Poster Abstract: Exposing Two Critical Myths about Correlation Aware Data Aggregation

I. Correlation Aware Data Gathering

In this paper, we consider one of the key tasks performed by wireless sensor networks (WSNs): the collection and transfer of sensor data from sensors in the field to the sink for processing. This task is referred to as *data gathering*.

In most sensor applications, data from different sensors are correlated with each other. This correlation can be leveraged in order to reduce the number of transmissions and hence energy consumption for the data gathering process. Using correlation unaware aggregation trees (CUATs) such as shortest path trees (SPTs), opportunistic aggregation can be achieved when paths from different sources overlap with each other. However, these structures do not necessarily maximize aggregation possible in the network and are hence correlation unaware. By explicitly constructing aggregation trees with the purpose of fusing data inside the network as early as possible and as much as possible, the correlation existing in sensor data can be fully exploited to reduce energy consumption. Many research works [1], [2], [3] have proposed solutions in this direction for determining such correlation aware aggregation trees (CAATs).

In this paper, we study the energy efficiency of CAATs from a new perspective: we consider data gathering applications with real-time requirements and explore how delay constraints and other network conditions affect the energy efficiency of a CAAT structure. Energy-delay tradeoff in wireless sensor networks has been explored in multiple dimensions. In this work, we consider energy-delay tradeoff from the perspective of aggregation tree structures. Intuitively, with a higher application delay tolerance, longer aggregation paths that support more en-route aggregation can be generated, thereby resulting in an aggregation tree with lower energy cost. Therefore, the energy improvement tends to increase monotonically with the application delay tolerance.

To study the energy-delay tradeoff, we consider a typical sensor network scenario with n sensors randomly distributed in a disk with radius R. All the sensors communicate using the same transmission range, which is slightly higher than that required for minimum connectivity [5]. Of the n sensors, k are randomly chosen as sources to report data to the sink located at the center of the disk. To evaluate the energy efficiency of a CAAT structure, we define energy improvement as the ratio of the cost of the CUAT structure to that of the CAAT structure for the same set of sensor nodes.

For ease of analysis, we assume perfect correlation in this paper, where two pieces of raw data packets can be combined and reduced to one packet of the same size as the original packets. In this case, Steiner Minimum Tree (SMT) over all source nodes and sink is the optimal aggregation structure [2]. Therefore, we choose SMT as the CAAT structure in this paper for evaluation because of its optimal energy cost. On the other hand, SPT is selected as the CUAT structure since it minimizes the delay required for data aggregations.

Notice that SPT is also the most efficient aggregation tree structure when there is no correlation between sensor data. Thus, it is expected that for partially correlated sensor data gatherings, the optimal aggregation tree structure are intermediate structures between that of SMTs and SPTs. Consequently, the energy improvement of SMT over SPT that we study in this paper serves as an upper bound on energy improvement for all other correlation models.

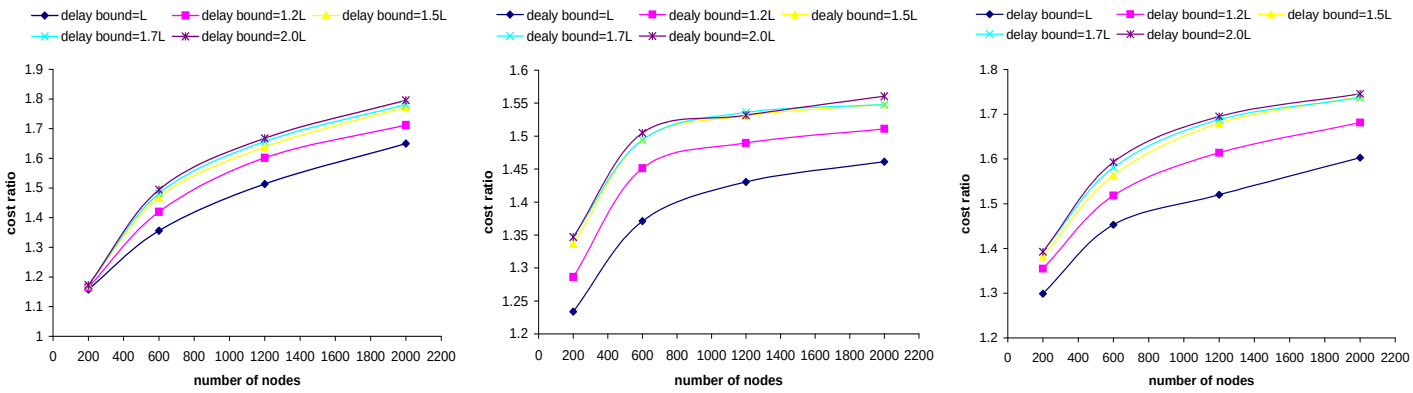
II. The Myths

In this section, we introduce the two common myths studied in this paper and explain the reasons for their belief.

