and contributions of this work.

II. Related Work

Literature has it that due to the inherent spatial correlation in the sensed data, aggregation (also called Data Fusion) techniques have to be incorporated into the routing protocols. Several aggregation-based data gathering techniques have been suggested in the literature. In [2] two strategies for optimal rate allocation using entropy-based coding, such as Slepian-Wolf Coding, as well as algorithms to enable optimize transmission using minimum energy have been suggested. In [3] Networked Slepian-Wolf is described in detail, and the author shows that finding an optimal transmission structure for an arbitrary traffic matrix is NP-hard. In the same article, an approximate optimization algorithm for the rate allocation in the single-sink data gathering case has been provided.

In [4], a clustering-based scheme for aggregation with a correlation model similar to ours was analyzed, and it was shown that there exists a single cluster size for which near-optimal aggregation can be achieved for a wide range of correlation coefficients. Here, it was assumed that all the sources are located inside clusters, each of which is at D hops

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on January 21, 2009 at 09:44 from IEEE Xplore. Restrictions apply.

away from the sink node. In contrast with [4], however, we assume a case wherein sensors are randomly "sprinkled" in the deployment area. Also, the Euclidean distance traversed by the transmitted bits is incorporated within the cost metric, which, to our knowledge is very realistic. This results in a complicated non-linear optimization problem to solve. Using heuristics, we then arrive at a scheme wherein we argue that near-optimal aggregation can be achieved.

In [1], a similar scheme wherein the deployment area is divided into annuli and sectors was considered. It was shown that the problem of finding the minimum routing scheme reduces to the problem of finding the minimum weight Steiner Tree on the network for a random choice of k vertices (corresponding to source sensors). The Steiner Tree problem can be summarized as follows [5]:

Let $G = (V, E)$ be a complete graph on n vertices and $m : E \rightarrow \mathbb{R}$ be a metric. Then we need to find a subtree $T \subseteq G$ on the vertex set S such that $\sum_{e \in E(T)} m(e)$ is minimum, for a given choice of $S \subseteq V$. In general, solving a Steiner Tree problem on a graph G is NP-hard, even for a 95/94 approximate solution [6], [7]. A Steiner Tree-like solution proposed for the above problem is termed "Semantic Correlation-aware Tree" (SCT) in [1]. The author has analyzed the case when there is perfect correlation in the sensed data, i.e., the correlation coefficient, $\rho = 1$, and has shown that this solution has the same order of magnitude as the global optimal solution. However, we have shown that this optimality property of this scheme breaks down when the correlation coefficient is strictly below 1. By judicious choice of the clustering scheme, we show that a near optimal solution can be achieved with our methodology, than the one described in [1].

III. Assumptions and Models

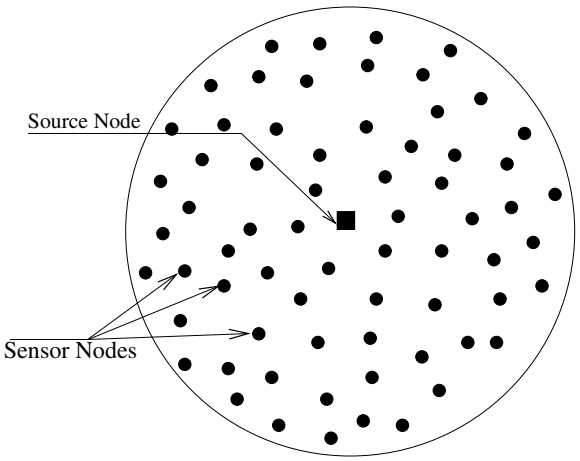

Fig. 1. Distribution of Sensor Nodes in the deployment area

Here, we present an abstract general model of a distributed sensor network, enabling our formulation to work for a wide range of applications. We assume a model similar to the one in [1]. We assume that the deployment area is in the form of a circular region of unit radius, where a single source is located at the center. n sensor nodes are placed uniformly at random in the deployment area. i.e., the expected number of nodes in any region of area A inside the circle is given by $\frac{A}{\pi}n$.

