Scattered Data Selection for Dense Sensor Networks

Lance Doherty
Berkeley Sensor and Actuator Center
497 Cory Hall
Berkeley, CA 94720
510-642-4571
ldoherty@eecs.berkeley.edu

Kristofer S. J. Pister
Berkeley Sensor and Actuator Center
497 Cory Hall
Berkeley, CA 94720
510-642-4571
pister@eecs.berkeley.edu

ABSTRACT

An evaluation methodology is presented for the performance of reporting node self-selection in wireless sensor networks. Five cost metrics are proposed along with several methods for self-selection that involve little or no collaboration with other nodes. These costs are used to evaluate how efficiently the various algorithms allow for node self-selection as simulated on different field complexities. Analysis of different methods over 100 test fields sampled by 2000 nodes indicates that there is no single method that is superior in all respects. Trade-offs in latency and overall energy consumption are revealed to be highly dependent on the selection method and the field complexity.

Categories and Subject Descriptors

H.1.1 [**Models and Principles**]: Systems and Information Theory – *General Systems Theory.*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Sensor networks, data recovery, energy measures.

1. INTRODUCTION

As sensor networks progress, the trend is to decentralized control and processing of information as Gupta and Kumar have shown that throughput decreases with network size [4]. Still, industrial interest in large-scale deployment of sensor nodes is focused on collecting data from a wide geographic region at a centralized database. The user in this paradigm is interested in visualizing a full representation of data gathered throughout the network and watching how it evolves in a periodic and timely fashion. This paper discusses one approach to analyzing the theoretical limits to obtaining such a representation in a multi-hop wireless network.

The application domain under consideration is one where topological changes occur seldom relative to the data retrieval period. This requirement is implemented so that the cost of network discovery and route management can safely be neglected. There have been a host of algorithms proposed in the literature for finding and maintaining routes and schedules in the network such as [1,5,7,13,15]. This paper also does not consider the challenges introduced by localization. With a sufficiently dense network and the implementation of promised RF time-of-flight distance measuring systems, node positioning should be possible to within acceptable levels of uncertainty [10,14]. This mechanism may be costly in terms of energy but the slowly-varying restriction allows amortization over several cycles of data collection. For the analysis that follows, nodes know both their positions and a multi-hop data path to send information back to a basestation for collection and processing.

While many types of algorithms are possible for fusing data within the network prior to sending it back to the basestation through several hops (see [2,6] for examples), this paper is concerned only with those that perform no data fusion at all. The algorithms considered are of the class that, with the allowance of a modicum of data exchange with nearest neighbors, allow nodes to self-select as part of a subset of nodes forwarding their data for collection. At the basestation, information consisting only of position-data pairs is received with the positions being irregularly placed throughout the monitoring environment. The reconstruction of such a data field is similar to the Scattered Data problem in image processing. This is a well-studied problem with optimized solutions discussed in the literature such as [8]. The contribution of this paper is to detail the energy costs in terms of atomic network operations that are required to provide the basestation with sufficient information on which to base a scattered data reconstruction of the field to a specified resolution. Several energy metrics are proposed and the variation in cost with the field complexity being monitored by the sensor network is detailed. The energy metrics considered focus on reducing communication as many authors suggest this as the biggest drain on network resources [3,9,11,12]. All results in this paper are in the context of a single snapshot in time. For multiple time steps, the data collection would occur multiple times.

This paper is organized as follows. In Section 2, details on the fields and network used for simulation are provided and parameters defined. Section 3 discusses various algorithms used by nodes to self-select as reporting nodes. Section 4 introduces five cost metrics to evaluate these algorithms. The simplest method is considered in detail in Section 5 and a summary of the important results for all methods is given in Section 6. Uncertainty in position and quantization is considered in Section 7 with conclusions and future research directions in Section 8.

IPSN'04, April 26–27, 2004, Berkeley, California, USA.
Copyright 2004 ACM 1-58113-846-6/04/0004...$5.00.

2. TEST FIELDS AND NETWORKS

2.1 Parameters

The symbols used to describe the fields and networks are as follows:

N – The number of nodes in the network
R – The communication radius of each node (a disc)
ρ – The spatial density of nodes in the network
C_{ave} – The mean connectivity of nodes (a function of R and ρ)
h_{ave} – The mean number of hops from nodes to the basestation

2.2 Field Generation

Fields used in the following simulation are of area $20R$x$20R$ and generated by a three-step process shown in Figure 1. The goal is a discrete approximation with $0.1R$x$0.1R$ resolution of a continuous field. First, data "sources" are placed at some of the 400 potential grid locations; no square of area R^2 has more than one source. Grid locations with sources generate fixed data identified by the value of the source. Second, all grid locations are assigned a data value based on a simulated multi-step diffusion process from the sources. Third, the grid is refined by a factor of ten in each dimension by interpolation to a finer grid of 200x200 units.

The number of sources is set before the field is generated with the data value of each source drawn independently from a uniform random distribution between 0 and 64. Likewise, x-y locations of each source are randomly chosen from the set of integers between 1 and 20. The number of sources in the field represents the complexity of the field. Ten test fields are generated for each number of sources under consideration: {4, 7, 10, 15, 20, 30, 50, 100, 200, 400}, resulting in 100 test fields.