- *Myth 1: Significant energy improvement can be achieved by using CAATs*
  As stated earlier, SMT proves to be the energy optimal aggregation structure for sensor applications involving perfect correlation. Compared to a naive SPT structure whose primary goal is to minimize delay, SMT structure explicitly maximizes data aggregation possible in the network, thereby ensuring minimum energy cost for the data gathering process. As a result, the cost ratio of SPT over SMT is always higher than one.
  Furthermore, in large scale sensor networks where sensor nodes are densely deployed and a fraction of the nodes are selected as sources, the energy cost of SPT as an aggregation tree is expected to be much worse than that of SMT. This can be intuitively explained as follows: when the node density is high, the different shortest paths connecting each source to the sink have a low probability of overlapping with each other. Hence, the cost of SPT over the fixed set of sources increases when node density increases. The extreme case is when node density tends to infinity ($\lambda_n \rightarrow \infty$). In this case, the resulting SPT structure becomes a SPT in Euclidean space, whose expected cost is $O(s)$, where s is the number of sources in the SPT structure. On the other hand, the cost of SMT over the same set of sources decreases as node density increases. This is because when node density increases, more candidates (nodes) are available to form Steiner points, thereby minimizing the cost of the Steiner tree. The extreme case is an Euclidean SMT structure, which results when node density tends to infinity. In this case, the expected cost is $\Theta(\sqrt{s})$ [6]. Since the cost of SPT is $O(\sqrt{s})$ worse than that of SMT in Euclidean space, the energy improvement of SMT can be expected to scale with node density. As a result, many heuristics for SMT structures have been proposed and used to improve the energy efficiency of data gathering process in wireless sensor networks.
- *Myth 2: High application delay tolerance is required in order to achieve significant energy improvement of CAATs*
  This common belief stems from the observation that on an average, paths on a SMT structure are longer than the paths on a SPT structure. To reduce the aggregation tree cost, longer paths that connect more sources en-route to sink are favored over shorter paths that connect each source to sink separately. In this way, sensor data from different sources can be combined well before they reach the sink to reduce the total number of transmissions required. Consequently, it is natural to expect that a SMT structure has longer average path length than a SPT structure. For sensor applications with a larger delay tolerance, aggregation paths with more hop count but higher degree of aggregation can be created to improve energy efficiency. Similarly, applications without delay constraints can always use SMTs more efficiently than those with delay constraints.

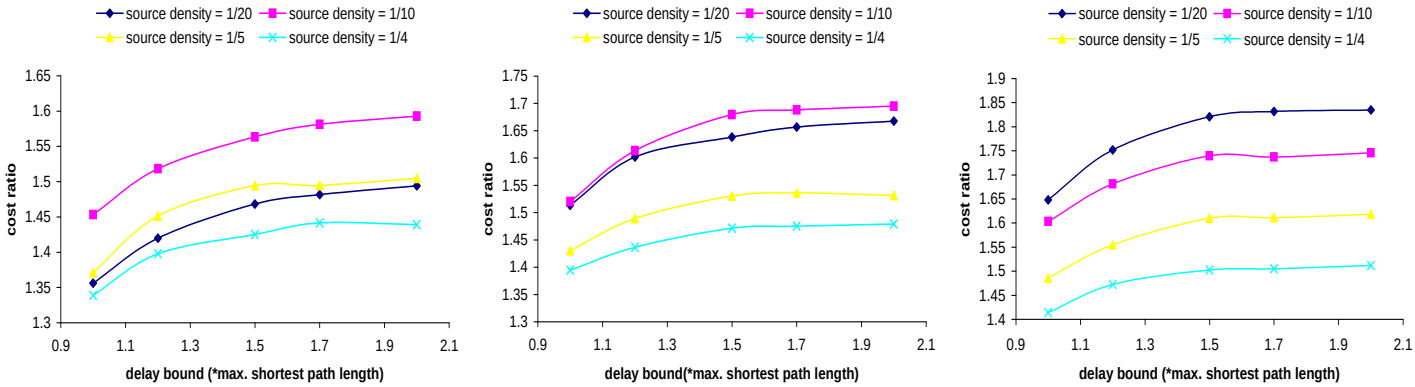
III. Studying the Myths

In this section we present simulation results to disprove the two commonly believed myths. We use a custom-built simulator written in C++ for all the simulations. Since determining the SMT structure is a NP-hard problem, a heuristic of SMT - namely the BSMA [4] (bounded-delay shortest multicast) algorithm is used to generate approximations for SMTs. This algorithm has been proven to be capable of constructing aggregation trees with an additional cost that is less than 7% that of the corresponding SMT cost. Using this algorithm, different SMT approximations can be obtained for different delay constraints. The