The sink node *broadcasts* queries to the source nodes, requesting for sensed data in the sensor. It is also assumed that on the average, k sensors respond to any single query. We assume that these k sensors are uniformly distributed at random in the circular area. Hence the expected number of source nodes in any region given by area A is also $\frac{A}{\pi}k$.

We assume that all the sensor nodes are identical, and so is the entropy of each source.

A. Correlation Model

Here, we describe a model similar to the one in [4] for the joint entropy of a collection of sources. However, the former is based on empirical results. Here, the joint entropy depends upon a correlation parameter, $\rho \in [0, 1]$. In general, depends upon the conditions of the environment during the sensing, and upon the nature of the sensed data. We assumed that ρ can be estimated, and is constant during one full query-response correspondence. Let H_0 be the entropy of any single source node, with a "correlated part" amounting to entropy ρH_0 and an "uncorrelated part" amounting to $(1-\rho)H_0$. The joint entropy of any two sources is calculated thus:

1) Uncorrelated part $= 2(1-\rho)H_0$,
2) Correlated part $= \rho H_0$.

Hence, the joint entropy = $H_0 + (1-\rho)H_0$. In general, if there are $q_1, q_2, \cdots, q_j$ sources at each of j levels of aggregation, then the joint entropy is given by

$$H_0 \{1 + (q-1)(1-\rho)\} \text{ where } q = \sum_{i=1}^{j} q_i.$$

B. Cost Function

It is evident that the solution to our general problem is also a Steiner tree. We need to define the cost function for each edge. Let $\xi_1, \xi_2, \cdots, \xi_k$ be the locations (i.e., 2-D vectors) of each of the responding nodes to any query in the deployment area. Define $G = (V, E)$ to be the complete graph on $V = \{\xi_1, \xi_2, \cdots, \xi_k\}$. Then, define the metric $\mathcal{M}$ mapping E to $\mathbb{R}$ as

$$\mathcal{M}(e) = \|u - v\| H_e.$$

where $e = \{u, v\}$ and H_e is the *expected entropy* of the part of the query-response that originates at that end of e farther from the origin. The reason for choosing this metric is that it presents a more sophisticated measure of the energy spent in the network, since distances are also taken into account with the number of hops, compared to previous work such as [4] which does not take the actual distances into account.

Let $T = (V_T, E_T)$ be a minimum Steiner-Tree solution to our problem. Then the cost function to be minimized is

$$C = \sum_{e \in E_T} \mathcal{M}(e). \quad (1)$$

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on January 21, 2009 at 09:44 from IEEE Xplore. Restrictions apply.

IV. Problem Formulation

Let the deployment area be divided into circular annuli of radii $\{r_j\}_{j=1}^m$, where r_j is the radius of the j^{th} annulus from the source. Each annulus corresponds to a *level* of aggregation. Furthermore, each annulus is divided into $\{S(j)\}_{j=1}^m$ equi-angular sectors. Sensors within a sector in an annulus now forms a cluster. The sensor at the geometric center of the inner circular arc declares itself as a "Steiener Node" [1]. The other nodes in any cluster tries to transmit the incoming message (or the sensed data itself, if the node is a source) through the shortest path to the Steiner Node of that cluster using Greedy Perimeter Stateless Routing (GPSR) [1],[8].

We note that the expected number of source nodes in any cluster is equal to $k\left\{r_j^2 - r_{j-1}^2\right\}/S(j)$. It can be easily verified that the worst case distance traveled by the message from a source to reach the corresponding Steiner Node is given by

$$f(j) = \sqrt{(r_j - r_{j-1})^2 + 4r_j r_{j-1}\sin^2\frac{\pi}{2S(j)}}. \quad (2)$$

Note that there are $S(j)$ Steiner Nodes in the j^{th} aggregation level. Thus, the *total entropy* $H(j)$ at the j^{th} level, given by the sum of the *joint entropies* of the Steiner Nodes, assuming each source sensor has unit entropy is given by:

$$\begin{aligned} H(j) &= \left\{\rho S(j) + k(1-\rho)\left(1 - {r_j}^2\right)\right\} \\ &+ \left\{k({r_j}^2 - {r_{j-1}}^2)\right\} \text{ for } j = 1, \cdots, m. \end{aligned} \quad (3)$$