A sequence of 100 time steps is simulated with each non-source grid location setting its value to the mean of its four nearest neighbors. Locations with sources keep the same value throughout the simulated diffusion. The number of time steps used is sufficient to allow the system to reach steady state. Following diffusion, all the data at the 400 grid locations is stored in a variable called *tempgrid*. This 20x20 grid is locally interpolated using the following MATLAB commands:

```
[xc,yc]=meshgrid(0:20)
[xf,yf]=meshgrid(0.1:0.1:20)
finetempgrid=griddata(xc,yc,tempgrid,xf,yf,'cubic')
```

Following interpolation, the discrete field now has 40,000 entries representative of $0.1R$ increments in either dimension. The entries of *finetempgrid* are used as the sample values for sensor nodes that lie within the appropriate areas.

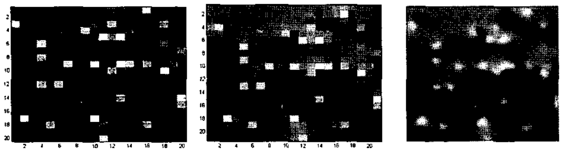

Figure 1. Field generation. Left: 100 sources shown on grid with brightness representing values and non-sources are the neutral background. Middle: values at 400 grid points after diffusion steps (*tempgrid*). Right: values interpolated to a 200x200 grid (*finetempgrid*). Nodes with positions inside a particular grid square sample only this data value.

The number of sources is the independent variable representative of complexity in this study. The goal is to relate the cost of representing a field by a sensor network data recovery algorithm to the complexity of this field, and hence it is important that the complexity be commensurate with the number of sources. By visually examining fields of different numbers of sources it is qualitatively obvious that more sources lead to fields requiring more sensor readings for reconstruction. This relationship is quantifiable through independent means by invoking the image processing analogy and using compression. Each of the 100 test fields is mapped to a 200x200 pixel monochrome image and highest-quality (100 level) jpeg compression is applied. The mean file size is a monotonically increasing function of the number of sources in the field and fields generated with the same number of sources have similar compressed file sizes (Figure 2).

The selection of the fields as originating from several sources is intended to represent a sensor network application where there are regions where data is changing rapidly with location and regions where sensor readings are similar over large areas, i.e. a variety of spatial frequencies. These two measures are relative to the number of nodes in the network and the node density. To represent the field with a small number of samples, it is speculated that nodes should be chosen in regions of high data variation, though the selection of these nodes must be balanced with the cost of local communication to determine the optimal nodes.

The number of sources chosen for the simulation spans the range from those that can be very effectively represented by a small number of nodes to those that are still poorly represented by the full number of samples that are permitted. The user in this application is interested in a full mapping of the field, not just determining the location of the data-generating sources. Each field as the one shown in Figure 1 is intended to represent a single moment in time: the temporal evolution and compression of data is not considered in this paper. In the following time step, the sources would potentially have changed in value and/or location; the field is assumed to be changing much more rapidly than the node topology.

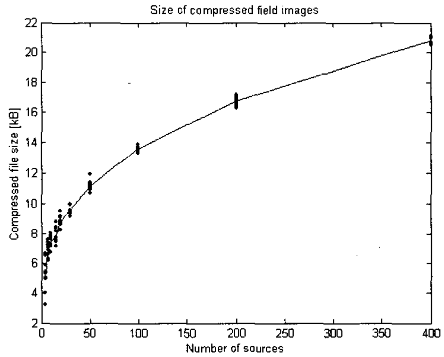

Figure 2. The compressed file size increases with the number of sources generating the field. Each dot represents one of the 10 test fields and the line interpolates the values.

2.3 Network Generation

The network used to study the algorithms in this paper consists of 2000 nodes placed at uniformly random real-valued locations in the 20*R*x20*R* region. The locations of the nodes and the number of hops to the basestation are depicted in Figure 3. The connectivity of the network is determined by the simple rule that two nodes are connected if and only if they lie within a distance *R*. By assigning the basestation role to the node closest to the center of the region, a flooding algorithm is used to count the number of hops from each node to the basestation, i.e. how many transmit and receive events must occur for information to be obtained from this node.

The network parameters are chosen to represent a network larger than those currently implemented. The large diameter ensures that trade-offs between local and basestation communication are clear from the simulation. If a node *h* hops away from the basestation is selected to relay its data, the overall energy consumption in the network is *h* times that required for a node a single hop away.

2.4 Field Reconstruction

Once data has been recovered at the basestation, the scattered data problem is solved by interpolating the received data to the 200x200 grid locations with 'griddata'. This interpolated field is compared to the original field. The absolute difference of these two fields is averaged over the 40,000 grid locations and this value is the mean error of the reconstruction. A sample reconstruction is shown in Figure 4. A host of node selection methods is detailed in the following section.

In reconstructing the field, the interpolation mechanism allows only for estimation within the convex hull spanned by the location of the data. To eliminate regions where no reconstruction is possible, four artificial nodes are placed at the corners of the network and always report data. Costs of these nodes reporting are accounted for in energy calculations.

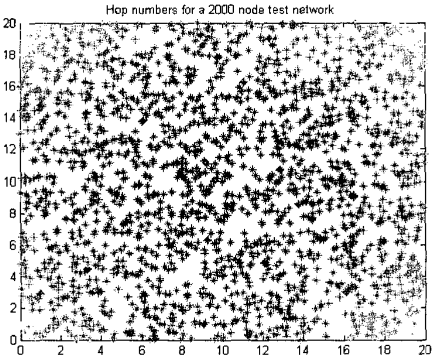

Figure 3. Hop numbers and locations of each of the 2000 nodes in the test network. Darker stars represent smaller hop numbers while lighter stars are nodes with more hops to the basestation at the center of the plot. Nodes at the corners are 21-22 hops away from the basestation.