(a) $source\ density = 1/20$ (b) $source\ density = 1/10$ (b) $source\ density = 1/5$

Fig. 1. Performance Improvement over SPT for Different Number of Nodes



(a) $n = 600$ (b) $n = 2000$ (b) $n = 2000$

Fig. 2. Performance Improvement over SPT for Different Delay Bounds

delay constrained SMT approximations are referred as DB-SMT (delay-bounded SMT) in the rest of the discussions. Node density and source density are the two major parameters in simulation study. For each network configuration, SPT and DB-SMT trees are generated, and the cost ratios between them is calculated as a measure of energy improvement. The maximum hop count in a shortest path tree is taken as the lowest possible delay constraint required by the application and the corresponding DB-SMT tree is computed initially. The same delay constraint is then relaxed to obtain DB-SMT trees with further reduced costs.

Figure 1 shows the cost ratio of SPT vs. DB-SMT when the source densities are $1/20$, $1/10$ and $1/5$ of the total number of nodes. It can be seen that the cost ratio increases as node density increases, which indicates that SMT is more energy efficient when sensor nodes are densely deployed in the field. As outlined in the previous section, when the node density is high, the probability of shortest paths overlapping with each other is low. Hence, the naive SPT has very low aggregation efficiency and there is higher potential for energy improvement using a CAAT structure.

But this improvement is not always high. According to the simulation results, the cost ratio is always a small constant (less than 2), and the rate of cost improvement tends to slow down as the node density increases. This is in contrast to the first myth considered in the earlier section, wherein the belief was that significant energy improvement can be achieved by adopting a correlation aware aggregation tree structure.

The key reason for this counter-intuitive result can be attributed to the natural path sharing ability of SPT structures in network graph that improves their cost performance. Since all the source nodes are connected to the sink via shortest paths, and there are finite number of nodes around the sink, all the shortest paths tend to start converging at a certain distance from the sink. This makes the total cost of SPT smaller than the sum of the individual path costs. If we consider the distance where the paths start to merge as the threshold distance, then within this threshold distance, almost all sensor nodes can be considered to be part of the SPT structure. For SMT, such a threshold distance also exists, because of the limited number of nodes around the sink, and is the same as that in the SPT structure. Thus, the main difference between the SMT and SPT costs is caused by the diverse tree structures beyond the threshold distance. For SPT, shortest paths beyond the threshold distance are typically independent of each other. While for SMT, the possible aggregations among paths are still maximized to ensure optimal cost. So the cost improvement of SMT over SPT is determined by the threshold distance. The smaller the threshold distance, the larger the difference between the two trees. Hence, the higher the room for aggregation in SMT, and hence higher the cost ratio. According to our analysis [1], the threshold distance is determined by both the node density and source density parameters, while increasing with the source density. The extreme case occurs when source density is equal to 1, wherein all the sensor nodes act as sources. In this case, the threshold distance is the radius of the entire network, and hence SMT structure is the same as SPT structure.

In summary, contrary to common belief, SMT does not always bring significant cost improvement over a naive SPT, and even for the cases where SMT does have a lower cost than SPT, the cost improvement is limited due to SPT's ability of natural path sharing in network graphs.

Figure 2 shows how the energy improvement of SPT over SMT changes with respect to delay tolerance. For low delay tolerances, the cost ratio increases steadily. But when the delay bound is higher than 1.5 times that of the maximum shortest path length, the increasing trend tends to slow down and the cost ratio tends to saturate. It is obvious from

[1]The theoretical analysis is omitted here due to space limitation.