Hence, the problem of determining the minimum cost routing scheme reduces to the following non-linear optimization problem:

Find m, and sequences $\{r_j\}_1^m$, $\{S_j\}_1^m$ such that

$$C \leq \sum_{j=1}^{m} H(j)f(j) \text{ is minimum.} \quad (4)$$

subject to the constraints:

$$m \in \mathbb{Z}_*^+, \quad (5)$$

$$S(j) \in \mathbb{Z}_*^+, \forall j \in \{1, 2, \cdots, m\} \text{ and} \quad (6)$$

$$0 = r_0 < r_1 < r_2 < \cdots < r_{m-1} < r_m = 1. \quad (7)$$

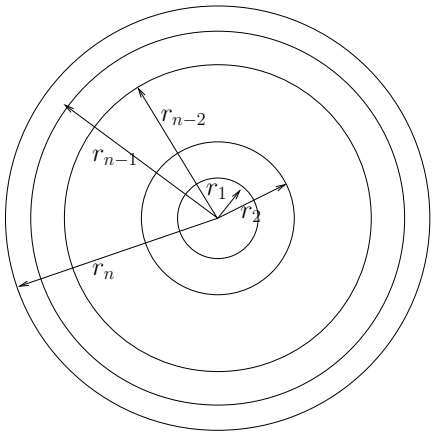

Fig. 2. Division of deployment area into Annuli in the generic scheme

For example, the proposed solution in [1], called the Semantic Correlation aware Tree or the SCT), is a feasible solution for our approach, wherein $r_j = \frac{j}{m}$ and $S(j) = \frac{2j-1}{m^2 s}$. In other words, the clustering is based on division of the deployment area into annular rings of equal radii, and each sector has a certain expected number (s) of sources, from which data is routed to the Steiner Node, which is at the center of the innermost arc region. The choice of the sequence $S(j)$ is legitimate for the particular choice of r_j since it ensures that the clusters are equal in area. The optimal m was determined for the case when $\rho = 1$. Also the message complexity for this solution was computed, and using Graph Theoretical results, it was shown that this solution has the same order of cost function as that of the optimal solution (determining which is computationally NP-hard), namely $O(\sqrt{k})$. The constant factor was found out to be $\frac{3}{2}\sqrt{3\pi}$ when $\rho = 1$.

The total entropy of the above scheme scheme when $\rho < 1$ can be determined as follows:

V. Analysis of the SCT as a function of the Correlation Factor

We shall first derive an expression for the total entropy at the i^{th} level. At this level, we note that the expected number of covered source nodes is equal to $\left(\frac{m^2-i^2}{m^2}\right)k$. We also note that the number of Steiner Nodes that are covered is equal to $\left(\frac{2i-1}{m^2 s}\right)k$. We assume that the sensor network uses Greedy Perimeter Stateless Routing (GPRS) [4] to transmit messages from a source to the next level. Let $f(s,m) = \sqrt{\left(\frac{\pi s m}{4k}\right)^2 + \left(\frac{1}{m}\right)^2}$ be the average path length from a source node to a Steiner node. The total entropy at this level is given by

$$\begin{aligned} H(i) &= \left\{k\frac{2i-1}{m^2} + \rho\frac{2i-1}{m^2 s}k\right\} \\ &+ \left\{(1-\rho)\left(\frac{m^2-i^2}{m^2}\right)k\right\}. \end{aligned} \quad (8)$$

Hence the cost function is

$$C \leq \sum_{i=1}^{m} f(s,m)H(i) \quad (9)$$

$$= kf(s,m)\left\{1 + \frac{\rho}{s} + (1-\rho)\left(\frac{4m^2 - 3m - 1}{6m}\right)\right\}. \quad (10)$$

The drawbacks of the above scheme are evident: this solution is conjecture-based, and does not describe how the annuli are aligned with respect to each other. A discussion on the performance of this scheme for $\rho < 1$ is also lacking. Also, the choice of the sequence $S(j)$ is such that messages that have already gone through several levels of aggregation are made to follow a circuitous path to the source, in order to facilitate aggregation of "fresh" sensed data from the sources within that cluster. This is unnecessary, and increases the energy cost function. However, in the general framework we described in Section IV, we can show that several of the above drawbacks can be avoided by a judicious choice of r_j and $S(j)$

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on January 21, 2009 at 09:44 from IEEE Xplore. Restrictions apply.

sequences. Furthermore, we can investigate the performance of the scheme for $\rho < 1$.