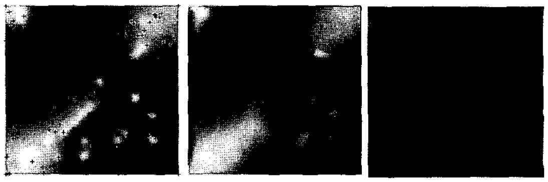

Figure 4. Field reconstruction from scattered data. Left: original field showing 212 randomly chosen sample locations. Middle: field reconstructed from the 212 scattered data points. Right: error field used for computation of the mean error.

3. METHODS UNDER CONSIDERATION

This section details algorithms used to allow nodes to self-select as reporting nodes. Different algorithms use various combinations of the following three categories of information: (i) the node's data, (ii) the node's neighbors' data, (iii) the node's distance from the basestation. Each method has a parameter that allows the number of nodes selected to be tuned from near zero reporting to all nodes reporting.

3.1 Random selection

The simplest approach to selecting nodes is through random selection. In this model, nodes independently compare a randomly generated number to a known network-wide parameter that specifies the fraction of nodes expected to communicate data back to the basestation. The node does not require any local communication or even to sample if not self-selected to report. This model is evaluated mainly as a basis to which the others can be compared and it is expected to exhibit poor per node performance but a low per node energy cost. The number of nodes reporting in this model is controlled directly through the reporting probability that would be disseminated through the network during initialization.

3.2 Hop-based selection

Assuming that each node has access to the number of hops to the basestation, this hop number is used to weight the probability that a node self-selects. In terms of the total number of messages required to obtain data, nodes more distant from the basestation require a greater expenditure of resources. Consequently, the probability of reporting should decrease monotonically with the hop number. Two examples are chosen for the rate of this decrease: linear and quadratic. The probability of reporting is a polynomial function f of the hop number at each node. The node does not need to communicate with its neighbors or sample prior to making a reporting decision. To tune the number of nodes reporting in this model, the function f representing the probability of self-selection can be scaled. Nodes whose hop counts lead to values of $f>1$ are always selected for data reporting. Aside from the hop number dissemination that should occur infrequently compared to data collection, this method requires no more communication than the random method.

3.3 Interval selection

Following sampling, each node self-selects based on its sensor reading. If the sensor reading is close to a certain periodic goal reading, it has a high probability of self-selecting. Periods of units and tens are used. For unit spacing, nodes sampling data slightly above each integer have a high probability of self-selecting while in the tens model, it is nodes with data close to {0,10,20,...,60}. This method selects nodes on the field contours suggesting an effective means of representing the field with fewer data points as shown in Figure 5. The decay rate of the probability as node readings become more distant from the goal readings is modified to vary the number of nodes reporting. Again, this method is no more costly per selected node than the random method.

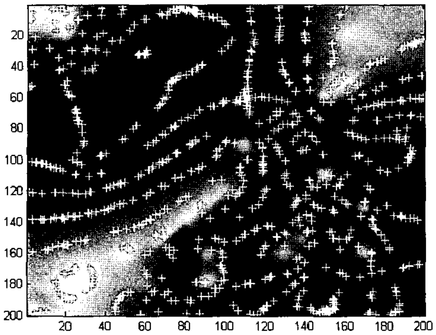

Figure 5. Nodes selected with the "interval = 10" method. In regions of slow spatial variation, the selected nodes are seen tracing contours in the field while the selection becomes random in regions of high variation.

3.4 Extremum selection

Nodes with sensor readings well above or well below the mean of their neighbors' readings self-select. Nodes are required to read their sensor and receive all neighboring data prior to making the decision rendering this method relatively costly compared to the previously discussed methods. The number of reporting nodes is controlled by setting the linear decay rate by which nodes with values close to their neighbors have less chance of reporting than those with extreme values. The per node selection cost of this method is higher than random selection, but this may be offset by the selection of more representative nodes.

3.5 Edge detection

An edge node is one where the field reading is changing rapidly in a spatial sense. Nodes with neighbors spanning a larger data range are given a higher probability of reporting data. Like the extremum method, the decay rate controls the number of nodes reporting and a similar per node selection cost is required. A modified edge algorithm, the "limited edge" selection, is also tested: nodes receive data from a maximum of four neighbors in this model.

3.6 Edge dilation

Following edge detection, this method selects only the minimum- and maximum-valued neighbors of edge nodes. The notification of neighbors adds another communication step to the edge detection algorithm making this selection the most costly per node. Each edge node makes two additional transmissions (to the selected dilation nodes) and each of the two dilation nodes has an additional receive event. The number of reporting nodes is controlled by the edge decay value but the actual selected may be up to twice the value of that in the edge detection model. As the edge nodes approach *N*, the dilation method might actually select fewer nodes than the edge method since with a low threshold, every node becomes an edge but not every node is an extreme nearest neighbor. The motivation behind this method is that accurate reconstruction requires a reference on either side of each discontinuity.

4. ENERGY COST METRICS

Five different cost metrics are proposed for evaluating the resource consumption:

1. Number of nodes reporting
2. Number of nodes sensing
3. Total number of transmit events in the network
4. Total number of receive events in the network
5. Average transmissions from one-hop nodes to the basestation

The sensing and reporting node counts are measures of general network activity. The number of transmit and receive events are indicators of the mean energy consumption. One-hop nodes are those that lie within a communication radius of the basestation and must serve as relays for all multi-hop messages originating in the network. The average number of transmit events at one-hop nodes is a measure both of the maximum energy consumption and the maximum bandwidth requirements as these nodes are both communication and battery lifetime bottlenecks since nodes are uniformly distributed and all information must pass through them. As we are dealing with a broadcast medium, the number of transmit events can be smaller than the number of receive events provided that the one-to-many communication is well scheduled.