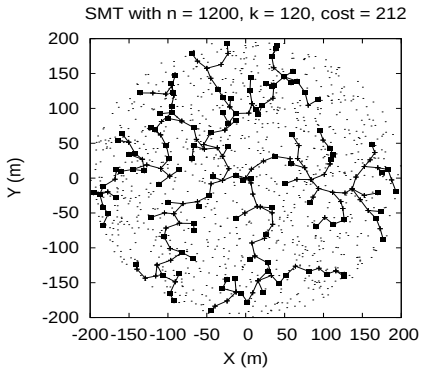


Fig. 3. Topology of SMT, n=1200, k=120

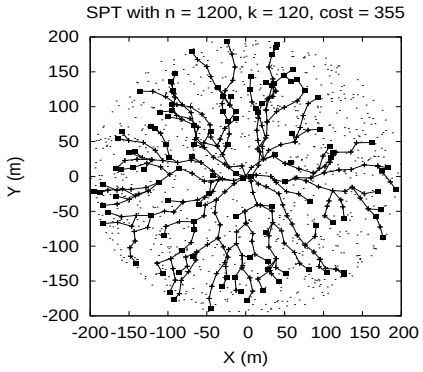


Fig. 4. Topology of SPT, n=1200, k=120

the results that after a certain threshold value, higher delay tolerance does not help increase the cost ratio anymore. This result in turn contradicts our second myth. To understand this phenomenon better, we plot the structures of SMT and SPT for $n = 1200$ in figures 3 and 4 respectively. It can be seen from the two structures that the "backbone" structure of SMT is similar to that of SPT, where several shortest paths tend to divide the network graph uniformly. But there are lesser number of shortest paths in the backbone of SMT. However, there are several "branches" connecting the sources to the backbone structure, although via shortest paths. Thus, while this structure is more efficient than SPT in terms of cost with the paths being combined as much as possible, the longest path length is not significantly higher than that in the SPT structure. Further, the maximum delay tolerance that is helpful in reducing the aggregation tree cost is given by the length of the longest path on the SMT structure. Hence, given the practical structural characteristics of SMT, the longest path in SMT tends to be only a small constant order that of the longest path in the SPT structure.

In summary, a relatively small delay tolerance is enough for a sensor application to guarantee a correlation aware aggregation tree with a near optimal cost.

IV. Practical Implications

A. Practical Implications of Energy Improvement Observation

As identified through simulation studies and analysis, the energy improvement of CAATs over that of CUATs is not always significant. We now discuss the practical implications of these observations.

Many research works have studied the optimization of data aggregation trees to maximize energy improvement. It is commonly believed that CAAT structures can always bring substantial energy savings. But energy saving of CAATs does not come for free. To set up the CAATs structures, explicit communication between sink and sensor nodes is required. Furthermore, the CAAT structures over different sets of sensor nodes are different from one another, which implies that a dedicated tree construction process is necessary for each data gathering round with different set of source nodes. This constitutes a non-negligible amount of energy consumption, and may even offset the cost savings resulting from optimization (data aggregation). Under these circumstances, CAATs are more energy efficient only when the extra cost incurred from the set-up of the structure itself can be compensated by energy savings due to aggregation over that of CUATs. But the observation clearly shows that the cost benefits of CAATs is limited and hence it might not be beneficial to consider them in all sensor network applications, taking into account the cost incurred in the setting-up process of the CAAT structure.

It is also possible that in some sensor network applications, the set of source nodes reporting data packets to the sink is not known a priori, in which case CAATs cannot be computed before the aggregation process and are hence not viable solutions for data gathering. On the contrary, CUATs such as shortest path trees, even for diverse sets of sources, can be obtained from the same shortest path tree over the sink and all sensor nodes by trimming branches that are entirely over non-source sensor nodes. Therefore, shortest routes can be programmed into sensor nodes before the data gathering process, thereby eliminating the cost of explicit tree construction.

Due to the above reasons CUATs such as shortest path trees may be a desirable data gathering structure when compared to CAATs under several circumstances.

B. Practical Implications of Delay Tolerance Observation

We also observed that increasing delay tolerance does not always help reduce aggregation tree cost. When the delay constraint is small, the cost of the SMT structure reduces with delay bound. But this is not always true: beyond a delay tolerance which is comparable to the longest shortest path length, the cost ratio improvement saturates. This is because the optimal aggregation tree can be constructed for this delay tolerance. Practically, this means that an application does not have to be designed with large delay tolerances to ensure energy efficiency.

V. Conclusions

In this paper, we study the energy efficiency of correlation aware data aggregation and the tradeoffs involved for the data gathering process in wireless sensor networks. Sensor applications with and without delay tolerance are considered. Through quantitative analysis and theoretical reasoning, we infer practical limitation on the achievable energy improvement in adopting a correlation aware aggregation structure as opposed to a correlation unaware structure, as well as practically maximum useable delay bound that can deliver the maximum achievable improvement.

References

[1] C. Intanagoniwawat, D. Estrin, R. Govindan, and J. Heidemann, "Impact of Network Density on Data Aggregation in Wireless Sensor Networks," in *International Conference on Distributed Computing Systems (ICDCS'02)*, Vienna, Austria, July 2002.

[2] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On Network Correlated Data Gathering," in *INFOCOM*, Hong Kong, Mar. 2004.

[3] Krishnamachari B. Pattem, S. 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*, April 2004, pp. 28 – 35.

[4] M.; Qing Zhu; Garcia-Luna-Aceves Parsa, "An iterative algorithm for delay-constrained minimum-cost," in *IEEE/ACM Transactions on Networking*, August 1998, vol. 6, pp. 461 – 474.

[5] Vikram Mhatre and Katherine Rosenberg, "Design Guidelines for Wireless Sensor Networks: Communication, Clustering and Aggregation," in *Elsevier Science Journal*, Aug. 2003.

[6] Konstantinos Kalpakis and Alan T. Sherman, "Probabilistic analysis of an enhanced partitioning algorithm for the steiner tree problem in r^d," in *Networks*, may 1994, vol. 24, pp. 147–159.