VI. Proposed Scheme - Annular Slicing-based Clustering

Each annulus in the proposed Annular Slicing-based Clustering (ASC) is divided into equal number of sectors, and the annuli are oriented in such a way that corresponding sectors overlap. To ensure equal area sectors in each annulus, we need to have $r_j = \sqrt{\frac{j}{m}}$. This ensures that the aggregation load on each Steiner Node is uniform, and none of the Steiner nodes is under utilized or over utilized.

Further, it is also legitimate to expect that the division of clusters in the optimal solution would be "matched" to the radial and angular marginal density functions of the original probability density function describing the distribution of the sensors. In this paper, we assume uniform distribution of sensors. Hence, the probability density function is given by:

$$p_{R,\Theta}(r,\theta) = \frac{1}{\pi}, \quad 0 \le r \le 1, 0 \le \theta \le 2\pi. \tag{11}$$

Hence, the marginal pdf's are given by:

$$p_R(r) = \int_0^{2\pi} p_{R,\Theta}(r,\theta) r d\theta = 2r \tag{12}$$

$$p_\Theta(\theta) = \int_0^{2\pi} p_{R,\Theta}(r,\theta) r dr = \frac{1}{2\pi}. \tag{13}$$

This legitimizes the division of sectors uniformly : since the marginal pdf on θ is independent of θ. Furthermore, the expected number of nodes within an annulus is proportional to

$$\int_{r_{j-1}}^{r_j} 2r dr = ({r_j}^2 - {r_{j-1}}^2) = 1/m. \tag{14}$$

which is independent of j, which is desirable.

VII. Analytical and Numerical Results

We now analyze the performance of the proposed Annular Slicing-based Clustering (ASC) scheme as follows. We need to define the following

$$\text{Let } f(x) = \frac{1}{\sqrt{m}} \sqrt{(2x-1) - 2\sqrt{x}\sqrt{x-1}\cos\frac{\pi}{l}}. \tag{15}$$

$$\text{Let } MC(r) = \left\{\rho l + k(1-\rho)\left(\frac{m-r}{m}\right)\right\} \times \left(\frac{\sqrt{r}-\sqrt{r-1}}{\sqrt{m}}\right) + \left(\frac{k}{m}\right) f(r) \tag{16}$$

We observe that $f(r)$ is the average path length for transmitting a message from a sensor to the Steiner node in the r^{th} level. Then the cost function is given by:

$$C \le \sum_{r=1}^{m} H(r) f(r). \tag{17}$$

$$= \rho l + \frac{k(1-\rho)}{m\sqrt{m}} \sum_{r=1}^{m} \{(m-r)(\sqrt{r}-\sqrt{r-1})\} + \frac{k}{m}\sum_{r=1}^{m} f(r) \tag{18}$$

$$\sum_{r=1}^{m} \{(m-r)(\sqrt{r}-\sqrt{r-1})\}.$$

$$= \sum_{r=1}^{m} \{m(\sqrt{r}-\sqrt{r-1}) - r(\sqrt{r}-\sqrt{r-1})\}. \tag{19}$$

$$\le m\sqrt{m} - \int_0^m \sqrt{u} du \tag{20}$$

$$= \frac{1}{3} m\sqrt{m}. \tag{21}$$

By Cauchy-Schwartz inequality, we have

$$\frac{1}{\sqrt{m}} \sum_{r=0}^{m} f(r) \le \sqrt{\sum_{r=0}^{m} f^2(r)}.$$

$$\le \sqrt{f^2(1) + \int_1^m f^2(x) dx}. \tag{22}$$

$$= \sqrt{\frac{1 + \int_1^m \left\{(2x-1) - 2\sqrt{x}\sqrt{x-1}\cos\frac{\pi}{l}\right\} dx}{m}}. \tag{23}$$