Table 1 summarizes the costs of the proposed methods. The first three rows of the table represent the methods without local communication to select nodes while the incremental effect of the local communication can be seen in the final three rows. These costs represent the lower bounds that could be obtained if all communication was synchronized and no collisions occurred.

In general, data collection can result in nonlinear relationships among each of the five costs, but in scattered data models the mean number of one-hop transmissions is a scaled version of the number of reporting nodes. The expected scaling factor is

Table 1. Summary of the costs associated with the different scattered data strategies. The probability of selection of any given node (which varies considerably among methods) is represented by P.

Method	*Reporting Nodes*	*Sampling Nodes*	*Tx Events*	*Rx Events*	*1-hop Tx per Node*
Random	PN	PN	PNh_{ave}	PNh_{ave}	$PN/\rho R^2$
Hops	PN	PN	PNh_{ave}	PNh_{ave}	$PN/\rho R^2$
Interval	PN	N	PNh_{ave}	PNh_{ave}	$PN/\rho R^2$
Extremum	PN	N	$PNh_{ave} + N$	$PNh_{ave} + NC_{ave}$	$PN/\rho R^2$
Edge	PN	N	$PNh_{ave} + N$	$PNh_{ave} + NC_{ave}$	$PN/\rho R^2$
Dilation	PN	N	$PNh_{ave}+ N + 2PN$	$PNh_{ave} + NC_{ave} + 2PN$	$PN/\rho R^2$

precisely the number of one-hop nodes available for relaying messages to the basestation. It is hence sufficient to use the number of reporting nodes (cost 1) as a measure of maximum individual energy and bandwidth requirements (cost 5). The sampling cost will be mainly ignored as an independent measure since sensing and reporting nodes are either the same or all nodes are sampling.

5. RANDOM METHOD RESULTS

A detailed description of the results is given for the random method as an illustration of the analysis procedure and to show that the test fields and networks are sufficient to span the intended set of behaviors. All methods are compared without the full detail in the next section.

Using the 100 test fields with varying numbers of sources, trials are completed with different probabilities of reporting data. The probability of each individual node reporting (*preport*) is varied logarithmically through 9 values between 0.01 (an expected 20+4 nodes reporting) and 1 (all nodes reporting). The mean error averaged over 10 trials for each field for the smallest and largest values of *preport* is shown in Figure 6. Three observations are of interest:

1. Each field having the same number of sources yields nearly the same mean error
2. The maximum mean error for a field with a smaller number of sources is rarely significantly higher than the minimum for a field with a larger number of sources
3. Aside from a few exceptions, fields that appear problematic for the *preport* = 0.01 case are not poorly handled in the *preport* = 1 case and vice versa

The combination of items 1 and 2 are further confirmation that the selection of the number of sources as an independent measure of field complexity is well founded. Item 3 suggests that the variation in error across fields of similar complexity is due mainly to statistical deviations in the node selection and not in fundamental differences among fields with the same number of sources.

These charts show a general decrease in mean error as more nodes are reporting, but the decrease is not proportionate across source numbers. The monotonic increase in error as either (a) fewer nodes report data or (b) more sources are added is salient in comparing these bar charts. Averaging over all fields at the same source number gives the results shown in Figure 7.

As anticipated, collecting data from more nodes by increasing *preport* allows for the field to be reconstructed with less error. There is, however, an increasing cost associated with larger values of *preport*. The five costs outlined in Table 1 are summarized in Figure 8. There are only three curves since only the nodes reporting need to sample their sensors and the number of receive and transmit events is symmetric.

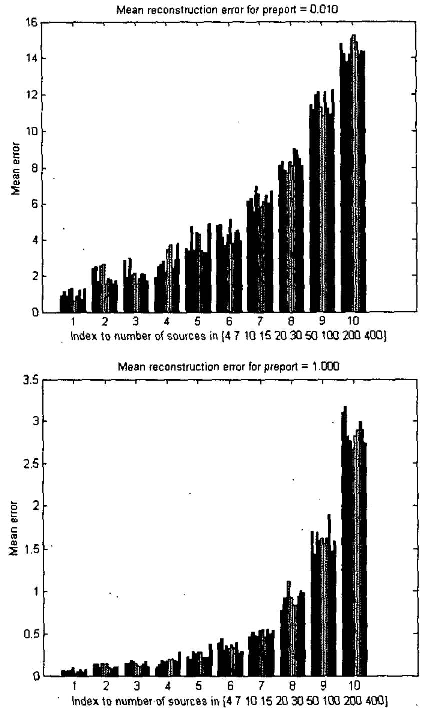

Figure 6. Bar charts showing the mean error for each field. Top: the expected number of reporting nodes is 24. Bottom: all 2004 nodes are reporting. Each bar represents the mean integrated error averaged over 10 trials for one of the 100 test fields. X-axis numbers correspond to the list below them.

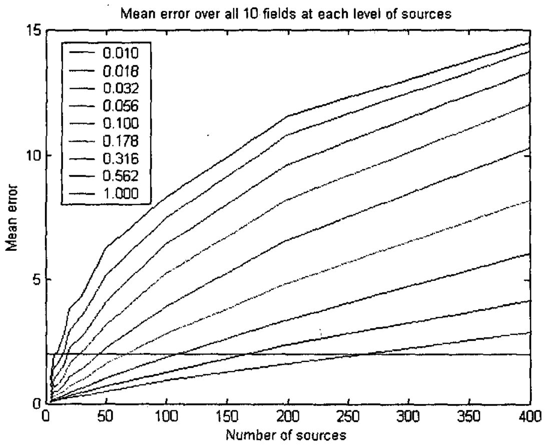

Figure 7. All trials of the random reporting method. Each curve represents an entire bar chart like in Figure 6. The legend shows values of *preport* in the vertical order they appear. Curve intersection points with the horizontal line denote the number of sources that can be monitored by each number of reporting nodes with a mean error of 2.

All costs are linear with *preport* in the random model. Suppose now that the interest is in designing a sensor network to obtain a particular level of performance. The intersection of a horizontal line like the one crossing Figure 7 with the curves allows for determination of the costs required to obtain a mean error of 2. As sources are added, the number of reporting nodes must increase. Figure 9 is obtained by plotting the cost at the x-values of the intersection of the horizontal and curved lines in Figure 7.

Values shown in this graph give expected costs to reconstruct the field to within an error of 2. For example, 579 sensing/reporting nodes, 5834 transmit/receive events and 41 relays from each of the one-hop nodes to the basestation are required to adequately reconstruct a field with 100 sources on the average. The one-hop node load can be translated to message latency if the particular communication parameters of the network are known. With limited bandwidth, only a certain throughput is possible from the interfering one-hop nodes to the basestation; the time to forward all messages varies linearly with the number of relays.

In the following section, all methods are compared using the same metrics as in Figure 7: a similar set of cost-complexity curves are generated through the same procedure. The curves in Figure 9 do not go beyond ~270 sources as beyond this even the collection of data from all nodes does not allow for reconstruction to resolutions below a mean error of 2. This same limit will be seen in all the selection methods as the best resolution is obtained with exactly the same data. The costs associated with this limit will however be larger for the local communication methods as they require more energy to determine that all nodes should self-select.

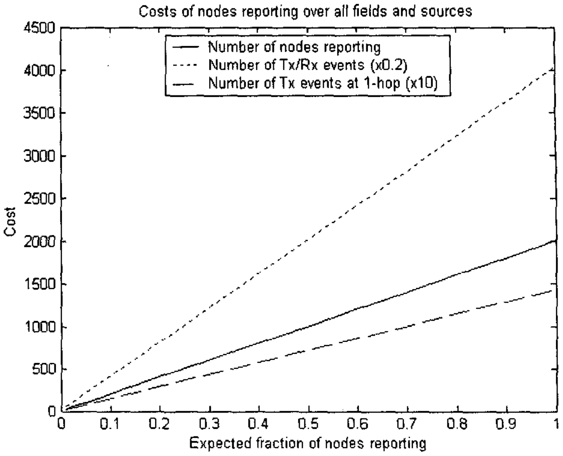

Figure 8. Costs associated with different values of preport. Curves are scaled to fit on the plot.

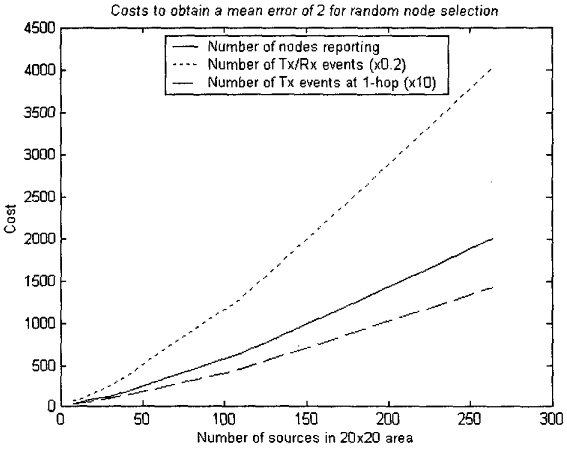

Figure 9. The costs associated with reconstructing the field to a mean error of 2 using nodes that are randomly self-selected. Costs are scaled as shown in the legend.

6. RESULTS FOR ALL METHODS

The same analysis of plotting mean error as the reporting parameters are changed is carried out for all the methods discussed earlier. It is of little use to compare the raw mean errors as functions of these reporting variables as each method has a different cost associated with it. The only way of justly comparing the methods is to plot each on the cost-complexity curve at a given mean error as done in Figure 9. As the plots are dense with information, three different plots are used to show how the costs vary with the method. The costs not explicitly plotted are (i) the sensing cost: this does not change with different reporting parameters and (ii) the number of one-hop relays: this is a scaled version of the number of reporting nodes. The first cost considered is the number of reporting nodes required to obtain a mean error of 2 and is shown in Figure 10.

The hop-based methods do not save many nodes from reporting – this is expected since these methods serve mainly to save cost in