$$\int_1^m \left\{(2x-1) - 2\sqrt{x}\sqrt{x-1}\cos\frac{\pi}{l}\right\} dx.$$

$$= m(m-1) - \left(m - \frac{1}{2}\right)\sqrt{m(m-1)}\cos\frac{\pi}{l}.$$

$$+ \frac{1}{4}\ln\left\{(2m-1) + 2\sqrt{m(m-1)}\right\}\cos\frac{\pi}{l}. \tag{24}$$

One can numerically show that the above expression approximates quite close to

$$f\left(\frac{m}{2}\right) = \sqrt{(m-1) - \sqrt{(m-1)^2 - 1}\cos\frac{\pi}{l}}. \tag{25}$$

$$\text{Hence } \frac{k}{m}\sum_{r=0}^{m} f(r)$$

$$\approx k\sqrt{\frac{m-1}{m}\left\{1 - \cos\frac{\pi}{l}\right\} + \frac{\cos\frac{\pi}{l}}{2(m-1)}}. \tag{26}$$

For minimizing the cost function, we need to minimize an expression of the form $\alpha\frac{m-1}{m} + \frac{\beta}{m(m-1)}$:

$$\frac{d}{dx}\left(\alpha\frac{m-1}{m} + \frac{\beta}{m(m-1)}\right) = 0 \tag{27}$$

$$\Rightarrow m = \frac{\sqrt{\alpha+\beta}}{\sqrt{\alpha+\beta}-\sqrt{\beta}} \tag{28}$$

$$\frac{d^2}{dx^2}\left(\alpha\frac{m-1}{m} + \frac{\beta}{m(m-1)}\right) = \frac{\beta}{(m-1)^3} - \frac{\alpha+\beta}{m^3} \tag{29}$$

$$= \frac{\beta}{(m-1)^3}\left\{1 - \frac{\sqrt{\beta}}{\sqrt{\alpha}+\sqrt{\beta}}\right\} > 0 \tag{30}$$

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on January 21, 2009 at 09:44 from IEEE Xplore. Restrictions apply.

Hence, the cost function has a minimum with respect to m. Therefore, this gives an expression for the optimum number of annuli as the optimum number of aggregation levels

$m_{opt} = \frac{\sqrt{\alpha+\beta}}{\sqrt{\alpha+\beta}-\sqrt{\beta}}$ is the optimum number of aggregation levels, where $\alpha = 1 - \cos\frac{\pi}{l}$ and $\beta = \frac{1}{2}\cos\frac{\pi}{l}$.

Hence the minimum energy cost C_{min} can be obtained as

$$C_{min} \leq \rho l + \frac{1}{3}k(1-\rho) + k\sqrt{\frac{m_{opt}-1}{m_{opt}}\left\{1-\cos\frac{\pi}{l}\right\} + \frac{\cos\frac{\pi}{l}}{2(m_{opt}-1)}} \quad (31)$$

We can then determine the optimal l, the number of sectors into which each annulus is 'sliced' into, which minimizes the energy cost.

The term $\frac{1}{3}k(1-\rho)$ corresponds to the uncorrelated components of the message. The corresponding term in the previous scheme, wherein the annuli are of the same width is $\frac{2}{3}mk(1-\rho)f_{avg} \geq \frac{2}{3}k(1-\rho)$. Hence we see that the suggested aggregation scheme gives considerable improvement over the scheme in [1] for the case when ρ is smaller than 1. Using numerical computation for the case when $\rho = 1$, we can show that our scheme still has an improvement of about 33% over [1]. A plot of the cost functions in this case is shown in Fig 3.