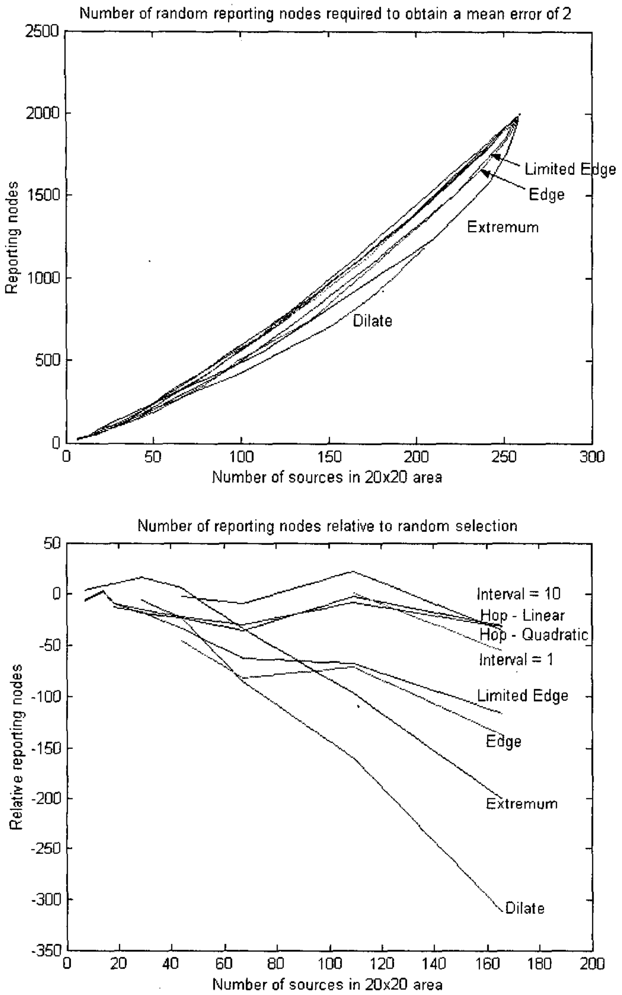

Figure 10. The number of reporting nodes for sufficient reconstruction of the fields for each method. Top: the cost of all methods plotted on the same axes is difficult to distinguish. Bottom: a detail view of the difference in cost between each method and random selection. Ten trials for each of the 100 test fields are used to generate this data.

the form of total transmit/receive events by preferentially selecting nodes closer to the basestation to report. The interval methods save up to 50 nodes from reporting but also perform worse than the random method under some circumstances. Since the interval methods do not consider hop number and do not require local communication, it is expected that the other costs for the interval methods will be the same relative to the random selection.

Substantial gains are possible when the simple local communication is used to select nodes that are better representatives of the field. Comparing the edge method with its limited counterpart, the restriction to four neighbor samples does impact the effectiveness of selecting representative nodes but both perform substantially and categorically better than the methods without local communication. The similar extremum method has both a higher and lower cost in different regions: it appears better than the edge method at representing complex fields. Finally, the additional communication required for the dilate method results in the best selection of nodes of the ones considered. Recall that this cost is a measure of the bandwidth and energy requirement at nodes near the basestation so these savings reflect bottleneck reduction near the basestation. All the curves plotted converge to the same point as the number of reporting nodes approaches the total size of the network. In fact, all methods considered report exactly the same data for reconstruction, namely all of it, in this limiting case.

The price of local communication is more apparent when evaluating the cost as the total number of receive events as shown in Figure 11. The method requiring the most local communication, the dilate method, actually remains on par with the edge and extremum methods through a better selection of random nodes for all but the more complex fields. As suggested by Figure 10, the additional local receive events are counterbalanced by less of a need to multi-hop selected data back to the basestation. The edge and extremum methods are equivalent at both ends of the plot but the regional superiority seen in Figure 10 is still apparent. All three of these methods require about 30,000 more receive events than random selection. The limited edge method reduces this gap significantly by limiting the allowed number of local receive events during the selection process and is a beneficial improvement if overall energy consumption is more of a constraint than bandwidth near the basestation.

The hop-based methods require fewer total receive events than the random method as they preferentially select nodes closer to the basestation and hence result in shorter multi-hop data collection paths. The savings in this scenario is up to ~12% with the quadratic penalty faring slightly better than the linear penalty. As

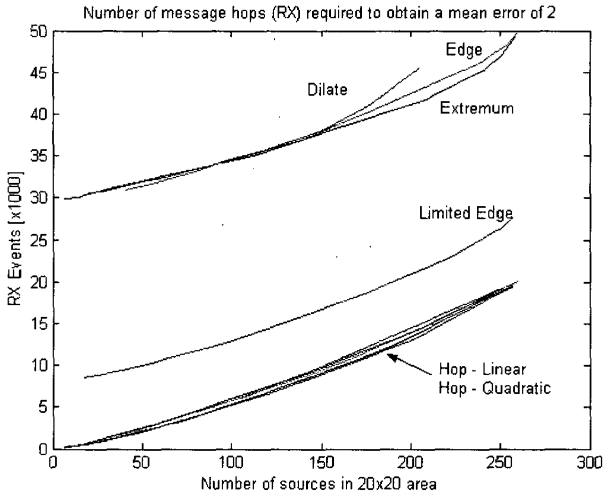

Figure 11. Total number of receive events for sufficient reconstruction of the fields.

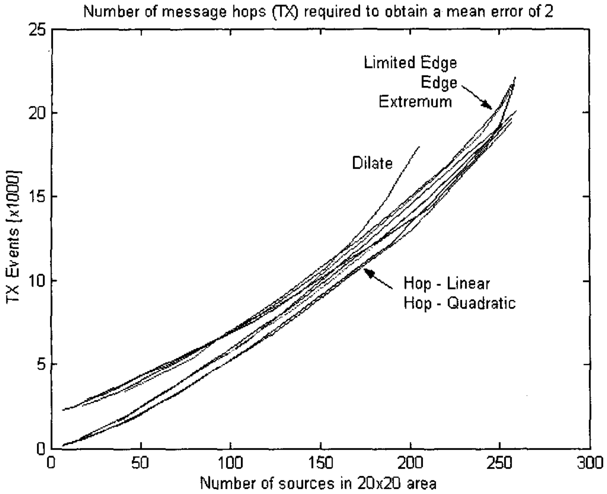

Figure 12. Total number of transmit events for sufficient reconstruction of the fields.

suggested in the discussion of Figure 10, the relative values for the interval and random methods remain the same for this cost as well.

Evaluation of the methods based on transmit events is shown in Figure 12. Since the transmit-receive pattern is symmetric in methods not requiring local communication, the same ordering of these methods as in Figure 10 is observed. The benefit of the limited edge method is only in reducing the number of receive events so it is back to the same level as the other local communication methods in this metric.

For more complex fields (>150 sources) the local communication methods do not require significantly more transmission events than the simpler ones. Again, the additional transmit events required for node selection is counterbalanced by a lesser requirement for multi-hopping messages back to the basestation.

7. PROPAGATION OF UNCERTAINTY

Given that the range of data lies in the interval [0, 2^6] and that mean error is unlikely to drop significantly below 1, it is ineffective to use many more than 7 or 8 bits to represent sensor data at each node. The quantization step q represents the resolution of the value reported by the node. For example, with q = 64 the node must send one of the symbols {0,64}, whichever is closer to the actual sampled value. This corresponds to using a single bit to represent the sensor data: a sending node either sends a high (value closer to 64) or a low (value closer to 0). The mean error of this method is expected to be (64-0)/4 = 16. Decreasing the quantization step, equivalently increasing the number of representative bits, reduces the error in the reconstructed field. This phenomenon is depicted in Figure 13.

As the quantization step is decreased below a certain threshold value, the mean reconstruction error ceases to be reduced. This threshold is a function of the number of nodes reporting data:

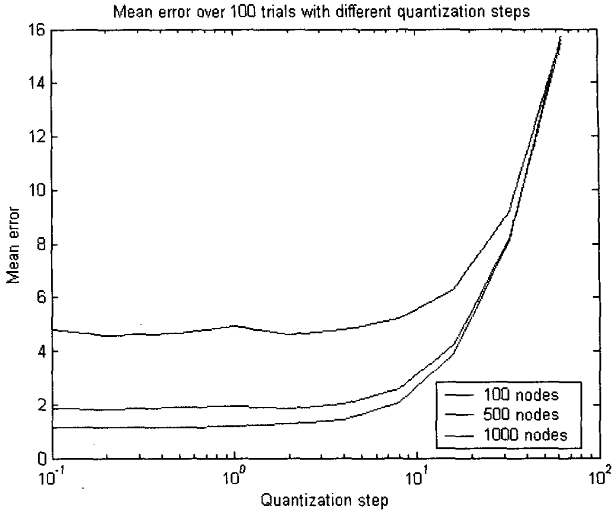

Figure 13. Mean error increases with the quantization step. Each curve represents the average of 100 random node samples. The curves all approach the point (64,16).

more nodes allow for the benefits of smaller quantization steps. It is inefficient to send a number of bits resulting in a quantization step value below this threshold. This also illustrates that using double-precision numbers in simulation does not falsely enhance the reconstruction – for one thousand nodes the majority of the error arises from mechanisms other than the quantization step.

For simulating uncertainty in node positions, instead of increasing the quantization step to induce errors in reconstructing the field, the x and y positions of each node are falsely reported to the basestation. The position error is drawn from a zero-mean normal distribution with a parametric variance. This variance is shown on the x-axis of Figure 14.

Certainly the slope of the curves in Figure 14 is highly-dependent on the complexity of the field. However, this plot does confirm qualitative intuition: imprecise knowledge of node locations obviates the benefits of more reporting nodes. A more detailed analysis of this trade-off could consider the resource cost of localizing nodes in the network and how this relates to the overall effectiveness of data recovery. The simulation in Sections 5 and 6 artificially introduces a mean position uncertainty of up to 0.05R in the earlier trials as each node samples only from a discrete set of grid values. According to Figure 14, this uncertainty does not impact the results.

8. CONCLUSIONS AND FUTURE WORK

Through simulation of 100 test fields without discontinuities and the resultant cost-complexity curves, the considered algorithms can be summarized in a few key comparisons. The addition of a hop-based penalty to the other methods is possible and should result in a similar 4-5% cost decrease of recovery without incurring any additional local communication. For example, the hop-based and extremum methods could be combined to reap the benefits of both. The optimization of the penalty is a worthy pursuit to further minimize costs: there is no reason why the

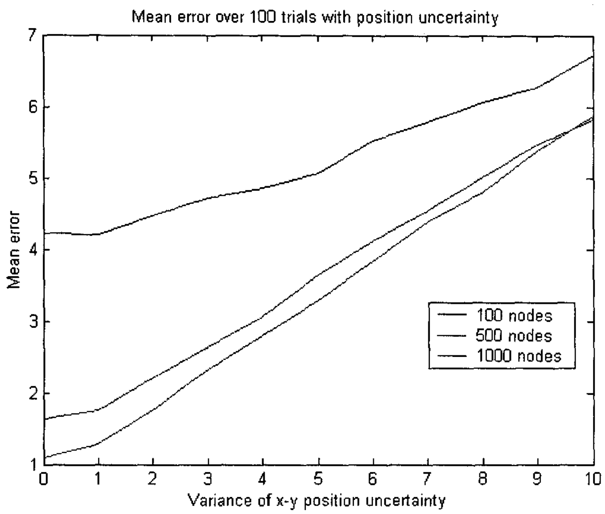

Figure 14. Adding position uncertainty to the reporting nodes increases the error in reconstruction. The variance is given in units of *R*. Values of the curves at the far left of the plot (those without uncertainty) correspond to the appropriate number of nodes in Figure 7.

quadratic penalty chosen arbitrarily for this study should be the optimal selection. However, a poorly selected penalty, even if monotonic, can adversely impact results as compared to completely random selection.