By plotting the energy cost as a function of the number of sectors for various ρ, we can see how the optimum l behaves as a function of the correlation factor. We can see whenever the correlation is above 0.5, the optimum value of l does not change much. Hence, l around 25-30 would be near optimal for a wide range of ρ. We have assumed in our derivations that the function bounding C_min is well-behaved, having only one extremum (with respect to both the number of aggregation levels, m and the number of sectors l). A plot of the function for various ρ as a function of the number of sectors l is given in 4. Clearly, the energy cost function has exactly one minimum. Again, the function has exactly one minimum. Hence, our assumptions that the bounding function is well behaved is proven by these plots.

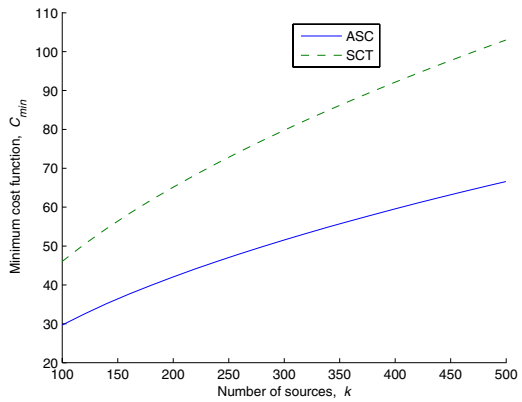

Fig. 3. Comparison of the two aggregation schemes for ρ = 1

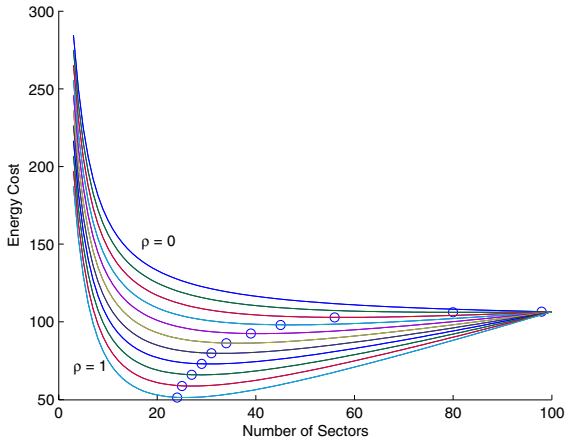

Fig. 4. Minimum Energy Cost as a function of ρ

VIII. Conclusions

In this paper, we have described a general framework for determining energy efficient data aggregation schemes for Distributed Sensor Networks in general. With a framework as in Section IV we have demonstrated how this framework works. We also described a near-optimal solution to the problem.

These heuristics are also expected to work as well for other interesting extensions (possibly application-specific), such as sensors distributed non-uniform at random, where multiple sinks and actors handle query responses, and so on. Also, more complicated correlation models for the data could be considered with the same approach. Our scheme gives a basic framework to handle such generalizations, and a similar clustering based on annuli and sectors can also be found.

References

[1] Y.Zhu and R.Sivakumar, "Enabling efficient aggregation in distributed sensor networks," in *Technical Report, Georgia Institute of Technology*.

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

[3] B. Beferull-Lozano R. Cristescu and M. Vetterli, "Networked slepian-wolf: Theory, algorithms and scaling laws," *Transactions on Information Theory*, submitted December 2003.

[4] B. Krishnamachari S. Pattem and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *International Symposium on Information Processing in Sensor Networks (IPSN)*, Berkeley, CA, April 2004, ACM/IEEE.

[5] D.R. Dreyer and M.L. Overton, "Two heuristics for the steiner tree problem," *Journal of Global Optimization*, vol. 13, pp. 95–106, 1998.

[6] M. Chlebik and J. J.Chlebikova, "Approximation hardness of the steiner tree problem on graphs," in *Proc. 8th Scandinavian Workshop on Algorithm Theory (SWAT)*. 2002, pp. 170–179, Springer-Verlag.

[7] S.S. Skiena, *The Algorithm Design Manual*, pp. 339–342, Springer-Verlag, 1997.

[8] B. Karp and H.T. Kung, "Gpsr: Greedy perimeter stateless routing for wireless networks," in *Proc. 6th Annual International Conference on Mobile Computing and Networking(MobiCom 2000)*, Boston, Massachusetts, August 2000, Sigmobile/ACM.

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on January 21, 2009 at 09:44 from IEEE Xplore. Restrictions apply.