Attempting to map contours as in the interval method requires that each node sample its sensor but does slightly improve the other costs. This tradeoff merits consideration if local communication is to be avoided but not if the energy cost of sensing is on the same order of communication. The interval method might also be better suited to a different reconstruction technique specific to contours or a different error metric.

The one-step local communication in the extremum and edge-based methods offer significant improvements in the selection of representative data for recovery but may require a higher total number of transmit and particularly receive events. One additional benefit not captured by the results presented is that the local communication methods are adaptive: by setting the selection threshold for the network, more complex fields will automatically result in more nodes reporting and a higher-resolution reconstruction of the field. The network parameter set in other methods is simply a probability of reporting and will result in the same number of data points regardless of the complexity of the field. If this parameter is set assuming a simple environment is to be monitored, there will be no feedback to indicate when this assumption has failed.

The analysis above applies only to networks of 2000 nodes with the particular communication distances considered. While a general quantitative assessment of the scaling properties of the scattered data collection as functions of the number of nodes and their connectivity patterns is beyond the scope of this paper, there are a few statements that can be made at this point. If the number of nodes increases while maintaining a constant node density (i.e. the field area is increasing appropriately or the communication radius is decreasing), the negative impacts of local communication as an become less taxing and the relative benefits of the edge/extremum algorithms become more attractive. This is because the mean reporting paths become longer from the self-selected nodes while the local communication costs remain the same. However, increasing the mean connectivity of the network while keeping other field and network parameters constant results in local communication becoming relatively more costly. The limited methods discussed can be used to circumvent this penalty. The size and density of the network to be implemented will determine the best approach for data recovery.

One particularly useful future direction is the determination of the optimal number of nodes reporting to meet a particular error requirement. For a given selected from the 100 considered, what is the minimum number of data points required to reconstruct the field? What are the properties of these optimal nodes relative to their close neighbors? Finding this bound would motivate further algorithm discovery and limit what is possible with local communication. The same analysis repeated with a maximum instead of a mean error metric might result in a different ordering of the methods.

It is apparent that no single scattered data recovery method is singularly superior to the others in consideration for all levels of complexity. A decision on which method to use is dependent on the complexity of the field as well as the real energy costs of the atomic operations considered herein: sensing, transmission, reception and potentially even computation. If receiving is cheap compared to transmission then the local communication methods provide better bandwidth properties for a modicum of extra energy cost to the network.

The next logical step is to compare the scattered data methods to those performing intra-network processing using the same cost and complexity metrics for evaluation. It may be that under certain conditions, the random method is the best method for recovering data from sensor networks even when compared to data fusion methods.

9. ACKNOWLEDGMENTS

Our thanks to Kannan Ramchandran for contributing some of the ideas found in this work.

10. REFERENCES

[1] K.A. Arisha, M.A. Youssef, M.F. Younis, "Energy-aware TDMA based MAC for sensor networks", IEEE IMPACCT 2002, New York City, NY, USA, May 2002.

[2] M. Chu, H. Haussecker, F. Zhao, "Scalable Information-Driven Sensor Querying and Routing for ad hoc Heterogeneous Sensor Networks," Int'l J. of High Performance Computing Applications, 2002.

[3] L. Doherty, B.A. Warneke, B.E. Boser, K.S.J. Pister, "Energy and Performance Considerations for Smart Dust," International Journal of Parallel Distributed Systems and Networks, vol. 4, no. 3, 2001, pp. 121-133.

[4] P. Gupta and P. R. Kumar, "The capacity of wireless networks," IEEE Trans. on Info. Theory, vol. IT-46, Mar. 2000.

[5] Z. Haas, J. Halpern, and L. Li, "Gossip-based ad-hoc routing," IEEE INFOCOM 2002, New York, NY, June 2002.

[6] J. M. Hellerstein, W. Hong, S. Madden, K. Stanek, "Beyond Average: Toward Sophisticated Sensing with Queries," Proc. IPSN 2003, Palo Alto, CA, Apr. 22-23, 2003.

[7] B. Karp and H.T. Kung, "GPSR: Greedy Perimeter Stateless Routing for wireless networks," MobiCom 2000, Boston, MA, Aug. 2000.

[8] O. Kreylos, B. Hamann, "On Simulated Annealing and the Construction of Linear Spline Approximations for Scattered Data", IEEE Trans. on Visualization and Computer Graphics, vol. 7, no. 1, Jan. 2001.

[9] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, J. Anderson, "Wireless sensor networks for habitat monitoring," WSNA 2002, Atlanta, GA, September 2002.

[10] D. Niculescu and B. Nath, "Ad hoc Positioning System (APS) using AoA", In Proceedings of INFOCOM 2003, San Francisco, CA.

[11] G.J. Pottie, W.J. Kaiser, "Wireless Integrated Network Sensors," Communications of the ACM, vol. 4, no. 5, May 2000.

[12] V. Raghunathan, C. Schurgers, S. Park, and M.B. Srivastava, "Energy-aware wireless microsensor networks," IEEE Signal Processing Magazine, vol. 19, no. 2, March 2002.

[13] E. M. Royer and C-K. Toh. "A review of current routing protocols for ad-hoc mobile wireless networks," IEEE Personal Communications, April 1999.

[14] C. Savarese, "Robust Positioning Algorithms for Distributed Ad Hoc Wireless Sensor Networks", Masters Thesis, UC Berkeley EECS, 2002.

[15] Y. Yu, R. Govindan, D. Estrin. "Geographical and Energy Aware Routing: a recursive data dissemination protocol for wireless sensor networks," UCLA/CSD-TR-01-0023, May 2001.